Heriot-Watt University

Automated Detection of Lung Fibrosis in Chest X-ray Images: A Machine Learning Approach

Author:
Mark Monaghan

Supervisor:
Dr. Michael Lones

A report submitted in fulfilment of the requirements for the degree of MSc.Artificial Intelligence

in the

School of Mathematical and Computer Sciences

August 2024

Declaration of Authorship

I, Mark Monaghan, declare that this thesis titled, 'Automated Detection of Lung Fibrosis in Chest X-ray Images: A Machine Learning Approach' and the work presented in it is my own. I confirm that this work submitted for assessment is my own and is expressed in my own words. Any uses made within it of the works of other authors in any form (e.g., ideas, equations, figures, text, tables, programs) are properly acknowledged at any point of their use. A list of the references employed is included.

Signed: Mark Monaghan

Date: 15/08/2024

Abstract

This project explores fine-tuning machine learning (ML) models to automate fibrosis detection in CXRs. Multiple Vision Transformer (ViT) models, trained on Posterior to Anterior (PA), Anterior to Posterior (AP) and combined view datasets were developed, to find the most effective approach.

The ViT PA model showed significant challenges in generalisation, with a notable drop in performance on the test dataset across all key metrics. This indicates overfitting, making the model unreliable for clinical use. In contrast, the ViT AP model demonstrated high specificity but struggled with low sensitivity, which could lead to missed fibrosis cases, a critical limitation for clinical applications.

Interestingly, the ViT AP model outperformed the PA model across several metrics, raising questions about the underlying reasons for this outcome, typically unexpected given the clearer imaging associated with the PA view. The combined view model showed improved sensitivity but decreased performance in other metrics, highlighting the challenge when merging CXR datasets from different views.

A lack of diversity metadata in the CXR dataset used presents a limitation, as the model's performance across ethnically diverse populations remains unexamined. This issue must be addressed in future work to enable real-world model deployment.

The foundation laid by this research provides a pathway for further advancements in the automated detection of fibrosis in CXRs, emphasising the importance of collaborating with expert clinicians to validate model predictions and define clinical goals and patient outcomes, to guide future model development.

Acknowledgements

Dr. Michael Lones for his invaluable guidance and support throughout this project. His expertise and feedback have been critical in navigating the challenges and I am hugely grateful for his time to make the project the best it could be.

Amit Parekh helped me massively in the first semester in particular, when I felt hopelessly out of my depth. Those Monday night study sessions really helped me, I do not know if you realise how much.

Harry Addlesee for his frienship and encouragement throughout this course. It has been a pleasure to get to know you and I am grateful for the countless hours spent discussing ideas, challenges, and finding solutions. Thank you for being such a big part in how much I have enjoyed this year.

My parents and family for the fantastic start to life I was lucky enough to enjoy. Your continued support means so much to me.

My wife Gemma for her unwavering backing and encouragement to sign up for the course in the first place. You are a constant source of inspiration and your determination to succeed has been hugely motivating for me to keep improving and be the best version of myself. I cannot thank you enough.

My son Isaac and daughter Amelia, who continue to amaze me everyday and are a constant source of joy in my life "You are my sunshine".

Contents

Declaration of Authorship i

Abstract ii

Acknowledgements iii

Contents iv

List of Figures vii

List of Tables viii

Abbreviations ix

1 Introduction 1

2 Literature Review 3
2.1 Background . . . 3
2.2 NIH Chest X-ray Dataset . . . 4
2.3 Importance of Gender and Age in Assessing Medical Imagery . . . 5
2.4 Computer Aided Diagnosis and its Challenges . . . 6
2.5 Building Effective Models for Computer Aided Diagnosis . . . 8
2.6 Fine-Tuning vs. Creating Models from Scratch . . . 9
2.7 Image Preprocessing . . . 11
2.8 Convolutional Neural Networks for Computer Aided Diagnosis . . . 12
2.9 Visual Transformers for Computer Aided Diagnosis . . . 14
2.10 CNNs or Vision Transformers? . . . 16

3 Requirements Analysis 19
3.1 MoSCoW . . . 19
3.2 Functional Requirements . . . 20
3.3 Non-Functional Requirements . . . 21

4 Methodology 23
4.1 Approach . . . 23

4.2 Tools and Development Environment . . . 23
4.2.1 Hardware Configuration . . . 24
4.2.2 Programming Environment . . . 24
4.2.3 Libraries and APIs . . . 24
4.3 How to Detect Fibrosis in a CXR . . . 25
4.4 CXR PA and AP Views . . . 26
4.5 Dataset Analysis . . . 28
4.5.1 Label Distribution . . . 29
4.6 Data Partitions and Stratification . . . 30
4.7 Data Augmentation . . . 31
4.8 Data Legalities and Ethics . . . 33
4.9 Models . . . 33
4.9.1 Selecting and Fine-Tuning Models . . . 33
4.9.2 Model Development for PA and AP Views . . . 34
4.9.3 Experiments . . . 34
4.9.4 Baseline Hyperparameters . . . 36
4.9.5 Training Protocol . . . 36
4.9.6 Model Architectures and Hyperparameters Assessment . . . 36
4.9.7 Incremental Tuning Strategy . . . 36
4.9.8 Maximising Model Performance . . . 37
4.9.9 Creating Models from Scratch . . . 37
4.10 Evaluation . . . 38
4.10.1 Evaluation Metrics . . . 38
4.10.2 Balancing Class Importance . . . 40
4.10.3 Visualisation . . . 41

5 Professional, Legal, Ethical, and Social issues 42
5.1 Professional . . . 42
5.2 Legal . . . 43
5.3 Ethical . . . 44
5.4 Social . . . 45

6 Project Plan 47
6.1 Project Plan . . . 47
6.1.1 Project Timetable . . . 48
6.1.2 Gantt Chart . . . 49
6.2 Risk Analysis . . . 49

7 Results and Discussion 52
7.1 Model Development and Evaluation Results . . . 52
7.2 Measuring Model Performance . . . 52
7.3 Training and Validation Results . . . 53
7.3.1 Experiment 1: Establishing Model Performance Capabilities Using the PA View CXR Data . . . 53
7.3.2 Experiment 2: Establishing Model Performance Capabilities Using the AP View CXR Data . . . 56
7.3.3 Experiment 3: Establishing Model Performance Capabilities with Combined PA and AP View CXR Dataset . . . 59

7.4 Final Results 60

8 Conclusions **64**
8.1 Conclusions 64

9 Future Work **67**
9.1 Future Work 67
9.1.1 Enhancing the Model's Diagnostic Effectiveness 67
9.1.2 Facilitating the Model's Use in Clinical Decision-Making 68
9.1.3 Related Models Which Could Provide Benefit to CF Patients . . . 69

Bibliography **71**

List of Figures

2.1 The figure above is a conceptual rendering outlining the process of developing, validating, and deploying tailored machine learning tools that support bespoke medicine and scientific discovery in healthcare. 4
2.2 Parameter-Efficient Fine-Tuning Methods for Pre-trained Vision Models . 10
2.3 Overview of the IEViT model 15

4.1 Fibrosis CXR 25
4.2 Fibrosis CXR - 20 months later 26
4.3 AP v PA Projection 27
4.4 PA (left image) v AP (right image) View Quality 27
4.5 Fibrosis Label Distribution 29
4.6 All Label Distribution 30
4.7 3-Fold Cross-Validation 30
4.8 Confusion Matrix 38
4.9 Sensitivity and Specificity 39
4.10 Matthews Correlation Coefficient 40

5.1 AI Framework Roadmap 44

6.1 Project Timetable 48
6.2 Gantt Chart 49

7.1 Final Results Graph 61

List of Tables

3.1 Functional Requirements 21
3.2 Non-Functional Requirements 22

4.1 Benchmark Comparison of ViT Models and Weights 34

6.1 Risk Assessment 51

7.1 Metrics Abbreviation and Interpretation Table 53
7.2 Experiment 1 - Full Results for PA Experiment 55
7.3 Experiment 2 - Full Results for AP Experiment 57
7.4 Experiment 3 - Full Results for Combined PA and AP View 59
7.5 Final Results 61

Abbreviations

AI	Artificial Intelligence
AP	Anterior to Posterior
API	Application Programming Interface
AUROC	Area under the Receiver Operating Characteristic Curve
BB	Bounding Box
CAD	Computer-Aided Diagnosis
CF	Cystic Fibrosis
CFTR	Cystic Fibrosis Transmembrane Conductance Regulator
CNN	Convolutional Neural Network
CT	Computed Tomography
CTFC	Convolution and visual Transformer for Few-shot Chest X-ray images
CXR	Chest X-ray
DeiT	Data-efficient image Transformers
DPA	Data Protection Acta
FEV1	Forced Expiratory Volume
GDPR	General Data Protection Regulation
GPUs	Graphics Processing Units
IDE	Integrated Development Environment
IEViT	Input Enhanced Vision Transformer
MCC	Matthews Correlation Coefficient
MHRA	Medicine and Healthcare Regulatory Agency
MIL	Multiple Instance Leaning
ML	Machine Learning
MLP	Multi-Layer Perceptron
MoSCoW	Must, Should, Could, Would

NIH	National Institutes of Health
PA	Posterior to Anterior
PEFT	Parameter-Efficient Fine-Tuning
PLES	Professional, Legal, Ethical, and Social
PVM	Pre-Trained Vision Model
ROC	Receiver Operating Characteristic
ROI	Region Of Interest
SaMD	Software as a Medical Device
SPT	Sensitivity-aware visual Parameter-efficient fine-Tuning
VDSNet	Visual Geometric Group Data Spatial Transformer with CNN
ViT	Visual Transformer
ZCA	Zero Component Analysis

For my son, Isaac. Our little warrior, smashing every challenge put in your way, all while making us smile and laugh. May research like this help you live the brilliant life you deserve.

Chapter 1

Introduction

Cystic Fibrosis (CF) is a genetic condition affecting more than 10,900 people in the UK.[1] It is primarily diagnosed in the UK via the blood spot test, also known as the heel prick test, in newborn screening.

CF is often thought of as a lung disease, as it causes many lung symptoms. However, in reality, CF affects several parts of the body. The mutated gene that causes CF prevents the body from effectively moving salt and water in and out of cells. This leads to a build-up of thick mucus, especially in the lungs and digestive system.[2] Typically, people with CF will undertake a daily physio regimen, to help loosen and remove mucus that builds up in the lungs.

This mucus concentration in the lungs makes people with CF more susceptible to lung infection. Repeated lung infections can lead to lung damage. It is common for people with CF to require intravenous antibiotic treatment in hospital to combat this. People with CF can also suffer from other problems that affect their breathing and lungs, such as asthma, gastro-oesophageal reflux disease and aspiration (accidentally breathing something into the lungs). Occasionally, people with CF may get other more serious complications, such as a full or partially collapsed lung.

People with CF have their lung function closely monitored, via regular hospital appointments with specialist CF clinicians. Lung function tests are performed to measure how much and how quickly air is inhaled and exhaled from the lungs. Other lung tests check the volume of oxygen in the lungs and blood, to establish whether mucus in the airways is interfering with breathing and the absorption of oxygen. Chest X-rays (CXRs) are commonly used, at least annually, to look for signs of inflammation, infection and other lung complications. They also help clinicians monitor lung health over time.

[1] https://tinyurl.com/what-is-cf
[2] https://tinyurl.com/cf-lungs-impact

The rapidly expanding field of ML has created new opportunities in healthcare for enhancing diagnostic accuracy and efficiency. Developers have been able to build systems capable of analysing medical images, taking advantage of ML algorithms to detect subtle patterns and abnormalities in images, helping clinicians diagnose medical conditions and improve patient outcomes.

This project looks at many ML techniques used in computer vision to perform Computer-Aided Diagnosis (CAD), such as Convolutional Neural Networks (CNNs) and ViTs. In addition, it analyses how these techniques can use state-of-the-art optimisation, to support radiologists, provide equivalent diagnostic performance to a human expert and in some cases, surpass human capability.

While there are ML tools that have been developed to benefit the CF community, as discussed later in the report, much of the focus has been on using non-image data from the UK CF Registry. This work will focus on one of the most commonly taken medical images for people with CF, the CXR, with the objective of creating an ML model capable of automatically diagnosing fibrosis from a CXR, with the potential for the model to be deployed in a clinical setting in the future.

My motivation for this project is my son Isaac, who was diagnosed with CF when he was born. I would like to make a meaningful contribution to the advancement of CAD, to enhance the accuracy and efficiency of fibrosis detection in CXRs, thereby facilitating timely interventions by clinicians.

The Report includes a literature review to understand the current state of knowledge in performing CAD, particularly on X-rays, and identify any research gaps and opportunities for advancement. The requirements of the project are detailed, as well as the methodology to deliver them. As this project concentrates on medical data, there is a discussion of the professional, legal and ethical and social considerations. An analysis of the risks involved, and a project plan is included. The results of the project are presented, followed by conclusions that summarise the key findings and implications. The Report concludes with suggestions for future work to build upon the findings of this work.

Chapter 2

Literature Review

2.1 Background

Much of the existing ML based work on CF, has focused on using the data contained in the CF Registries in various countries, such as the UK.[1]

These registries hold static and longitudinal data for patients, including demographic information, Cystic Fibrosis Transmembrane Conductance Regulator (CFTR) genotype, and disease-related measures including infection data, comorbidities and complications, lung function, weight, intravenous antibiotic usage, medications and transplantations.

The Van der Schaar lab[2] has developed a number of ML tools using the UK CF Registry as its data source. Research examples include Prognostication and Risk Factors for Cystic Fibrosis via Automated Machine Learning [Alaa and van der Schaar, 2018], Dynamic-DeepHit: A Deep Learning Approach for Dynamic Survival Analysis With Competing Risks Based on Longitudinal Data [Lee et al., 2020] and Attentive State-Space Modelling of Disease Progression [Alaa and van der Schaar, 2019].

The lab has developed an excellent diagrammatic summary[3], shown in Figure 2.1, of the process for creating, validating and deploying ML tools that can yield new insights into the nature of conditions, such as CF.

