Deep Learning For The Classification of Lung Diseases Using Chest X-Rays

Seyedeh Zahra Khamesi

A Thesis

in

The Department

of

Computer Science and Software Engineering

Presented in Partial Fulfillment of the Requirements

for the Degree of

Master of Applied Science (Computer Science) at

Concordia University

Montréal, Québec, Canada

November 2023

© Seyedeh Zahra Khamesi, 2023

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: **Seyedeh Zahra Khamesi**

Entitled: **Deep Learning For The Classification of Lung Diseases Using Chest X-Rays**

and submitted in partial fulfillment of the requirements for the degree of

Master of Applied Science (Computer Science)

complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the Final Examining Committee:

_____ Chair
Dr. Marta Kersten-Oertel

_____ Examiner
Dr. Muna Khayyat

_____ Supervisor
Dr. Ching Yee Suen

Approved by _____
Hovhannes Harutyunyan, Chair
Department of Computer Science and Software Engineering

_____ 2023

_____
Mourad Debbabi, Dean
Faculty of Engineering and Computer Science

Abstract

Deep Learning For The Classification of Lung Diseases Using Chest X-Rays

Seyedeh Zahra Khamesi

The discovery of X-rays marked a significant milestone in the field of medicine. One of the most common types of X-rays, the chest X-ray (CXR), allows doctors to examine an individual's internal structure without surgery. Over the years, deep learning methods and algorithms have been developed to automate lung disease detection and identification. This paper introduces RADIA, a project that combines multiple deep learning techniques to identify abnormal areas and abnormalities in chest X-rays. RADIA builds upon previous studies conducted by the Stanford ML group, such as ChexNet and ChexPert. Our team utilized the ConvNeXt-Large, a deep learning convolutional model, implemented with a pre-trained ConvNext algorithm on the ImageNet database to classify various pathologies from public datasets like ChestX-ray14, CheXpert, MIMIC-CXR, PadChest, and VinDr-CXR, as well as a private dataset obtained from the Picture Archiving Communication System (PACS) at Verdun and Notre Dame Hospitals in Montreal in the collaboration with CIUSSS (Centre Integre Universitaire de Sante et de Services Sociaux du Centre-Sud-de-l'Ile-de-Montreal) and valuable consultants from the radiology team at Notre Dame Hospital contributed to the project's success. Our team employed image enhancement and augmentation techniques to create various image versions. We used different and novel approaches to address the challenges, and the results were evaluated using metrics such as AUC, F1, and G-Means to analyze performance with imbalanced input data. It is essential to note that the project's development extends beyond the creation of a web tool based on deep learning techniques. Our future plans involve building a decision helper that combines inference models and web tools to assist healthcare professionals.

Acknowledgments

This thesis reflects the current state of the RADIA project which is currently under development in collaboration with the Research Group at CIUSSS(Centre intégré universitaire de santé et de services sociaux du Centre-Sud-de-l'Île-de-Montréal).

I'm extremely grateful to the following people who have helped me undertake this research. This endeavor would not have been possible without their generous support, collaboration and consultants, invaluable patience, encouragement, and feedback:

Dr. Ching Yee Suen, Co-Director of CENPARMI(Centre for Pattern Recognition and Machine Intelligence), Computer Science and Software Engineering, Concordia University

Mathieu Mailhot, Directeur adjoint des ressources informationnelles du Centre de services régional,Direction des ressources informationnelles, CIUSSS du Centre-Sud-de-l'Ile-de-Montréal

Dr. Jean-Francois Samson, Chef de la recherche et ingénieur en informatique, CIUSSS du Centre-Sud-de-l'Ile-de-Montréal

Dr. Martin Chandonnet, Radiologist at Centre Hospitalier de l'Université de Montréal

Annabelle Cambron-Premont, Chef de service informatisation clinique, DARI-CIUSSS at CIUSSS du Centre-Sud-de-l'Ile-de-Montréal

Tania Asselin-St-Arneault, Agente de planification, de programmation et de recherche, CIUSSS du Centre-Sud-de-l'Ile-de-Montréal

I remain also forever grateful to anyone who taught me anything throughout my journey, especially my dear friend and mentor Michelle Khalife, and also thankful to my parents and friends, who kept my spirits and motivation high during this process.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 An Overview of X-rays

The development of non-invasive methods for exploring the internal structure of the body is one of the most important milestones in medical history. One of the most significant advances in this field was the discovery of X-rays. As a result of this groundbreaking innovation, a wide range of ailments could be diagnosed without the need for surgical intervention. The invention of this revolutionary tool can be traced back to Wilhelm Conrad Röntgen's experiments in December 1895. During his study of a novel type of radiation, Röntgen discovered that it was capable of penetrating screens with substantial thickness. This radiation's properties were revealed in merely seven weeks of relentless research. As a result of their enigmatic nature, Röntgen appropriately named them 'X-rays', emphasizing their unknown characteristics [75]. Within a short period of time after its discovery, X-ray machines were widely used in medical diagnostics. During World War I, these devices proved invaluable in pinpointing external objects, such as bullets, within the body. Furthermore, they played a significant role in detecting fractures, thus revolutionizing the field of orthopedics.

Today, clinical evaluations, especially those related to pulmonary disorders, frequently involve the use of magnetic resonance imaging (MRI), computed tomography (CT), and X-rays. In addition to facilitating the visualization of anatomy, these diagnostic instruments also facilitate an understanding of potential pathological signs and symptoms [2]. X-rays are pivotal diagnostic tools that

allow medical professionals to examine a wide variety of anatomical structures, including bones, joints, and soft tissues, with a special focus on internal organs. As a result of their relatively soft consistency, tissues such as blood, skin, fat, and muscle appear as dark gray shades on an X-ray image. Alternatively, structures with higher density, such as bones and tumors, tend to obstruct the X-ray beam, resulting in a white representation. The X-ray beam penetrates the break, delineating it as a conspicuous dark line in the case of a fracture. Chest X-ray (CXR) images have become ubiquitous and popular in the medical field for a variety of reasons. The procedure is characterized by its cost-effectiveness, its judicious use of radiation, and its commendable sensitivity to a range of diseases and disorders. Consequently, CXR is often employed as the initial imaging modality and continues to play a pivotal role in the screening, diagnosis, and therapeutic management of a wide range of medical conditions [62].

In the context of chest X-ray imaging (CXR), there are three predominant orientations. First, the posteroanterior (PA) perspective involves the trajectory of the X-ray beam from the posterior to the anterior of the patient, rendering it the standard approach that provides a comprehensive representation of thoracic structures. the lateral view provides a perspective from a side angle that clarifies the depth and relative placement of the thoracic components. Lastly, an anteroposterior (AP) perspective refers to the beam's path from the anterior to the posterior of the patient (Fig1.1).

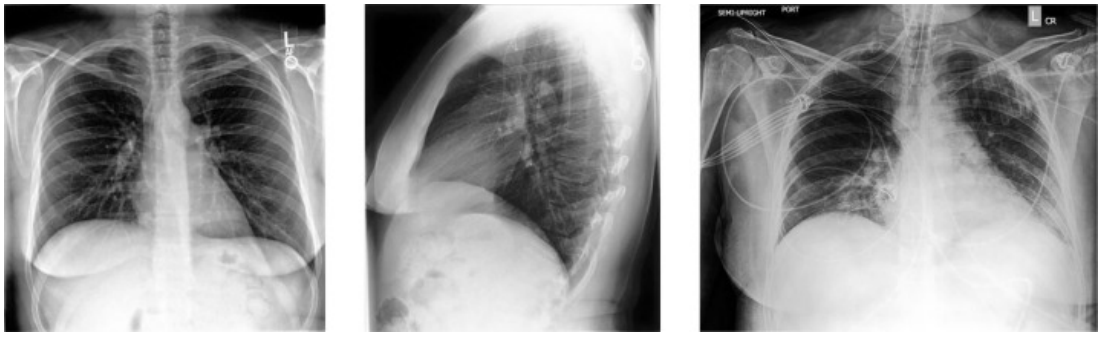

Figure 1.1: Left: posterior-anterior (PA) frontal chest radiograph. Middle: lateral chest radiograph. Right: Anterior-posterior (AP) view chest radiograph [13]

1.2 Motivation

Chest X-ray images are typically interpreted effectively by skilled radiologists. However, radiograph interpretation can be obscured by a number of external and internal factors, resulting in potential misinterpretations. Even experienced radiologists may find it challenging to read X-rays because of subtle variations in tissue densities, overlapping structures, or the inherent complexity of disease. Misinterpretations or oversights, although inadvertent, can have critical implications for patient care. In addition, it is important to note that a significant number of fatalities are caused by delayed detection, which prevents patients from receiving optimal therapeutic interventions. When pulmonary mortality is detected early and treated appropriately, mortality can be markedly reduced. Thus, early detection of lung diseases is imperative for enhancing preventive measures, curtailing the spread of disease, and reducing death rates. Further, the urgent need for rapid detection can increase the risk of misinterpretation, especially during emergency situations, when healthcare professionals often have a limited window to make accurate detection, affecting patient outcomes.

Our motivation is a desire to enhance patient care, improve the detection process, and assist radiologists through the use of advanced techniques and tools to interpret X-rays, particularly in emergency situations where the margin for error is minimal and the risk of incorrect interpretation is high.

1.3 Automated Disease Diagnosis

Deep learning methodologies have contributed significantly to the advancement of automated disease diagnosis over the past decade. In spite of inconsistencies such as resolution disparities, variable illumination, and unpredictable camera positioning, this advanced computational approach has proven to be extremely effective in analyzing images. The deep learning algorithms have demonstrated exemplary performance across a wide range of tasks, from classification to recognition. Computer vision, speech recognition, natural language processing, audio discernment, and bioinformatics are examples of these areas [12].

In recent years, deep learning has emerged as the preferred method of addressing medical challenges due to its unparalleled ability to extract salient features autonomously. This proficiency ensures its

applicability across a wide range of scenarios, enabling accurate and rapid data analysis. A primary focus of image processing techniques is to emphasize pertinent information, enhance sharpness, and make the image more computationally feasible. A significant component of augmented performance is the representation of data effectively [7].

1.4 Research Challenges

In the field of medical imaging, the automated analysis of chest X-rays remains an intricate domain warranting meticulous investigation. Chest X-rays present significant challenges for automatic classification, detection, auto-labeling, segmentation, and feature extraction. It is necessary to address these issues by accessing extensive datasets, which are kindly provided by research groups and healthcare professionals who are committed to this area of research. For detection to be accomplished precisely, it is imperative that pertinent features are selected and extracted meticulously.

In addition, medical image data, particularly those of high resolution, are continually expanding. This has necessitated the development of faster methods to extract and index features that are both numerically and semantically meaningful [65, 50]. Despite the fact that this domain provides numerous advantages, it is imperative to acknowledge its limitations. In particular, there are numerous challenges associated with the large amount of medical data that is available. There can be significant variations in the quality of data, often as a result of the diverse equipment used for data acquisition [12].

Furthermore, the process of comparing findings across different datasets can be challenging for researchers. Methodologies used in data processing and training can have a significant impact on the results of such comparisons [15]. In addition, the process of labeling medical images poses a distinct challenge, as the annotation process necessitates the expertise of a medical professional [78].

Moreover, integrated image processing techniques and automatic textual analysis are emerging challenges in the field of medical informatics [65]. One promising approach to addressing some of these challenges is through the use of natural language processing (NLP). By utilizing natural

language processing (NLP), information can be extracted from medical records in an alternative manner. Medical diagnostics and treatment modalities could be enhanced by integrating image content analysis with natural language processing techniques [65, 50].

Subsequently, in certain clinical scenarios, the inclusion of healthcare contexts or expert annotations is essential to enhance confidence in a model's ability to detect or categorize diseases. For these contexts to mitigate the potential for incorrect clinical decisions, it is imperative to differentiate among three labels: uncertain, positive, and negative. The term 'uncertain' refers to a state of ambiguity or indeterminacy, which indicates that a diagnosis or categorization cannot be made with certainty [39].

Furthermore, identifying and analyzing outliers presents a significant challenge due to their potential deviations from expected behavior. In deep learning, data distribution remains static; therefore, the input during training resembles that during testing. Real-world scenarios and upstream data are dynamic, and models may encounter unfamiliar input concepts as a result of these factors. When there is a divergence in data distributions during training and testing, this discrepancy increases the risk of model failure. These distributional variations can affect both the training and testing datasets or only the training dataset, a phenomenon referred to as in-distribution and out-of-distribution detection [24, 22].

This thesis aims to address some of the existing challenges in interpreting and analyzing chest X-ray images which are mentioned. In response to these a variety of challenges, our research is directed towards the development of advanced deep learning model and variety of methods. These methods are systematically designed to enhance image enhancement in X-rays, and address the problem of imbalanced datasets.

1.5 Research Objectives

Radiologists are most often requested to perform imaging studies by a variety of health care professionals, including general physicians, oncologists, and surgeons. The radiologist is required to thoroughly review the patient's comprehensive medical records prior to undertaking an interpretative assessment, including symptomatic presentations, previous imaging studies, and relevant

medical records. The imaging scans are subjected to a thorough examination following the complete assessment to detect any potential abnormalities, whether they are tumors, skeletal fractures, or other pathological conditions. Radiologists then prepare analytical reports based on their interpretations of these imaging studies. The clarifying reports are then transmitted to the requisitioning physician, who will formulate a diagnosis and formulate a treatment plan. It may be necessary to consult with radiologists and other specialists, as well as experienced radiologist peers and colleagues in order to determine the best treatment strategy for a given patient. It is especially important to note that radiologists face intense pressure during emergency situations to provide accurate interpretations. Radiological reports provided during intense on-call periods may be subject to diagnostic inaccuracies, which may be the result of increased operational demands,[34]. A radiologist has a wide range of tasks to perform, and working with images is among the most challenging tasks that have been discussed. The goal of this project is to provide radiologists with better tools to assist them by utilizing advanced technology.

Therefore, the primary objective of this thesis is to develop an automated classification tool to assist radiologists in analyzing chest images. Due to the complexity and importance of accurate interpretation in chest X-rays, this tool is intended as an essential assistant for radiologists. Using the proposed system, the radiologists will be able to focus on complex cases and make more informed decisions by automating the detection process. Particularly in emergency situations, when every second counts, this becomes particularly valuable. The overall goal of the project is to effectively integrate deep learning algorithms with medical professionals' expertise by RADIA web tool, thereby improving detecting procedures and improving patient care.

1.6 Contributions

Providing new insights and understandings is an integral part of academic research, not just the assimilation of existing knowledge. As a result, it becomes necessary to differentiate between the various contributions made within this thesis.

- The fisrt contribution is the process of systematically employing a variety of chest X-ray datasets from public to private. Medical imaging data is required to be both adequate and

sufficiently diverse to allow a comprehensive and robust analysis to be conducted.

- The second unique contribution of this research is the integration of multiple datasets. Through the integration of these datasets, the research does not simply gather data sequentially. Rather, it combines various information elements into a coherent framework. In particular, these datasets, which come from a variety of sources and contexts, reveal the depth and complexity, characteristics, and layers of the information being analyzed.

- As the Third contribution, the ConvNeXt model is employed as an innovative and pioneering methodology for training extensive datasets. By adeptly integrating global and local attributes, the approach offers an innovative approach to extracting features. Consequently, while the broader context of an image is preserved, the precise details crucial for detection purposes are also preserved. ConvNeXt's scalability is one of its most significant features, making it suitable for our extensive datasets. Although it is capable of handling complex tasks, it maintains a reasonable number of parameters, enhancing its computational efficiency and making it adaptable for a variety of hardware configurations.

- The fourth major contribution to this research comes from a comprehensive set of experiments. In these experiments, parameters and hyperparameters and datasets were systematically varied to evaluate their respective effects on the outcomes. Our experiments did not focus on a single dataset but were conducted across a variety of datasets, thus enhancing the robustness of our findings and ensuring their utility across a wide range of settings. Through this extensive experimental setup, we have gained a deeper understanding of the interaction between different computational elements and datasets and their impact on the research outcome.

- The fifth significant contribution is transitioning from these foundational elements to their practical implementation in the real world, we introduce RADIA. This web-based tool, engineered with precision, is designed to assist radiologists. By automating the detection process, RADIA stands as a source of support, reducing radiologists' workload. It is especially beneficial during emergencies when time is of the essence and precision is of the utmost importance.

The following is a brief description of how this thesis is structured:

Chapter 1 provides a general overview of this thesis, including background and motivations, deep learning and automated diagnosis of disease, challenges, thesis objectives, and contributions.

Chapter 2 discusses the previous research done on medical image classification as the literature review.

Chapter 3 describes the X-ray datasets that are used in this study.

Chapter 4 is an investigation of image processing, feature extraction, and deep learning algorithms.

Chapter 5 discusses the methodology that was employed for this study.

Chapter 6 provides details about our implementation, experiments, results, conclusion, and limitations.

Chapter 7 covers the ethics subject and other content.

Chapter 2

Literature Review

Over the past few years, deep learning has gained significant recognition for its unparalleled ability to address complex technical challenges. In particular, this can be seen in areas such as image classification, machine translation, and speech recognition. Healthcare has identified a wide range of potential applications for deep learning, including medical imaging analysis, waveform sleep analysis, inpatient outcome prediction, outpatient disease risk prediction, and more advanced applications such as treatment recommendation, clinical trial matching, and even the generation of molecules for drug discovery [82].

Medical research relies heavily on image analysis. It is essential that both the detection and analytical processes rely on sufficient evidence derived from the healthcare domain in order to facilitate precise diagnoses or predictions based on images. Among the wide range of medical imaging techniques available, Magnetic Resonance Imaging (MRI), Ultrasound (US), Computer Tomography (CT), and X-rays stand out as the most commonly used. Considering that Deep Learning methodologies are capable of seamlessly adjusting to variations in image scale and resolution, medical practitioners must select the most appropriate imaging technology and technique with care. Taking into account the distinct characteristics of each technology and technique. In order to ensure that the selected imaging modality is aligned with the clinical requirements and provides accurate diagnostic information, it is important to consider each of these factors. [78]. Furthermore, imaging techniques have the distinct advantage of requiring a low level of invasiveness. The purpose of this is to ensure that the bodily integrity of the patient is not compromised while also facilitating the acquisition of

high-quality therapeutic and detection information. Non-invasive imaging of various organs, such as the brain, heart, chest, lung, kidney, and liver, can provide detailed insight into lesions and the anatomical details of various organs. As a result of their precision, these techniques are beneficial across a wide range of medical conditions.

The use of deep learning models has become increasingly prevalent in the last decade as a method to diagnose diseases based on the analysis of diverse medical images such as MRI, US, CT, and X-rays [72].

The field of medical imaging has been extensively explored in terms of various classification and extraction methods based on the inherent characteristics of the datasets being analyzed as well as their potential categorization. It is important to note that the research in this area utilizes a variety of methods. Other researchers conducted a comprehensive analysis of datasets that include every possible type of lung disease, a methodological approach referred to as 'multilabel classification'. There are also some studies that are specifically focused on detecting specific diseases, for example, studies that are designed to detect pneumonia and tuberculosis. Furthermore, a subset of these studies investigated the impact of varying parameters and datasets on classification results.

As an example, Iqbal et al.[27] propose a method for the detection of indications of lung diseases using the National Institutes of Health (NIH) Chest X-Ray dataset,as well as the COVID-19 X-ray images. This method uses deep learning techniques to address the challenges posed by datasets with limited annotations and class imbalances in COVID-19. For image classification, the proposed methodology integrates a class-balancing algorithm and a Support Vector Machine (SVM). In addition, a parallel Convolutional Neural Network (CNN) model, called 'Hybrid Feature Fusion,' was introduced to allow for the consideration of multiple models. This study employed the Anisotropic filter based on Perona and Malik's equation to eliminate noise from medical datasets. The following three methods were applied to address class imbalance issues: the algorithmic level method, which modified the learner through a weighting and costing schema, the data level method, such as Random Oversampling and Undersampling, and the Hybridized method, which combines both algorithmic and data sampling approaches. The comprehensive approach used resulted in an F1 score of 0.968 and an Area Under the Curve(AUC) of 0.967 on average.

Furthermore, Sharma et al. [66] use deep learning with Visual Geometry Group 16 (VGG16)

to classify pneumonia dataset on Kaggle, balancing and normalization procedure was implemented during the data preprocessing phase. This technique was used to standardize data within a range of 0 to 255, which achieving resulted an F1 score of 0.954. While Liz et al.[41] worked on multilabel classification and presented a deep learning methodology for imbalanced, multilabel ChestX-ray, CheXpert, and PadChest datasets, using an ensemble of five state-of-the-art individual models and the ensembles such as Densenet201, EfficientNet, Inception, InceptionResnet, Xception and a novel explainable Artificial Intelligence (AI) technique based on heatmaps. Several strategies are employed to address class imbalance in datasets, including oversampling, data augmentation, histogram equalization, and generative adversarial networks (GANs), in addition to the generation of synthetic chest X-rays. The model's learning process was also adjusted by increasing minority class weights and applying weighted loss functions. On average, they achieved an AUC of 0.825.

In conclusion, this research broadly categorizes research findings into two overarching approaches in X-rays investigations: the binary distinction of Normal/Abnormal classes and the more specific multilabel Pathology Classes.

2.1 Normal / Abnormal Classes

In clinical practice, labeling considerations play the most important role. A particular point of interest is the 'No Finding' label. Based on the National Institutes of Health (NIH) Dataset, it is evident that a definitive label to define 'normal' or 'without disease' does not exist. Instead, the dataset is referred to as 'No Finding'. As a result of this particular terminology, questions are raised regarding whether or not it denotes a clear absence of disease or the inability to detect any disease.

In pursuit of clarifying this ambiguity, Wong et al. [81] proposed that radiologists delineate images of the ChestX-ray14 dataset that are free from pathological signs. This initiative aimed to enhance radiologists' precision in distinguishing between normal and abnormal images. It may be possible for radiologists to increase their confidence in detecting if the system flawlessly identifies merely the normal categories. As a result, the proposed solution deploys a deep neural network configured as a binary classifier. This classifier is systematically structured to impose heavy penalties for false positives identified within the normal categories. The image size was 128x128 and

normalized by using contrast limited adaptive histogram equalization to enhance contrast. Various image augmentation techniques, including rotation, shifting, and scaling, were used. Leveraging features derived from the ImageNet pre-trained Inception-ResNet-v2 model, the classifier demonstrates the capacity to correctly identify approximately 0.50 of normal images, defined as images that are disease-free.

In order to enhance the field of medical image diagnosis, Nahida et al. [46] propose a nuanced approach to diagnosing pediatric pneumonia based on chest X-ray images 224 × 224. The method involves the combination of two distinct CNN models, CheXNet and VGG-19, that have been meticulously fine-tuned. Using a 121-layer DenseNet architecture, the first system, CheXNet, incorporates pre-trained ImageNet weights and undergoes further refinement using chest X-ray images obtained at the Guangzhou Women and Children's Medical Center. They used a variety of image enhancement techniques on X-ray images, including adaptive histogram equalization (AHE) to enhance image contrast, and rescaling the images to a range of 0–1. Geometric translations, such as shearing, zooming, and flipping, were used for data augmentation. The training process for VGG-19, on the other hand, is based on the ImageNet dataset. Given the disparate image counts relating to pneumonia and general health within the dataset, techniques such as Random Under Sampler (RUS), Random Over Sampler (ROS), and Synthetic Minority Oversampling Technique (SMOTE) become instrumental in harmonizing feature distribution. Upon feature selection, algorithms including Random Forest (RF), Adaptive Boosting (AdaBoost), and K-Nearest Neighbors (KNN) are enlisted for precise pneumonia detection and differentiation from normal imagery. These computational studies have resulted in an impressive 0.989 prediction performance.

When it comes to severe lung ailments, pneumonia stands out as one of the most formidable diseases. For the detection of this disease, Rajasenbagam et al. [57] utilize the robust 16-layer VGG19Net. Tested on previously unseen Chest X-ray8 dataset. A Deep Convolutional Generic Adversarial Network (DCGAN) was used to generate the augmented images. In order to avoid duplicate images, a content-based image retrieval technique was used, their model has an impressive accuracy rate of 0.995.

In addition, Tuberculosis (TB), which is primarily a lung disease, poses substantial health risks. Ayaz et al. [9] develop a fully automated Computer-Aided Detection (CAD) system capable of

efficiently detecting TB using two public datasets, Montgomery and Shenzhen. In this system, pre-trained CNN architectures are combined with supervised learning techniques. Additionally, Ensemble Learning seamlessly merges the results of each classifier to produce a composite result. In order to extract features efficiently from images, they normalized, downscaled, and then applied the Gabor filter to the images. Image sizes were 300 x 300. Their reported results indicate an accuracy rate of 0.975.

2.2 Multilabel Classes

Chest X-ray interpretation requires a thorough understanding of anatomy and pathology. This complexity emphasizes the challenge of formulating an automated method that consistently interprets chest X-ray images across all thoracic disorders. As a result of this background, it is important to consider the complexity of making precise clinical decisions with regard to lung disease classification and detection from chest X-rays.

Expanding upon this concept, Rajpurkar et al. [58] developed an algorithm, CheXNet, which is a 121-layer convolutional neural network trained on ChestX-ray14 dataset, which comprises over 112,120 frontal view X-ray images encompassing 14 distinct disease classes, with an emphasis on lung disease detection. The mean and standard deviation are used to normalize the images.. Additionally, they enhance the training data using random horizontal mirroring and the size of the images was 224 x 224. The performance of their model achieved an AUC of 0.84.

In order to enhance the accuracy of computer-assisted diagnosis of chest X-rays Yao et al. [84] proposed 2D ConvNets as encoders and decoders on chest x-ray8 dataset with size of 256 × 256 which randomly translated, rotated and scaled without pre-training,and evaluated them. A Long-Short Term Memory Network (LSTM) approach is used to capitalize on the connectivity of target labels for predicting 14 pathologies using their formulas. With this method, multi-label classification is viewed as sequence forecasting using a consistent length decoder.

To enhance clarity in classification and reduce label ambiguity particularly for the classes 'No Finding' Irvin et al. [28] provided it as an automatically labeler using uncertainty labels from radiology reports by Convolutional Neural network(CNN) training, which outputs the probability

of 14 observations. As result, they developed CheXpert dataset (Chest eXpert), which is a large set of data for interpreting chest radiographs.

In their study, Rakshit et al. [59] use transfer learning, ResNet18 to classify lung disease with Chest X-ray14 dataset, which has been trained on ImageNet. The model described here has a shallow level of deep learning compared with the most advanced models. A shallow model is used because deep models require more time and resources to train due to many parameters. The image size was 224 × 224. Images were normalized based on the mean and standard deviation of the dataset. Image cropping and horizontal flipping were used to augment the data. Their model was enhanced slightly by using the TenCrop method during testing, which improved AUC result on average from 0.841 to 0.849.

According to Yuan et al. [88], the proposed (Deep Auc Maximization) DAM method on CheXpert dataset significantly improves cross-entropy loss optimization performance by a new margin-based measure of loss, resulting in an AUC of 0.930. There were five ensemble networks that they trained with: DenseNet121, DenseNet161, DensNet169, DensNet201, and Inception-renset-v2 with image size 320 × 320.

Additionally, Meedeniya et al. [44] and Çalli et al. [13] present a comprehensive analysis of chest X-rays using deep learning that provides detailed and insightful information based on different data sets, methods, and their performance.

Comparison of the methods present in literature. Table 2.1

Our research team has developed a new approach that utilizes Transfer Learning techniques. Specifically, we have used the ConvNeXt model, which was originally trained on the ImageNet dataset, in order to facilitate the classification of a variety of lung pathologies. By leveraging previously trained deep neural network models, Transfer Learning is capable of addressing novel problems, especially when there is a similarity or commonality between two domains. A methodology such as this is particularly advantageous in situations where there is a lack of large amounts of labeled data. We encourage readers to refer to the forthcoming chapters for a comprehensive understanding of the methodology, including details regarding the datasets used, the intricacies of the models, as well as the parameters used.

Author	AUROC (Average)	Method	Dataset	Dataset Size	Image Size
Normal/ Abnormal Classes					
Wong et al. [81]	0.90	Transfer Learning / Inception-ResNet-v2	ChestX-ray 14	112,120	128 ×128
Rajasenbagam et al. [57]	0.99	VGG19Net	ChestX-ray 14	112120	—
Nahida et al. [46]	0.99	DenseNet-121,VGG19Net	Guangzhou	5232	224 × 224
Ayaz et al. [9]	0.98	Ensemble	Montgomery Shenzhen	138 , 662	300 × 300
Pathology Classes					
Yao et al. [84]	0.80	2D ConvNets	ChestX-ray 14	112120	256 × 256
Rajpurkar et al. [58]	0.84	DenseNet-121	ChestX-ray 14	112120	224 × 224
Rakshit et al. [59]	0.85	Transfer Learning / ResNet18	ChestX-ray 14	112120	224 × 224
Liz et al. [41]	Different results for different Methods	Individual models and ensembles	ChestX-ray14 CheXpert PadChest	112120 224316 160000	512 × 512
Yuan et al. [88]	0.930	Ensembles models	CheXpert	224316	320 × 320

Table 2.1: A comparison of the methods presented in the literature.

Chapter 3

Chest X-ray Datasets

Our experiments were conducted using a variety of datasets, which are discussed in this chapter. Specifically, it provides information on the number of samples included, the range of pathologies represented, as well as the specific research labs or hospitals from which these datasets were procured. The following is an overview of these datasets.

3.1 ChestX-ray14

ChestX-ray14 represents an advancement over ChestX-ray8, introduced by the National Institutes of Health (NIH). This dataset is derived from radiological reports by applying natural language processing (NLP) techniques to patient records collected between 1992 and 2015. It consists of a collection of 112,120 frontal chest X-ray images, formatted in PNG, each with an image size of 1024×1024. These images are sourced from 30,805 distinct patients. There are 14 diseases associated with the thorax, including Atelectasis, Consolidation, Infiltration, Pneumothorax, Edema, Emphysema, Fibrosis, Effusion, Pneumonia, Pleural Thickening, Cardiomegaly, Nodule, Mass, and Hernia. The term 'No finding' indicates that these diseases were not detected [5, 79].

3.2 CheXpert

CheXpert is a valuable public resource for chest radiography, containing radiographic examinations performed at Stanford Hospital between October 2002 and July 2017. These examinations span both inpatient and outpatient facilities, with the addition of corresponding radiology reports. The comprehensive dataset comprises 224,316 chest radiographic images, both frontal and lateral views, obtained from 65,240 distinct patients. Taking into account the dataset's inclusion of uncertainty labels: unmentioned instances are left blank, negative findings are labeled as '0', uncertain findings as '-1', and positive findings as '1'. These labels have been determined through automated, rule-based labeling of extracted observations from radiology reports. Each image within the CheXpert dataset is in JPEG format with a size of 256 x 256. A pathology label is assigned to each image, including Atelectasis, Cardiomegaly, Consolidation, Pneumonia, Pneumothorax, Enlarged Cardiomediastinum, Edema, Lung opacity, Lung Lesion, Pleural Effusion, Pleural other, Fracture, Support devices, and No findings. In accordance with the Fleischner Society's recommended glossary, these 14 classes have been selected based on the frequency of occurrence in the reports and clinical relevance [28].

3.3 MIMIC-CXR-JPG

Including 377,110 chest X-rays and 227,827 imaging studies, the MIMIC-CXR Dataset consists of a substantial collection of chest radiographic images. During the period of 2011 to 2016, these images were obtained from the Beth Israel Deaconess Medical Center (BIDMC) located in Boston, Massachusetts. This dataset was designed to facilitate research in automatic chest radiograph analysis. Radiology images in MIMIC-CXR are available in Digital Imaging and Communications in Medicine (DICOM) format, is the most common format used by clinicians. It is important to remember that DICOM images contain a lot of meta-data as well as the values of the pixels. In order to reduce the size, JPEG images derived from DICOM files are standardized by MIMIC-CXR-JPG to provide a standard reference with a size of 256 x 256. The images are also labeled using open-source labelers derived from free-text radiology reports [30, 54, 28].

3.4 PadChest

The PadChest(Pathology Detection in Chest radiographs) is the largest and most comprehensive public chest X-ray dataset with Spanish sources. The data set contains more than 160,000 images from 67,000 patients providing six different views of a position and extra information interpreted and reported by radiologists at the San Juan Hospital (Spain) between 2009 to 2017. Regardless of the language, the images are labeled using standard terminology from the Unified Medical Language System (UMLS). The patient's demographics, projection type, acquisition parameters for an imaging study, anatomical labels, as well as extensive information regarding the imaging study. An annotation strategy based on deep neural networks and label resolution rules was developed to extract labels from the remaining reports. This pipeline tackles some common NLP challenges for annotating large quantities of chest X-ray reports from the Spanish healthcare system. As a result, the dataset included 174 different radiographic finding labels, 19 pathological diagnoses, and 104 anatomical locations [11].

3.5 VinDr-CXR

The VinDr-CXR comprises both training and test sets labeled by radiologists, obtained from two major Vietnamese hospitals, Hanoi Medical University Hospital (HMUH) and Hospital 108 (H108), between 2018 and 2020. There are 18,000 labeled images released from 100,000 raw chest X-rays, including 22 local labels that indicate abnormalities and 6 global labels of possible diseases, all manually annotated by 17 qualified radiologists. The dataset contains 15,000 training sets and 3,000 test sets. Each scan in the training set was independently labeled by three radiologists, while each scan in the test set was individually labeled by five radiologists. Images are available in DICOM format with size 2788×2446, along with the labels of both the training set and test sets [51].

3.6 Private Dataset

A collection of diagnostic imaging data was collected by the Répertoires d'imageage diagnostic (RID) between the years of 2018 and 2021. The datasets are derived from two prominent medical

institutions in Montreal: Notre-Dame Hospital and Verdun Hospital. The collection comprises 377,000 radiographic images, which have been annotated via a Natural Language Processing (NLP) algorithm developed within our research laboratory. The corresponding radiological reports, written by expert radiologists, are predominantly in the French language.

In terms of orientation, approximately 0.2 of these radiographic images are lateral views, while 0.8 are frontal views. Initially available in the DICOM format, these images were subsequently converted to the JPEG format to facilitate their integration as model inputs. For training purposes, the image size was standardized to 244 x 244. Furthermore, each image has been systematically labeled to indicate one of 27 distinct thoracic diseases.

It's imperative to note that to ethical considerations and to protect patient confidentiality, any individually identifying information contained within the dataset has been eliminated or replaced with arbitrary values.

Table 3.1 is an overview of datasets.

Dataset	Research Lab or Hospital	Release Year	Pathology	Samples	Labeled by
ChestX-ray8	NIH	2017	8	108,948	NLP algorithm
ChestX-ray14	NIH	2017	14	112,120	NLP algorithm
CheXpert	Stanford Hospital	2019	14	224,316	NLP algorithm
Padchest	Hospital San Juan Hospital (Spain)	2019	19	160,868	NLP algorithm / Radiologist
MIMIC-CXR-JPG	Beth Israel Deaconess Medical Center (BIDMC) in Boston, MA	2019	14	377,110	NLP algorithm
VinDr-CXR	Vietnam's Hospitals	2020	28	18,000	Radiologist
Private Dataset	Notre-Dame and Verdun Hospitals in Montreal	-	27	377,000	NLP algorithm

Table 3.1: A comprehensive overview of datasets discussed in this chapter.

Chapter 4

Image Processing, Feature Extraction, and Deep Learning Algorithms

In recent years, deep learning algorithms, a subset of machine learning, have revolutionized the world of computing capacity and pattern recognition. Their ability to automatically generate hierarchical representations from raw data makes them particularly valuable in the field of image processing [38]. With the use of deep learning, neural networks consisting of multiple layers have been proven to be effective in the analysis of large datasets in order to assist in the modeling of complex concepts [71]. Moreover, image processing includes a variety of techniques designed to enhance, analyze, and extract valuable information from visual data, serving as the foundation for the subsequent application of deep learning algorithms. There is a mutual relationship between image processing techniques and deep learning algorithms, whereby image processing techniques preprocess and refine visual data so that it can be analyzed by advanced analytics algorithms [37]. The enhanced images are then fed into deep learning models, which are capable of learning hierarchical features and making accurate predictions or classifications. As a result of this foundational understanding, the following sections provide an overview of image processing, feature extraction, various algorithms, deep learning models overview, and the functions of these processes.

4.1 Image Processing

As humans, we are inherently capable of interpreting complex visual cues. However, it is very challenging to provide machines with the capability of analyzing visual content derived from images and videos [1]. A digital image consists of a matrix of data points or picture elements, commonly referred to as pixels, stored as numerical arrays in computing systems [52]. For identification purposes, it is crucial to evaluate distinct features, such as points, edges, textures, or objects by analyzing these pixel values and their distributions. For the extraction of features from medical datasets characterized by complex features, deep convolutional neural networks are widely utilized [29]. Consequently, the following section discusses the details of this feature extraction technique followed by deep learning algorithms.

4.2 Feature Extraction

Machine learning problems often involve a large number of features or variables, all of which are necessary for the establishment of a final prediction. An increasing number of features, or dimensionality, further complicates the task of processing the training dataset. This complexity makes dimensionality reduction algorithms essential, especially when the features exhibit redundancy or correlation.

The features of image recognition systems can be broadly classified into two categories: global and local. Local features offer a more detailed perspective by focusing on discrete regions or patches within an image, while global features provide a holistic description of an image by representing whole objects with a single vector [63] . Therefore, when using local features, an image may have a variable number of feature vectors. By focusing on the most significant features, this method does not only reduce computational resources but also aids in the elimination of irrelevant data. In this way, machine learning can be performed much more quickly while providing accurate and original descriptions of the data set during the process of learning and generalization [32]. This context underscores the importance of feature extraction as a powerful tool. There will be further discussion of the intricacies of these global and local features in the following sections.

4.2.1 Global Features

In most object recognition systems, global features describe the entire image. High-dimensional feature spaces allow images with these features to be represented as points. However, global features can be affected by clutter and occlusion. This category includes most shape and texture descriptors. Invariant Moments and Histogram Oriented Gradients (HOGs) are among the global features extracted.

- **Invariant Moments**

  Shape recognition uses invariant moments primarily to extract global features from images as a robust tool. As a result of their inherent invariant characteristics, these characteristics remain unchanged irrespective of transformations such as translation, rotation, scaling, and reflection. As a result of this distinct property, they are a prominent choice for obtaining shape attributes from binary images. The invariant moments provide a quantitative measure of the pixel distribution relative to the centroid of the character, thereby providing comprehensive information about the object's overall shape [60]. Different types of invariant moments have been developed and introduced to the scholarly community since 1962. In the field of image recognition and classification, these moments and their associated invariants have proven valuable. Application areas involving image analysis, including facial recognition systems, fingerprint recognition systems, and palm print recognition systems, have taken advantage of these moments. Further, such systems demonstrate the capability and importance of invariant moments in a wide range of fields, such as personal identification and medical imaging [48].

- **Histogram Oriented Gradients (HOG)**

  In the fields of computer vision and image processing, the Histogram of Oriented Gradients (HOG) represents a pivotal feature extraction technique. First introduced by Dalal and Triggs in 2005 [16], the HOG algorithm captures the local gradient information inherent in images. For the purpose of calculating gradient orientations, the image has been systematically segmented into smaller cells. Following this segmentation, each discrete cell's gradient orientation is captured in a histogram. The method involves grouping these histograms into distinct orientation bins, thereby providing a structured representation of gradient orientation.

Using this binning process, gradient orientations are separated into discrete categories that facilitate subsequent analysis and feature extraction.

4.2.2 Local Features

To determine local features, multiple interest points are computed within specific local neighborhoods of an image. This approach, however, poses a number of challenges, such as the variation in the number of features across different images, which can make it difficult to compare the images directly, particularly when considering local features. The extraction of local features can be accomplished in a variety of ways, including Scale-Invariant Feature Transforms (SIFTs), and Speeded-Up Robust Features (SURFs) [3].

- **Scale-Invariant Feature Transform (SIFT)**

  A Scale-Invariant Feature Transform (SIFT) is a technique in computer vision, designed to identify and extract key points within an image, which represent local features. Their invariance to both scale and rotation makes them particularly suitable for tasks such as image matching, object detection, and scene identification. In contrast to other features, such as edge and HOG, SIFT features are not affected by changes in the size and orientation of the image. An object can be identified within a novel image by comparing its characteristics to a pre-existing database using the SIFT algorithm. The comparison is based on the calculation of the Euclidean distance between the respective feature vectors.

  The initial step in this process involves the use of a collection of reference images in order to identify SIFT key points of relevant objects. The optimal location of these key points in both the spatial and frequency domains results in enhanced resistance to disruptions resulting from obstructions, clutter, or noise. A number of features can be extracted from a typical image using efficient algorithms. A further advantage of these features is that they are inherently unique, facilitating their accurate matching against a wide range of features, thereby increasing the likelihood of accurate object and scene recognition. By applying appropriate filtering,

it is possible to identify clusters of key points that correspond to the actual location, scale, and orientation of the object in the image[42].

- **Speeded-Up Robust Features (SURF)**

  It is considered one of the most powerful and expedient algorithms within the field of computer vision, designed to detect and describe local features as well as facilitate similarity-invariant representations and comparisons of images. As described by Bay et al. [10], the SURF algorithm can be divided into two sequential stages: Feature Extraction and Feature Description. SURF can be applied to a wide variety of computer vision tasks, including object detection, image identification, image classification, and 3D image reconstruction. The SURF descriptor includes elements from the Scale-Invariant Feature Transform (SIFT) technique, which makes it useful for real-time applications, such as object tracking.

  In general, input data is interpreted by algorithms using global and local feature extraction as the link between raw data and algorithmic capabilities. The capabilities of deep learning algorithms, particularly convolutional neural networks (CNNs), are inherent in their ability to extract features. For instance, in CNNs, the initial layers often capture low-level features like edges and textures, while deeper layers identify more complex patterns. Deep learning models learn and extract hierarchical features automatically from the input data through their layered architectures. In diverse applications, such as image recognition and natural language processing, deep learning models are known for their ability to independently detect and extract significant features from vast and complex datasets.

4.3 Deep Learning Algorithms

In deep learning algorithms, Artificial Neural Networks (ANNs) are utilized, inspired by the human brain's complex structure and function. Having the ability to learn autonomously from vast datasets, these algorithms are extremely useful in addressing complex problems. As a result of its architecture, Conventional Neural Networks are designed to resemble the computational mechanisms of the brain, enabling them to perform pattern recognition in a

remarkable manner. Their inherent ability to identify underlying patterns in data facilitates the extraction of features efficiently. In deep learning, feature extraction plays a key role in obtaining essential information from raw data [35]. There are several algorithms used in deep learning models and a variety of algorithms suited to different types of tasks. In spite of this, none of these algorithms can be considered perfect, since some algorithms are better suited to accomplish specific tasks than others. Understanding the primary algorithms is necessary for selecting the appropriate one.

In supervised deep learning algorithms, annotated datasets are used throughout the training process. As a result of this approach, the algorithm is trained to apply knowledge from the training dataset to new scenarios[49].