[1] https://tinyurl.com/uk-cf-registry (accessed April 2024)
[2] https://tinyurl.com/vds-cf-projects (accessed April 2024)
[3] Source: https://tinyurl.com/spotlight-on-cf-projects (accessed April 2024)

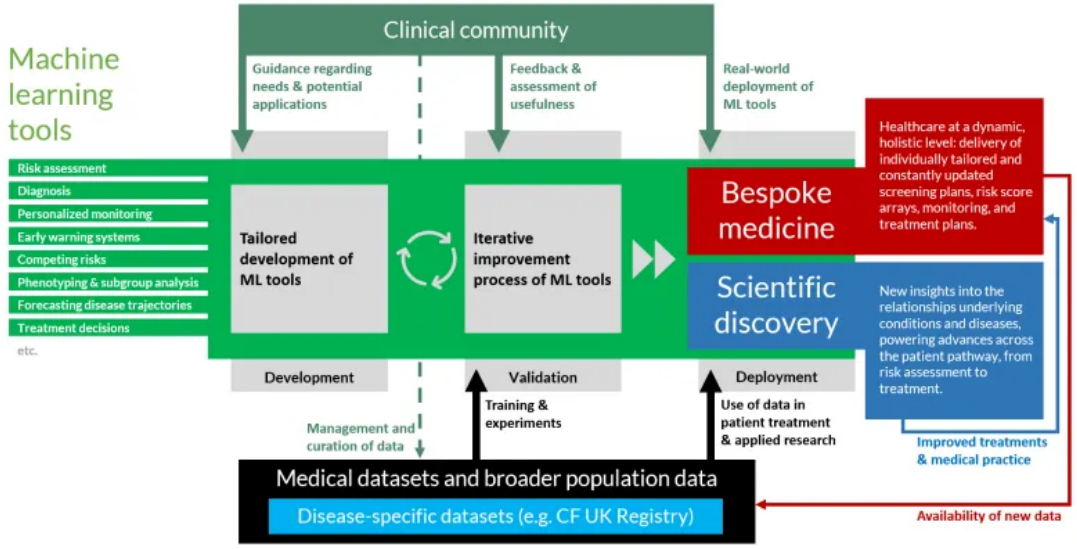

FIGURE 2.1: The figure above is a conceptual rendering outlining the process of developing, validating, and deploying tailored machine learning tools that support bespoke medicine and scientific discovery in healthcare.

While there has been work to diagnose lung fibrosis using ML and medical imaging, including [Bak et al., 2019], [Walsh et al., 2018], [Watadani et al., 2013] and [Lynch et al., 2005], it tends to have been conducted using computed tomography (CT) images, rather than CXRs. I know, from experience with my son who has CF, CXRs are used regularly, at least annually, by CF clinicians to gain an overall picture of lung health and look for deterioration over time. This is because fibrosis in the lungs is one of the most common causations impacting people with CF.

2.2 NIH Chest X-ray Dataset

In 2017, the National Institutes of Health (NIH) released the most comprehensive CXR dataset to date, comprising 112,120 X-ray images with diagnosis labels from 30,805 unique patients. Prior to the release of this dataset, Openi was the largest publicly available source of CXR images, with 4,143 images available [Wang et al., 2017].

Wang et al. facilitated research into CAD, described in more detail in the subsequent section. The dataset used eight different classes, corresponding to clinical symptoms that might be symptomatic of multiple conditions (definitions found at www.nih.gov, accessed March 2024):

1. Atelectasis - occurs when one or more areas of your lungs collapse or do not inflate properly.

2. Cardiomegaly - an enlarged heart.
3. Effusion - build up of fluid.
4. Infiltration - descriptive term for an abnormality on a CXR, giving little to no information about the diagnosis.
5. Mass - spot on a CXR.
6. Nodule - lump on a CXR.
7. Pneumonia - infection causing inflammation in one or both lungs.
8. Pneumothorax - collection of air outside the lung but within the pleural cavity.

In 2018, a further six classes were added to the image database:

1. Consolidation - the alveolar airspaces being filled with fluid, cells, tissue or other material.
2. Edema - caused by the extravascular movement of fluid into the pulmonary interstitium and alveoli.
3. Emphysema - enlargement of the airways distal to the alveoli due to the destruction of the alveolar walls.
4. Fibrosis - the thickening and/or scarring of connective tissue.
5. Pleural Thickening - typically involves the apex of the lung, which is called 'pulmonary apical cap'. On CXRs, the apical cap is an irregular density located at the extreme apex and is less than 5mm in width.
6. Hernia - happens when part of an internal organ or tissue bulges through a weak area of muscle.

One additional class existed for "No findings". Images were classified as "No findings" or one or more classes.

2.3 Importance of Gender and Age in Assessing Medical Imagery

Gender and age are critical in assessing medical imagery, particularly for respiratory conditions such as fibrosis. This is highlighted in a number of papers, including, Assayag

et al. [2020], Baratella et al. [2023], Kalafatis et al. [2019], Somayaji and Chalmers [2022], Tomoto et al. [2024]. Accounting for their observations, this work needs to be mindful of:

- Lung fibrosis may be more common or severe in older adults, due to the progressive lung degeneration experienced by people with CF. Age-related changes can also mimic or obscure fibrosis, affecting diagnosis and model performance.
- Lung fibrosis may have different prevalence rates or presentations between genders, which might affect how models generalise across genders.
- Ensuring training data represents different age groups and genders helps ML models learn to identify fibrosis across diverse populations. If a model is trained on data that is not representative, it may perform poorly on underrepresented groups.
- Upholding diversity in the validation and test datasets is essential for assessing the model's performance and generalisability. A balanced dataset helps ensure that performance metrics are not skewed by over- or under-represented groups.

2.4 Computer Aided Diagnosis and its Challenges

In CAD, ML models are used to analyse imaging and non-imaging data from patients to associate the extracted information with certain symptoms or a condition. The developed model is then used to predict the outcome of a new, unknown case when data from a new case is provided [Chan et al., 2020a].

If effectively trained and validated, CAD predictions may be used to provide supporting information in a clinicians decision-making process, or where appropriate, as a second opinion. The approach of using ML models to analyse patient data could be applicable to many patient care processes, such as disease or lesion detection, image classification, treatment planning and response assessment, and prognosis prediction. Frequently, imaging data can play a major role in each of these areas, and thus image analysis is a main component in CAD.

Chan et al. [2020a] discuss how one of the primary challenges in developing robust ML models for a given task, such as image classification, is collecting a sufficiently large and representative set of patient data of each class. This is necessary to allow the model to correctly identify the statistical properties of the images and assess any new unknown case from the same population. For robust training, the proportion of the classes should ideally be balanced, and thus classes of rare events will require even more effort to collect.

In the NIH dataset, of the fifteen available classification labels, fibrosis is one of the more rare labels, which may make it challenging to create a robust ML image classification model to reliably identify fibrosis.

Complementary approaches to using pure ML have been considered to diagnose abnormalities from CXRs. An example is the research conducted by Taylor et al. [2001] which developed a system to diagnose mammographic abnormalities from CXRs. It incorporated a knowledge base developed alongside radiologists and subject-matter experts to describe the characteristic features of benign and malignant calcifications. These descriptors were encoded as a set of image processing measures that characterise calcifications on digitised mammograms.

The research states the potential performance benefits for users if the rule or rules used in the classification of calcifications can be made explicit. While this is true, the study does not appear to consider the potential benefits of having a "chain of thought" on the system decision-making process to increase the likelihood of regulatory and clinical acceptance of such a system.

Using ML to act on medical data and guide clinical decisions poses challenges and raises important questions, such as, 'How will regulators evaluate medical ML to ensure its safety and effectiveness?' and 'What additional considerations should be taken into account in the international context?' [Minssen et al., 2020].

Great care must be taken to ensure ML can provide accurate results regardless of the subject's age, disabilities, ethnic origin, skin colour, or gender. False diagnoses or improper treatment recommendations present a real threat to patient health.

Having a system, as described by Taylor et al. [2001], which can trace the decision-making process, would facilitate fine-tuning of models which reach incorrect decisions or display bias. This would potentially give them an advantage in achieving regulatory and clinical acceptance over "black box" systems that provide no traceability.

It is not only the technical challenges of building a robust model that need to be considered. To take a model from a theoretical exercise and integrate it into a clinical setting there are many important other factors [Chan et al., 2020b].

Workflow efficiency is important in clinical practice, and it is crucial any efficiency gains from using ML tools are not lost when clinicians use CAD tools in a clinical environment as part of their day-to-day role.

Appropriate user training is necessary to ensure medics understand the capabilities and limitations of any CAD system. Clinicians' experience and level of enthusiasm for CAD

also strongly impact on whether they will accept a CAD tool and how they may respond to its recommendation.

Chan et al. also stress that performance standards and acceptance testing should be established to ensure the CAD tool can meet certain criteria before routine clinical use. These standards need to be monitored over time via quality assurance to ensure the consistency and accuracy of the CAD tool, preventing improper use that could impact patient safety.

Some of the key challenges for the future of CAD in medicine have been summarised by Yanase and Triantaphyllou [2019]:

1. Data collection and quality assessment.
2. Developing advanced segmentation approaches for medical imaging.
3. Developing advanced feature extraction selection approaches for image and signal analysis.
4. Developing better classification and other data mining approaches.
5. Dealing with big data.
6. Developing standardised performance assessment for CAD systems.
7. Adopting CAD systems for clinical practice.

2.5 Building Effective Models for Computer Aided Diagnosis

A survey on deep learning in medical image analysis was conducted by Litjens et al. [2017], summarising 308 papers applying deep learning to different applications, including image classification. One of its most interesting conclusions is the exact architecture of an ML model is not the most important factor in having an effective solution. Many researchers use the exact same model architectures, but have widely varying results. The paper references the Kaggle Diabetic Retinopathy Challenge, where participants were tasked with developing models to identify signs of diabetic retinopathy in eye images, as an example.

Another important factor identified in this wide-ranging survey by Litjens et al. is expert knowledge about the task to be solved can provide advantages over simply adding more layers to an ML model. This could be domain-specific understanding that humans

possess about image classification, which could be leveraged to improve the performance of ML models by informing feature selection, algorithm choice, data augmentation, and data and image preprocessing.

Groups and researchers that obtained good performance when applying deep learning algorithms often differentiated themselves in aspects outside the deep network, like novel data preprocessing or augmentation techniques.

Several researchers have showed that designing architectures incorporating unique task-specific properties can obtain better results than straightforward ML models. Two examples which Litjens et al. encountered several times in their extensive survey were multi-view and multiscale networks and the network input size and receptive field (i.e., the area in input space that contributes to a single output unit). Factors such as hyper-parameter optimisation provided additional, albeit smaller, performance gains.

2.6 Fine-Tuning vs. Creating Models from Scratch

The decision on whether to fine-tune an existing model or build a new model from scratch depends on factors such as the availability of data, computational resources, domain expertise, and the specific task, such as image classification.

Fine-tuning involves transfer learning, where you take an existing model that has been trained on a dataset, often on a different task, and retraining it on a smaller dataset or a related task [Matsoukas et al., 2022]. For this project, this involved taking an existing model trained on one of the largest (nonmedical) image datasets available, ImageNet, and then retraining it specifically for CXR image analysis.

Large-scale Pre-trained Vision Models (PVMs) have grown to billions or even trillions of parameters, as such, the standard full fine-tuning method is becoming unsustainable due to high computational and storage demands. Consequently, researchers have developed Parameter-Efficient Fine-Tuning (PEFT), which seeks to exceed the performance of full fine-tuning with minimal parameter modifications [Han et al., 2024]. The study by Han et al. explored a number of PEFT methods[4], shown in Figure 2.2.

[4]The diagram says "PEFT Methods for PLMs", I believe it should be "PEFT Methods for PVMs"

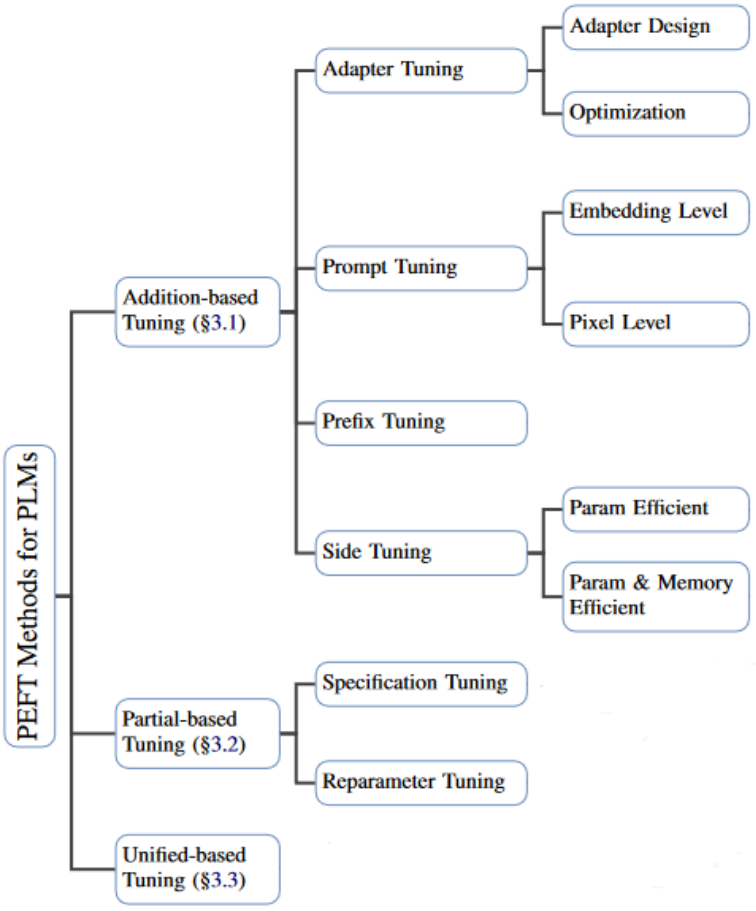

FIGURE 2.2: Parameter-Efficient Fine-Tuning Methods for Pre-trained Vision Models

The paper [Han et al., 2024] highlights AdaptFormer by Chen et al. [2022] as an example of how adapter design can be successfully applied to fine-tune pre-trained ViTs. AdaptFormer introduced lightweight modules that added less than 2% extra parameters to a ViT, and without updating the ViT's original pre-trained parameters, outperformed fully fine-tuned models on benchmark tasks. These modules proved they can be plugged into different ViTs and scaled to many visual tasks, such as image classification.

An extension to PEFT was developed by He et al. [2023], which studied where to introduce and how to allocate trainable parameters via Sensitivity-aware visual Parameter-efficient fine-Tuning (SPT). SPT identified the sensitive parameters that required tuning and boosted the representational capability for the weight matrices whose number of sensitive parameters exceeded a pre-defined threshold. During experimentation, He et al. showed SPT is complementary to existing PEFT methods and largely boosts their performance. In one experiment SPT improved Adapter mean Top-1 accuracy with supervised pre-trained ViT-B/16 by 4.2%.

Building a model from scratch enables complete customisation of the model [Weidman, 2019]. This may be particularly important when a dataset is highly specialised or unique,

to capture relevant features in the data or image effectively. However, it requires expertise in model architecture design, hyperparameter tuning, and optimisation techniques. Building and training a model from scratch often requires a larger dataset compared to fine-tuning, and can be computationally intensive, especially for complex models or large datasets. Fine-tuning is often preferred when computational resources are limited or when pre-trained models are available and suitable for the task, and a customised model is not necessary.