Among the various types of supervised models, recurrent neural networks (RNNs) are characterized by their ability to handle sequential data through their internal memory system as a result of their recurrent connections and hidden states [31]. RNNs are extremely useful for text translation and natural language processing applications.

In contrast, in unsupervised learning, machines operate based on unlabeled data to enhance algorithm performance by making data-driven decisions. Among the unsupervised models, Generative Adversarial Networks (GANs) stand out for their ability to generate realistic images, distinguishing between real patterns and fake ones.

Furthermore, in semi-supervised learning, the model is trained using both a large dataset of unlabeled data and a small dataset of labeled data. Based on supervised learning principles, the algorithm refines its model weights for the labeled data points. In addition, it attempts to minimize the difference between predictions derived from analogous training instances and approximates the overall distribution of the unlabeled data [76]. Among the semi-supervised learning models, the combination of Generative Adversarial Networks (GAN) and Convolutional Neural Networks (CNN) is of particular interest.

In image processing, deep learning models have demonstrated remarkable efficacy in distinguishing patterns, making predictions, and providing insights from a large array of data.

Computer vision, medical imaging, and multimedia analysis have all benefited from the integration of deep learning models in image processing [77]. The purpose of this overview is to provide readers with a brief overview of the various deep learning models used in image processing throughout the literature review.

In addition, our proposed ConvNeXt model, utilizing Vision Transformer methods, is an update of ResNet-50 has been developed. This narrative begins with a foundational understanding of Vision Transformers, followed by CNN models before proceeding into the details of ConvNeXt.

4.3.1 Vision Transformer (VIT)

In a Convolutional Neural Network (CNN), images are analyzed pixel by pixel. As the network identifies features such as corners and lines, it gradually builds up an understanding of the image from local features to a global one. Transformers, on the other hand, use a self-attention mechanism that facilitates the establishment of connections between distant locations of images from the very beginning of the information processing process.

While CNNs use arrays of pixels to represent images, the Vision Transformer (ViT) segments the image into visual tokens. In more detail, the Vision Transformer divides an image into patches, embeddings appropriate to each patch, and incorporates positional embeddings as input to the encoder. Using this approach demonstrates how CNNs and Transformers process spatial hierarchies differently, illustrating the different methodologies employed by each architecture in the recognition and analysis of images [17].

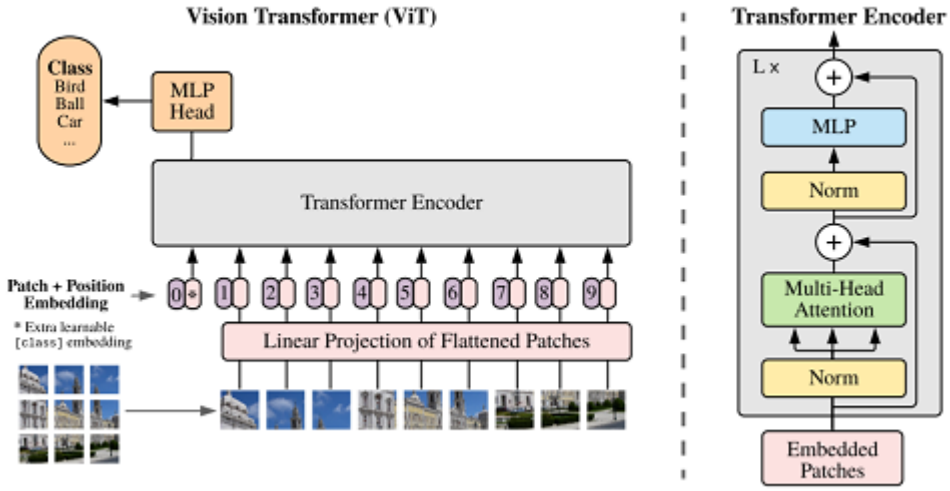

Figure 4.1: Vision Transformer architecture,[17]

4.3.2 Convolutional Neural Networks (CNNs)

Convolutional Neural Networks (CNN or ConvNet) are deep neural networks that compute the weighted sum of inputs using convolutional functions. As a result of its architecture, CNNs can learn spatial hierarchies of features automatically and adaptively through back-propagation [83]. In the field of computer vision, CNNs have gained significant recognition, proving to be useful for a variety of tasks. These networks employ a layered approach to discern features in images, integrating outputs from previous layers with training images across multiple resolutions. This hierarchical organization facilitates the transformation of basic features into more complex ones, eventually uncovering distinctive object characteristics. Activation functions and dropout layers also contribute to optimizing and regularizing the network, as illustrated in Figure 4.2. The following is a concise description:

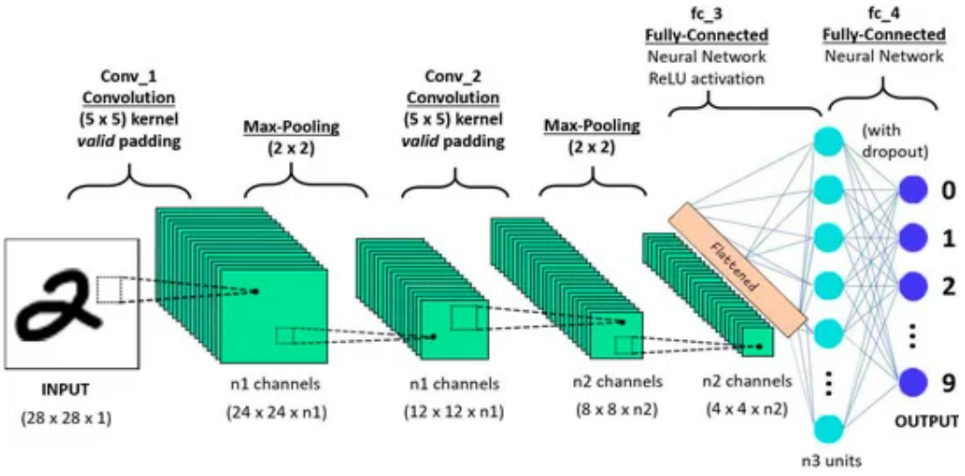

Figure 4.2: A Convolutional Neural Network (CNN) Architecture: Illustrating the layered structure [36].

- **Input Layer:** As the initial layer of the network, it primarily serves as an input layer, accepting images and forwarding them to later layers for processing.

- **Convolutional Layer:** A convolution layer is conceptualized as an image filtering mechanism based on a mathematical formula that calculates the weighted sum of multiple input values. As part of the architecture of a Convolutional Neural Network (CNN), this layer facilitates the generation of feature maps. Feature detectors or kernels may also be used to describe these maps, which are produced by convolving the input data with specific filters.

- **Pooling Layer:** As a result of applying the kernel to a designated segment of the image, the optimal pixel value can be extracted as a representative feature of the image. During this extraction process, parameters and dimensions are reduced, thereby preventing noise from entering into the dataset.

- **Dropout Layer :** During the training phase, dropout serves as a regularization technique, in which certain neurons are randomly deactivated. As a result of this intentional

deactivation, the selected neurons are excluded from both forward and backward propagation processes. As a result of this strategic omission, the risk of over-fitting within the model is reduced.

- **Activation Function :** As part of the network's predictive process, activation functions use threshold-based mathematical operations to determine the importance of a neuron's input. Several activation functions have been developed and widely utilized, including the Rectified Linear Unit (ReLU), the sigmoid, the softmax, and the hyperbolic tangent (Tanh).

- **Fully Connected Layers (FC Layers) :** In the Fully Connected (FC) layers, neurons are organized as layers in which each neuron maintains a comprehensive connection with all neurons in the adjacent layer. The architecture employs distinct weights to modulate connectivity, which in turn influences the dimensions and configurations of the subsequent output layers. The structure of basic CNN architecture is depicted in the corresponding Figure 4.3.

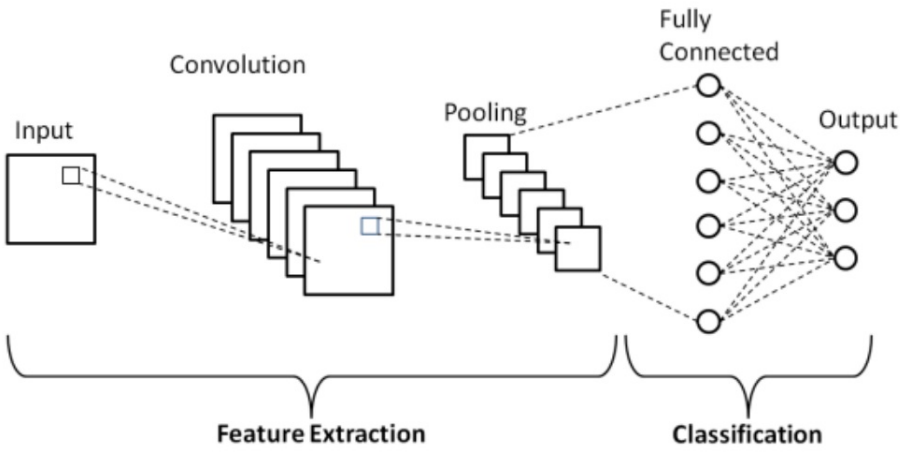

Figure 4.3: Basic Architecture of a Convolutional Neural Network (CNN): This schematic represents the fundamental components of a CNN [56].

DenseNet

A significant increase in complexity is evident in Convolutional Neural Networks (CNN), as evidenced by the numerous layers through which input information must be processed. The challenge arises when this information, when it reaches the deeper layers of the network, is at risk of disappearing. It is necessary to integrate diverse network topologies and training methodologies in order to address this issue. Through the use of dense blocks and transition layers, DenseNet refines this approach, resulting in a simplified but highly effective connectivity pattern.

There are inherent challenges associated with the training of deep networks, primarily due to the complexities associated with the flow of information and gradients. This complication is overcome by DenseNet by ensuring that each layer has direct access to the gradients derived from the loss function and the originating input image. A dense block consists of numerous layers, each exhibiting dense connections and consistent feature maps. As a result of this structure, all subsequent layers within a block receive inputs from all preceding layers, which in turn, contribute their own unique feature maps. A transition layer is implemented to pool features, thus connecting two adjacent dense blocks, in order to reduce the size of the feature map.

Due to DenseNet's densely connected architecture, the model effectively retains features from previous layers, eliminating the need to learn redundant features, which enhances its performance. DenseNet is composed of three dense blocks, each of which contains convolutional layers equipped with Batch Normalization (BN), ReLU activation functions, and convolutions for feature extraction. A transition layer consisting of BN, a convolutional layer of 1x1, and a pooling layer of 2x2 is appended after the initial three dense blocks. The purpose of these layers is to modify the channel dimension and downsample data by connecting two dense blocks together [25].

The architecture of DenseNet is concluded by a global average pooling layer, a fully connected layer, and a softmax activation function, strategically placed to determine the final output of the model. The design of DenseNet-121 is so intricate and comprehensive that it

comprises 121 convolutional and fully connected layers in total. Using this architecture, one can demonstrate the advancements in CNNs, as well as the innovative solutions developed to overcome the challenges associated with information vanishing and deep learning. The architecture of DenseNet is illustrated in the reference Figure 4.4.

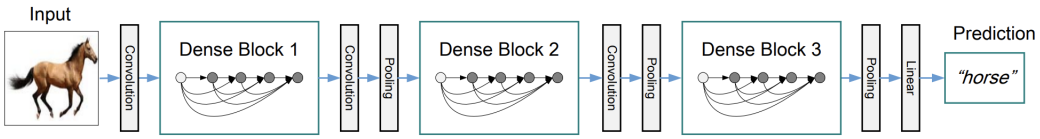

Figure 4.4: DenseNet Architecture Model : This figure shows the architecture consists of multiple dense blocks with layers that are interconnected [25].

EfficientNet

Google AI pioneered the introduction of EfficientNet, a methodology that enhances performance while maintaining high levels of efficiency. In general, Deep Learning models are characterized by either being overly broad, excessively deep, or displaying high resolution, thereby leading to overparameterization. Amplification of these parameters rapidly results in the model transitioning from a state of efficiency to one of overload, leading to an overabundance of parameters. EfficientNet achieves its enhancements by utilizing a compound coefficient, which facilitates the scalability of the model. According to Tan et al. [74], the width, depth, and resolution of the network were adjusted to ensure a harmonious and gradual enhancement across all dimensions.

EfficientNet is fundamentally focused on optimizing the efficiency of the system. Regardless of the type of network, the initial phase is characterized by the stem, followed by exploratory modifications to the architecture and the terminal layers. Each network consists of seven distinct blocks, which are encapsulated within varying subblocks, the quantity of which increases from EfficientNetB0 to EfficientNetB7. The EfficientNetB0 consists of 237 layers, while the EfficientNet-B7 consists of an impressive 813 layers. EfficientNet's design and systematic scalability demonstrate its commitment to balancing efficiency and performance in deep learning models. The architecture of EfficientNet is shown in the following Figure 4.5.

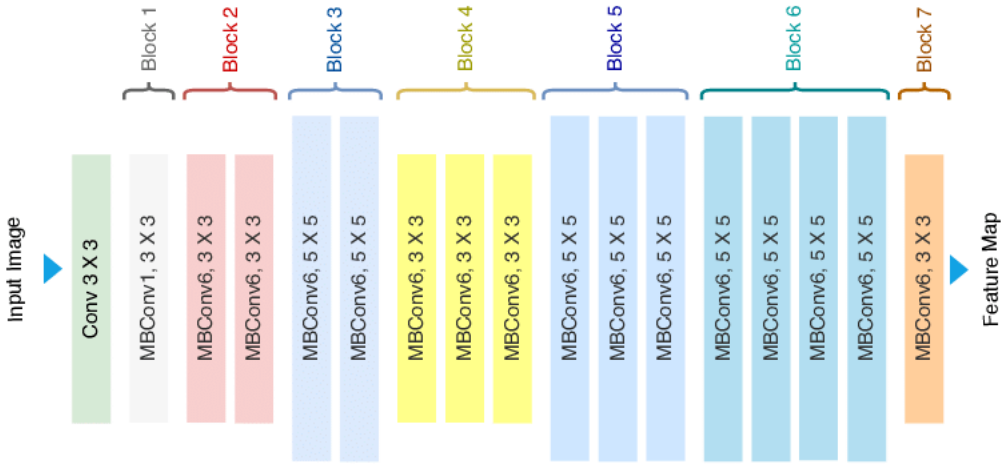

Figure 4.5: EfficientNet Architecture Model: This diagram highlights the model's progression from the input layer through a series of MBConv blocks [4].

VGG

The Visual Geometry Group, also known as VGGNet, provides a focal point for visual geometry analysis, utilizing a multi-layered architecture of deep Convolutional Neural Networks (CNNs). An object recognition model based on the VGG architecture has marked a paradigm shift in the field. A deep neural network was conceptualized with either 16 or 19 layers, denoted as VGG16 and VGG19, respectively, depending on the number of layers [85]. The VGG16 network encompasses approximately 138 million parameters. VGGNet is comprised of two fully connected layers, each containing 4096 channels, as well as a third layer with 1000 channels. A total of 1000 channels are covered by these three layers, yielding predictions across 1000 different labels [47]. A VGG-16 architecture is illustrated in Figure 4.6.

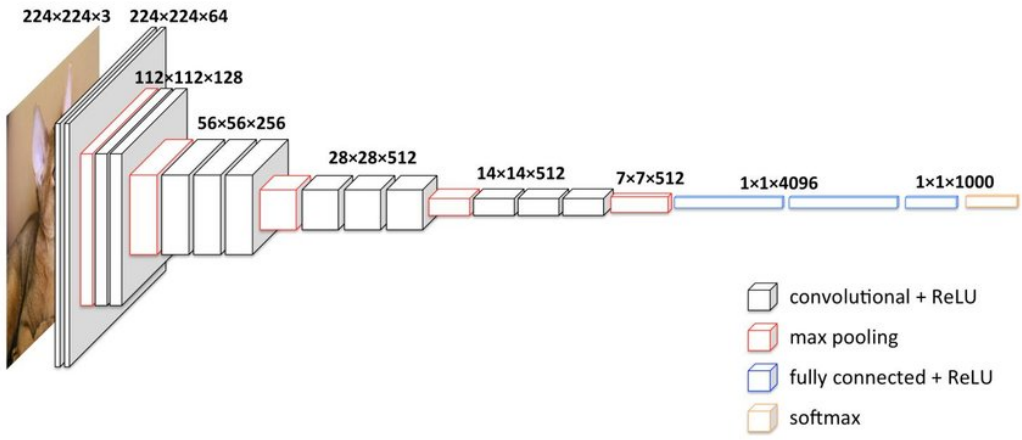

Figure 4.6: The illustration details the VGG-16 model with an input layer, followed by a series of convolutional layers. [47].

Inception

The GoogleLeNet algorithm utilizes a deeper and wider convolutional neural network, which results in a slight increase in evaluation time. Moreover, it uses Inception Modules to facilitate optimal sparse structures within localized regions of the image model. In contrast to a single filter size, these modules employ a variety of filter sizes, allowing the amalgamation of various filter types within one image block. Following this, the consolidated data is concatenated and transmitted to the following layer. As images within a single domain may exhibit substantial variations in the size of their key blocks, the choice of convolution kernel size is rendered extremely difficult due to enormous variations in information localization [86]. As a result of the distribution of information, the kernel size is determined by its size; a more global distribution requires a larger kernel, while a more local distribution requires a smaller kernel. The algorithm performs convolution on input by using three different filter sizes (1x1, 3x3, 5x5) [33] as shown in Figure 4.7. The resultant outputs are amalgamated and forwarded to the subsequent Inception Module [73].

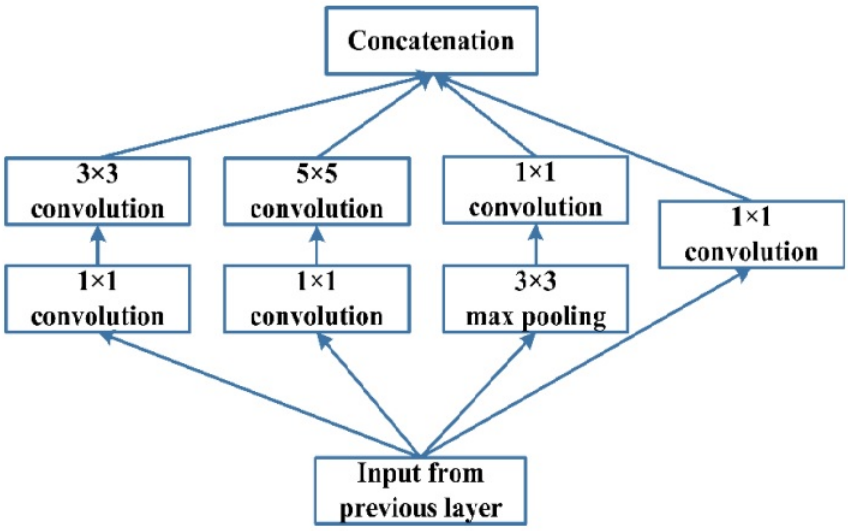

Figure 4.7: This diagram provides an overview of the basic Inception architecture. The input layer is followed by multiple Inception modules arranged in parallel. [33]

ResNet

One of the most influential and groundbreaking Deep Learning models is the Residual Network (ResNet). ResNets are characterized by what is known as a 'skip connection'. By skipping connections across layers and utilizing ReLU and batch normalization, ResNets provide enhanced performance due to their ability to maintain efficacy despite increased depth, provide efficient computations, and enhance network training capabilities [64].

As a result of these skip connections, ResNet is able to train networks consisting of hundreds or even thousands of layers while maintaining exemplary performance [20]. As a result of skip connections, deep neural networks are less susceptible to the vanishing gradient problem that is associated with them. Therefore, higher layers are able to learn feature attributes more adeptly than lower layers, matching their proficiency. By combining the outputs from additional layers, it is possible to train increasingly complex networks, thereby improving the ability of a residual block to learn identity functions [67].

In deep neural networks, ResNet exemplifies optimization by simultaneously minimizing errors and maximizing efficiency. Despite the differences in the number of layers, ResNet's underlying principle remains the same. As an example, ResNet-50 refers to a variant of ResNet that incorporates 50 layers. In contrast to the archetypal ResNet, which consists of 34 layers and uses 2-layer blocks, ResNet-50 uses 3-layer bottleneck blocks in order to enhance accuracy and reduce the duration of training [67]. A ResNet-50 architecture is illustrated in Figure 4.8.

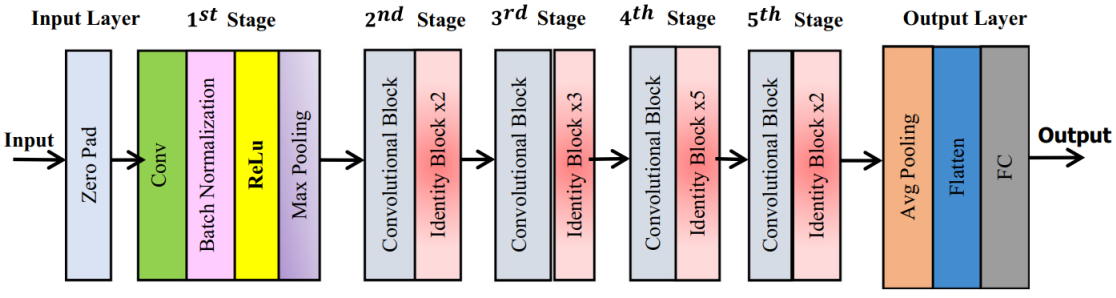

Figure 4.8: This illustration the ResNet-50 architecture, which consists of residuals and identities blocks.[67].