2.7 Image Preprocessing

Data preprocessing is an integral step in ML. A commonly cited concept in computer science is "garbage in, garbage out". This also applies to CAD, as the quality of the images will have a direct impact on the ability of a model to be trained effectively and produce an accurate diagnosis.

Commonly used techniques in image preprocessing are resizing, noise reduction, normalisation, binarisation and contrast enhancement. The CAD system developed by Heidari et al. [2020] built and tested a new model to detect COVID-19 infected pneumonia in CXRs. Researchers applied two image preprocessing steps to remove the majority of diaphragm regions and process the original image using a histogram equalisation algorithm as well as a bilateral low-pass filter. Then, the original image and two filtered images were used to form a pseudo colour image.

This newly created image was fed into three input channels of a transfer learning-based model to classify CXR images into three classes of COVID-19 infected pneumonia, other community-acquired non-COVID-19 infected pneumonia, and normal (non-pneumonia) cases. Using the two prepossessing steps improved the accuracy of the CAD model by 12%.

Additional preprocessing techniques for image classification were explored by Pal and Sudeep [2016]. The CIFAR10 dataset was used for these experiments; 60,000 coloured images divided into 10 classes: airplane, automobile, bird, cat, deer, dog, frog, horse, ship, and truck.

Mean normalisation, standardisation and zero component analysis (ZCA) were explored, with ZCA achieving the best performance increases for CNNs. ZCA transformation makes the edges of objects more prominent, and the convolutional layers detect various features through the feature maps based on these edges.

The variability of anatomical features in CXRs, such as varying lung shape, strong edges of the rib cage and visible shape of the heart, make the detection of the lung regions challenging. The study by Candemir et al. [2012] presented a graph cut based lung segmentation method that has two main stages - average lung shape model calculation, and lung boundary detection based on graph cut. By constructing a graph where pixels are represented as nodes and edges connecting neighbouring pixels, graph cut algorithms can efficiently find the optimal partitioning of the image into segments by minimising the cut cost.

A set of training masks were used to learn the lung shape model. Rather than using all the training masks, the training set was based on a shape similarity score in order to increase the lung shape model accuracy. Once the training masks were chosen, the approximate shape model was calculated by taking the average of the selected masks. The computed lung shape model was a probabilistic model in which each pixel intensity was the probability of the pixel being part of the lung field.

The second stage of the system detected the lung boundary of the CXR using image properties and the lung shape model from the first step. Image segmentation was performed using a cut graph method, to find the global minimum which corresponds to foreground and background labelling of the pixels.

However, using a static shape model to establish the lung regions is limited, given the variability of different lung shapes in people. Even with this limitation, when evaluated on a data set of frontal CXRs from the Japanese Society of Radiological Technology, the system achieved around 91% segmentation accuracy. This compared favourably with state-of-the-art algorithms (95%) and human-observer scores (94%).

2.8 Convolutional Neural Networks for Computer Aided Diagnosis

A CNN is a type of deep learning neural network architecture commonly used in computer vision. The network consists of multiple layers, namely the input layer, convolutional layer, pooling layer, and fully connected layers. The convolutional layer applies filters to the input image to extract features, the pooling layer downsamples the image to reduce computation, and the fully connected layer makes the final prediction. The network learns the optimal filters through backpropagation and gradient descent.[5]

A study by Candemir et al. [2018] used CNN models for automatic detection of cardiomegaly. Two CXR datasets were used, the same one I will use (NIH CXRs) and a

[5]https://www.geeksforgeeks.org/introduction-convolution-neural-network (accessed April 2024)

CXR collection from Indiana University, which was significantly smaller. A two-stage approach was employed. Firstly, applying a pre-trained CNN and then fine-tuning with cardiomegaly CXRs. In the second part, a deep CNN architecture (VGG-16) was trained with the full NIH CXR dataset and fine-tuned with cardiomegaly CXRs.

In the first experiment, the pre-trained models used were AlexNet, fine-tuned VGG-16, fine-tuned VGG-19 and InceptionV3. Fine-tuned VGG-16 was the most accurate (correctly predicting cardiomegaly in the CXR), fine-tuned VGG-19 had the highest sensitivity (finding abnormal as abnormal) and InceptionV3 has the highest specificity (finding normal as normal).

In the second experiment, using a VGG-16 CNN architecture, a model trained with ImageNet achieved a higher sensitivity, compared with the CXR-based pre-trained model, However, the CXR-based model achieved a higher accuracy and specificity.

This experiment was particularly interesting as one of the datasets used, NIH CXRs, is the one I will use in my study. The VGG-16 CNN architecture proved the most effective in this study, which is supported by other papers showing impressive ML image classification results in healthcare using a VGG-16 CNN [Kaur and Gandhi, 2019a, Sharma and Guleria, 2023].

Another study which used the NIH CXR dataset (and the COVID-19 Image Data Collection [Cohen et al., 2020]) to detect COVID-19 using pre-trained CNNs was by Horry et al. [2020]. They found both VGG-16 and VGG-19 classifiers provided good results within the experimental constraints of the small number of available COVID-19 X-Ray image samples at that time.

Kaur and Gandhi [2019b] successfully performed brain image classification using the same model architecture and transfer learning. Studies by Belaid and Loudini [2020] and Singh et al. [2021] used pre-trained VGG-16 CNNs to effectively classify different types of brain tumours. Outside healthcare, an experiment by Tammina [2019] used the VGG-16 CNN model and transfer learning successfully for image classification on images of cats and dogs.

Multiple studies have been conducted to compare CNN architectures for disease detection based on medical image classification. One such study, by Bressem et al. [2020] used two datasets, the CheXpert and COVID-19 image data collection, to compare different deep learning architectures for classification of CXRs. They concluded deeper CNNs do not necessarily perform better than shallow networks. Instead, an accurate classification of CXRs was achieved with comparably shallow networks, such as AlexNet (8 layers), ResNet-34 or VGG-16.

In the comparison of multiple CNN algorithms for feature extraction on images to detect glaucoma, by Sunanthini et al. [2022], ResNet-101, VGG-19, AlexNet and GoogleNet were the most accurate, with Inception ResNet-V1 providing the highest sensitivity. In this study, VGG-16 showed comparatively modest results.

Yadav and Jadhav [2019] found the best results came from the transfer learning of VGG-16 (with one retrained ConvLayer), when compared with the InceptionV3 model when applying CNN-based classification to a small CXR dataset.

While VGG-16 has proved an effective CNN model across multiple studies, its use is not the only factor that contributes to an effective image classification model. Other considerations, such as how an existing model is fine-tuned using transfer learning and image preprocessing, are vital in optimising a model for best results.

From this, I conclude that using a pre-trained VGG-16 CNN architecture would be appropriate for performing image classification to detect fibrosis in the NIH CXR dataset.

2.9 Visual Transformers for Computer Aided Diagnosis

Alternative models to CNNs are the ViT and the Data Efficient Image Transformer (DeiT). These models modify the fully transformer-based models that have revolutionised the field of Natural Language Processing (NLP) to computer vision tasks.

ViTs have been used to diagnose lung-related pneumonia and COVID-19 from CXRs. A study by Ukwuoma et al. [2022] used an ensemble technique to derive rich features from the CXR, then second-order pooling was used to derive higher global features of the images. The images were then separated into patches with positional embedding before analysing the patches individually via a ViT.

The model proved highly successful, yielding 96.01% sensitivity, 96.20% precision, and 98.00% accuracy for the COVID-19 dataset. For the Covid-ChestX-ray-15k dataset, 97.84% accuracy, 96.76% sensitivity and 96.80% precision was achieved.

The model compared favourably to state-of-the-art models from Ibrahim et al. [2021], which used a pretrained AlexNet model for automatic detection of COVID-19 pneumonia, non-COVID-19 viral pneumonia, and bacterial pneumonia, using COVID-19 datasets from Kaggle, GitHub and images made available by Kermany et al. [2018].

It was also more effective than the deep CNN model for detecting the presence of pneumonia in CXRs by Naralasetti et al. [2021], which used the Kaggle Pneumonia Identification Challenge dataset and a new CXR dataset contributed by Ge et al. [2019].

In addition to the new ensemble technique, the effectiveness of the Ukwuoma et al. [2022] model highlighted the relevance of deep learning feature extraction and importance of hyperparameter adjustment in processing new data, before adding more complicated architectures to the model.

Another interesting element of the study by Ukwuoma et al. [2022] was using the model's explainability-driven heatmap visualisation to emphasise the key aspects influencing the prediction decision. These heatmaps could be used to help explain the decisions the model makes and benefit radiologists in diagnosis.

Okolo et al. [2022] developed an improved ViT architecture for CXR classification, the Input Enhanced Vision Transformer (IEViT)[6], shown in Figure 2.3.

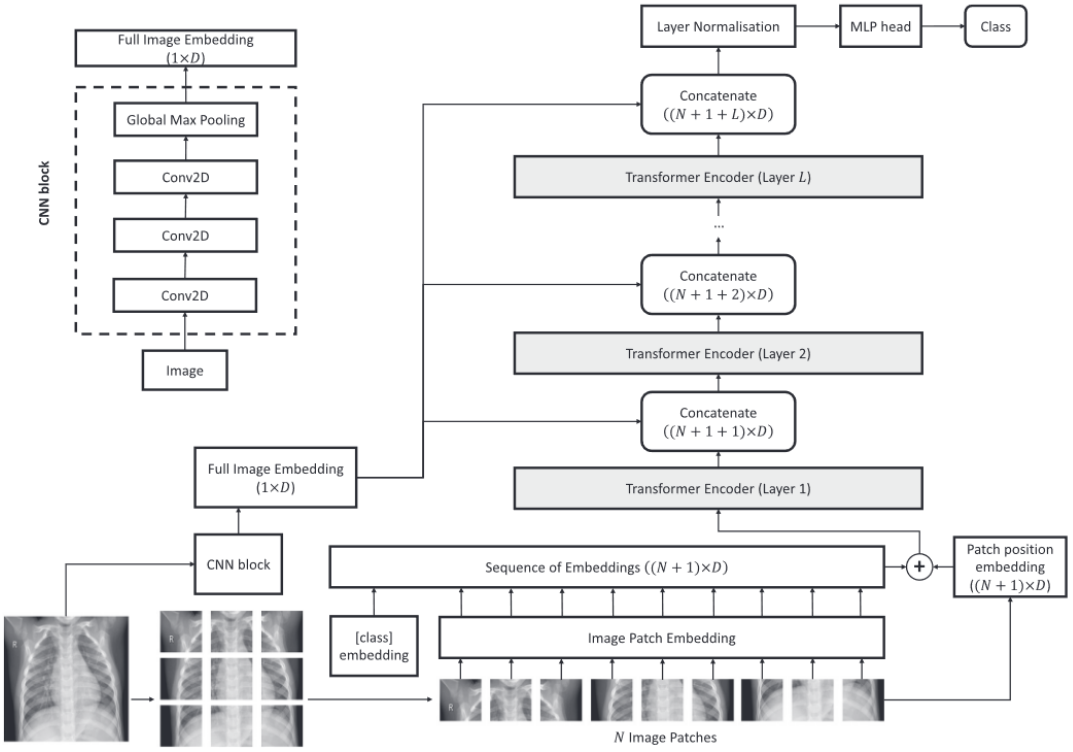

FIGURE 2.3: Overview of the IEViT model

IEViT was partially motivated by the ResNet architecture, which introduced the skip connection, essentially adding the original input to the output of each convolution block. In IEViT a representation of the original input image is iteratively added to the output of each transformer encoder layer.

This was achieved by first designing a convolutional block in parallel with the ViT network. As an input, the CNN block took the whole input image and output an embedding of the whole image, which is then iteratively concatenated to the output of

[6]https://tinyurl.com/IEViT-model (accessed April 2024)

each transformer encoder layer. Thus enabling the network to always "remember" the full image at the end of each transformer block output.

The proposed IEViT models, as well as the original ViT models, for the B/16, B/32, L/16, and L/32 variants, were evaluated using several CXR datasets [Chowdhury et al., 2020, Rahman et al., 2020, 2021, Wang et al., 2020] and `https://tinyurl.com/OCT-CXR-images` (accessed April 2024).

The results from the experiments show the addition of the convolutional block in parallel with the ViT network led to a consistently improved classification performance of IEViT over ViT for all the examined data sets and variants.

To provide additional assurance of the effectiveness of the new proposed model design, experiments were run with a model with a CNN block and Multi-Layer Perceptron (MLP) head. This provided further evidence that CNN variants performed significantly worse than the respective IEViT and ViT variants.

ViTs have shown robust performance in CAD, including image classification, particularly when compared to CNNs. The IEViT proved even more capable, incorporating a convolutional block in parallel with the ViT network, thus enabling the network to always "remember" the full image at the end of each transformer block output.

2.10 CNNs or Vision Transformers?

CNNs have reigned for over a decade as the de facto approach to automated medical image diagnosis. ViTs have emerged as an alternative to CNNs, yielding similar or better levels of performance while possessing several interesting properties, such as attention, that could prove beneficial for medical imaging tasks [Matsoukas et al., 2021].

The study by Matsoukas et al. [2021] investigated whether ViTs can be substituted as a plug-and-play alternative for CNNs for medical diagnosis tasks. To ensure a fair and interpretable comparison, RESNET50 was chosen as the representative CNN model, with DEIT-S with 16*16 tokens as the ViT.

It found that while CNNs performed better when trained from scratch, off-the-shelf ViTs using default hyperparameters are on par with CNNs when pretrained on the ImageNet dataset, and outperform their CNN counterparts when pretrained using self-supervision.

Two benefits of ViTs the research highlighted are that as the number of data samples grew, the margin between ViTs and CNNs is also expected to grow, and, built-in high-resolution saliency maps can be used to better understand the model's decisions.

This explainability could be a crucial differentiator in achieving clinical and regulatory acceptance for ViTs in clinical practice.

A study by Murphy et al. [2022] for Visual Transformers and CNNs for disease classification on CXRs, compared performance, sample efficiency and hidden stratification. It used fine-tuned Data-efficient image Transformers (DeiT) and CNN classification models pretrained on ImageNet using the NIH CXR and MURA datasets.

DeiT-B and DeiT-Ti (Tiny) both performed slightly worse than all CNNs for CXR classification tasks, but had similar sample efficiency. DeiT was less susceptible to certain hidden stratification. Ultimately, the report summarised, there was no clear performance advantage in using DeiT-B over DenseNet121 for classification tasks.

Chen et al. [2021] provided a comprehensive review of multiple image classification methods using the ImageNet and CIFAR datasets. The models reviewed ranged from the AlexNet CNN model in 2012 to the ViT-G/14 Vision Transformer and CoAtNet CNN + Transformer in 2021.

The general trend of results was the earlier CNNs and models have lower accuracies, when compared with modern ViTs such as ViT-H/14 and ViT-G/14. The most effective model from this study was CoAtNet, which interestingly used mixed methods, both convolutional and transformer.