ConvNeXt

Vision Transformers, as well as pure convolutional models, are both inspirations for the ConvNeXt model. It combines previously established methodologies - specifically ConvNeXt and transformers - that have been simultaneously refined and restructured. On the ImageNet-1K dataset, ConvNeXt has demonstrated superior performance compared to the Swin Transformer, one of the most potent transformer models in existence, by meticulously adjusting training parameters over several epochs, utilizing the AdamW optimizer, modifying convolution kernel sizes, and applying robust data augmentation and regularization techniques.

ConvNeXt is characterized by the implementation of layer stacking strategies and internal modifications. Stage stacking numbers have been changed from 3:4:6:3 to 3:3:9:3. As part of ConvNeXt's architecture, the bottleneck is responsible for a variety of functions, ranging from feature extraction to feature reduction to subsequent dimension increase. There have also been significant modifications made, such as the enlargement of the convolution kernel from three to seven, the substitution of the ReLU activation function for GELU, and a change in the normalization method from batch to layer. As a result of these diligent explorations, ConvNeXt models have been gradually developed [40].

An overview of the differences between a ResNet, a Swin Transformer, and a ConvNeXt can be found in Figure 4.9, which provides a visual representation of the block designs characteristic of each. In Chapters 5 and 6, the ConvNext model will be discussed in further detail, with Chapter 5 exploring into its methodology and Chapter 6 focusing on the specifics of its implementation.

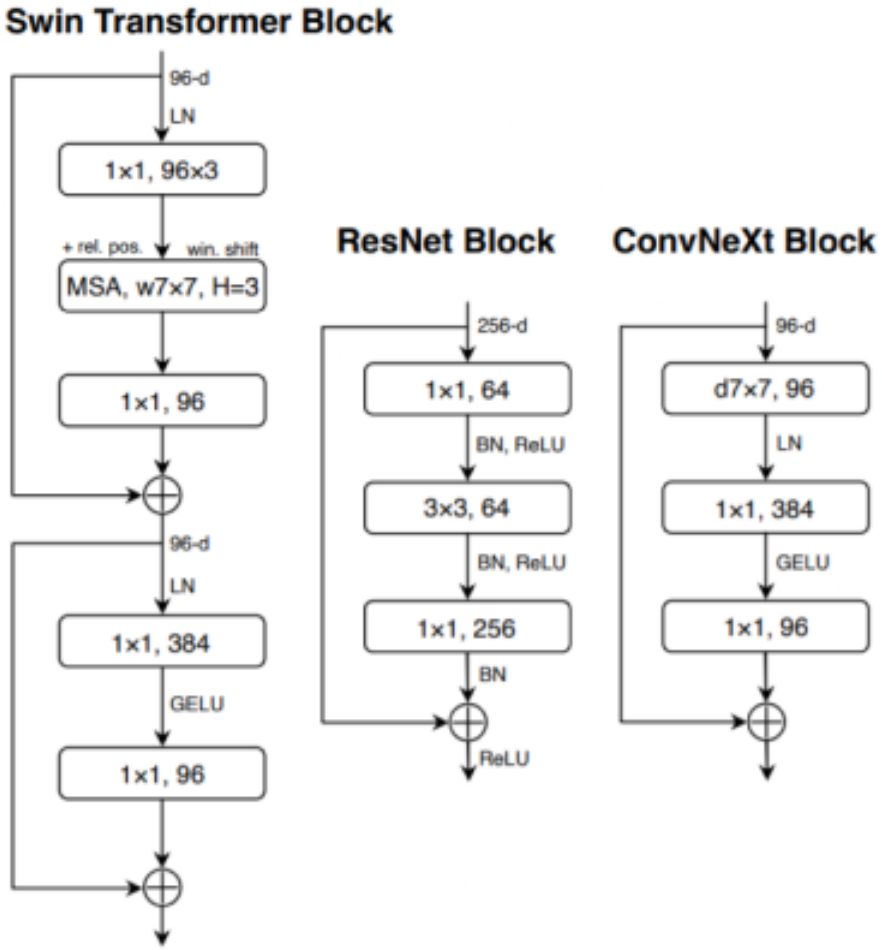

Figure 4.9: This figure displays the block structures of a ResNet, a Swin Transformer, and a ConvNeXt [40].

Chapter 5

Methodology

The use of chest radiography is an important diagnostic tool in modern medicine to determine the condition of a patient's thoracic area. Although this is the case, radiologists often experience difficulties in interpreting radiographic images as a result of factors such as overlapping chest tissues, lack of detail, and common symptoms of abnormalities. During the past few years, advances in computational methods have led to the development of deep learning algorithms that can identify features inherent within these images, thus reducing the difficulties described above. As a result, pulmonary diseases can be detected and also categorized more accurately. Deep learning has proven highly effective in image classification.

For the purpose of ensuring robust results, a systematic methodology was employed in this study. The process began with a comprehensive data collection phase, during which datasets were curated from a variety of reliable sources to ensure diversity and comprehensiveness. To optimize the dataset for training purposes, subsequent data preprocessing involved cleaning, normalization, and augmentation. To achieve optimal performance, an advanced deep learning model was evaluated and hyperparameters were tuned. Analyses of results were conducted to improve the models based on performance metrics.

An overview of the research design, data preprocessing, image preprocessing, the proposed model architecture, and performance metrics is provided in the following sections. A detailed implementation of all methods is provided in Chapter 6.

5.1 Research Design

One of the objectives of this research project is to develop and implement a computational model that is capable of interpreting chest X-ray images. The model is trained on several publicly available and private datasets. Data preprocessing, image enhancement, and image augmentation are utilized to ensure dataset integrity and robustness. This will be followed by training, evaluating, and testing the dataset. A specialized web application known as RADIA is used to automatically integrate the trained model following its training. Specifically, radiologists can use this tool as a helpful assistant in emergency situations. The entire research process is outlined into a comprehensive workflow to facilitate understanding. workflow begins with data pre-processing and enhancing a dataset using CLAHE before balancing it to ensure uniform class distribution. Training and validation sets are created, followed by various augmentation techniques to improve the ConvNeXt model's robustness. The process concludes after training the model and evaluating its performance. The flowchart provides a visual overview of the sequential steps in the research process as shown in Figure 5.1

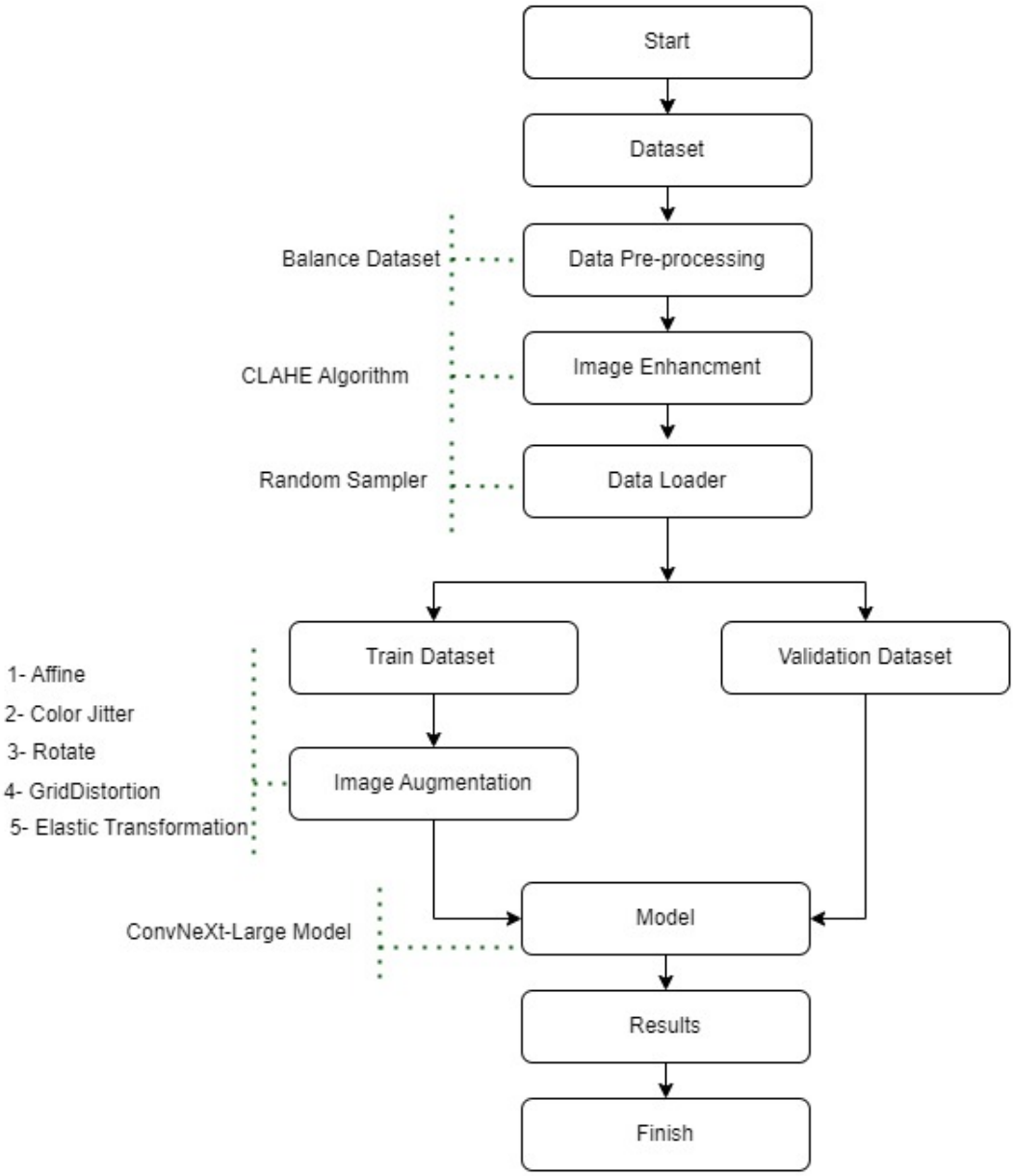

Figure 5.1: This figure provides a comprehensive visual representation of the process.

5.2 Data Preprocessing

The preprocessing of data is an essential component of deep learning research. A raw dataset may contain anomalies or inconsistencies that may adversely affect the performance of a model. To optimize model performance, it is essential to invest significant effort to thorough preprocessing and dataset preparation. It involves a series of steps designed to refine, select, and structure datasets to train deep learning architectures.

This study utilized several publicly available datasets that comprise chest X-ray images: ChestX-ray14, CheXpert, MIMIC-CXR, PadChest, and VinDr-CXR. As part of the preparation for model training, the team performed specific preprocessing steps.

For the purpose of analyzing the collected data, it is essential to understand the types of categorization that are suitable for training. There are different forms of this categorization. Binary Classification deals with two distinct classes, like spam and not spam. Multi-Class Classification involves more than two classes but ensures each item falls into only one class, for instance, identifying digits from 0 to 9. Multi-Label Classification allows for items to belong to several classes simultaneously, such as assigning multiple topics to a single blog post. Binary vectors are used in situations requiring multi-class and multi-label classification to determine whether a class label is present or absent. As part of our research, these vectors were particularly beneficial in capturing representations of multiple diseases.

Furthermore, Imbalanced Classification points to situations where items are unevenly distributed across classes. In the medical field, datasets often exhibit imbalances due to the varied distribution of diseases, from common to rare conditions [61]. In this research, the issue of imbalance within datasets has been addressed through various balancing strategies, including collecting more data, adjusting weights within classes, and using label smoothing techniques which are discussed in the following sections.

The following section provides a brief overview of dataset balancing techniques and binary vector encoding. Chapter 6 will provide a more comprehensive discussion and practical applications of these concepts.

5.2.1 Balanced Datasets

Balanced datasets are characterized by approximately equal representations of samples across all classes or categories. On the other hand, an imbalanced dataset is characterized by a marked difference in the distribution of classes. As an example, consider a dataset consisting of 100 observations divided into two classes: A and B. It is evident that there is a significant bias towards Class A in this sample because 90 instances belong to Class A, whereas only 10 instances belong to Class B. When there is such an imbalance in the dataset, models have a greater tendency to integrate insights from the overrepresented class (Class A in this example) over the underrepresented class. It is possible for this skewed learning to result in models overlooking the minority class, thus compromising the accuracy of future predictions. Strategies such as collecting more datasets, downsampling, adjusting class weights, and employing label smoothing techniques are essential for reducing such imbalances [61, 19]. To achieve a balance between datasets, our research team integrated multiple public datasets, enriched minority classes, and downsampled majority classes. Downsampling is a preferred technique for balancing datasets due to its simplicity and efficiency. To equalize the class distribution, samples from the majority class are randomly removed, which can be implemented quickly and without requiring complex algorithms.. As a regularization technique, we also used label smoothing to reduce the effects of imbalances, ensuring the model does not overestimate its predictions and increasing the number of generalizable features [45]. While deep learning training aims to minimize loss by optimizing parameters, imbalanced datasets may cause this to be skewed, potentially leading to the model overvaluing the majority class. However, a balanced dataset facilitates even learning across classes. The use of class weighting can also be beneficial in adjusting the loss function, penalizing the misclassification of the minority class more, and ensuring that all classes receive the same amount of attention.

5.2.2 Binary Vectors

The binary vector, also known as the binary code, is a representation of data that is only consisting of 0s and 1s. Categorical variables or features can be represented using vectors of this type. In

multiclass and multilabel classification scenarios, each distinct label is represented by a binary vector. For each input sample, the resultant output is a binary vector whose length equals the number of classes. It is important to note that each element in this vector represents the presence or absence of the associated class label. We have employed binary vectors to represent disease feature columns in practical applications, particularly when datasets contain multiple diseases, up to eight, in one column.

5.3 Image Preprocessing

Image preprocessing is a fundamental step in the field of image processing and computer vision, with the objective of enhancing an image's quality and characteristics to make it more suitable for a particular application or further analysis. The purpose of image preprocessing is to optimize raw image data before performing complex image processing tasks such as segmentation, feature extraction, and pattern recognition [68]. These techniques include image enhancement, normalization, and image augmentation. The following section provides a brief explanation of several of the terms discussed above.

5.3.1 Image Normalization

In order to ensure that pixel intensities correspond to a specified range, such as 0 to 1, normalization of data is an important step. Scaling these pixels to this defined range is a commonly used preprocessing step. In this process, the pixel intensities are converted from their inherent range, which is typically 0 to 255, to the desired normalized interval, which might be between 0 and 1 or -1 to 1. Since deep learning algorithms are highly sensitive to the scale of input features, scaling is critical. Pixel values that are incorrectly scaled may lead to the model struggling to determine optimal weights for the input features, resulting in insufficient performance. As a result, suitably scaled pixel values can enhance the effectiveness of the optimization phase, thereby accelerating the training process [70].

5.3.2 Image Enhancement

Enhancing an image is a key technique within the domain of image processing that aims to enhance the visual appearance and quality of the image. As a result of manipulating the image data, certain attributes, such as contrast, brightness, and sharpness, are enhanced. These modifications are intended to enhance the visual clarity of an image, facilitate its interpretation, and provide necessary information. Increasing contrast between tissues and organs in medical imaging, such as X-rays and CT scans, is crucial to the accurate diagnosis of disease. For the enhancement of images, CLAHE (Contrast Limited Adaptive Histogram Equalization) has been employed. This method is described in detail in Chapter 6.

5.3.3 Image Augmentation

Image augmentation is a pivotal technique used to enhance training datasets by generating multiple versions of existing images. The purpose of this approach is to provide models with a more comprehensive understanding of the potential variations a given image subject may display under a variety of environmental conditions in real-world. Among these modifications are rotations, brightness adjustments, and scaling changes, each presenting a distinct perspective on the original subject.

In addition to significantly enhancing the performance metrics of a model, image augmentation can also improve the accuracy of the model. During the training phase, the model was exposed to a wide array of variations on the same image that enhanced its ability to generalize across a variety of images. In addition to enhancing performance, such augmentation also reduces model errors. As a result of this extensive training, the model is able to distinguish features that remain consistent across a wide variety of images. The most widely used techniques used for image augmentation include cropping, zooming, flipping, skewness, translation, rotation, and shear for geometric invariance, as well as sharpening, embossing, and color jittering [43]. A detailed discussion of image enhancement techniques can be found in Chapter 6.

5.4 Proposed Model Architecture

In a comprehensive study, Lui et al. [40] developed a modernization of the ResNet-50 model, by taking inspiration from Vision Transformer techniques, which are discussed in Chapter 4. Using the ImageNet-1K dataset for evaluation, they discovered that incorporating elements from Transformers could enhance the performance of traditional convolutional networks. The results of this comprehensive research led to the conceptualization of a novel model called ”ConvNeXt”. This research examined diverse design modifications, including macro-level modifications, the adoption of ResNeXt methodologies, the incorporation of an inverted bottleneck structure, kernel enlargement, and intricate micro-level design modifications. As a result of their modified training strategy, the number of epochs increased from 90 to 300, as well as the switch from the Adam to the AdamW optimizer. The method also incorporated advanced data augmentation techniques such as Mixup, Cutmix, RandAugment, and Random Erasing, and sophisticated regularization techniques such as Stochastic Depth and Label Smoothing. Furthermore, the study highlighted the importance of stage-specific ratio computations and the nuanced structure of stem cells within the model. Swin Transformers' modifications to this distribution provided insights into potential optimizations. Swin Transformers process data at a variety of resolutions, as demonstrated by the 1:1:3:1 stage ratio of Swin-T, whereas its larger version employs a 1:1:9:1 stage ratio. Accordingly, the team adapted ResNet-50's block distribution from its original (3, 4, 6, 3) configuration to a (3, 3, 9, 3) scheme. Furthermore, the study examined the initial image processing protocols; contrasting downsampling techniques of traditional ConvNets like ResNet with the ”patchify” methodology characteristic of Vision Transformers. Moreover, depthwise convolution was smoothly integrated into the revised design, which is a variant of grouped convolution. Based on the Transformer models, this depthwise convolution layer was further enhanced by adopting a 7x7 kernel size after experiments with other kernel sizes, including 3, 7, 9, and 11. At the micro-level, the research examined ConvNeXt activation functions, normalization protocols, and spatial downsampling methodologies, leading to pivotal changes, such as the reduction of GELU activations and the substitution of BatchNorm with LayerNorm. The ConvNeXt architecture is shown in Figure 5.2. Model implementation is discussed in more detail in Chapter 6.

Overview of the ConvNeXt Model Architecture :

Input: The model receives an image as an input, similar to other CNN architectures.

Number of Channels: The number of input channels defaults to 3.

patch size: Patch size to use in the patch embedding layer defaults to 4.

Number of stages: The number of stages in the model defaults to 4.

Hidden Sizes: Dimensionality (hidden size) at each stage defaults to [96, 192, 384, 768]).

Depths: Depth (number of blocks) for each stage defaults to [3, 3, 9, 3].

Hidden Activation: The non-linear activation function (function or string) in each block. If string, gelu, relu, selu and gelu-new are supported defaults to gelu.

Initializer Range: The standard deviation of the reduced normal initializer for initializing all weight matrices defaults to 0.02.

Layer Normalization: The epsilon used by the layer normalization layers defaults to 1e-12.

Layer Scale Initial Value: The initial value for the layer scale defaults to 1e-6.

Drop Path Rate: The drop rate for stochastic depth defaults to 0.0.

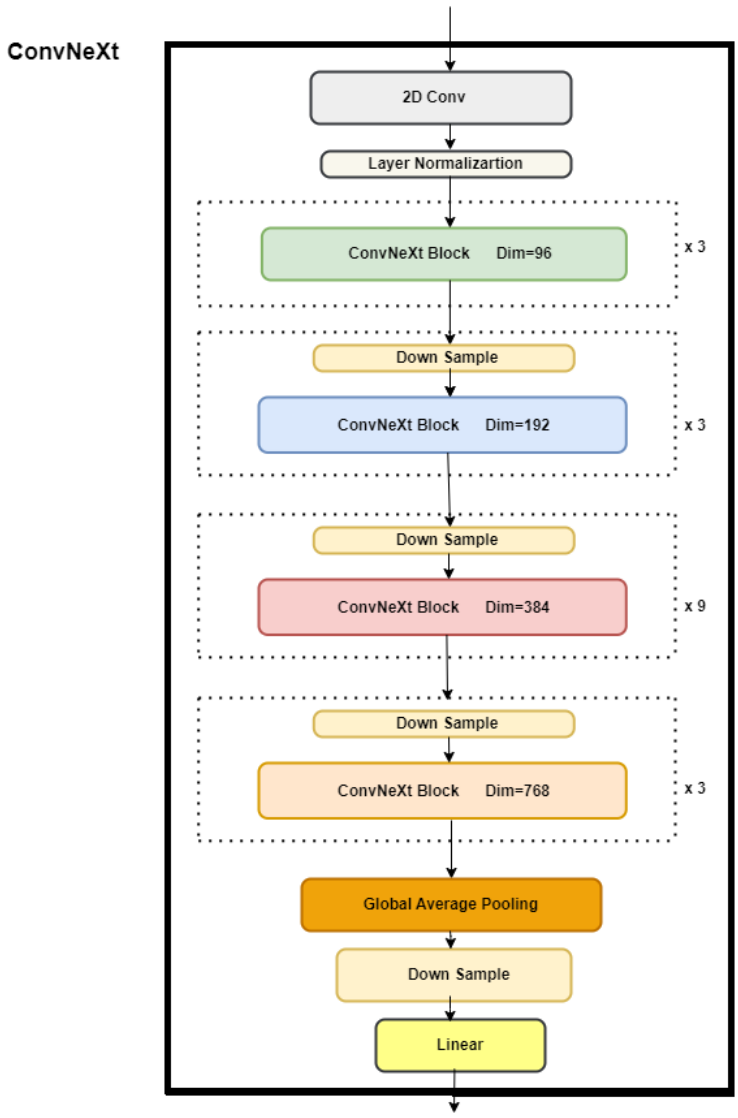

Figure 5.2: This figure illustrates the ConvNeXt network design, highlighting its improvements in depth, layer normalization, and expanded convolutional layers.

5.5 Transfer Learning

Transfer learning is a valuable approach in the field of machine learning. By employing this methodology, an established model that is initially trained for a specific task, referred to as ”task A”, is repurposed for a different but related task, referred to as ”task B”. For subsequent applications, the knowledge acquired in the initial task is efficiently transferred across different domains. The significance of transfer learning becomes increasingly evident when target training data is limited by its availability, substantial acquisition costs, or challenges with accessibility. Furthermore, the collection of comprehensive datasets is time-consuming and may even prove impossible in some real-world situations. As deep learning models require significant computational resources, transfer learning in deep learning has become increasingly popular because of the inherent challenges associated with training them. There are several benefits to implementing transfer learning, including a reduction in the training time, an increase in the effectiveness as a result of their ability to process information faster, or generalize better to unseen data, and a decrease in the need for extensive datasets [80, 6].

5.6 ImageNet Database for Pre-trained

With over 15 million labeled images, the ImageNet dataset plays an important role in the field of large-scale image processing. In terms of medical image analysis-based transfer learning, ImageNet has emerged as the benchmark dataset. In addition, pre-trained models—particularly those trained on large datasets such as ImageNet, which contains approximately 14 million images across 20,000 categories—have proven invaluable in the classification of images. A model can be complex and time-consuming to construct from scratch. The use of a pre-trained model, on the other hand, offers considerable benefits in terms of time and resource efficiency. The inherent features of such models can be effectively adapted to address new challenges as a result of extensive training. The research team decided to employ pre-trained ImageNet models as a foundation for the training procedures in this study as a result of these advantages [87, 26].

5.7 Methods and Experiments

This experimental framework presents a comprehensive method for developing and evaluating a model designed specifically for medical imaging, with a particular focus on radiology detection of lung abnormalities. The final goal is to simulate the real-world clinical environment, where the detection predictions made by the model are subject to verification by radiologists at Notre Dame and Verdun Hospitals in Montreal, thus ensuring the model's reliability.

As part of the construction of the dataset, a wide range of patient images encompassing both frontal and lateral views are included. As a result, various medical conditions can be analyzed more comprehensively. This dataset is representative of the population diversity and pathological diversity found in the real world, as provided by research laboratories and hospitals, as explained in detail in Chapter 3. Over 100 experiments have been conducted to optimize the model. These experiments range from adjusting parameters and hyperparameters to implementing varied training datasets. To understand the model's capabilities and limitations, each change is systematically recorded, with 12 of the most varied changes being described in more detail in Chapter 6. The model is trained and validated on diverse datasets to ensure it generalizes well and remains robust across varying statistical characteristics. Once the model has been thoroughly trained and validated, it undergoes a final examination against an independent test set annotated by radiologists. Try a test set that annotation by radiologists is more than just an evaluation of model quality; To bridge the gap between technological predictions and their practical application in medicine, this interaction is essential.

Chapter 6 provides a detailed of experiments and insight into the results. The following is a brief explanation of the methods.

5.7.1 Training and Validation Methodology

The experiments are structured around the application of labels corresponding to each exam at the output stage of the model. This process facilitates the accurate classification of results. To efficiently manage computational resources which ensures the system is not just limited to a specific set of images but is robust and versatile enough to handle the diverse range of medical images it will

face in everyday clinical use, thereby providing an unbiased evaluation of the model's generalizability and performance in varied scenarios. The integrated dataset is divided into smaller segments or batches. These batches are then sequentially loaded onto GPU memory, enabling dynamic adjustment of parameters after each training loop.

Following the training loop, a validation loop is initiated using a separate set of exams. This loop functions in parallel with the training process, with activation occurring at the end of each batch. Its primary purpose is to provide real-time assessments of both training and validation losses, thereby enabling continuous monitoring and adjustment of the model's performance.

5.7.2 Final Testing Phase

Concluding the training phases, which consist of multiple epochs, a test loop is initiated. This loop is critical for assessing the model's effectiveness and the robustness of the parameters. The system operates on a set of images distinct from those datasets annotated only by radiologists. This approach is designed to evaluate the system's performance in real-world scenarios, as opposed to controlled or experimental settings. Testing it on new, unseen data, allows us to gain a greater understanding of its ability to adjust to the variety and unpredictability of clinical environments. Such testing is crucial because it moves beyond theoretical accuracy, addressing the system's reliability and effectiveness in a practical situation, thus confirming its utility and potential for integration into routine medical practice.

5.7.3 Metric Evaluation and Parameter Measurement

A notable aspect of the experiment is the incorporation of specific metrics that facilitate the evaluation of parameter performance during a single training session. This setup allows for subsequent comparisons between various training sequences, each employing different parameter values. The tracking of metrics and losses is facilitated through storage on the "Weights and Biases" server, where they are readily accessible for review and display. The next section explains metrics in more detail.

5.8 Performance Metrics

The research team assessed the model's performance using three metrics: Area Under the Curve (AUC), F1 Score, and Geometric Mean (G-mean). All of these metrics have unique advantages and disadvantages and are frequently used to evaluate classification models [23]. Using a combination of these metrics was intended to provide a more comprehensive understanding of the model's performance.

In addition, we used G-mean as an alternative method for evaluation to capture a more balanced measure between sensitivity and specificity due to its robustness in handling imbalanced datasets. By incorporating G-mean, we aimed to provide a more comprehensive understanding of the model's performance, especially in situations where a balance between correctly identifying true cases and avoiding false identifications is crucial.

5.8.1 AUC

AUC plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at different threshold values. It measures a model's ability to distinguish between positive and negative classes. AUC ranges from 0 to 1, where the ideal classifier has an AUC score of 1, and a random classifier has an AUC score of 0.5. Higher AUC values indicate better performance.

5.8.2 F1 score

F1 score is the harmonic mean of precision and recall. It considers both precision and recall, which are the number of true positive predictions divided by the number of all positive predictions. In addition, it considers the number of true positive predictions divided by the number of actual positives, respectively. F1 scores range from 0 to 1, where an ideal classifier has an F1 score of 1. The formula for the F1 Score is presented below:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{1}$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{2}$$

$$F1Score = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

5.8.3 G-mean

G-mean is the square root of the product of the True Positive Rate (TPR) and True Negative Rate (TNR) of a model. It is a measure of model performance that balances sensitivity and specificity. By taking the square root of the product of TPR and TNR, G-mean provides a balanced measure of model performance across both positive and negative classes. A higher G-mean score indicates better overall model performance.

$$\text{TPR (Sensitivity)} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{4}$$

$$\text{TNR (Specificity)} = \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Positives}} \tag{5}$$

$$\text{G-mean} = \sqrt{\text{TPR} \times \text{TNR}} \tag{6}$$

5.9 Web Tool

The RADIA web tool is a software application currently being developed by the Research Group at CIUSSS (Centre intégré universitaire de santé et de services sociaux du Centre-Sud-de-l'Île-de-Montréal), intended to assist radiologists in detecting lung abnormalities with consultants from Notre-Dame and Verdun Hospitals in Montreal. The model receives inputs in the form of chest X-ray images from the hospital's database. Results from the deep learning analysis are then presented through a user-friendly Web User Interface. The interface displays an image viewer for the X-rays and a likelihood of disease panel.

Overall, RADIA is designed to facilitate collaborative decision-making and improve the quality

of patient care through a more comprehensive and representational analysis process. Chapter 6 provides a detailed explanation of the RADIA webtool.

Chapter 6

Implementations, Experiments, and Results

6.1 Implementations

In this section, we examine the steps that have been taken to develop our theoretical framework into practical. An overview of our research framework is presented, emphasizing the importance of data preprocessing and image processing in enhancing raw data for further feature extraction. Following these foundational steps, we move on to model implementation. Next, the discussion turns to an in-depth examination of the experiments conducted, the results obtained, and the subsequent analytical discussions. The final step in our research will be to reflect on the conclusions obtained from our research, address the limitations discovered along the way, and propose directions for future research to improve the results.

6.1.1 Data Preprocessing

There are several datasets that we chose to study, such as CheXpert, MimicCxrJpg, PadChest, ChexNet, and VinDr-CXR, each of which offers unique characteristics and challenges such as limited diversity, noisy annotations, class imbalance, and complex annotations. These datasets allow us to test how well our model handles different types of images, including images from varying sources and levels of complexity. Several of the datasets we worked with presented challenges related to

uncertain class labels, in addition to positive and negative classes. Uncertain class labels refer to instances where the presence or absence of a particular pathology in a chest X-ray image is ambiguous or difficult to determine with certainty. The presence of uncertain class labels underscores the real-world complexity of medical diagnosis based on chest X-rays. This reflects the challenges healthcare professionals or machines face in interpreting images exhibiting subtle or inconclusive signs of disease. Dealing with uncertain class labels added an extra layer of complexity to our AI research. It required us to develop specialized strategies and algorithms to handle these instances appropriately during training and evaluation. The calculation and process were simplified in this stage by eliminating uncertain classes.

Observation is crucial to analyze data. The unbalanced distribution of classes in datasets can be observed in a simple chart as seen in Figure 6.1. While the model can be trained better with balanced data, several solutions have been considered for balance datasets by our team, such as collecting more data and merging datasets, the class weighting method, label smoothing, and oversampling. This step ensures that each dataset class has sufficient data to avoid bias towards a specific class. This step is crucial to ensure that the model can learn from all classes fairly.

Collecting and Combining Datasets

For a comprehensive and unbiased analysis of data, it is essential to achieve a balanced distribution within datasets. As part of our efforts to achieve a balanced dataset, we enhanced our initial dataset by systematically sourcing samples of under-represented classes from various public data sources. Given this diversity, our team undertook a systematic approach to feature selection, which means choosing which attributes or columns from datasets are relevant for analysis, to ensure our integrated dataset was both comprehensive and relevant to our analytical objectives. The combination of data from diverse sources, however, introduces certain challenges such as column inconsistencies, representation differences, and variance in detail. The following describes the specific steps and challenges we experienced during our integration process, along with the solutions we implemented to overcome each challenge.

1: Column consistencies

As the first step, our team identified varied structuring of columns across datasets. For instance, some datasets might combine multiple entities or formats, like disease names into one column while others keep them separated into distinct columns. The solution involves taking a data harmonization approach to address inconsistencies in column structuring across datasets. Specifically, where a single column contains multiple entities, it should be split into separate columns to ensure that each entity is represented separately.

2: Standardization of labels

The second challenge relates to the varied ways data is represented across datasets. This can be in the form of numerical entries versus string data. Examples include how disease names or gender representations are presented. Some datasets use abbreviations while others use complete terminologies or numerical values. For example, numerical representations such as ”1” are used in one dataset, while alphabetic representations such as ”One” are used in another dataset. As well, gender may be represented differently in various datasets: some use full terms, such as ”female” or ”male”, while others refer to abbreviations or numerical codes, such as ”f”, ”m”, ”1”, or ”0” and the term ”TB” or ”Tuberculosis” to identify Tuberculosis diseases. When all these different representations are converted into a uniform format, it becomes easier to integrate, analyze, and interpret the data consistently across various sources and systems. To label diseases uniformly, our team utilized binary vectors. A binary vector, also known as a binary code or binary representation, represents data using a sequence of 0s and 1s. The binary vector is used in computer science and machine learning to represent categorical variables or features. Each label is represented as a binary vector in multi-class and multi-label classification. Based on this approach, each input sample is converted into a binary vector of length equal to the number of classes, where each element represents the presence or absence of the corresponding class label. Because some datasets contain eight distinct diseases in a single column separated by a specific character, our team applied binary vectors to disease feature columns following column consistency checks.

3: Eliminate unnecessary details

This relates to the depth and specificity of information in datasets. There can be notable differences in how detailed datasets are. It is crucial to ensure that when combining datasets with different levels of detail, the combined dataset maintains a coherent level of detail that is consistent with the analytical objectives. An integrated dataset can be more insightful, focused, and relevant by finding the right balance of level of detail, ensuring that it serves its intended purpose without becoming overloaded with unnecessary details or lacking essential information. This is as seen where some datasets might provide detailed Digital Imaging and Communications in Medicine (DICOM) specifics while others offer a more general overview.

4: Removes ambiguity

Using the CheXpert dataset, conditions are classified according to three distinct labels: 0, 1, and -1. Labels '0' and '1' indicate the absence and presence of specific conditions, respectively. On the other hand, the '-1' label indicates that the presence of the condition is uncertain or ambiguous. CheXpert's multi-labeling system distinguishes it from many other medical datasets that rely on binary labeling. To ensure compatibility and uniformity when integrating CheXpert with other datasets, we removed all instances labeled as '-1'. Our goal was to harmonize the labeling system across all datasets, thus providing a more straightforward and consistent method for analyzing data. Furthermore, disease classification and terminology can be inconsistent in various datasets, resulting in potential ambiguities in data interpretation, particularly if there is no clear explanation or standardization. When a dataset lacks a clear distinction between "Edema" and "Pulmonary Edema", it remains unclear whether "Edema" encompasses all forms or is exclusive of "Edema." Another example is the amalgamation of terms such as "nodule" and "mass" into one class, while other datasets use these terms to differentiate between two different classes. For accurate analysis, we consulted domain experts due to the lack of uniformity in these classifications across datasets, which resulted in the creation of separate or integrated classes.

5: Evaluation of disease classes

A list of all disease classes was extracted and compiled for each dataset. Once compiled, identify overlaps by comparing these lists. The intersecting set of disease classes will represent the common classes present across all datasets. By doing so, we intend to identify disease classes that appear in every list, indicating their presence across all datasets. The outcome consists of the same disease classes consistently recognized and categorized across all datasets. Based on the collection of datasets used by our team, 57 distinct disease classes have been identified. 34 of these classes are unique, meaning they appear in only one dataset and are not shared with any other. Furthermore, 16 classes appear in several datasets, but not all, indicating a semi-common occurrence. Finally, 7 of these classes are universally present, and found across every dataset we analyzed. These classes are shown in Table 6.1. It resulted in a diversity of resources, a sufficient volume of data, and consistency across all datasets.

Unique Classes Across All Datasets	Common Classes Present Across Some Datasets	Common Classes Across All Datasets
Calcification	Enlarged Cardiomediastinum	Atelectasis
Clavicle fracture	Pleural Effusion	Cardiomegaly
COPD	Emphysema	Consolidation
Enlarged PA	Lung Lesion	Edema
Interstitial Lung Disease	Pleural Other	No Finding
Lung Cavity	Support Devices	Pneumonia
Lung Cyst	Fracture	Pneumothorax
Lung Tumor	Emphysema	
Mediastinal shift	Infiltration	
Other Diseases	Mass	
Other Lesion	Nodule	
Pulmonary Fibrosis	Pleural Thickening	
Rib Fracture	Effusion	
Tuberculosis	Fibrosis	
Pleural	Hernia	
Nasogastric Tube	Lung Opacity	
Endotracheal Tube		
Central Venous Catheter		
COPD signs		
Asbestosis Signs		
Atypical Pneumonia		
Bone Metastasis		
Heart Insufficiency		
Lepidic Adenocarcinoma		
Lung Metastasis		
Lymphangitis Carcinomatosa		
Post Radiotherapy Changes		
Pulmonary Artery Hypertension		
Pulmonary Edema		
Pulmonary Fibrosis		
Pulmonary Hypertension		
Respiratory Distress		
Tuberculosis		
Tuberculosis Sequelae		

Table 6.1: This table summarizes the unique and shared classes across multiple datasets.

Balanced Datasets

Observation is essential to data analysis. A significant challenge faced in this process is the unequal distribution of classes within datasets. This concept is demonstrated in Figure 6.1. For the

training of predictive models to be effective, an appropriate balance of data distribution is essential. To address the imbalance, our research team has proposed a comprehensive approach that includes strategies such as data collection as discussed. It also includes the implementation of downsampling, class weighting techniques, and, label smoothing. Following is a description of the methods has been used.

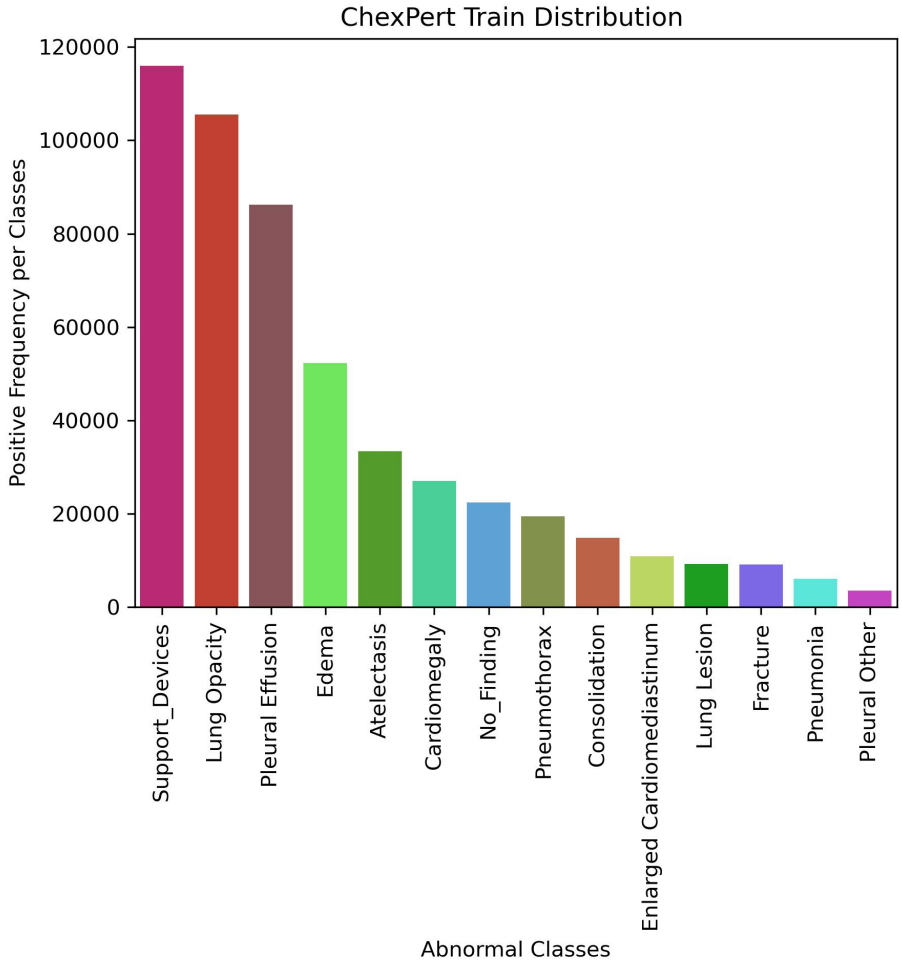

Figure 6.1: This chart demonstrates an unbalanced data distribution in the ChexPert training dataset.

- **Downsampling**

  The downsampling technique equalizes the distribution of classes within a dataset by reducing the number of instances of overrepresented classes. In our research, this method was applied

specifically to the ”No-Finding” class by removing samples randomly, which was found to be significantly larger than other classes. Our purpose was to achieve a more balanced dataset by specifically downsampling this class. This would prevent the model from being negatively affected by the dominant ”no-finding” class. This approach ensures that the model is not biased towards this class during training and provides a more comprehensive representation of all classes during training. In Figure 6.2, the distribution of data before and after the implementation of downsampling is illustrated.

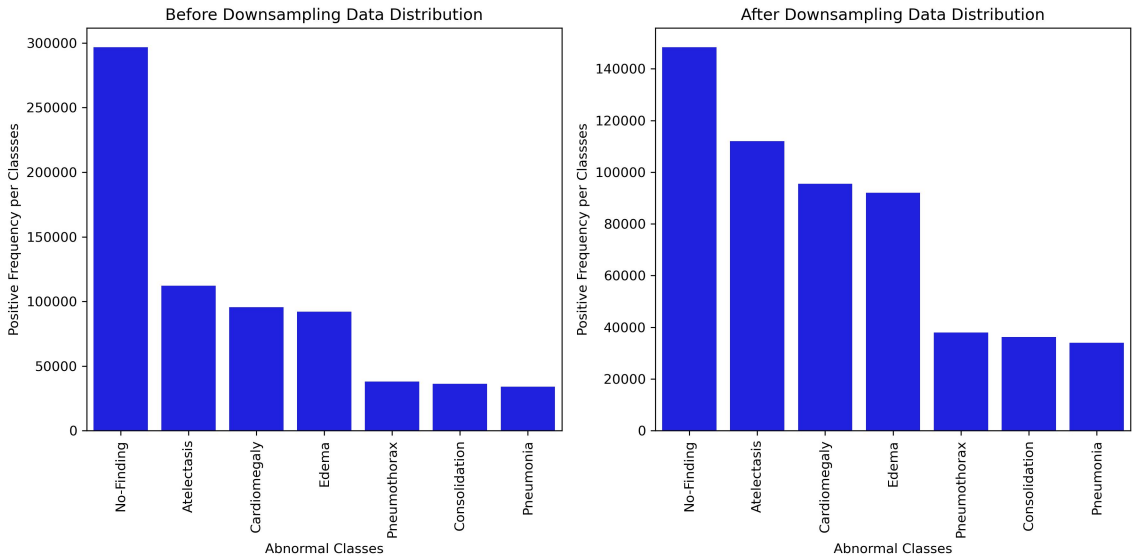

Figure 6.2: These figures compare the class frequency distributions before and after downsampling techniques.