Jin et al. [2021] also supported a mixed convolution and ViT model, Convolution and visual Transformer for Few-shot Chest X-ray images (CTFC). The paper proposed a new feature extractor structure, specifically used to extract the features of few-shot sample CXR images from the NIH CXR dataset. The feature extractor combined the advantages of convolution and vision transformers to make the extracted features contain more information. In comparison with a pure CNN or ViT, the CTFC model obtained the highest accuracy across three sets of comparative experiments.

Krishnan and Krishnan [2021] proposed an approach using pretrained models, fine-tuned for detecting the presence of COVID-19 disease on CXRs. The COVID-19 X-ray database and COVID-19 pneumonia normal CXR PA dataset from Kaggle were used. The DenseNet, InceptionV3, WideResNet101 CNNs and, ViT-B/32 models used a transfer learning process, where the model was pre-trained on the ImageNet dataset.

In the experiments the highest average precision, average recall, average F1 score and test accuracy were achieved by the ViT-B/32 model. The ViT was also much quicker to train, taking around ten minutes, compared with around thirty-five minutes for CNNs.

Uparkar et al. [2023] conducted a particularly interesting study as it focused on detecting lung diseases using the NIH CXR dataset, the same dataset being used in this project.

The study used an off-the-shelf ViT-based approach and compared it with a CNN-based hybrid deep learning model that outperformed existing deep learning techniques. The hybrid deep learning model used for comparison was called Visual Geometric Group Data Spatial Transformer with CNN (VDSNet). The ViT model, after appending additional parameters to the last layer, performed better compared to VDSNet.

These experiments highlighted a number of different considerations when comparing the performance of ViTs and VDSNet. While ViTs performed better as the patch size of the input image decreased and the number of internal layers increased, the higher the number of layers led to more parameters to train, which increased the training time of the network. Consequently, the network required higher computational power, and the additional performance gains had to be considered against these potential downsides. Similar to Matsoukas et al. [2021], this research highlighted the benefit of built-in high-resolution saliency maps for ViTs that can be used to better understand the model's decisions.

Explainability of ViTs to detect COVID-19 in CXRs was also explored by Chetoui and Akhloufi [2022] and Mondal et al. [2022]. They visualised the most important signs of COVID-19 and the opacity of the lungs detected by the ViT using attention maps.

This facet is a clear positive differentiator for ViTs, compared with CNNs, as the features learned by transformer networks are explainable, helping to focus on meaningful regions in the images for detection and aiding in localisation of the infected area.

Chapter 3

Requirements Analysis

In any development project, requirements analysis is crucial in identifying, documenting and validating what the project needs to deliver to meet the intended purpose effectively. The requirements are essential for steering all phases of the project, including design, development and evaluation.

Functional requirements outline the specific features, capabilities, and operations expected from the system to automate detection of lung fibrosis in CXRs using ML. They describe what the system should do, how it should behave, and what actions it should enable.

Non-functional requirements address aspects such as performance, security and usability. Additionally, any time constraints need to be considered to ensure the system is feasible within the required timeframe.

3.1 MoSCoW

To assess the value of project requirements, they can be ranked using the MoSCoW (Must, Should, Could, Would) analysis technique [Kravchenko et al., 2022].

- **Must** - must be satisfied in order for the final solution to be considered a success.
- **Should** - high-priority items that should be included in the solution if it is possible.
- **Could** - considered desirable but not necessary.
- **Will not** - will not be implemented, but may be considered in the future.

3.2 Functional Requirements

Requirement	**MoSCoW**
The system **must** be able to process the CXRs in the NIH dataset	M
The system **must** must preprocess the NIH CXR images to enhance their quality, remove noise, and standardise image resolution and orientation	M
The system **must** accurately distinguish between normal lung tissue and fibrotic lesions within the ROI	M
The system **must** extract relevant features from the segmented lung regions to represent key characteristics indicative of fibrosis	M
The system **must** train an ML model using annotated CXR images to learn the patterns associated with fibrosis	M
Any ML model built or fine-tuned as part of the project **must** must be evaluated to establish its effectiveness in detecting fibrosis in CXRs	M
The evaluation of any model **must** use appropriate performance metrics, depending on whether the classes in the dataset are balanced	M
The final version of a model's performance **must** be validated using an independent (holdout) dataset	M
The dataset **must** have the same patient in the training, validation or test dataset	M
The project **must** fine-tune existing ML models to establish which are the most effective at detecting fibrosis in CXRs	M
The system **must** generate automated outputs summarising its findings from analysing the CXR, the presence or absence of fibrosis	M
The project **should** fine-tune existing ML models, specifically trained on X-rays, to establish which are the most effective at detecting fibrosis in CXRs	S
The project **should** compare fine-tuned ML models, trained on nonmedical images and trained on X-rays, to establish which are the most effective at detecting fibrosis in CXRs	S
The project **should** compare fine-tuned models trained on nonmedical images, fine-tuned models trained on X-rays and a ViT created from scratch, to establish which is the most effective at detecting fibrosis in CXRs	S

The system **should** be able to segment lung regions to identify the region of interest (ROI) containing potential fibrotic abnormalities	S
The system **could** provide visualisation tools, such as saliency maps, that highlight the fibrotic lesions identified by the model, to gain insight into the model's behaviour and allow clinicians to explore and interpret model predictions	C
The project **could** create a ViT to perform CAD to detect fibrosis in CXRs, training it on the full NIH dataset, and then fine-tuning it on the fibrosis labelled images	C
The system **could** provide output metrics, such as probability scores or confidence levels, to assess the reliability of the model's predictions	C
The system **will not** integrate with existing healthcare systems	W
The system **will not** incorporate semi-supervised learning to iteratively improve model performance using unlabelled data	W
The system **will not** implement mechanisms for continuous model monitoring following changes to the dataset	W

Table 3.1: Functional Requirements

3.3 Non-Functional Requirements

Requirement	**MoSCoW**
The system **must** comply with relevant healthcare data privacy and security regulations in the region in which it is deployed, such as the General Data Protection Regulation and Data Protection Act, for the UK	M
The system **must** provide accurate results regardless of the subject's age, disabilities, ethnic origin, skin colour, or gender	M
The system **must** ensure cross-platform compatibility, allowing seamless deployment on different hardware platforms, operating systems, and cloud infrastructure	M
The system **must** have minimal latency when processing CXRs and making its diagnosis	M
At the start of the project, a holdout dataset **must** be created by partitioning a subset of the data	M

The holdout dataset **must not** be used, other than to measure the generality of the final model once all experiments are complete	M
The system **should** be scalable to accommodate increasing CXR volume over time	S
The system **should** achieve high accuracy in detecting fibrosis in CXRs, with low false positive and false negative rates to minimise diagnostic errors	S
The system **should** be accessible to users with diverse abilities, ensuring compliance with accessibility standards	S
The system **should** be well-documented to enable system support and future development, covering system architecture, components and their interactions, dependencies, data preprocessing techniques, model training processes, evaluation metrics, model deployment, troubleshooting and usage instructions	S
The system **should** have an intuitive and user-friendly interface, enabling clinicians to upload CXRs easily, and interpret model predictions effectively	S
The system **could** handle a large volume of CXRs concurrently, processing multiple images in parallel to meet clinical demand	C
The system **will not** have the ability to scale resources dynamically	W
The system **will not** have mechanisms in place to automatically recover from errors, crashes, or data inconsistencies	W

Table 3.2: Non-Functional Requirements

Chapter 4

Methodology

4.1 Approach

The project aimed to build a state-of-the-art ML model that can be used to detect fibrosis in CXRs. While mixed method CCNs and ViTs have shown effective performance in some studies (Chen et al. [2021] and Jin et al. [2021]), there is greater evidence that ViTs are more effective in medical image classification when compared with CNNs or a hybrid approach. This is particularly the case when an IEViT model is employed.

ViT and IEViT models also have the additional benefit of enhanced explainability, when compared with CNNs, from built-in high-resolution saliency maps that can be used to better understand a model's decisions. This in turn increases the likelihood of clinical and regulatory acceptance.

This research prioritised building ViT ML models to identify fibrotic abnormalities, distinguishing between normal lung tissue and fibrotic lesions. Pre-existing foundation models, trained on nonmedical images, were fine-tuned to determine which are most effective at detecting fibrosis in CXRs.

4.2 Tools and Development Environment

The development and experimentation was conducted using Google Colab, with a Colab Pro+ subscription, a cloud-based platform for running Python code and executing ML workflows. The following is a summary of the tools, hardware configuration, programming environment, and Application Programming Interfaces (APIs) used throughout the project.

4.2.1 Hardware Configuration

Google Colab resources are not guaranteed, with dynamic hardware availability. The following hardware options are indicative of ones assigned during the project.

- GPUs: A100, L4 and T4
- System RAM: 83.5GB
- GPU RAM: 40GB
- Disk: 200GB

4.2.2 Programming Environment

- Python Version: 3.10.12.
- Integrated Development Environment (IDE): Google Colab served as the primary IDE, offering a Jupyter notebook interface for code execution and visualisation in a single environment.

4.2.3 Libraries and APIs

In addition to the default libraries and versions provided by Google Colab, the following libraries and APIs were used in this project to enhance functionality and support the development and evaluation of ML models.

- Matplotlib v3.7.1: visualisation tool.
- NumPy v1.26.4: package for numerical computing in Python, providing support for arrays, matrices, and a large library of mathematical functions.
- Pandas v2.1.4: data manipulation and analysis library, offering data structures like DataFrames, commonly used in ML.
- Pathlib v1.0.1: module for object-oriented filesystem paths.
- PyTorch v2.3.1: ML library based on the Torch library, used for applications such as computer vision. Provides tools for tensor computation.
- Torchmetrics v1.4.1: library that provides a standardised interface for calculating and aggregating various ML metrics.

- Torchvision v0.18.1: an extension of PyTorch that provides tools specifically for computer vision tasks, including pre-trained models, image datasets, and utilities for image processing.
- Tqdm v4.66.5: provides fast, extensible progress bars for loops and other iterable objects
- WandB v0.17.6: Weights & Biases is an external API for experiment tracking, model monitoring, metrics logging and results visualisation.

4.3 How to Detect Fibrosis in a CXR

To better understand the way in which the model will need to identify fibrosis in CXRs, some research was conducted to understand how radiologists perform this task. The Radiology Masterclass website[1] provides online medical imaging education resources for medical students, doctors, and health care professionals.

The image below, in Figure 4.1, highlights reticular (net-like) shadowing as symptomatic of fibrosis in a CXR, where there is shadowing of the lung peripheries, typically more prominent towards the lung bases. This can also cause the contours of the heart to be less distinct or "shaggy".

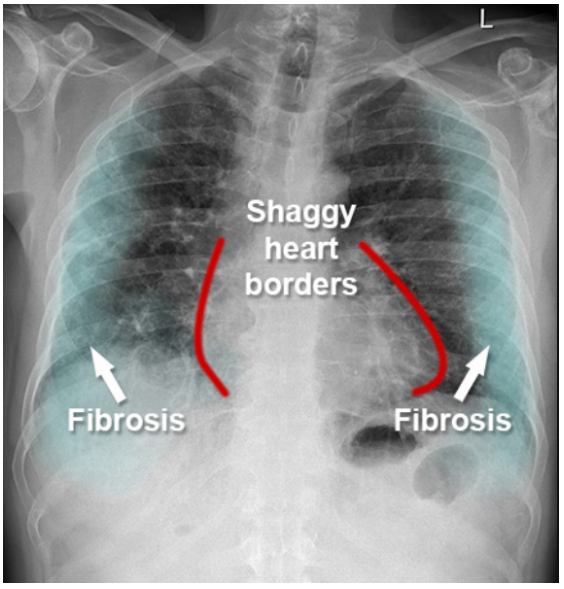

FIGURE 4.1: Fibrosis CXR

The image in Figure 4.2 is of the same patient's lungs, twenty months later, showing disease progression as the fibrosis becomes more widespread, leading to lung volume loss.

[1] https://www.radiologymasterclass.co.uk/

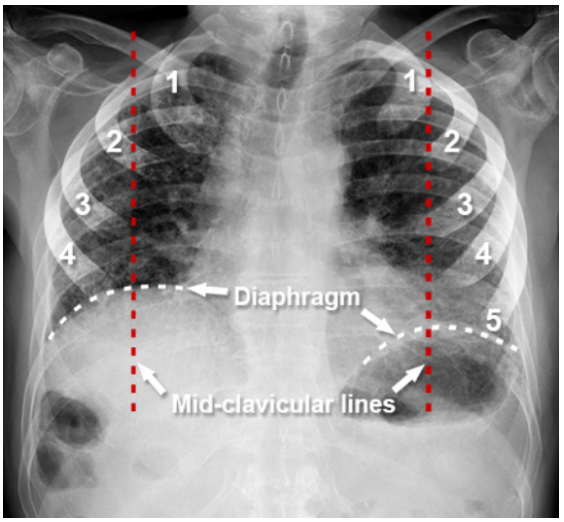

FIGURE 4.2: Fibrosis CXR - 20 months later

While the models developed as part of this project do not specifically take into account progression of fibrosis over time, it highlights two different example fibrosis CXRs, which the model needs to identify.

4.4 CXR PA and AP Views

The standard CXR is acquired with the patient standing up, and with the X-ray beam passing through the patient from Posterior to Anterior (PA). The image produced is viewed as if looking at the patient from the front, face-to-face. The heart is on the right side of the image as you look at it. PA views are of higher quality and more accurately assess heart size than Anterior to Posterior (AP) images. However, sometimes it is not possible for radiographers to acquire a PA CXR. This is usually because the patient is too unwell to stand. Figure 4.3 shows the difference in the PA and AP projections.

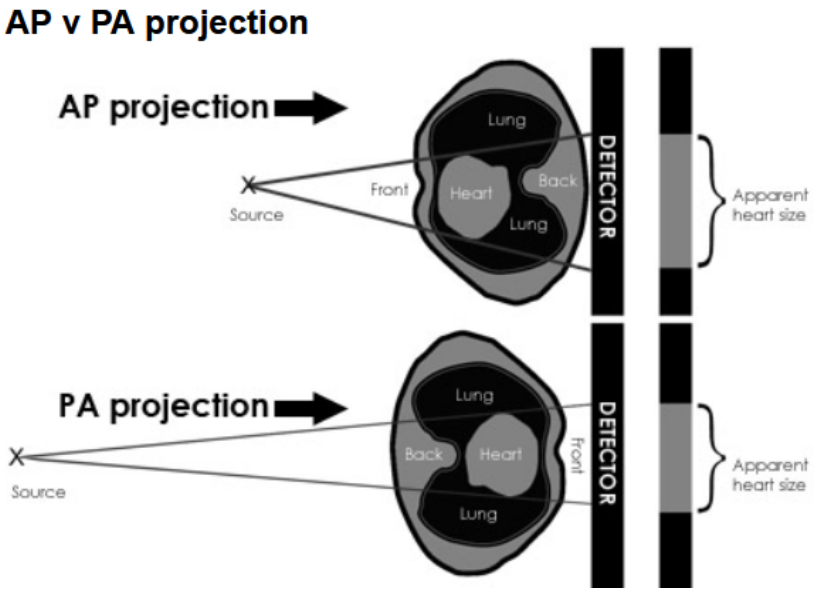

FIGURE 4.3: AP v PA Projection

In order to take a PA view the patient places their arms around the side of the detector plate, or stands with hands on hips. This ensures the scapulae (shoulder blades) are rotated laterally and do not overlap the lungs. In AP views the scapulae are not retracted laterally, and remain projected over each lung. This obstructed AP view of the lung may hinder a model's ability to detect fibrotic lung regions. The two images below in Figure 4.4 highlight the challenge.