- **Label Smoothing**

  Label smoothing is a regularization technique employed in deep learning to improve model generalization. The purpose of label smoothing is to prevent the model from being overconfident about its predictions and encourage it to learn more robust and generalizable features. In a balanced dataset, where each class has an equal number of examples, label smoothing may not be necessary. However, in an imbalanced dataset, where some classes have significantly fewer examples than others, label smoothing can help to reduce the effect of class imbalance by assigning a small probability to the class labels. This means that instead of assigning a fixed 0 or 1 label to each class, label smoothing assigns a probability distribution between 0

and 1 to the label for all classes. This is where the probability of the positive class is slightly higher and the probability of negative classes are slightly lower. This means that the model is less confident in its prediction and more likely to consider other classes. As a result, the model is encouraged to learn more general features that are not specific to one class but shared across multiple classes, which can help to reduce overfitting and improve the model's generalization ability.

- **Class-weighted**

Deep learning aims to learn the most efficient set of parameters that accurately predict output for a given input by minimizing the loss function. During training, the model evaluates each sample equally, calculates the loss function for each input, and determines the average loss for the entire dataset. An increase in dataset size can result in a smaller average loss, which is beneficial. By using a large and balanced dataset, the model is exposed to an equal number of samples from each class, allowing it to learn equally from each class. A model may learn to predict a majority class more accurately than a minority class in an imbalanced dataset. The minority class may suffer a greater loss as a result. The model can focus more on the majority group and neglect the minority group. A balanced dataset helps reduce class imbalance's impact on the loss function. This prevents the model from overemphasizing the majority class and improves its ability to predict the minority class, resulting in better model performance.

The class weights are calculated as the inverse of the frequency of each class in the dataset. This means that the weight for each class is higher for the classes that are underrepresented in the dataset and lower for the overrepresented classes. The goal of using class weights is to give equal weight to each class during training, even if the number of samples in each class is imbalanced. This can lead to better performance in the underrepresented classes.

n-samples / (n-classes * np.bincount(y)) calculates class weights in a balanced dataset.

n-samples : refers to the total number of samples in the dataset.

n-classes : refers to the number of classes in the dataset.

np.bincount(y) : returns an array of size n-classes, where each element represents the number of occurrences of each class in the dataset.

6.1.2 Image Preprocessing

The preprocessing of images is an essential component of various image processing applications. Several techniques are applied to the image before it is subjected to the actual processing. The purpose of image preprocessing is to enhance features that will be relevant to subsequent processing steps, remove noise, and improve the quality of the image. Image preprocessing techniques include image augmentation, normalization and enhancement. In addition to rectifying any defects present in the image, these techniques can also improve its overall quality, which can lead to improved processing results. Described below are the techniques have been used in our study.

Image Normalization

The normalization of images is a fundamental step in the pre-processing of images in computer vision. It may be more beneficial to convert an image to grayscale when color differentiation is not necessary, since it simplifies the task by focusing on only one channel instead of the usual three. It is imperative to determine the mean and standard deviation of pixel values for each channel (e.g., Red, Green, and Blue in color images). By subtracting the mean and dividing by the standard deviation for each pixel value within the respective channels, normalization achieved. To achieve pixel values confined to a specific range, such as [0, 1], a subsequent scaling step applied, multiplying the normalized value by the range difference and adding the minimum desired value.

Image Enhancement

By utilizing the Contrast Limited Adaptive Histogram Equalization (CLAHE) method, the cumulative distribution function (CDF) is determined through histogram equalization. Through a mapping process, pixel intensity values are distributed more uniformly within an image, enhancing the contrast and revealing more intricate details. Unlike conventional histogram equalization, CLAHE restricts the enhancement of contrast to individual pixels. It is necessary to divide the

image into smaller segments called ”tiles,” and to perform the histogram equalization on each segment individually. As a result, the pixel values within these tiles are normalized and scaled within the interval [0, 1]. Depending on tile dimensions, a clip limit is applied to each tile, serving as a limit on the amount of contrast augmentation applied. When pixel intensity values exceed this threshold, they are considered outliers and are reallocated among diminished intensity values in the corresponding tile, thus constraining the maximum post-normalization pixel value. As a result of the employing of this clip limit, CLAHE ensures moderate contrast enhancement and prevents excessive noise amplification [53]. Additionally, CLAHE's ability to adapt to the unique contrast characteristics of each tile prevents over amplification in areas of high contrast, thus reducing the risk of oversaturation and detail loss.

In Figure 6.3, the image enhancement resulting from CLAHE is shown.

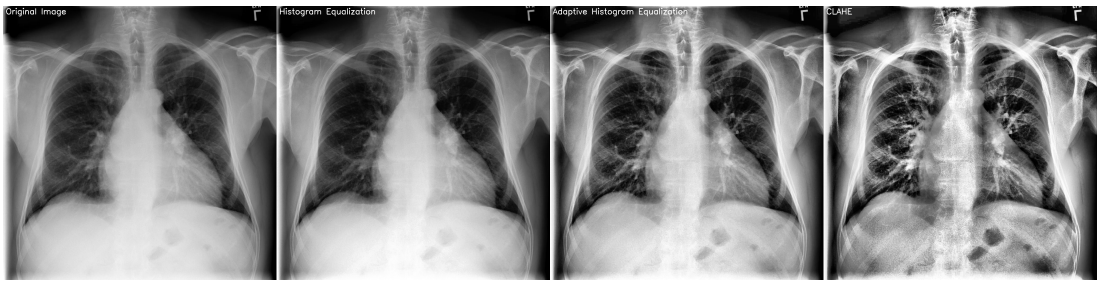

Figure 6.3: Left to Right: Original Image, Histogram Equalization, Adaptive Histogram Equalization, and CLAHE Algorithm

Image Augmentation

To improve the model's robustness and performance, a variety of image augmentation techniques are employed. Specifically, Elastic Transformation is used to simulate natural deformations in images, and ColorJitter is used to adjust brightness, and contrast, enabling the model to generalize to varying lighting conditions and color distributions. Moreover, Gaussian Noise is introduced to simulate sensor noise and other disturbances, and GridDistortion is applied to distort the input image, ensuring the model remains invariant to these changes. The following is a description of the image augmentation techniques we employed:

- **Affine Transformations**

  An affine transformation is a geometric transformation applied to an image. These include operations such as rotation, translation, scaling, shearing, and flipping. To increase the diversity of the training dataset, affine transformations change the position and orientation of objects within images. As a result, machine learning models are made more robust. When X-ray images are rotated randomly, the model can identify X-rays from different orientations, potentially improving its ability to classify real-world images that may not be perfectly oriented. The Affine Transformation has been set of 0.9 to 1.1 and a probability of 0.5 in our study . For example for scaling, images can either be reduced to 0.9 of their original size or increased to 1.1. Further, the probability factor of 0.5 indicates that there is a 0.5 chance that the scaling transformation will be applied to any given image, making it a random event.

- **GridDistortion**

  GridDistortion modifies an image by randomly shifting a predefined grid of points imposed upon it. After these grid points are distorted within a defined range, thin-plate spline interpolation is applied to the distorted grid. As a result, the image is warped, simulating various forms of distortion, including lens or perspective distortions. It is possible to modulate the magnitude and nature of the deformation by manipulating the magnitude and vector of the grid point displacements. The distortion value of this technique is 0.1, and the probability of its application is 0.5 in our study.

- **Gaussian Noise**

  Contrary to human cognition, algorithms frequently experience difficulties when it comes to ignoring noise. During image processing, noise is often introduced by purposefully altering pixel values, resulting in a deviation from the original image.

  In deep learning, an overfitted tends to incorporate inconsequential high-frequency features. Using Gaussian noise in this context is an effective method of altering these features at high frequencies by affecting data points at all frequencies. In response to this intentional distortion, the neural network decreases its focus on lower-frequency components. Thus, the model

can emphasize more relevant features, resulting in enhanced generalization capabilities. In our study, this technique has a value of 0.1 and a probability of 0.5.

- **ColorJitter**

  As part of the ColorJitter method, random changes are made to an image's attributes, including brightness, contrast, intensity, and hue. As a result of such adjustments, derivative images are generated from the foundational image, which enhances the diversity of the dataset. As a result of this improvement, deep learning models are both more effective and more resilient. Particularly, ColorJitter has proven effective for training models intended for real-world data characterized by diverse illumination conditions, color variations, and additional visual anomalies. Our study uses a set of 0.1 and a probability of 0.5 to alter an image's brightness, contrast, intensity, and hue.

- **Elastic Transformation**

  Elastic transformation is used to introduce random deformations into an image, simulating the inherent variances in shape and orientation that result from the motion or deformation of an object. In this method, a grid of control points is applied to the image, followed by random displacement of these points by a Gaussian distribution. This is followed by the formulation of a continuous curve or surface between two or more defined data points. The result of such a procedure is a smooth deformation of an image, resulting in a realistic representation of the object's shape and location. With values defined by parameters alpha 0.1, sigma 20, and alpha affine 10.0, along with a 0.5 probability. This process also utilizes interpolation between three pixels to calculate distortion.

6.1.3 Model

Our experimental design is adjusted to the project's unique requirements by using public default training, validation, and testing, with an intent to address data imbalances and provide sufficient training to strengthen the model's foundation. To enhance model performance and achieve comprehensive results, we expanded our training dimensions by incorporating diverse views rather than

engaging in layer tuning. We conducted several experiments incorporating dataset integrations and parameter modifications.

During experiments, the ConvNeXt-L backbone model pretrained on ImageNet-1K was used, a subset of ImageNet-22K, for training as feature extraction, with input image size 384x384 and JPEG format with pixel values of 0-1 with lateral and frontal images in parallel for input.

The PyTorch framework is used for building the classification model; the main parameter values in the model are mini-batch size 32 and 64, the learning rate of 0.0001, Weight decay of 0.05 and 0.01, the dropout rate of 0.5, sample per Epoch-1 is 120000 images. The optimizer is AdamW. During training, the OneCycle scheduler adjusts the learning rate. Training stops early when no validation loss improvement is measured. Label smoothing is applied with a very limited value (from 0 to 0.05), and we clip the gradient norm to 1. A random sampler was applied to select data randomly from the training dataset with initial weights from ImageNet-1K By using the pre-trained weights as the initial starting point, the model can benefit from the learned representations and generalizations learned from a diverse set of images. This transfer learning technique allows the model to leverage the knowledge captured during ImageNet-1K training and adapt it to our task, resulting in a reduction in the training duration. With the help of these initial weights, the model can perform better, especially when the target dataset at the beginning of the experiment does not contain enough labeled samples. During training, class weights were used to balance class distribution since the datasets were imbalanced and early stopping which is a technique used to avoid overfitting. It stops the training when validation loss starts to increase.

In addition to these factors, we varied several other factors to conduct an exhaustive analysis. Different numbers of training epochs, augmentation probabilities with different percentages, and batch sizes were adjusted across the experiments to observe their impact on model performance. The model's performance was also examined with different loss functions and thresholds, as well as dynamic thresholds instead of static thresholds.

Finally, each experiment uses different data sets for training and validation, processed through customized pipelines to fine-tune the model's parameters via stochastic gradient descent with each completed batch and epoch. By doing so, we aim to provide a more efficient learning mechanism and simplify the training and validation data loading processes. The ultimate goal is to capture

the main pathologies and abnormalities in chest X-rays, applying consistent labels to images from the same exam, whether they contain multiple sets of lateral and frontal images, and to effectively evaluate how our model performs in detecting and classifying these conditions across varied data sources.

Below is a table 6.2 that summarizes the model's implementation and configuration details.

Element	**Detail**
Model Backbone	ConvNeXt-L, pretrained on ImageNet-1K
Input Image Specifications	384x384 size, JPEG format, pixel values 0-1
Framework	PyTorch
Mini-batch Sizes	32, 64
Learning Rate	0.0001
Weight Decay	0.05, 0.01
Dropout Rate	0.5
Epoch	Value range: 50-400
Samples per Epoch-1	120,000 images
Optimizer	AdamW
Learning Rate Scheduler	OneCycle scheduler
Label Smoothing	Value range: 0-0.05
Gradient Norm Clipping	Clipped to 1
Sampling Technique	Random sampler with initial ImageNet-1K weights
Augmentation Probabilities	Different percentages during experiments for Scale, rotate, flip, elastic transform

Table 6.2: This table outlines the key parameters and settings used in model implementation and configuration.

6.2 Experiments and Results

In this study, various datasets, parameters, and metrics were integrated to develop a model dedicated to analyzing chest X-ray images, to understand the unique features of various medical conditions and the impact of dataset diversity on the results. Using common pathological labels to obtain more data, and having the test data annotated by radiologists, the complexity of clinical care was reflected. Our systematic approach involved using standardized public datasets for training and validation, while the VinDr-CXR dataset, annotated by radiologists, was used for testing to ensure real-world applicability. The comprehensive validation phase, consisting of over 100 different experiments, was essential in enhancing the model's performance and reliability and preparing it for the final test of its suitability for clinical deployment.

Among over 100 experiments, 12 are specifically listed in Table 6.3, chosen because they demonstrate a range of positive or negative responses to each modification. As a result of this detailed organization, we can obtain a clear understanding of datasets, classes, and the model by showcasing their performance under a variety of different parameters, and conditions, as well as gradually adding more datasets. This comprehensive approach results in a more accurate testing phase.

As a result, the model performance was evaluated by AUc, F1, andG-mean metrics, the results of which are presented in Tables AUC6.4, F1 6.5 , andG-mean 6.6. In these tables, each column corresponds to a distinct medical condition and each row represents an experiment.

In addition, this study did not use benchmarks as a direct comparison tool. Our study examines the impact of diverse datasets and their combination during training, validation, and testing on performance and results, which is an rare objective and method in benchmarking studies. In contrast to typical benchmark criteria, the innovative methods we have developed for the combination of datasets are customized to our specific research needs, making comparisons of them irrelevant. Due to the nature of our research and our objective, we have concluded that standard benchmarks cannot be used to evaluate the findings of our study.

An explanation of various experiments is provided in continue, along with their configurations and results.

Exp	Augment_prob	Batch size	Epoch	Train_dataset	Weight decay
1	[1,1,0.5,0.8,0.8]	32	200	[”CheXpert”]	0.05
2	[0.5,0,0.5,0.5,0.5]	32	400	[”CheXpert”]	0.05
3	[0.5,0,0.5,0.5,0.5]	32	400	[”MimicCxrJpg”]	0.05
4	[0.5,0,0.5,0.5,0.5]	32	400	[”ChexNet”]	0.05
5	[0.5,0,0.5,0.5,0.5]	32	400	[”vinBigData”]	0.05
6	[0.8,0,0.8,0.8,0.8]	68	100	[”CheXpert”,”ChexNet”]	0.05
7	[0.8,0,0.8,0.8,0.8]	68	100	[”PadChest”,”ChexNet”]	0.05
8	[0.5,0,0.5,0.5,0.5]	68	50	[”CheXpert”, ”ChexNet”]	0.05
9	[1,1,0.5,0.8,0.8]	32	400	[”CheXpert”,”MimicCxrJpg”,”PadChest”]	0.01
10	[1,1,0.5,0.8,0.8]	32	200	[”CheXpert”,”MimicCxrJpg”,”PadChest”]	0.01
11	[1,1,0.5,0.8,0.8]	32	400	[”CheXpert”,”MimicCxrJpg”,”PadChest” ,”ChexNet”]	0.01
12	[0.8,0,0.8,0.8,0.8]	68	50	[”CheXpert”,”MimicCxrJpg” ,”PadChest” ,”ChexNet”,”VinDr-CXR”]	0.05

Table 6.3: This Table details the specific configurations employed during the experimental phase.

Exp	Atelectasis	Cardiomegaly	Consolidation	Edema	No_Finding	Pneumonia	Pneumothorax
1	0.730	0.689	0.715	0.841	0.708	0.706	0.752
2	0.771	0.805	0.809	0.895	0.870	0.723	0.828
3	0.829	0.822	0.764	0.880	0.912	0.631	0.869
4	0.772	0.823	0.813	0.754	0.646	0.799	0.740
5	0.756	0.696	0.795	0.685	0.895	0.775	0.596
6	0.724	0.872	0.880	0.895	0.871	0.554	0.826
7	0.766	0.686	0.805	0.726	0.644	0.670	0.612
8	0.748	0.771	0.823	0.900	0.898	0.314	0.875
9	0.739	0.761	0.787	0.910	0.857	0.708	0.924
10	0.753	0.734	0.780	0.924	0.870	0.755	0.937
11	0.825	0.662	0.765	0.901	0.850	0.685	0.978
12	0.795	0.925	0.595	0.872	0.488	0.695	0.828

Table 6.4: The AUC performance results are presented in this table for each specific condition.

Exp	Atelectasis	Cardiomegaly	Consolidation	Edema	No_Finding	Pneumonia	Pneumothorax
1	0.343	0.362	0.394	0.393	0.931	0.125	0.500
2	0.690	0.618	0.514	0.716	0.941	0.137	0.400
3	0.692	0.628	0.296	0.653	0.935	0.113	0.429
4	0.667	0.633	0.000	0.000	0.935	0.000	0.000
5	0.582	0.562	0.516	0.000	0.939	0.000	0.000
6	0.642	0.684	0.525	0.644	0.585	0.106	0.600
7	0.592	0.556	0.450	0.368	0.379	0.143	0.066
8	0.641	0.621	0.515	0.675	0.610	0.364	0.714
9	0.632	0.610	0.362	0.541	0.941	0.114	0.240
10	0.635	0.590	0.335	0.590	0.935	0.103	0.203
11	0.638	0.554	0.370	0.550	0.935	0.131	0.400
12	0.416	0.500	0.140	0.291	0.718	0.160	0.382

Table 6.5: In this table, F1 Scores performance results are presented according to each condition.

Exp	Atelectasis	Cardiomegaly	Consolidation	Edema	No_Finding	Pneumonia	Pneumothorax
2	0.707	0.703	0.783	0.844	0.666	0.656	0.574
3	0.738	0.479	0.708	0.810	0.635	0.589	0.698
4	0.669	0.000	0.712	0.000	0.277	0.000	0.000
5	0.657	0.677	0.572	0.000	0.800	0.000	0.000

Table 6.6: This table presents the performance results in terms ofG-mean Scores for each specific condition evaluated.

6.2.1 Experiments and Results 1

- Conducted the first experiment (Exp 1) on the ChexPert dataset with specific configurations.
- Adjusted the data augmentation probability distribution to [1,1,0.5,0.8,0.8] for diverse training samples.
- Training for 200 epochs with a batch size of 32.
- Set label smoothing to zero and weight decay to 0.05 to prevent overfitting.
- Applied CLAHE method with a clip limit of 2 for image enhancement on chest X-rays.
- AUC results were consistent across a variety of medical conditions.
- There are remarkable AUC values for Edema at 0.841, No Finding at 0.708, and Cardiomegaly at 0.689, suggesting that conditions can still be detected more efficiently.
- No Finding achieved a high F1 score of 0.931, but Pneumonia achieved a low F1 score of 0.125, suggesting the need for further optimization.
- The configurations were found to be effective for selecting No Findings and detecting Edema, however improvements may be required in other areas to increase AUC and F1 scores.

6.2.2 Experiments and Results 2

- Conducted the second experiment (Exp 2) on the"ChexPert dataset with different configurations.
- Modified data augmentation probability to [0.5,0,0.5,0.5,0.5] for increased training sample diversity.
- Extended training to 400 epochs with a batch size of 32.
- Maintain label smoothing at zero and weight decay at 0.05.
- Set the CLAHE clip limit to 0.12 for image enhancement.

- Implemented a dynamic threshold for classification, enhancing detection rates across all conditions.
- Introduced theG-mean metric in Exp 2 to provide a more thorough evaluation of the model's performance.
- Observed improvements in AUC and F1 scores across all classes from Exp 1 to Exp 2.
- A significant rise in AUC for Edema to 0.895 and Cardiomegaly to 0.805.
- The AUC for No Finding was 0.941, demonstrating high classification accuracy.
- The F1 score for Pneumonia remained low at 0.137, suggesting that further model modification is necessary.
- A significantG-mean score of 0.844 was obtained for Edema.
- There was a significant improvement in Exp 2 over Exp 1, suggesting that the configuration settings were effective for same dataset.

6.2.3 Experiments and Results 3

- The third experiment (Exp 3) was conducted using a different training dataset, called Mimic-CxrJpg.
- The configuration was the same as the previous one.
- An evaluation of the model's adaptability and effectiveness was conducted using data sources from a variety of sources.
- AUC results showed variability, with high scores for Atelectasis 0.829, Cardiomegaly 0.822, and Pneumothorax 0.869, with a significant decrease in Pneumonia from 0.723 to 0.631 compared to Exp 2.
- F1 scores varied among conditions, with improvements in No Finding and Pneumothorax, but drops in Pneumonia and Edema.

-mean metrics were introduced, providing scores across conditions such as Atelectasis 0.738, and Pneumothorax 0.698.

- The comparison across experiments with consistent validation datasets but different training datasets highlighted the impact of training data quality on model performance.
- The experiment demonstrated that it is imperative to select parameters that align with the characteristics of the training data, and despite the change in the dataset, the model was able to achieve significant results.