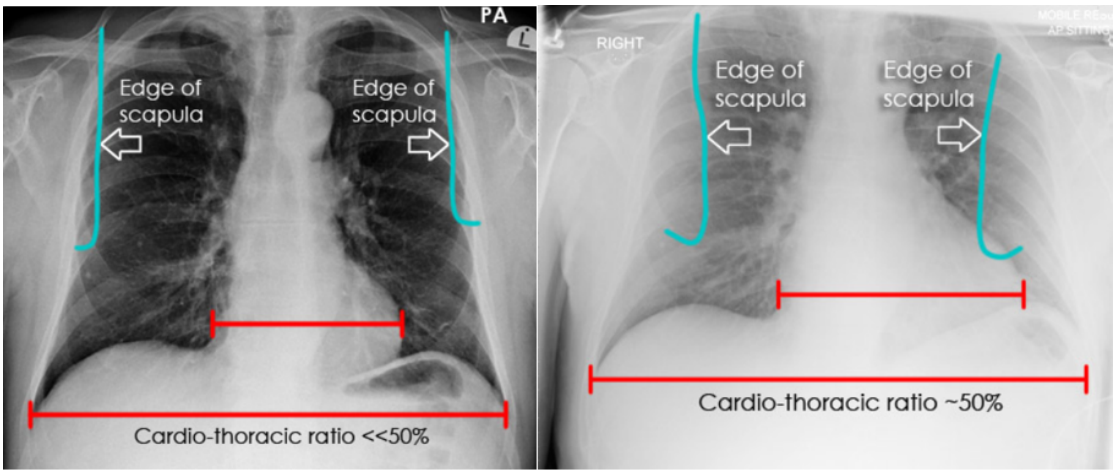

FIGURE 4.4: PA (left image) v AP (right image) View Quality

The clear difference in quality of the two images shows the challenges the model may encounter when trying to perform fibrosis diagnosis on the AP CXRs in particular. Creating two different models, one for each view, allows for specialised optimisation with tailored feature extraction and reduced complexity, as each model only needs to learn features relevant to a single view. It also facilitates incremental improvements, where enhancements can be made to one model without impacting the other, allowing

for more controlled experimentation and improvement. This comes at a cost of increased computational resources and training time, increased complexity in deployment and maintenance, and the need for sufficient training data for both views.

The potential limitations of having two models could be overcome by various techniques, such as transfer learning to reduce training time and computational resources, a unified inference pipeline that can handle both models and using MLOps tools and frameworks for continuous integration and deployment to automate the deployment, monitoring, and maintenance of the models. Data could be augmented to ensure both models have sufficient and balanced data for training.

More advanced techniques could also be employed, for example, multitask learning, where a single model is trained to handle both views but with separate heads or branches for PA and AP view-specific tasks, like performing fibrosis diagnoses. This can leverage shared representations while still specialising for each view. Knowledge distillation could be implemented, where a single "teacher" model trained on both views can transfer knowledge to smaller, specialised "student" models for each view, reducing the need for extensive separate training. Ensemble methods could enhance robustness and accuracy, where predictions from both models are combined to make a final decision.

This project will explore the impact of having separate PA and AP models to perform fibrosis diagnosis.

4.5 Dataset Analysis

In the NIH CXR dataset there are fifteen classifications, fourteen clinical diagnoses and one for "no findings". One image can have multiple labels applied to it. Only 1,686 of the 112,120 images in the NIH dataset are labelled fibrosis (1.5%). While there are strategies that can be used to address the small number of instances of the fibrosis class, there are a number of reasons this could make building an effective binary classification model challenging.

- The small number of images for the fibrosis class may not capture the relationships in the data and nuances of that class. This could prevent the model from learning the distinguishing features effectively, resulting in poor generalisation to unseen data.
- The model may become biased towards the non-fibrosis class, reducing the ability to identify the fibrosis class.

- With limited fibrosis images, the model may overfit to the small number of training examples for the fibrosis class, detrimentally impacting generalisation.
- ViTs typically require a substantial amount of data to train effectively. With few images for the fibrosis class, the training process may become unstable, leading to issues such as high variance in the model's performance across different training runs.

Having balanced, or unbalanced classes will impact other areas of the project, such as model evaluation, discussed later.

4.5.1 Label Distribution

The following Figure, 4.5, shows how fibrosis labels are distributed across the dataset, by age and gender.

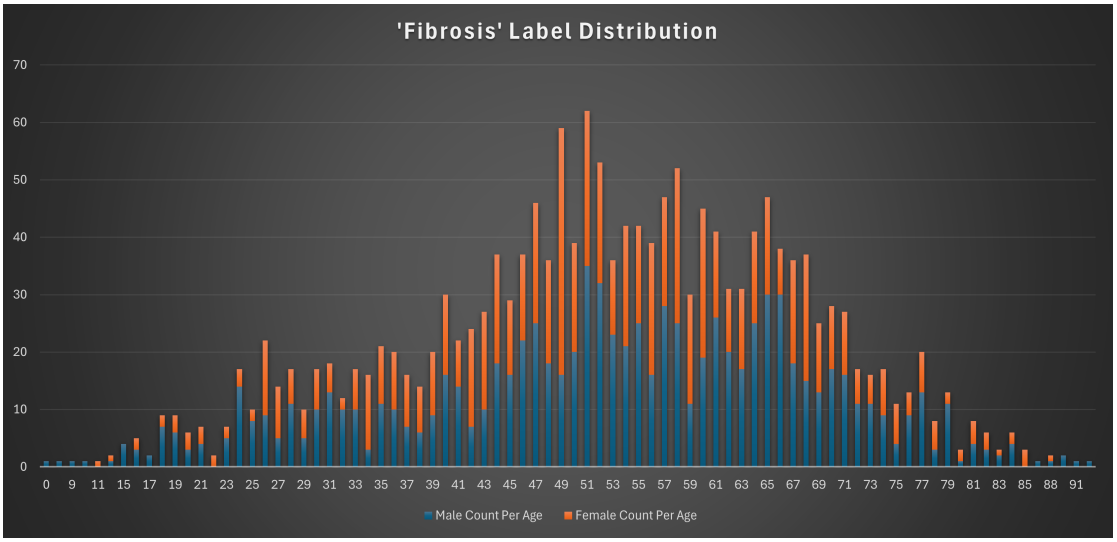

FIGURE 4.5: Fibrosis Label Distribution

Figure 4.6 shows how all labels are distributed across the dataset, by age and gender.

The study by de Bruijne [2016] highlights the importance of considering the makeup of the dataset when interpreting and evaluating ML model output. It highlights that depending on the training, diagnosis decisions may not be driven by signs of the disease, rather a confounding factor correlated with the disease in the training set. As an example, if a disease has a higher prevalence in men than in woman, a model might decide the size of certain structures is a good indicator for the risk of the disease.

de Bruijne et al. proposed collecting a balanced training set for confounding factors, with age and gender matching between control groups or, preferably, incorporating possible

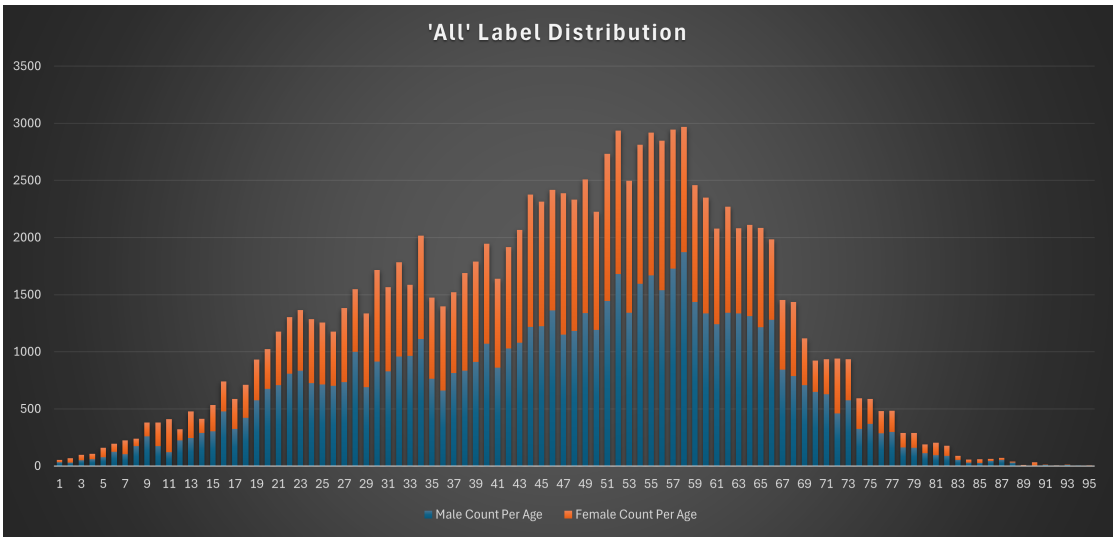

FIGURE 4.6: All Label Distribution

other predictors in the learning and thus learning the joint relation between confounders and image appearance. Factors such as these will need to be taken into account when creating the input dataset, which may be a subset of the full NIH CXR dataset.

For the model input data, an image with multiple labels, fibrosis being one, will be considered the same as images with just a fibrosis label. This will allow the classifier to be usable when a patient has multiple clinical signs caused either by the same disease or by comorbidities.

4.6 Data Partitions and Stratification

Data splitting is commonly used in ML to split data into train, validation and test datasets, as shown in Figure 4.7, for 3-fold cross-validation.[2]

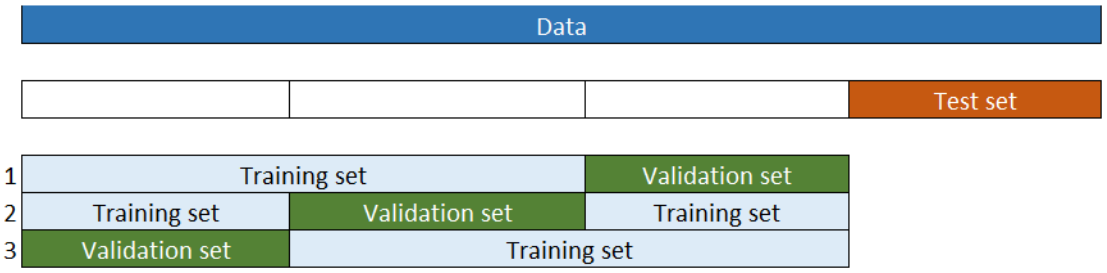

FIGURE 4.7: 3-Fold Cross-Validation

[2]https://tinyurl.com/data-partitions (accessed April 2024)

The model is initially fit on a training dataset. Successively, the fitted model is used to predict the classification on a second dataset, the validation dataset. At the start of the project, a holdout dataset will be created by partitioning a subset of the data. This independent test dataset will only be used once, to measure the generality of the final model once all experiments are complete [Lones, 2021].

As the experiments will use medical data, where in some cases the same patient has multiple CXR images in the full dataset, it is crucial to keep the same patient in either the training, validation or test set, to prevent data leaks. The effect of which is overfitting training data, where a model performs well on the data it was trained on but poorly on new, unseen data [JM et al., 2018].

Stratification is an important technique in data splitting. It ensures the age and gender distributions are consistently applied across the training, validation, and test sets, reflecting real-world scenarios. This consistency helps in building robust models that generalise well to unseen data, ensuring performance metrics are reliable and representative of the entire dataset.

Preventing data leaks by having a single patient ID in only one of the train, validation and test datasets means that the ages and genders cannot be perfectly distributed across the data splits. Analysis was conducted to ensure the datasets were both leak-free and adequately stratified to reflect the diversity of the population in the NIH CXR dataset, helping ensure a robust and generalisable model.

4.7 Data Augmentation

Data augmentation is a pivotal technique in ML, particularly for medical imaging tasks. In the medical domain, acquiring a substantial and diverse dataset can be challenging due to privacy and legal concerns, high costs and the rarity of certain conditions. This scarcity of data can lead to overfitting, where models fail to generalise to new, unseen data. Data augmentation can help mitigate this problem by artificially expanding the dataset, helping improve model robustness and performance.

It can also help address class imbalance, a common issue in medical datasets, such as the NIH CXR images, where there are considerably more non-fibrosis images than fibrosis. By selectively augmenting fibrosis images, the dataset can achieve a more balanced representation, preventing the model from being biased towards the non-fibrosis class.

Certain augmentation methods can be tailored to medical imaging and CXRs. The paper by Cossio [2023] outlines a comprehensive catalogue of sixty-five techniques for

augmenting medical imaging. The fact there are so many diverse techniques, each with unique characteristics and applications, means some experimentation is required to understand the most appropriate techniques to perform fibrosis binary classification on the NIH dataset and to help the model generalise to other datasets.

Manually reviewing the NIH CXRs and taking into account the guidance on Radiology Masterclass, emphasises the importance of contrast in identifying fibrosis in the images. The "Color and Contrast Adjustment Based" section in the Cossio [2023] paper provides some areas to be explored to help the model identify subtle fibrotic lesions.

It also provides direction on techniques to help simulate variations and imperfections in real-world CXR imaging, simulating different exposure levels of CXRs and cropping, rotating and flipping the image to help assist the model take variables, such as patient positioning during the CXR capture, into account.

The research by Elgendi et al. [2021] provides insight into appropriate techniques for image augmentation for detecting COVID-19 using CXR datasets (albeit, not the NIH dataset). As part of this work the opinion of radiologists was sought, providing expert insight into appropriate X-ray augmentation techniques.

- Reflection in the x-axis is not recommended.
- Rotation could be helpful, depending on the range. Between -5°and 5°is seen in clinical practice, severe rotations such as -90°and 90°is not recommended.
- An equal scaling in tehe x-axis and y-axis is possible, however, scaling in only the x-axis or y-axis is not recommended clinically.
- Shearing (slanting) is not recommended as it produces images that do not exist clinically.
- Translation or "shifting" X-ray images up, down, left, or right, could be a useful augmentation step. This is because the X-ray images do not always produce lungs in the centre of the image. This can depend on the patient's position, as well as the radiographic unit itself, such as if it is portable. Having X-ray images where the lungs are centred could lead to a more robust COVID-19 detector. As such, this step seems to be acceptable clinically as it is observed. However, there is no clearly recommended range for translation.

Having this input from radiologists on which data augmentation techniques are useful for CXRs is helpful as it ensures the chosen methods accurately reflect real-world clinical variations and enhance the model's ability to detect medical conditions effectively.

Taking account of available literature and the guidance above, multiple experiments were conducted to establish which data augmentation hyperparameters were optimal for binary classification of fibrosis on the NIH CXR dataset.

- RandomAffine(degrees=15, translate=(0.1, 0.1))
- RandomResizedCrop(scale=(0.9, 1.0)
- GaussianBlur(kernel_size=(5, 9), sigma=(0.1, 5.0))

These hyperparameters were used in experiments where data augmentation was added, alongside image resizing and normalisation, according to the input requirements of the model.

4.8 Data Legalities and Ethics

To fully consider the legal and ethical implications of using this dataset, consultations were held with the MACS Ethics Committee. They highlighted there is no evidence of informed consent for the NIH dataset, which contains patient data collected from 1992 to 2015. Specifically, patients from the mid-90s could not have given informed consent for the online publication of their medical data because they could not have understood the potential consequences.

After discussions with the committee, it was concluded the dataset could be used for the project due to mitigating factors, such as data anonymisation. Additionally, approval to use the NIH dataset for this project was obtained from the registered data owner at NIH, Ronald Summers, and the UK's independent regulator for data protection and information rights, the ICO.

4.9 Models

4.9.1 Selecting and Fine-Tuning Models

As discussed in the Literature Review, a common approach to building medical ML models is to pre-train a model using nonmedical datasets and then refine the model on a target medical task, such as image classification on CXRs.

For the initial experiments, the PyTorch ViT-B/16 model architecture was selected due to the familiarity of the PyTorch framework and having a relatively small ViT, which

enabled initial experiments to run quickly with the computational resources available. The ViT-B/16 architecture, with a patch size of 16x16 pixels, has a higher resolution of feature maps compared to ViT-B/32, allowing the model to capture more fine-grained details in CXRs. This higher resolution is expected to enable better classification of fibrosis. In addition, in benchmarking tests to evaluate the performance of image classification models on the ImageNet-1K dataset, ViT-B/16 performed better than ViT-B/32.

Model Architecture	**Weights**	**acc@1 (on ImageNet-1K)**
ViT-B/16	IMAGENET1K_SWAG_E2E_V1	85.304
ViT-B/32	IMAGENET1K_V1	75.912

TABLE 4.1: Benchmark Comparison of ViT Models and Weights

In subsequent experiment runs and where time allowed, alternative and larger models, such as ViT-H/14, were employed to assess whether they could achieve better performance and generalisation.

4.9.2 Model Development for PA and AP Views

Given the distinct anatomical perspectives, potential distortions, clinical contexts, and feature sensitivities inherent in the AP and PA views in CXRs, separate models were developed for each. Multiple experiments were conducted to ensure optimal and accurate performance for both image views and to assess whether there is an optimal dataset structure on which an ML model should run.

4.9.3 Experiments

A series of experiments were designed and conducted to train ML models on the NIH CXR data, or subsets of it, and evaluate the performance of the models under different conditions and dataset structures.

Experiment 1: Establishing Model Performance Capabilities Using the PA View CXR Data

This experiment aimed to assess the performance capabilities of the ML model when trained, validated and tested solely on PA view CXR data. The objective was to understand how well the model could generalise and perform with this specific type of data.

In the full NIH CXR dataset, there are 1,408 fibrosis images for the PA view, compared with 65,902 non-fibrosis images. To balance the classes at the start of the experiments and allow the models to run quickly at the beginning, the number of non-fibrosis images

were down sampled to 1,408, giving a total dataset size of 2,816. This was then split into train (1,898), validation (421) and test (497) datasets and stratified.

Experiment 2: Establishing Model Performance Capabilities Using the AP View CXR Data

Similar to Experiment 1, this experiment focused on evaluating the model's performance with AP view CXR data. The goal was to determine the model's effectiveness and accuracy when dealing with the different anatomical perspective and potential distortions present in AP view images.

In the full NIH CXR dataset, there are 278 fibrosis images for the AP view, compared with 44,532 non-fibrosis images. To balance the classes at the start of the experiments the number of non-fibrosis images were down sampled to 278, giving a total dataset size of 556. This was then split into train (319), validation (134) and test (103) datasets and stratified.

In the initial experimental runs, the small dataset led to the model overfitting after a small number of epochs and challenges in calculating accurate training and validation performance metrics. To help mitigate this, the number of AP fibrosis images were upsampled from 278 to 1,408. To balance the classes, the number of non-fibrosis images were downsampled to 1,048 also, giving a total dataset size of 2,816. This was then split into train (1,744), validation (376) and test (696) datasets and stratified.

Experiment 3: Establishing Model Performance Capabilities with Combined PA and AP View CXR Dataset

In this experiment, the performance of the ML model was tested using a combined dataset consisting of both PA and AP view CXR images. The objective was to explore whether a combined dataset could improve model performance and provide better generalisation across different both CXR views.

In the full NIH CXR dataset, there are 1,686 fibrosis images for the combined PA and AP view, compared with 110,434 non-fibrosis images. To balance the classes at the start of the experiments and allow the models to run quickly at the beginning, the number of non-fibrosis images were down sampled to 1686, giving a total dataset size of 3372. This was then split into train (2,258), validation (519) and test (595) datasets and stratified.

4.9.4 Baseline Hyperparameters

Baseline hyperparameters were established through initial experimentation, computational resource availability, and by selecting options best suited to the binary classification problem at hand. The chosen hyperparameters included:

- Batch Size: 128
- Optimizer: AdamW
- Scheduler: ExponentialLR
- Loss Function: Cross entropy loss
- Learning Rate: 1e-4
- Device: Cuda

4.9.5 Training Protocol

Training runs were executed for 30 epochs, striking a balance between providing sufficient training time and efficient use of computational resources. An early stopping mechanism was implemented based on validation loss, where training was halted after five consecutive increases in the validation loss, to prevent overfitting, and maintain efficient use of time and computing resources.

4.9.6 Model Architectures and Hyperparameters Assessment

Various model architectures and hyperparameters were assessed within each experiment to establish the optimal model configuration. The goal was to determine the best combination of model complexity, learning parameters, and resource efficiency to achieve high accuracy and reliability in CXR image classification.

4.9.7 Incremental Tuning Strategy

An incremental tuning strategy was employed to build insight into the problem and maximise model improvements. In each experiment run, all hyperparameters were fixed except for one "scientific hyperparameter", whose impact on the model's performance was measured. If the scientific hyperparameter positively impacted performance, it was converted into a fixed hyperparameter for subsequent runs. If it did not positively

impact performance, it was either removed or its value was adjusted for future experimentation. This strategy allowed for systematic refinement and optimisation of the model configurations.

4.9.8 Maximising Model Performance

The lack of specific guidance on how to optimise ML models stems from the fact that there is a vast diversity in applications, algorithms, data, hyperparameters, evaluation metrics, computational resources, and the evolving nature of the field. Tailoring optimisation strategies to the unique aspects of each ML project is essential for achieving the best performance.

The Google Research Tuning Playbook [Godbole et al., 2023], developed by researchers and engineers from the Google Brain research team and Harvard University, is a resource designed to help engineers and researchers optimise deep learning models. The primary focus of the playbook is on hyperparameter tuning, which is critical for maximising model performance. It also covers other areas such as choosing a model architecture, optimizer selection, initial configuration, training duration and pipeline optimisation. It has been used throughout this project to provide a systematic framework to enhance the performance of the models, using time and resources efficiently.

4.9.9 Creating Models from Scratch

As established in the Literature Review, building a model from scratch is a time-consuming and complex endeavour, involving model architecture design, hyperparameter tuning, and optimisation techniques. It is also computationally intensive, especially for complex models or large datasets.

Creating and training a ViT from scratch was considered, training it on the full NIH CXR dataset, and fine-tuning on the fibrosis labelled images. However, given the computational resources required to do this, experimenting with foundation models without any pre-trained weights was considered a more efficient use of project time and resources.

4.10 Evaluation

4.10.1 Evaluation Metrics

Model evaluation is a crucial stage in ML that assesses the performance of a trained model and establishes how effectively it will perform if deployed in a real-world environment.

As explained previously, the data was split into train, validation and test datasets. Assessing the model's performance will be done on the holdout test dataset only, as it will be representative of CXR data the model is anticipated to encounter should it be deployed, and it is also data the model has never seen before.

The models being developed were binary classifiers, to detect whether fibrosis is present in the CXR sample or not. A confusion matrix[3], as shown in Figure 4.8, is a way to record how many times a classification model correctly or incorrectly classifies things into the corresponding buckets.

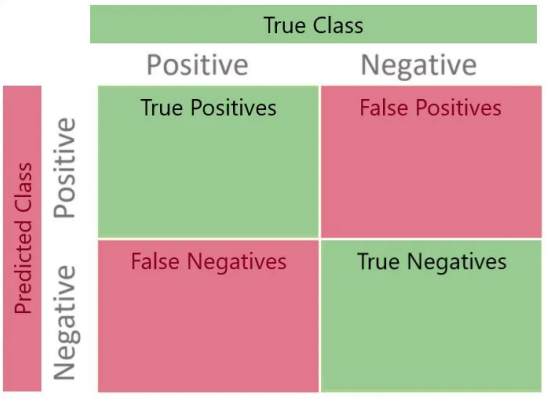

FIGURE 4.8: Confusion Matrix

In the case of classification models, the most commonly used metric is accuracy, which is the proportion of samples in the data set that were correctly classified by the model. This works sufficiently if the classes are balanced, i.e., if each class is represented by a similar number of samples within the data set. But many data sets are not balanced, and in this case accuracy can be a very misleading metric [Lones, 2021].

The paper by Bradley [1997] recommends that area under the receiver operating characteristic (ROC) curve (AUC), AUC-ROC, be used in preference to overall accuracy for "single number" evaluation of ML models. It found AUC exhibits a number of desirable properties when compared to overall accuracy: increased sensitivity in analysis of variance tests, a standard error that decreased as both AUC and the number of test samples

[3]https://tinyurl.com/confusion-matrix-diag (accessed April 2024)

increased, decision threshold independence, and it is invariant to priori class probabilities (the model will solely rely on the information in the image to make a classification).

An excellent model has AUC near to 1 which means it has a good measure of separability. A poor model has an AUC near 0 which means it has the worst measure of separability. In fact, it means it is reciprocating the result. It is predicting 0s as 1s and 1s as 0s. When AUC is 0.5, it means the model has no class separation capacity whatsoever[4].

It takes a pair of complementary indicators, such as sensitivity and specificity, to adequately measure the performance of a classifier based on its confusion matrix in a class-sensitive way [Cichosz, 2011].

Sensitivity of a test (also called the true positive rate) is defined as the proportion of cases that are correctly identified as "positive" by the test. The specificity of a test (also called the true negative rate) is the proportion of cases that are correctly identified as "negative" by the test[5]. Figure 4.9 shows how sensitivity and specificity are calculated.

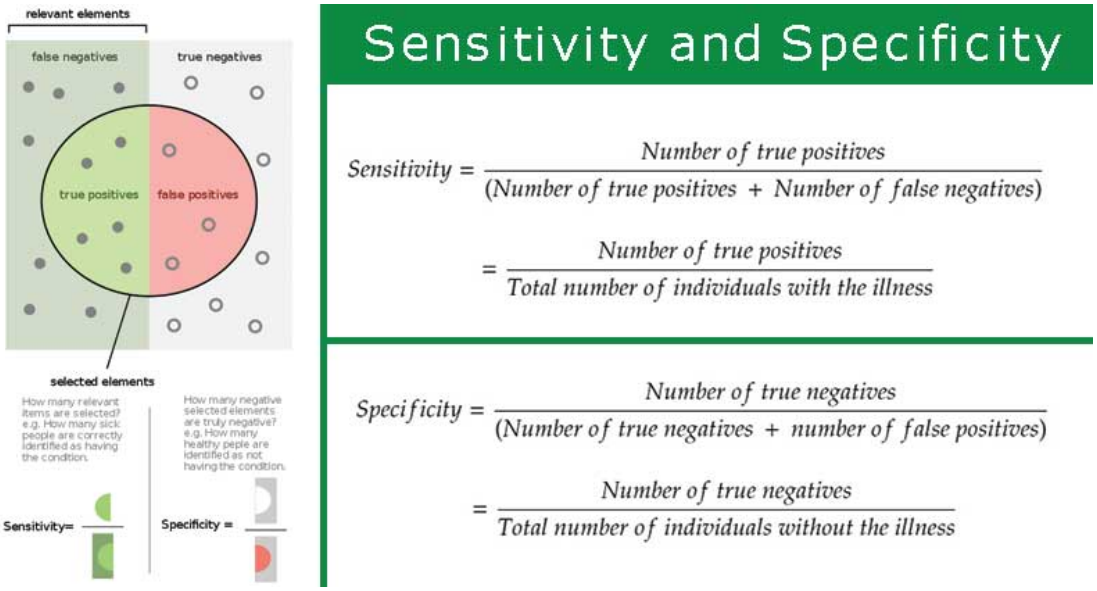

FIGURE 4.9: Sensitivity and Specificity

The F1 score assesses the predictive ability of a model by examining its performance on each class individually, rather than considering overall performance like accuracy does. It combines two competing metrics, precision and sensitivity[6]. Precision quantifies the proportion of correct positive predictions made by the model. The F1 score blends precision and recall using their harmonic mean. Maximising for the F1 score implies simultaneously maximising for both precision and recall. Generally, in a binary classification model, an F1 score of 1 indicates excellent precision and sensitivity, while a

[4]https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5
[5]https://tinyurl.com/sensitivity-specificity (accessed April 2024)
[6]https://tinyurl.com/f1-score (accessed August 2024)

low score indicates poor model performance. However, what what constitutes a "good" or "acceptable" F1 score varies based on factors such as the domain, application, and consequences of errors.

Where experiments used a balanced set of classes, accuracy was a key evaluation metric. With imbalanced classes, the Matthews Correlation Coefficient (MCC)[7] was more appropriate, due to it being relatively insensitive to class size imbalance [Lones, 2021]. The MCC is a single number metric that takes into account true positives, true negatives, false positives, and false negatives, similar to sensitivity and specificity. Figure 4.10 shows how MCC is calculated.

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

FIGURE 4.10: Matthews Correlation Coefficient

The MCC ranges from -1 to +1, where +1 indicates perfect classification, 0 indicates random classification, and -1 indicates total disagreement between classification and observation.