6.2.4 Experiments and Results 4

- Experiment 4 (Exp 4) continued using ChexNet for training and ChexPert for validation, following the established framework.
- The AUC results for Experiments 3 to 4, Atelectasis' score slightly decreased to 0.772, Cardiomegaly's score increased to 0.823, Pneumothorax's score declined to 0.740, and Pneumonia's score improved significantly from 0.631 to 0.799.
- The F1 scores for Atelectasis, Consolidation, and Edema dropped dramatically to 0.000 in Experiment 4. The No Finding F1 score remained at 0.935. In contrast, Pneumonia and Pneumothorax dropped dramatically to 0.000 from their previous scores in Experiment 3. As a point of clarification, for some classes, the result is rounded down to zero if the result falls below the threshold, since the value was so small and consider it as zero, the system did not update them for these specific classes, which it has been solved for other experiments.
- There were significant variations in G-mean scores between Exp 4 and Exp 3, with improvements in Cardiomegaly but declines in other conditions.
- As a result, the study demonstrated that the dataset was highly impactful on performance and that there were variations between conditions despite using similar configurations.

6.2.5 Experiments and Results 5

- Experiment 5 (Exp 5) includes the same augmentation as before, but with a different training dataset referred to as vinBigData.
- There was an average decrease of 0.2-0.3 in AUC scores for conditions such as Atelectasis, Consolidation, Cardiomegaly, Pneumonia, and Pneumothorax between Exp 4 and Exp 5. Additionally, the AUC score for Edema decreased from 0.754 to 0.685. In contrast, the AUC score for No Finding increased significantly from 0.646 to 0.895.
- F1 results displayed inconsistent patterns, with moderate scores for Atelectasis dropping from 0.667 to 0.582 and Cardiomegaly dropping from 0.633 to 0.562 from Exp 4 to Exp 5. However, Pneumonia and Pneumothorax remained at 0.000 in Ex 5, while Consolidation improved from 0.000 to 0.516, and No Finding remained at 0.935 to 0.939.
- Despite identical configurations, the performance differences highlight the different characteristics and limitations of each dataset. These characteristics have implications for disease identification performance.

6.2.6 Experiments and Results 6

- Experiment 5 (Exp 6) employed a combination of ChexPert and ChexNet datasets to improve AUC scores, especially after ChexPert exceeded 0.80 in all classes except Pneumonia in Exp 2, suggesting potential benefits from combining datasets for this class.
- Some configurations remain the same as before, including the validation dataset, weight decay, and image enhancement.
- Augmentation probability and batch size have been increased.
- There was a reduction in the Epoch from 400 to 100.
- The label smoothing was changed from 0 to 0.1.

- AUC result for Atelectasis was 0.724, while Cardiomegaly achieved 0.872. The Consolidation score was 0.880, as well as the edema score of 0.895 and the No Finding score of 0.871. Pneumonia, however, showed a 0.554. For Pneumothorax, the AUC was 0.826.

- A F1 score of 0.642 was achieved for Atelectasis, whereas an F1 score of 0.684 was obtained for Cardiomegaly, Consolidation rate was 0.525, and Edema 0.644. No Findings 0.585, Pneumonia 0.106, and Pneumothorax had a score of 0.600.

- The team's strategic modifications resulted in noticeable improvements in the model's overall performance. The dataset combination and parameter tuning improved the AUC for most conditions for the majority of classes.

6.2.7 Experiments and Results 7

- A similar configuration is used in Experiment 7 (Exp 7), with an augmentation probability of [0.8, 0, 0.8, 0.8, 0.8], a batch size of 68, 100 epochs, and a label smoothing value of 0.1. Both experiments use the ChexPert dataset for validation, a weight decay of 0.05, and an image enhancement clip limit of 0.12.

- The main difference between Exps 6 and 7 is the source of training data: Exp 6 uses ChexPert and ChexNet, while Exp 7 uses PadChest combined with ChexNet.

- The AUC scores for Atelectasis and Pneumonia in Experiment 7 improved significantly from 0.724 to 0.766 and 0.554 to 0.670, respectively. On the other hand, Cardiomegaly decreased from 0.872 to 0.686, Consolidation from 0.880 to 0.805, Edema from 0.895 to 0.726, No Finding from 0.871 to 0.644, and Pneumothorax declined from 0.826 to 0.612.

- F1 scores across a variety of conditions. Atelectasis' F1 score decreased from 0.642 to 0.592, Cardiomegaly from 0.684 to 0.556, Consolidation from 0.525 to 0.450, Edema from 0.644 to 0.368, No Finding from 0.585 to 0.379, and Pneumothorax from 0.600 to 0.066. However, Pneumonia improved from 0.106 to 0.143.

- The use of a different combination of datasets in Experiment 7 appears to have negatively affected the model's ability to accurately identify most conditions. These findings highlight the crucial role that choosing the right dataset plays in the training of machine learning models, in which the unique characteristics of each dataset can significantly alter the effectiveness of the model.

6.2.8 Experiments and Results 8

- Experiment 8 mimics Experiment 6's configuration with the same batch size, label smoothing, datasets, weight decay, and CLAHE image enhancement settings.
- Two main differences are noted: Exp 8 uses lower data augmentation probabilities ([0.5, 0, 0.5, 0.5, 0.5]) and trains for fewer epochs 50 vs. 100 in Exp 6.
- As a result of previous experiments, adjustments were made to achieve a regularizing effect by using less intensive data augmentation techniques and shorter training durations.
- The AUC score between Exp 6 and Exp 8 for atelectasis, pneumothorax, edema, and no findings has increased to 0.748, 0.875, 0.900, and 0.898, respectively. However, the model did not perform as well in detecting Cardiomegaly, Consolidation, and Pneumonia, with AUC scores falling to 0.771, 0.823, and 0.314, respectively.
- The F1 score for Atelectasis dropped to 0.641, Cardiomegaly to 0.621, and Consolidation to 0.515. However, the model shows improvement in detecting Edema, which increased to 0.675, No Findings to 0.610, Pneumonia to 0.364, and Pneumothorax to 0.714
- These results suggest that while the model's overall robustness may have improved, the impact of the experimental modifications on its performance has been mixed. This requires a review of the configuration for a more uniformly enhanced performance.

6.2.9 Experiments and Results 9

- Experiment 9 (Exp 9) combined datasets from ChexPert, MimicCxrJpg, and PadChest with a highly aggressive data augmentation strategy with probabilities like [1, 1, 0.5, 0.8, 0.8], maximizing training data diversity.
- The batch size is 32, and the model trains run for 400 epochs.
- Label smoothing is not used, weight decay is set at 0.01, and CLAHE with a clip limit of 0.12 is employed for image enhancement.
- The AUC metrics demonstrate a high level of detection efficiency for Edema 0.910 and Pneumothorax 0.924, indicating that the model is very effective for these conditions. Scores for Atelectasis were 0.739, Cardiomegaly 0.76, Consolidation 0.787, Pneumonia 0.708, and No Finding with an AUC of 0.857.
- F1 scores for Atelectasis 0.632,r Cardiomegaly 0.610, Consolidation 0.362, Edema 0.541, No Finding 0,941, Pneumonia 0.114, and Pneumothorax 0.240.
- The combination of large datasets and enhanced performance by modifying configurations demonstrates a significant improvement in results in most conditions.

6.2.10 Experiments and Results 10

- In Experiment 10 (Exp 10), the only change was the training duration, which was reduced to 200 epochs, the datasets and configurations were the same as in Experiment 9.
- AUC scores increased for Edema to 0.924, No Finding to 0.870, and Pneumonia and Pneumothorax to 0.755 and 0.937, respectively. However, there were slight decreases in AUC for Atelectasis to 0.753 and for Consolidation to 0.780.
- F1 scores also showed mixed results: an improvement for Atelectasis to 0.635 and Edema to 0.590 but decreases in several other conditions, such as Cardiomegaly decreased to 0.590, Consolidation to 0.335, No Finding to 0.935, Pneumonia to 0.103, and Pneumothorax to 0.203.

- Changing the training approach in experiment 10, specifically reducing the number of epochs, led to mixed results. Some conditions resulted in improved metrics, while others indicated that further optimization is required. However, although some aspects of the model's detection capabilities have been enhanced, the balance between training intensity and duration may need to be adjusted to achieve consistent improvements across all conditions.

6.2.11 Experiments and Results 11

- Exp 11 Expanded training datasets including ChexPert, MimicCxrJpg, PadChest, and ChexNet and the same configuration as previous.
- The number of epochs has been increased by 400.
- For AUC Atelectasisvand Pneumothorax improved to 0.825 and 0.978, respectively. Edema achieved at AUC 0.901.
- Detection performance for cardiomegaly, consolidation, and no findings decreased with AUCs of 0.662, 0.765, and 0.850, respectively; pneumonia performance decreased with an AUC of 0.685.
- The F1 scores for Atelectasis, Consolidation, Pneumonia, and Pneumothorax have improved to 0.638, 0.370, 0.131, and 0.400, respectively. No Findings were detected with an F1 score of 0.935. Cardiomegaly and edema decreased to 0.554 and 0.550, accordingly.
- We observed trade-offs with improvements in detecting certain conditions but declines in others after adding datasets and changing the epoch. As a result of the varied effects of adjusting training complexity, this study demonstrates the need for further optimization of the model's diagnostic accuracy across a variety of conditions using a balanced approach.

6.2.12 Experiments and Results 12

- Experiment 12 used a combination of 5 datasets ChexPert, MimicCxrJpg, PadChest, ChexNet, VinBigData , and ChexNet for validation with different configurations.

- A uniform augmentation was used.
- Increased batch size to 68, and reduced epochs to 50, prioritizing speed and efficiency.
- Used label smoothing with a value of 0.1 and weight decay increased to 0.05.
- For AUC results, Cardiomegaly Pneumonia detection is improved by experiment 12 with levels of 0.925 and 0.695, whereas No Findings and Consolidation are poorly detected with levels of 0.488 and 0.595, respectively, with an Acceptable result for Atelectasis with levels of 0.795 and Pneumothorax with levels of 0.828 and 0.828, as well.
- As a result of the F1 scores of 0.416 and 0.500 for detecting atelectasis and cardiomegaly, respectively. Consolidation received a score of 0.140, while Edema and Pneumothorax were rated at 0.291 and 0.382, respectively. 'No Finding,' achieved 0.718, and Pneumonia, with an F1 score of 0.160.
- Experiment 12 demonstrates a strategic shift in the training approach, by utilizing an extensive combination of five datasets and streamlining the process with different configurations. In general, the performance spectrum ranges from good in certain detections to needing significant improvement in others.

6.2.13 Test Phase

The table below 6.7 provides a comprehensive overview of the model's performance during the Test Phase. It includes both the AUC and F1 scores, offering insights into the model's for each evaluated condition, and follows with a brief discussion of findings.

Abnormalities	**AUC**	**F1**
Atelectasis	0.575	0.364
Cardiomegaly	0.828	0.946
Consolidation	0.75	0.836
Edema	0.727	0.969
No Finding	0.506	0.558
Pneumonia	0.861	0.980
Pneumothorax	0.756	0.963

Table 6.7: This table presents AUC and F1 Results for the Test Phase.

- Testing phase configured as per Experiment 2 parameters: with an augmentation probability of [0.5,0,0.5,0.5,0.5], mini-batch size of 32, and 400 epochs.
- Utilized the ChexPert dataset for both training and validation with zero weight decay and label smoothing of 0.05.
- Applied CLAHE image enhancement with a clip limit range of 0 to 12.
- Utilized the VinDr-CXR dataset, which has been annotated by radiologists, to align closely with clinical examination conditions.
- The AUC results for Cardiomegaly demonstrate strong diagnostic ability with an AUC of 0.828, followed by Pneumonia with 0.861, and Pneumothorax with 0.756. AUCs of 0.75 for consolidation and 0.727 for edema demonstrate fair accuracy. In contrast, Atelectasis shows an AUC of 0.575, while "No Finding" achieved 0.506.
- With an F1 score of 0.980, pneumonia and edema indicate outstanding performance. Cardiomegaly at 0.946 is followed by pneumothorax at 0.963. Consolidation also displays a good score of 0.836. In comparison, "No Finding" has a moderate F1 score of 0.558 and Atelectasis has the lowest score of 0.364.

- It is important to emphasize that a different dataset was used for testing, the VinDr-CXR, rather than the ChexPert dataset used for training and validation. The purpose of this strategy is to ensure that the model's performance is assessed on data that it has never seen before, thus minimizing the risk of overfitting and providing a better representation of its use in real-life situations.

6.3 RADIA Web Tool

The RADIA (Radiologie Intelligence Artificielle) project integrates a web tool, which is currently under development by the Research Group at CIUSSS(Centre intégré universitaire de santé et de services sociaux du Centre-Sud-de-l'Île-de-Montréal) to detect lung abnormalities with consultants of the Notre-Dame and Verdun Hospitals in Montreal. By integrating adaptable web tools with deep learning, radiologists can gain a significant advantage. With this combination, radiologists receive automated analysis results in real-time via a user-friendly web interface, resulting in increased productivity. Adaptable to the specific needs and workflows of various healthcare institutions, the web tool effortlessly integrates with existing EMR systems and accommodates the latest deep-learning algorithms and models.

A primary objective of the project is to develop a tool to provide radiologists with a safety net, preventing them from neglecting potential anomalies in chest X-rays. This tool is implemented as a Web User Interface using the Blazor framework, making it accessible via smartphones. A tool such as this is designed to assist in decision-making during periods of high workload by providing an additional perspective on X-rays. In emergency situations, radiologists can customize the tool's options according to their specific detection needs, resulting in more efficient and effective care for patients.

RADIA aims to enhance the efficiency and performance of the analysis process, supporting radiologists in making well-informed decisions. It is important to note that RADIA does not aim to replace human expertise but rather complement it. The RADIA web tool comprises an image viewer and a confidence probability panel Figure6.4. The panel presents the likelihood of disease based on the medical team's requirements, showcasing categories and subsets of

anomalies. A colorful bar is utilized to indicate the presence or absence of an anomaly, utilizing a predetermined threshold. For each image, the viewer allows medical teams to provide feedback and information about anomalies. The feedback provides new data for the training process, thereby improving the parameters of the model. The medical team can view both raw and transformed images as well as the probability of each anomaly using web-connected stations.

Healthcare teams can benefit from a more comprehensive and accurate analysis process by combining the expertise of radiologists with RADIA's technological capabilities. The tool serves as a valuable asset in the medical field, enabling collaborative decision-making and ultimately improving patient care outcomes.

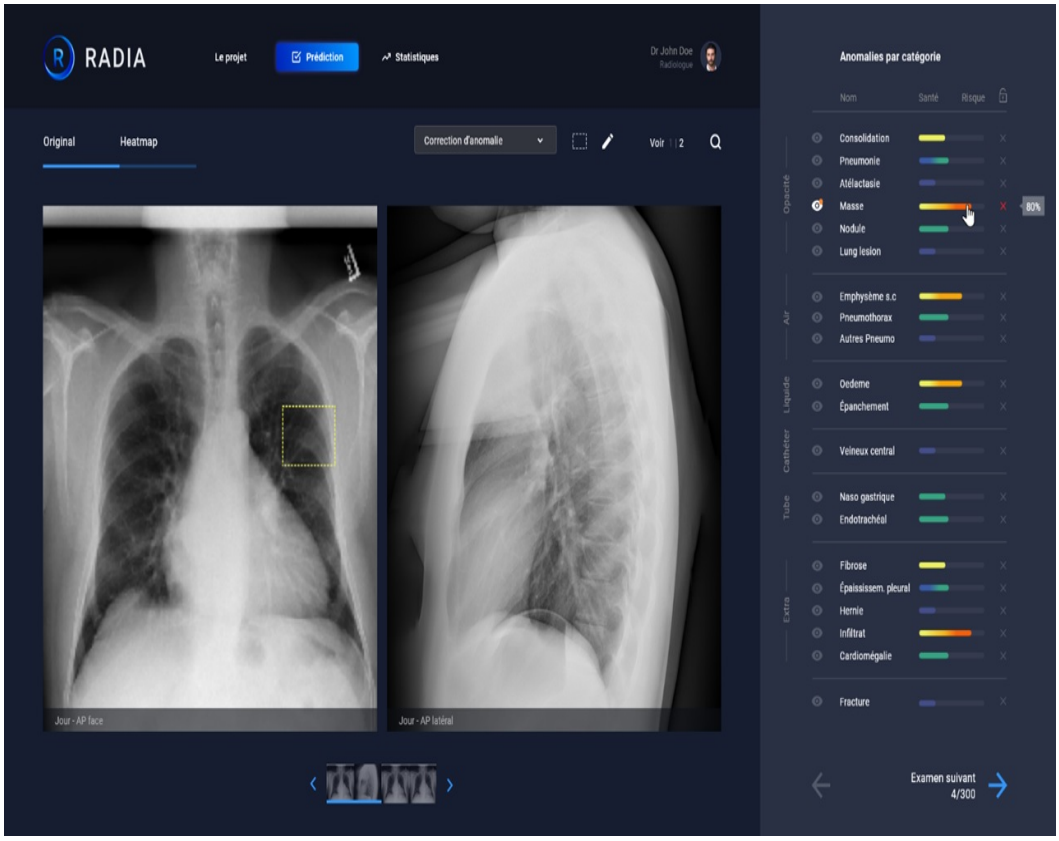

Figure 6.4: RADIA WebTool interface for anomaly detection in chest X-ray images.

6.4 Conclusion

As a result of these experiments and process, we can identify the model's strengths and weaknesses. Understanding the model's strengths is just as critical as identifying its weaknesses. Strengths might include high performance at detecting specific types of abnormalities or robust performance under certain imaging conditions. Recognizing these areas helps us leverage the model's strengths and apply them effectively in clinical settings. When the model can detect pneumothorax with outstanding performance, this ability can be emphasized in clinical applications in which rapid and accurate detection of this condition is critical.

Considering the results presented, experiment 8 performs better in most classes and among all classes, Atelectasis is the second highest performer, followed by No Finding. For most classes and experiments the F1 score yielded poor results, possibly because it is directly associated with imbalanced data, threshold selection, misclassification of positive and negative cases, and complex decision boundaries, especially for classes that have similar symptoms, feature quality which decreases after reducing the size of the input images. It acknowledges that different sources of data might vary in quality, balance, or representation of classes, which in turn could affect the accuracy, reliability, and generalizability of the model's predictions. It is necessary to have the radiologists consultant to describe the specific characteristics and features of the data in these situations.

In addition, the above experiments have provided comprehensive insights into the effects of various training datasets, data augmentation techniques, image enhancement, batch sizes, weight decays, label smoothing, and epoch counts on the performance of ConVNext in detecting pathological features in chest X-rays.

- Exp1 established a baseline with significant AUC scores, particularly for Edema and No Finding conditions.
- Exp 2's modifications in augmentation probabilities and CLAHE enhancement led to improved AUC and F1 scores, suggesting the potential benefits of fine-tuning image preprocessing techniques.

- As the experiments progressed, with Exp 3 introducing a new training dataset, MimicCxrJpg, Exp 4 using ChexNet, and Exp 5 using the VinBigData, the results demonstrated the importance of diverse training data in achieving robustness. The results also revealed valuable insights into the distinct characteristics of each dataset. By examining their performance under the same experimental conditions, we can gain insight into the behavior of specific classes within these datasets. It is important to understand these characteristics so that algorithms can be refined and models can be adapted to handle the variability in clinical presentations across different datasets.

- Moreover, Exp 6 took this further by combining datasets and optimizing the augmentation probability and batch size, leading to generally improved AUC and F1 scores for several conditions.

- Exp 7 and Exp 8 continued to adjust datasets and training parameters. This analysis demonstrated the delicate balance between data diversity and model performance, particularly in detecting Atelectasis and Pneumonia.

- Exp 9 and 10 explored the effects of intensive data augmentation and epoch count adjustments. While they achieved high AUC scores for conditions like Edema and Pneumothorax, the F1 scores for other conditions indicated a need for model optimization.

- Exp 11's inclusion of an additional dataset and extended training epochs resulted in significant improvements in Pneumothorax detection. However, the performance for other conditions such as Cardiomegaly and Pneumonia varied, illustrating how challenging it is to achieve uniform improvements across different pathological features.

- Exp 12 underscores the intricate balancing act required in model training, where optimizing for certain conditions can impact the detection efficacy of others.

- In addition, the test phase indicates that it is crucial to use a different dataset to evaluate the model's performance on unseen data, reduce overfitting risks, and ensure the model applies to real-world scenarios.

In summary, The experimental results underscored the close interplay between dataset characteristics, model architecture, and training hyperparameters in medical image analysis. It highlighted how the choice of the dataset, the architecture, and the fine-tuning of training hyperparameters and parameters such as augmentation probability, epoch, and image enhancement together affect the performance of models in detecting medical conditions from images. In addition, it is important to recognize the role datasets play in training, validating, and testing, particularly in terms of their source and quality. Datasets collected from different medical institutions may differ in imaging techniques, quality of the data, patient demographics, and pathology rates, all of which can have a significant effect on detection performance.

6.5 Limitations

During our research, we encountered a variety of limitations that impacted the results and process. Insufficient representation of rare diseases within our datasets was caused by the uniform nature of patient populations. Model training and validation could have been affected by the use of potentially unverified annotations. Even under identical configurations, results varied across experiments, especially when different datasets and classes were analyzed.

In some datasets, more than three images are available for each patient, including lateral, posteroanterior (PA), and anteroposterior (AP) views. Other datasets provide only one type of image per patient, in contrast.

In addition, there are diseases that are inherently uncommon or rare, occurring infrequently in the general population. As a consequence of this infrequency, it is difficult to collect comprehensive and representative samples of these conditions. Thus, despite the use of balancing techniques, datasets often lack sufficient data regarding these rarer diseases, resulting in a lack of comprehensive information.

Moreover, During this study, we were unable to utilize our private dataset due to the fact that Notre Dam and Verdun Hospitals in Montreal exam reports are in French, and labeling was generated using an NLP algorithm and we had no source to compare our labeling results for

the French language. On the other hand, based on hospital policy the reports tend to be more general, not mentioning specific diseases by name.