4.10.2 Balancing Class Importance

The clinical setting where an ML model is deployed guides the balancing of class importance. If a binary classification fibrosis model were to be deployed in a primary care setting where the objective is (often) early detection and referral, sensitivity would be vital as in primary care, early detection of fibrosis is critical to ensure patients are referred to specialists promptly. Missing a case of fibrosis could delay treatment and worsen patient outcomes.

Implementing the model in a specialist respiratory clinic, such as the CF clinics my son attends, where the primary objective is typically accurate diagnosis and management, the focus would likely shift to specificity and management of already suspected cases. False positives in this environment could lead to unnecessary treatments, increased healthcare costs, as well as patient anxiety.

A population level fibrosis screening program would aim to identify cases in a cost-effective manner. High sensitivity in this context would be crucial to capture all potential cases, but the model would also need to be efficient to handle large volumes of cases and data. The ability of an ML model to provide instant decision-making could be advantageous, facilitating rapid response and efficient resource allocation.

[7] https://tinyurl.com/mcc-diagram (accessed April 2024)

For any ML models to be deployed in a clinical setting, close collaboration with a clinician(s) would be required to define the clinical goals and specific outcomes to be improved. This collaborative approach would ensure the model is not only accurate and robust, but also clinically effective and tailored to the specific needs of the deployment setting.

4.10.3 Visualisation

Ideally, the saliency maps generated by the ViT would be validated to confirm the actual fibrosis lesions are being detected. The NIH CXR dataset includes a file "BBB_List_2017.csv", which shows bounding box (BB) coordinates for the finding label. Unfortunately, the fibrosis labelled images were added after this BB list was created and therefore the BB coordinates for the fibrosis images are unknown. If this model were to be employed in clinical practice, an appropriate validation exercise and further testing on external data would have to take place to provide additional assurance of the model's accuracy.

Chapter 5

Professional, Legal, Ethical, and Social issues

In developing ML models to diagnose fibrosis in CXRs, it is important to explore the Professional, Legal, Ethical, and Social (PLES) considerations, particularly as this work sits in the medical domain.

There are a significant amount of aspects to consider, including legal frameworks concerning data privacy, liability, and regulatory compliance, professional medical standards regarding accuracy, reliability, and interpretability, ethical principles for patient consent, fairness, and bias, as well as social implications covering equity of access, impact on healthcare workforce roles, and the broader societal impact of integrating ML systems into medical diagnostics.

By carefully looking at the PLES factors, the aim is to facilitate the responsible development of ML models, upholding the professional, legal, ethical and social standards required in a sensitive domain like healthcare.

5.1 Professional

In receiving medical care, patients are familiar with and have grown accustomed to human contact. Heyen and Salloch [2021] highlight how clinicians have the ability to pay special attention to facts of an individual patient's case, that ML models will not have the benefit of. For example, the patient's personality, life situation or cultural background.

In addition, something as simple as a doctor noticing a patient limp, or how they cough, can provide vital clues to help determine the correct diagnosis and treatment. It must be acknowledged that in virtually every circumstance, ML not having access to this type of information may lead to potentially poorer healthcare results, when compared to human experts.

Heyen et al. highlight the importance of the focus and commitment of a healthcare professional to provide the best patient care. They are often described as "going the extra mile" to ensure patients get the best possible experience and care. This is not something that ML, and AI more generally can replicate.

Errors in ML can cause negative consequences and pose risks to patient safety. Drabiak et al. [2023] point out that ML errors are different from others in medicine, as they can potentially impact thousands of patients where ML has been used as part of their care. This underscores the need for ongoing validation and performance assessment of ML models, which is also discussed by Vollmer et al. [2020].

5.2 Legal

In February 2024 the UK Government released its regulatory approach to AI.[1] It proposed five cross-sector principles for existing regulators to interpret and apply to drive safe, responsible AI.

1. Safety, security and robustness.
2. Appropriate transparency and explainability.
3. Fairness.
4. Accountability and governance.
5. Contestability and redress.

Regulators will implement the framework in their domains by applying existing laws, such as the Data Protection Act (DPA) and UK General Data Protection Regulation (GDPR), and issuing supplementary regulatory guidance.

The release of the UK AI regulatory approach is the initial step. Additional steps have been outlined as part of the roadmap for a final AI framework[2], shown in Figure 5.1.

[1] https://tinyurl.com/uk-ai-regulatory-approach
[2] https://tinyurl.com/ai-framework-roadmap

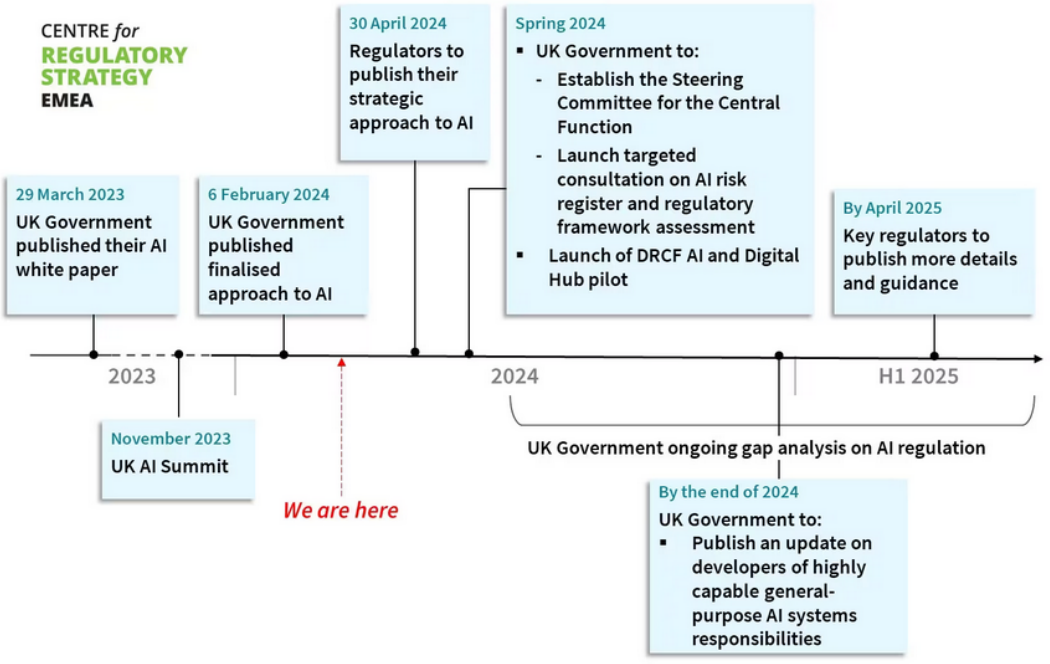

FIGURE 5.1: AI Framework Roadmap

One of the primary legal considerations for my project is the use of the NIH dataset. An analysis of the metadata of the NIH CXR dataset[3] has been conducted, the only personal metadata present is gender and age. There is a Patient ID column, but this relates to an ID within the dataset only, rather than a personal ID, such as NHS number. As the NIH data has been fully anonymised, this mitigates many of the legal challenges in using the dataset.

The ML model development and fine-tuning will use multiple existing libraries and code. The use of such artefacts will be done under strict adherence to the licensing terms of the used components.

5.3 Ethical

Developments and applications of ML are occurring at a rapid pace, and pose a challenge to medical ethics. The Hippocratic Oath has been one of the cornerstones of medical ethics, and is traditionally taken by physicians upon entering the medical profession. One of the key principles is the recognition of patients' right to make their own healthcare decisions, with physicians providing information and guidance to help patients make informed choices.

[3]https://nihcc.app.box.com/v/ChestXray-NIHCC/file/219760887468 (accessed April 2024)

Introducing ML can further complicate this, where models vary in terms of transparency, depending on their architecture and complexity. This puts ML in conflict with clinicians trying to provide guidance to help a patient make an informed decision, if they cannot interpret how an ML model has arrived at its output.

For ViTs, saliency maps can help provide insight into a model's decision-making process. These maps highlight the regions of the CXR the model deems the most important for making its diagnostic decision. Clinicians can compare these regions with their own observations and experience in detecting fibrosis in CXRs, helping validate the effectiveness of the model and increasing confidence in its ability.

Saliency maps can also reveal where the model is focused on unexpected or irrelevant areas of the image. This could indicate potential errors or biases in the model's training data or highlight areas where the model may need further refinement.

Bias is one of the major ethical challenges highlighted by Vollmer et al. [2020] in using ML for patient benefit. Any ML models must be trained on diverse and representative datasets to avoid biases that could disproportionately impact specific demographic groups. It also emphasises how vital the quality of training data is to produce accurate diagnoses for patients.

5.4 Social

This project considers the various social challenges of integrating ML tools into the healthcare sector and patient care. Some of the primary social questions to be considered are 'Does the possible benefits of deploying an ML model to automatically diagnose fibrosis in CXRs outweigh the potential risks?', and 'Are CF patients and clinicians likely to support their adoption?'.

While there may be job displacement or losses as a result of ML tool integration into healthcare, there may are potential benefits too. The NHS has a well publicised recruitment and retention difficulties, ML models such as these may help reduce the burden on already stretched radiologists and specialists to interpret X-rays. It could also help alleviate so called "postcode lotteries", where the models could be deployed in places where patients are receiving comparatively substandard care due to the lack of available specialist staff.

Establishing CF patient and clinical acceptance of ML systems such as the ones developed in this project, requires upfront and honest communication about their capabilities,

limitations and possible risks. By close engagement with patients and clinicians, the responsible deployment of ML could be achieved by ML tools initially being to supplement and support human experts, with humans making final clinical decisions. Over time, acceptance may increase as these technologies prove their value.

Chapter 6

Project Plan

6.1 Project Plan

A project plan outlines the roadmap and key milestones for completing project objectives. A well devised project plan is essential for finishing a project on time. The final deliverable date for the project is Thursday 15th August. The project plan makes two key assumptions: “Duration” includes weekdays only and no work is planned for “Holiday”. Depending on how the project progresses this may change, particularly weekend working.

6.1.1 Project Timetable

Task	Start Date	Duration	End Date
Preprocess NIH dataset	06-May	10	17-May
Select existing ML models trained on non-medical and medical images as baseline models	20-May	2	21-May
Fine-tune baseline models on CXR dataset to diagnose fibrosis	22-May	5	28-May
Optimise, re-train and track model performance on training and validation sets	29-May	5	04-Jun
Test final baseline models using holdout dataset	05-Jun	1	05-Jun
Complete final baseline models evaluation and reporting	06-Jun	2	07-Jun
Develop ViT	27-May	20	21-Jun
Train ViT on NIH dataset	24-Jun	5	28-Jun
Fine-tune ViT on fibrosis labelled images	01-Jul	5	05-Jul
Optimise, re-train and track ViT performance on training and validation sets	08-Jul	5	12-Jul
Test final ViT using holdout dataset	15-Jul	1	15-Jul
Complete ViT evaluation and reporting	16-Jul	2	17-Jul
Contingency	18-Jul	7	26-Jul
Holiday	27-Jul	7	03-Aug
Final Report Review & Submission	05-Aug	10	15-Aug

FIGURE 6.1: Project Timetable

6.1.2 Gantt Chart

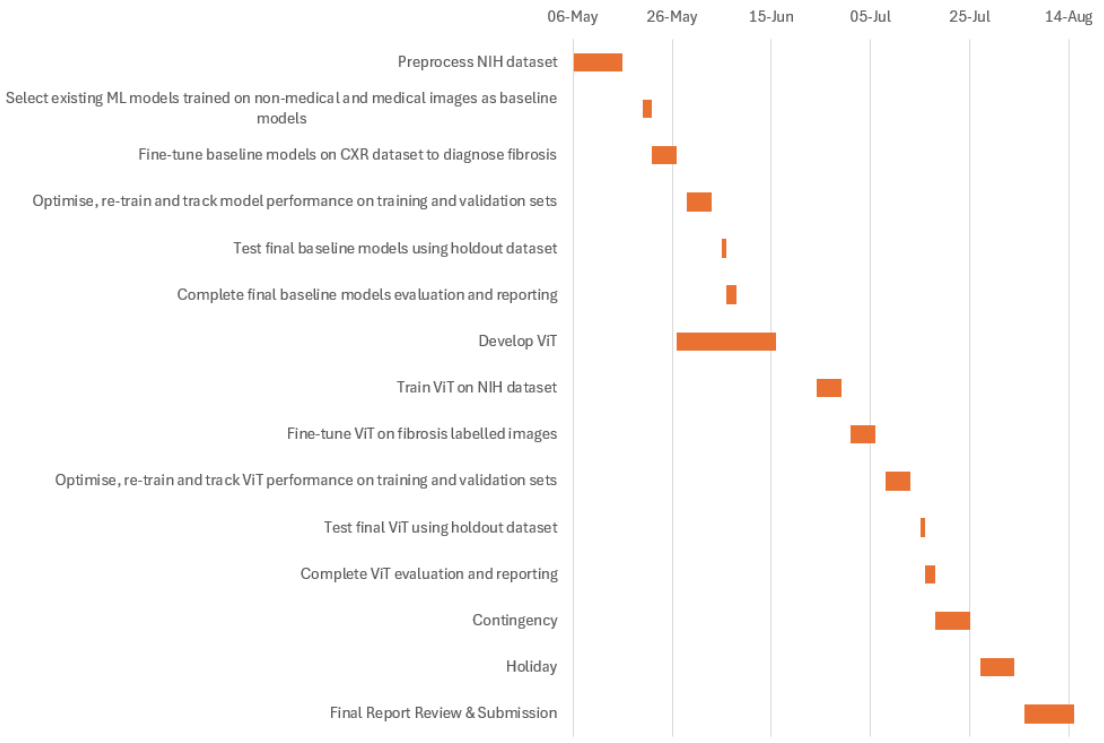

FIGURE 6.2: Gantt Chart

6.2 Risk Analysis

Various risks can arise that may impact the successful completion of the project. Each risk identified has been given a likelihood and impact level (low, medium or high) and suggested steps to deal with each risk should it arise. Possible risks are summarised in Table 6.1 below.