Another reason is that our test data will contain the most recent images provided by PACS with radiologists' comments during testing the final product, and the team is currently tuning the training model to produce the best possible results and preparing the model for the final evaluation. As a means of reducing the risks of generating errors through our AI systems, it was necessary to incorporate robust label verification processes, including expert radiological review. Due to a high level of workload, we were unable to utilize their consultant.

In addition to ethical considerations regarding patient privacy, we were also restricted in our ability to access a wide range and depth of data for private datasets. It would be beneficial to incorporate supplementary data, such as blood test results and patient medical histories, to enhance disease detection capabilities, as radiologists typically consider this additional information when diagnosing patients. It was however difficult for us to access these auxiliary data, including detailed patient histories, for both public and private datasets. Due to this limitation, our model was unable to simulate a doctor's detection process, which may affect its ability to detect diseases in a real clinical environment.

Although we are optimizing the model using the latest annotated data with expert insights, the absence of additional patient information poses challenges in mirroring a radiologist's diagnostic process. As a result of the lack of radiologists directly involved in the project, a critical limitation has been addressed. A radiologist's deep understanding of diseases contributes to the development of the model to detect complex pathologies and ensure high-quality results. Their insights into error analysis provide crucial feedback for model improvement and ensure clinical relevance.

Overall, we encountered several limitations in our research, including the insufficient representation of rare diseases due to uniform patient populations. In addition, there were challenges related to the fact that some patients received multiple types of images while others only received one in dadatsets. The inability to fully integrate radiologists' expertise in label verification, as well as the limitations on accessing additional patient data, limited the model's

ability to simulate a doctor's detection process. This affected its ability to detect disease in a clinical environment, highlighting the need for expert radiological involvement to improve accuracy and clinical relevance.

6.6 Future Work

In future work, our team aims to address several key areas of focus.

Firstly, we recognize the importance of data labeling for private datasets to ensure accurate and reliable model training and invest effort in developing robust labeling processes and techniques to maintain privacy while achieving high-quality annotations.

Furthermore, we will study parameter and hyperparameter selection and optimization in more detail, such as considering different losses. By conducting extensive experimentation and analysis, we aim to fine-tune our models to achieve even higher levels of accuracy and performance. This will involve exploring various techniques, such as advanced optimization algorithms and automated hyperparameter tuning. The research team is actively exploring these methods to ensure that the RADIA tool maintains its effectiveness across different techniques and all classes, including the rare ones.

In addition, we aim to expand the RADIA tool's capabilities by integrating additional pipelines for generating segmentation, heat maps, and bounding boxes. These visualizations can provide valuable insights into the regions of interest and help radiologists focus on specific areas during the diagnostic process.

Moreover, merging datasets, it's also beneficial to group different classes based on their primary symptoms, reasons, or medical categories. The purpose of this is to understand how different classes interact with one another through their pattern and feature correlation between diseases.

In addition, we recognize the importance of customization to meet the unique needs of radiologists in different local hospitals. We plan to collaborate closely with medical professionals to understand their specific requirements and characteristics of different diseases and classes and adapt the disease classification framework accordingly. This customization will ensure that the RADIA tool effectively supports radiologists in their daily practice while aligning with each healthcare institution's specific challenges, needs, and characteristics.

Lastly, future developments of the RADIA project will focus on integrating medical information recorded by other healthcare professionals into the tool's evolution. We will also expand clinical testing, incorporate additional imaging modalities, and provide multilingual support. In addition, efforts will be made to ensure AI transparency.

By focusing on these areas in future work, we aim to continuously improve the RADIA tool and deep learning model to enhance its capabilities to contribute to the advancement of AI-assisted radiology for more accurate and efficient diagnoses and the RADIA web tool's final phase includes developing automated reports.

Chapter 7

Ethics & Concerns

7.1 Ethical in Clinical Research

Clinical research began in the early 18th century and suffered tragedies in the 19th century. In the history of clinical research, a number of tragedies have resulted in the creation of some rules and ethical principles [21].

Participants' identities and records are kept as confidential as possible (except when legally required). This is to avoid any form of hardship, discrimination, or stigmatization as a consequence of having participated in the research [14].

Ethics in clinical research are governed by four fundamental principles: autonomy, non-maleficence, beneficence, and justice [8]. It refers to a patient's right to choose what is right for him or her, not to harm the patient in any way, and all choices and decisions need to be based on the patient's benefit. The subset of privacy and confidentiality is one of the most important principles of ethics in clinical research, having a direct impact on our work. For example, obtaining patients' clinical histories and information is essential for diagnosing and finding lung abnormalities. Certain lung diseases are directly related to age and sometimes to gender. It is common for aging individuals to suffer more lung damage and have higher risk factors for certain types of lung abnormalities[69]. In addition to aging, environmental stressors include cigarette smoke, cooking and heating fumes, pollen, bacteria, and viruses, as well as toxins and pathogens from the workplace and home environments. Additionally, dust, wood or metal particles, and welding fumes can also cause lung abnormalities

when they are inhaled from a factory or industrial work environment. Information such as this may contribute to the classification of diseases by adding them as features or probabilities. The use of these types of data, however, is not permitted in order to protect the privacy of patients.

In summary, clinical research has a long history marked by the development of ethical standards to protect participants' confidentiality and adhere to principles of patient autonomy, non-harm, benefit, and fairness. As a result, these ethical guidelines are crucial for the responsible collection and use of sensitive data, ensuring that patient privacy is protected to the greatest extent possible.

7.2 Concerns

In applying AI to a real-world problem there are many concerns, and the medical domain is not an exception to face them too.

The first concern relates to the trustworthiness of AI-based detection and diagnosis by patients.

The second question is who is responsible for the consequences of an incorrect diagnosis or detection.

The third question is whether artificial intelligence will replace radiologists? By introducing AI into the medical world, especially in the field of radiology, has had some effects on different roles and tasks, leading to a growing concern among radiologists that AI will replace the radiologist, and the future of these roles are uncertain.

It is important to note, a radiologist's role is more complex than performing single routine task at a time, and AI can improve the performance of these routine tasks, and radiologists can collaborate in more vital and emergency situations where decision-making is needed. As a matter of fact, automating has not resulted in roles being lost, instead, responsibilities or priorities will change [18].

In addition, automating systems, starting with archiving images or reporting and continuing with the detection of difficult diseases, aim to provide assistance to radiologists [55].

In summary, radiologists are currently facing challenges related to AI integration. To address

these concerns requires close collaboration between AI technologies and radiologists without displacing medical professionals' valuable expertise. A partnership such as this is essential for unlocking the full potential of artificial intelligence in improving patient care and advancing medical practice.

Appendix A

My Appendix

Appendix figure example is shown in A.1 below

Figure A.1: An figure example in Appendix A.

Bibliography

[1] Acharya, T. and Ray, A. K. (2005). *Image processing: principles and applications*. John Wiley & Sons.

[2] Agrawal, T. and Choudhary, P. (2022). Segmentation and classification on chest radiography: a systematic survey. *The Visual Computer*, pages 1–39.

[3] Ahmed, K. T., Irtaza, A., and Iqbal, M. A. (2017). Fusion of local and global features for effective image extraction. *Applied Intelligence*, 47(2):526–543.

[4] Ahmed, T. and Sabab, N. H. N. (2022). Classification and understanding of cloud structures via satellite images with efficientunet. *SN Computer Science*, 3(1):1–11.

[5] Allaouzi, I. and Ahmed, M. B. (2019). A novel approach for multi-label chest x-ray classification of common thorax diseases. *IEEE Access*, 7:64279–64288.

[6] Alzubaidi, L., Duan, Y., Al-Dujaili, A., Ibraheem, I. K., Alkenani, A. H., Santamaría, J., Fadhel, M. A., Al-Shamma, O., and Zhang, J. (2021a). Deepening into the suitability of using pre-trained models of imagenet against a lightweight convolutional neural network in medical imaging: An experimental study. *PeerJ Computer Science*, 7:e715.

[7] Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M. A., Al-Amidie, M., and Farhan, L. (2021b). Review of deep learning: Concepts, cnn architectures, challenges, applications, future directions. *Journal of big Data*, 8(1):1–74.

[8] Avasthi, A., Ghosh, A., Sarkar, S., and Grover, S. (2013). Ethics in medical research: General principles with special reference to psychiatry research. *Indian Journal of Psychiatry*, 55(1):86.

[9] Ayaz, M., Shaukat, F., and Raja, G. (2021). Ensemble learning based automatic detection of tuberculosis in chest x-ray images using hybrid feature descriptors. *Physical and Engineering Sciences in Medicine*, 44(1):183–194.

[10] Bay, H., Tuytelaars, T., and Van Gool, L. (2006). Surf: Speeded up robust features. *Lecture notes in computer science*, 3951:404–417.

[11] Bustos, A., Pertusa, A., Salinas, J.-M., and de la Iglesia-Vayá, M. (2020). Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical image analysis*, 66:101797.

[12] Cai, L., Gao, J., and Zhao, D. (2020). A review of the application of deep learning in medical image classification and segmentation. *Annals of translational medicine*, 8(11).

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

[14] CIOMS and WHO (2017). International ethical guidelines for health-related research involving humans. *International ethical guidelines for health-related research involving humans.*

[15] Cohen, J. P., Viviano, J. D., Bertin, P., Morrison, P., Torabian, P., Guarrera, M., Lungren, M. P., Chaudhari, A., Brooks, R., Hashir, M., et al. (2021). Torchxrayvision: A library of chest x-ray datasets and models. *arXiv preprint arXiv:2111.00595*.

[16] Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 886–893. Ieee.

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

[18] ESR (2019). What the radiologist should know about artificial intelligence–an esr white paper, european society of radiology communications@ myesr. org neri emanuele de souza nandita brady adrian bayarri angel alberich becker christoph d. coppola francesca visser jacob. *Insights into imaging*, 10(1):44.

[19] Fotouhi, S., Asadi, S., and Kattan, M. W. (2019). A comprehensive data level analysis for cancer diagnosis on imbalanced data. *Journal of biomedical informatics*, 90:103089.

[20] Gouda, N. and Amudha, J. (2020). Skin cancer classification using resnet. In *2020 IEEE 5th International Conference on Computing Communication and Automation (ICCCA)*, pages 536–541. IEEE.

[21] Grady, C. (2018). Ethical principles in clinical research. In *Principles and practice of clinical research*, pages 19–31. Elsevier.

[22] Hendrycks, D. and Gimpel, K. (2016). A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*.

[23] Hossin, M. and Sulaiman, M. N. (2015). A review on evaluation metrics for data classification evaluations. *International journal of data mining & knowledge management process*, 5(2):1.

[24] Hsu, Y.-C., Shen, Y., Jin, H., and Kira, Z. (2020). Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10951–10960.

[25] Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708.

[26] Huh, M., Agrawal, P., and Efros, A. A. (2016). What makes imagenet good for transfer learning? *arXiv preprint arXiv:1608.08614*.

[27] Iqbal, S., Qureshi, A. N., Li, J., Choudhry, I. A., and Mahmood, T. (2023). Dynamic learning for imbalance data in learning chest x-ray and ct images. *Heliyon*.

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

[29] Islam, M. R. and Nahiduzzaman, M. (2022). Complex features extraction with deep learning model for the detection of covid19 from ct scan images using ensemble based machine learning approach. *Expert Systems with Applications*, 195:116554.

[30] Johnson, A. E., Pollard, T. J., Greenbaum, N. R., Lungren, M. P., Deng, C.-y., Peng, Y., Lu, Z., Mark, R. G., Berkowitz, S. J., and Horng, S. (2019). Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*.

[31] Kanagachidambaresan, G., Ruwali, A., Banerjee, D., and Prakash, K. B. (2021). Recurrent neural network. *Programming with TensorFlow: Solution for Edge Computing Applications*, pages 53–61.

[32] Khalid, S., Khalil, T., and Nasreen, S. (2014). A survey of feature selection and feature extraction techniques in machine learning. In *2014 science and information conference*, pages 372–378. IEEE.

[33] Khan, A., Sohail, A., Zahoora, U., and Qureshi, A. S. (2020). A survey of the recent architectures of deep convolutional neural networks. *Artificial intelligence review*, 53(8):5455–5516.

[34] Lantsman, C. D., Barash, Y., Klang, E., Guranda, L., Konen, E., and Tau, N. (2022). Trend in radiologist workload compared to number of admissions in the emergency department. *European Journal of Radiology*, 149:110195.

[35] LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *nature*, 521(7553):436–444.

[36] LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551.

[37] Li, Y. (2022). Research and application of deep learning in image recognition. In *2022 IEEE 2nd International Conference on Power, Electronics and Computer Applications (ICPECA)*, pages 994–999. IEEE.

[38] Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., Van

Der Laak, J. A., Van Ginneken, B., and Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88.

[39] Liu, H., Seedat, N., and Ive, J. (2022a). Modeling disagreement in automatic data labelling for semi-supervised learning in clinical natural language processing. *arXiv preprint arXiv:2205.14761*.

[40] Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., and Xie, S. (2022b). A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986.

[41] Liz, H., Huertas-Tato, J., Sánchez-Montañés, M., Del Ser, J., and Camacho, D. (2023). Deep learning for understanding multilabel imbalanced chest x-ray datasets. *Future Generation Computer Systems*, 144:291–306.

[42] Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110.

[43] Maharana, K., Mondal, S., and Nemade, B. (2022). A review: Data pre-processing and data augmentation techniques. *Global Transitions Proceedings*.

[44] Meedeniya, D., Kumarasinghe, H., Kolonne, S., Fernando, C., De la Torre Díez, I., and Marques, G. (2022). Chest x-ray analysis empowered with deep learning: A systematic review. *Applied Soft Computing*, page 109319.

[45] Müller, R., Kornblith, S., and Hinton, G. E. (2019). When does label smoothing help? *Advances in neural information processing systems*, 32.

[46] Nahida, H. et al. (2020). Ensemble of chexnet and vgg-19 feature extractor with random forest classifier for pediatric pneumonia detection. sn comput sci 1: 359.

[47] Nash, W., Drummond, T., and Birbilis, N. (2018). A review of deep learning in the study of materials degradation. *npj Materials Degradation*, 2(1):1–12.

[48] Nasrudin, M. W., Yaakob, N. S., Rahim, N. A. A., Ahmad, M. Z. Z., Ramli, N., and Rashid, M. S. A. (2021). Moment invariants technique for image analysis and its applications: A review. In *Journal of Physics: Conference Series*, volume 1962, page 012028. IOP Publishing.

[49] Nasteski, V. (2017). An overview of the supervised machine learning methods. *Horizons. b*, 4:51–62.

[50] Névéol, A., Deserno, T. M., Darmoni, S. J., Güld, M. O., and Aronson, A. R. (2009). Natural language processing versus content-based image analysis for medical document retrieval. *Journal of the American Society for Information Science and Technology*, 60(1):123–134.

[51] Nguyen, H. Q., Lam, K., Le, L. T., Pham, H. H., Tran, D. Q., Nguyen, D. B., Le, D. D., Pham, C. M., Tong, H. T., Dinh, D. H., et al. (2022). Vindr-cxr: An open dataset of chest x-rays with radiologist's annotations. *Scientific Data*, 9(1):1–7.

[52] Nixon, M. and Aguado, A. (2019). *Feature extraction and image processing for computer vision*. Academic press.

[53] Omarova, G. and Starovoitov, V. (2022). Application of the clahe method contrast enhancement of x-ray images.

[54] Peng, Y., Wang, X., Lu, L., Bagheri, M., Summers, R., and Lu, Z. (2018). Negbio: a high-performance tool for negation and uncertainty detection in radiology reports. *AMIA Summits on Translational Science Proceedings*, 2018:188.

[55] Pesapane, F., Codari, M., and Sardanelli, F. (2018). Artificial intelligence in medical imaging: threat or opportunity? radiologists again at the forefront of innovation in medicine. *European radiology experimental*, 2:1–10.

[56] Phung, V. H. and Rhee, E. J. (2019). A high-accuracy model average ensemble of convolutional neural networks for classification of cloud image patches on small datasets. *Applied Sciences*, 9(21):4500.

[57] Rajasenbagam, T., Jeyanthi, S., and Pandian, J. A. (2021). Detection of pneumonia infection in lungs from chest x-ray images using deep convolutional neural network and content-based

image retrieval techniques. *Journal of Ambient Intelligence and Humanized Computing*, pages 1–8.

[58] Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., et al. (2017). Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*.

[59] Rakshit, S., Saha, I., Wlasnowolski, M., Maulik, U., and Plewczynski, D. (2019). Deep learning for detection and localization of thoracic diseases using chest x-ray imagery. In *International Conference on Artificial Intelligence and Soft Computing*, pages 271–282. Springer.

[60] Ramteke, R. (2010). Invariant moments based feature extraction for handwritten devanagari vowels recognition. *International Journal of Computer Applications*, 1(18):1–5.

[61] Ramyachitra, D. and Manikandan, P. (2014). Imbalanced dataset classification and solutions: a review. *International Journal of Computing and Business Research (IJCBR)*, 5(4):1–29.

[62] Raoof, S., Feigin, D., Sung, A., Raoof, S., Irugulpati, L., and Rosenow III, E. C. (2012). Interpretation of plain chest roentgenogram. *Chest*, 141(2):545–558.

[63] Salau, A. O. and Jain, S. (2019). Feature extraction: a survey of the types, techniques, applications. In *2019 international conference on signal processing and communication (ICSC)*, pages 158–164. IEEE.

[64] Sarwinda, D., Paradisa, R. H., Bustamam, A., and Anggia, P. (2021). Deep learning in image classification using residual network (resnet) variants for detection of colorectal cancer. *Procedia Computer Science*, 179:423–431.

[65] Scholl, I., Aach, T., Deserno, T. M., and Kuhlen, T. (2011). Challenges of medical image processing. *Computer science-Research and development*, 26(1):5–13.

[66] Sharma, S. and Guleria, K. (2023). A deep learning based model for the detection of pneumonia from chest x-ray images using vgg-16 and neural networks. *Procedia Computer Science*, 218:357–366.

[67] Shehab, L. H., Fahmy, O. M., Gasser, S. M., and El-Mahallawy, M. S. (2021). An efficient brain tumor image segmentation based on deep residual networks (resnets). *Journal of King Saud University-Engineering Sciences*, 33(6):404–412.

[68] Shyni, H. M. and Chitra, E. (2022). A comparative study of x-ray and ct images in covid-19 detection using image processing and deep learning techniques. *Computer Methods and Programs in Biomedicine Update*, 2:100054.

[69] Silveyra, P., Fuentes, N., and Rodriguez Bauza, D. E. (2021). Sex and gender differences in lung disease. In *Lung Inflammation in Health and Disease, Volume II*, pages 227–258. Springer.

[70] Singh, D. and Singh, B. (2020). Investigating the impact of data normalization on classification performance. *Applied Soft Computing*, 97:105524.

[71] Suganyadevi, S., Seethalakshmi, V., and Balasamy, K. (2022). A review on deep learning in medical image analysis. *International Journal of Multimedia Information Retrieval*, 11(1):19–38.

[72] Sultan, H. H., Salem, N. M., and Al-Atabany, W. (2019). Multi-classification of brain tumor images using deep neural network. *IEEE access*, 7:69215–69225.

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

[74] Tan, M. and Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR.

[75] Tubiana, M. (1996). Wilhelm conrad röntgen and the discovery of x-rays. *Bulletin de l'Académie nationale de médecine*, 180(1):97–108.

[76] Van Engelen, J. E. and Hoos, H. H. (2020). A survey on semi-supervised learning. *Machine learning*, 109(2):373–440.

[77] Vargas, R., Mosavi, A., and Ruiz, R. (2017). Deep learning: a review.

[78] Wang, J., Zhu, H., Wang, S.-H., and Zhang, Y.-D. (2021). A review of deep learning on medical image analysis. *Mobile Networks and Applications*, 26(1):351–380.

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

[80] Weiss, K., Khoshgoftaar, T. M., and Wang, D. (2016). A survey of transfer learning. *Journal of Big data*, 3(1):1–40.

[81] Wong, K. C., Moradi, M., Wu, J., and Syeda-Mahmood, T. (2019). Identifying disease-free chest x-ray images with deep transfer learning. In *Medical Imaging 2019: Computer-Aided Diagnosis*, volume 10950, pages 179–184. SPIE.

[82] Xiao, C. and Sun, J. (2021). *Introduction to Deep Learning for Healthcare*. Springer Nature.

[83] Yamashita, R., Nishio, M., Do, R. K. G., and Togashi, K. (2018). Convolutional neural networks: an overview and application in radiology. *Insights into imaging*, 9(4):611–629.

[84] Yao, L., Poblenz, E., Dagunts, D., Covington, B., Bernard, D., and Lyman, K. (2017). Learning to diagnose from scratch by exploiting dependencies among labels. *arXiv preprint arXiv:1710.10501*.

[85] Yaseliani, M., Hamadani, A. Z., Maghsoodi, A. I., and Mosavi, A. (2022). Pneumonia detection proposing a hybrid deep convolutional neural network based on two parallel visual geometry group architectures and machine learning classifiers. *IEEE Access*, 10:62110–62128.

[86] Yoo, H.-J. (2015). Deep convolution neural networks in computer vision: a review. *IEIE Transactions on Smart Processing & Computing*, 4(1):35–43.

[87] Yousfi, Y., Butora, J., Khvedchenya, E., and Fridrich, J. (2020). Imagenet pre-trained cnns for jpeg steganalysis. In *2020 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–6. IEEE.

[88] Yuan, Z., Yan, Y., Sonka, M., and Yang, T. (2021). Large-scale robust deep auc maximization: A new surrogate loss and empirical studies on medical image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3040–3049.