Risk	**Likelihood**	**Impact**	**Mitigation**
The NIH CXR dataset does not get ethics approval	Medium	High	Work with the ethics committee to understand concerns, contact data owner to find any additional assurance to gain ethics approval. Find an alternative dataset

Delays in ethics approval, ViT model development, may threaten the project timeline	Medium	Medium	Establish a realistic project plan and regularly verify progress. Reduce the scope of the project where required
The NIH CXR dataset is of poor quality and may lead to suboptimal model performance	Low	High	Conduct a data quality assessment on a sample of the dataset. Reduce the size of the dataset, removing sub-quality images. Augment the dataset or if the quality issue is on a large scale, find an alternative dataset
The model overfits and performs well on the training data but fails to generalise well on unseen data	Medium	High	Use regularisation, cross-validation, and hyperparameter tuning to help prevent and reduce overfitting
Model takes too long or too much memory to train with the computational resources available	Low	High	Reduce the image resolution. Train the model on fewer samples. Use simpler models. Use platforms like Google Colab and Graphics Processing Units (GPUs)
NIH CXR images were included in the X-ray training dataset for any pre-trained models I use in my experiments, such as the Google CXR Foundation, causing potential data leaks	Medium	High	Contact the developers to see if they used the NIH CXR dataset to train their model. Find alternative pre-trained models

Illness of myself or my supervisor	Low	High	If my supervisor is ill for an extended period, try to find alternative support. If I become ill for an extended period, apply for a mitigating circumstance to seek an extension or take this into consideration for marking

Table 6.1: Risk Assessment

Chapter 7

Results and Discussion

7.1 Model Development and Evaluation Results

In this project, ML models were trained and evaluated for the binary classification of CXRs to detect the presence of lung fibrosis. Using the NIH CXR dataset with annotated CXR images, state-of-the-art ViT algorithms were employed to create predictive models.

7.2 Measuring Model Performance

The performance of models was assessed using metrics such as accuracy, precision, sensitivity, specificity, F1 score and AUROC. A metrics abbreviation table is provided below in Table 7.1.

Abbreviation	**Metric**	**Range**	**Interpretation**
Acc	Accuracy	0-1	1 indicates perfect accuracy, 0 indicates no correct predictions
Prec	Precision	0-1	1 indicates every predicted positive is positive, 0 indicates that none of the predicted positives are actually positive
Sen	Sensitivity	0-1	1 indicates that all actual positives are correctly identified, 0 indicates that no actual positives are identified

Spec	Specificity	0-1	1 indicates that all actual negatives are correctly identified , 0 indicates that no actual negatives are identified
F1	F1 Score	0-1	1 indicates the best possible balance between precision and sensitivity, 0 indicates that either precision or sensitivity is zero
AUR	AUROC	0.5-1	1 indicates perfect discrimination, where the model can perfectly distinguish between positive and negative classes, 0.5 indicates no discrimination ability, equivalent to random guessing

TABLE 7.1: Metrics Abbreviation and Interpretation Table

7.3 Training and Validation Results

The following tables show the results obtained from each model for each experiment. Metrics from the training run are shown, with validation metrics also shown in brackets underneath. The second column shows the scientific hyperparameter, which has been updated from the previous run, and whose impact on the model's performance was measured.

The starting model for all three experiments was the ViT-B/16, initialised with the IMAGENET1K_V1 weights. Data augmentation was added according to the hyperparameters discussed in section 4.7. Section 4.9.4 outlines the baseline hyperparameters used as a starting point for the models on their first run.

7.3.1 Experiment 1: Establishing Model Performance Capabilities Using the PA View CXR Data

The following table shows the results for every experiment designed and run using the PA dataset.

ID	**Scientific Hyperparameter**	**Acc**	**Prec**	**Sen**	**Spec**	**F1**	**AUR**
1	ViT-B/16, 15 training epochs	0.633 (0.541)	0.528 (0.556)	0.667 (0.526)	0.609 (0.556)	0.589 (0.541)	0.638 (0.541)
2	30 training epochs	0.556 (0.541)	0.542 (0.556)	0.510 (0.526)	0.6 (0.556)	0.525 (0.541)	0.555 (0.541)
3	Model updated to ViT-H/14, 15 training epochs	0.556 (0.568)	0.471 (0.579)	0.545 (0.579)	0.565 (0.556)	0.505 (0.579)	0.555 (0.567)
4	15 additional training epochs	0.698 (0.595)	0.644 (0.611)	0.776 (0.579)	0.632 (0.611)	0.703 (0.595)	0.704 (0.595)
5	Data augmentation added	0.651 (0.649)	0.679 (0.666)	0.643 (0.632)	0.660 (0.666)	0.661 (0.649)	0.651 (0.649)
6	DA removed, added weight decay (1e-4)	0.698 (0.595)	0.644 (0.611)	0.776 (0.579)	0.631 (0.611)	0.704 (0.595)	0.704 (0.595)
7	Weight decay removed, both image classes sampled to 10000	0.634 (0.391)	0.585 (0.236)	0.649 (0.263)	0.622 (0.481)	0.615 (0.263)	0.635 (0.372)
8	Both images classes sampled to 5000	0.635 (0.370)	0.782 (0.5)	0.562 (0.352)	0.75 (0.4)	0.655 (0.414)	0.656 (0.376)
9	5000 of each class, data augmentation added	0.558 (0.26)	0.539 (0.385)	0.56 (0.294)	0.556 (0.2)	0.549 (0.333)	0.558 (0.247)
10	Balanced classes dataset, scheduler updated to ConstantLR	0.679 (0.621)	0.612 (0.619)	0.837 (0.684)	0.544 (0.555)	0.707 (0.650)	0.690 (0.620)
11	Weight decay added (1e-4)	0.679 (0.676)	0.627 (0.684)	0.824 (0.684)	0.545 (0.666)	0.712 (0.684)	0.684 (0.675)
12	Weight decay updated to 1e-5	0.679 (0.621)	0.612 (0.619)	0.837 (0.684)	0.544 (0.555)	0.707 (0.650)	0.690 (0.620)
13	ViT-H/14 model without pretrained weights	0.462 (0.514)	0.422 (0.519)	0.8 (0.737)	0.23 (0.278)	0.551 (0.608)	0.511 (0.507)

14	ViT-H/14 model with pre-trained weights, 75 training epochs	0.604 (0.595)	0.58 (0.611)	0.58 (0.579)	0.625 (0.611)	0.58 (0.594)	0.602 (0.595)
15	Data augmentation, weight decay 1e-4, 30 training epochs	0.509 (0.432)	0.537 (0.455)	0.518 (0.526)	0.5 (0.333)	0.527 (0.488)	0.509 (0.429)
16	Scheduler updated to ConstantLR	0.642 (0.568)	0.628 (0.579)	0.551 (0.579)	0.719 (0.555)	0.587 (0.579)	0.635 (0.567)

TABLE 7.2: Experiment 1 - Full Results for PA Experiment

The training and validation metrics for experiment one suggest that switching the scheduler from ExponentialLR to ConstantLR does not consistently lead to improvements across all performance metrics. However, the training and validation metrics are closer together with ConstantLR, which suggests better generalisation. This is particularly evident when comparing run 6, which uses ExponentialLR, and run 11, which uses ConstantLR.

Using ConstantLR generally led to higher sensitivity, especially in the training phase, where sensitivity scores exceeded 80%. However, there was a slight decrease in the corresponding validation metrics, indicating a potential trade-off between training performance and validation performance. Having higher sensitivity to detect positive fibrosis cases could be particularly beneficial to recognise cases promptly, allowing for timely treatment.

Although runs 11 and 12 exhibit broadly similar performance metrics, run 11 demonstrates slightly better generalisability, as indicated by the closer alignment of training and validation metrics. Therefore, run 11 is considered the optimal configuration for experiment one.

- Model: ViT-H/14
- Weights: IMAGENET1K_SWAG_E2E_V1
- Batch Size: 128
- Optimizer: AdamW (weight decay 1e-4)

- Scheduler: ConstantLR
- Loss Function: Cross entropy loss
- Learning Rate: 1e-4
- Data Augmentation: None
- Device: Cuda

This model is represented in the Final Results Table 7.5 as "ViT PA".

7.3.2 Experiment 2: Establishing Model Performance Capabilities Using the AP View CXR Data

The following table shows the results for every experiment designed and run using the AP dataset.

ID	**Scientific Hyperparameter**	**Acc**	**Prec**	**Sen**	**Spec**	**F1**	**AUR**
1	ViT-B/16, 15 training epochs	0.555 (0.167)	0.435 (0.333)	0.4 (0.25)	0.658 (0.566)	0.416 (0.286)	0.529 (0.125)
2	30 training epochs	0.444 (0.167)	0.476 (0.333)	0.294 (0.25)	0.62 (0.567)	0.363 (0.286)	0.457 (0.125)
3	Model updated to ViT-H/14, 15 training epochs	0.635 (0.313)	0.592 (0.315)	0.571 (0.5)	0.686 (0.171)	0.582 (0.388)	0.629 (0.336)
4	15 additional training epochs	0.683 (0.1667)	0.666 (0)	0.666 (0)	0.697 (0.5)	0.666 (0)	0.682 (0.25)
5	Both images classes sampled to 1408, data augmentation added	0.75 (0.575)	0.767 (0.533)	0.639 (0.302)	0.841 (0.791)	0.7 (0.386)	0.74 (0.546)
6	DA removed, added weight decay (1e-4)	0.712 (0.575)	0.676 (0.533)	0.657 (0.302)	0.756 (0.791)	0.667 (0.386)	0.706 (0.546)

7	Weight decay removed, both images classes sampled to 10000	0.788 (0.393)	0.777 (0.352)	0.833 (0.444)	0.737 (0.352)	0.805 (0.393)	0.785 (0.397)
8	Both images classes sampled to 5000	0.710 (0.347)	0.687 (0.346)	0.821 (0.545)	0.588 (1.90)	0.748 (0.424)	0.705 (0.368)
9	5000 of each class, data augmentation added	0.71 (0.387)	0.667 (0.386)	0.76 (0.667)	0.667 (0.167)	0.71 (0.489)	0.713 (0.417)
10	Balanced classes dataset, scheduler updated to ConstantLR	0.75 (0.542)	0.756 (0.471)	0.676 (0.302)	0.814 (0.731)	0.714 (0.368)	0.745 (0.517)
11	Weight decay added (1e-4)	0.738 (0.55)	0.72 (0.485)	0.563 (0.302)	0.854 (0.746)	0.632 (0.372)	0.708 (0.524)
12	Weight decay updated to 1e-5	0.7 (0.583)	0.852 (0.555)	0.535 (0.283)	0.892 (0.821)	0.657 (0.375)	0.713 (0.552)
13	ViT-H/14 model without pre-trained weights	0.7 (0.508)	0.882 (0.364)	0.405 (0.151)	0.953 (0.791)	0.555 (0.213)	0.679 (0.471)
14	ViT-H/14 model with pre-trained weights, 75 training epochs	0675 (0.567)	0.676 (0.515)	0.605 (0.321)	0.738 (0.761)	0.639 (0.395)	0.672 (0.541)
15	Data augmentation, weight decay 1e-5, 30 training epochs	0.75 (0.575)	0.767 (0.533)	0.639 (0.302)	0.84 (0.791)	0.697 (0.386)	0.74 (0.564)
16	Scheduler updated to ConstantLR	0.775 (0.574)	0.75 (0.533)	0.75 (0.302)	0.795 (0.791)	0.75 (0.386)	0.773 (0.546)

Table 7.3: Experiment 2 - Full Results for AP Experiment

The training and validation metrics for experiment two highlights the model's ability to identify negative cases of fibrosis, as indicated by the relatively high specificity scores in

runs 5, 10, 11, 12, 13, 14, 15, and 16. The fact that the training and validation scores in these runs are broadly similar suggests solid generalisability of the model.

The training and validation results suggest the PA view may be more attuned to identifying positive cases of fibrosis, while the AP view appears to be more effective at identifying negative cases of fibrosis. However, to draw definitive conclusions, further testing with larger and more diverse datasets would be necessary.

A general trend across all the results in experiment two, aside from the analysis mentioned above, indicates overfitting. This is evidenced by consistently higher performance on the training data compared to the validation data across all metrics, except for specificity. While run 13 has the highest specificity score on the training data, run 12 demonstrates better robustness in the non-fibrosis class, as indicated by the closer alignment of training and validation specificity scores. Therefore, run 12 is considered the optimal configuration for experiment two.

- Model: ViT-H/14
- Weights: IMAGENET1K_SWAG_E2E_V1
- Batch Size: 128
- Optimizer: AdamW (weight decay 1e-5)
- Scheduler: ConstantLR
- Loss Function: Cross entropy loss
- Learning Rate: 1e-4
- Data Augmentation: None
- Device: Cuda

This model is represented in the Final Results Table 7.5 as "ViT AP". The optimal configuration for the AP view is very similar compared with the PA view, with the only difference being the smaller weight decay for the AP view.

In both experiments one and two, the optimal model configurations do not include any data augmentation. This could be considered surprising given the relatively small number of fibrosis images, where the number of non-fibrosis images were downsampled to match. This may be because with the small number of fibrosis images, the model may already have a tendency to overfit the available data. Introducing augmentation could sometimes exacerbate this by causing the model to memorise specific transformations

rather than learning general patterns. Introducing augmentation may also have introduced noise that confused the model, where certain augmentations created data points that are not representative of the underlying patterns in the data. In small datasets, even a small amount of noise can impact performance because the model has fewer examples to learn the underlying patterns in the data and distinguish between the classes.

7.3.3 Experiment 3: Establishing Model Performance Capabilities with Combined PA and AP View CXR Dataset

The following table shows the results for every experiment designed and run using the combined PA and AP dataset.

ID	**Scientific Hyperparameter**	**Acc**	**Prec**	**Sen**	**Spec**	**F1**	**AUR**
1	15 training epochs	0.671 (0.429)	0.591 (0.4)	0.743 (0.667)	0.617 (0.25)	0.658 (0.5)	0.679 (0.458)
2	Data augmentation added	0.561 (0.143)	0.6 (0.2)	0.488 (0.333)	0.641 (0)	0.539 (0.25)	0.565 (0.167)
3	Both image classes sampled to 10000	0.643 (0.235)	0.857 (0.25)	0.6 (0.205)	0.75 (0.270)	0.706 (0.225)	0.675 (0.237)
4	Weight decay (1e-4) added to optimizer	0.857 (0.296)	0.833 (0.333)	0.833 (0.296)	0.875 (0.297)	0.833 (0.313)	0.854 (0.296)
5	Weight decay increased to 1e-3	0.786 (0.309)	0.6 (0.342)	0.75 (0.295)	0.8 (0.324)	0.666 (0.317)	0.774 (0.31)
6	Weight decay restored to 1e-4, ColorJitter added to data augmentation	0.786 (0.321)	0.8 (0.351)	0.667 (0.3)	0.875 (0.352)	0.727 (0.321)	0.771 (0.323)
7	Added normalization to transforms	0.643 (0.407)	0.625 (0.458)	0.714 (0.5)	0.571 (0.297)	0.667 (0.478)	0.643 (0.399)
8	Both image classes sampled to 20000	0.7262 (0.625)	0.737 (1)	0.683 (0.35)	0.767 (1)	0.709 (0.4)	0.725 (0.625)

Table 7.4: Experiment 3 - Full Results for Combined PA and AP View

Having the PA and AP views as a combined dataset resulted in poor generalisation. This is perhaps unsurprising given the differences in the PA and PA views, leading to variations in anatomical presentation, image quality and the overall appearance of the CXRs. These differences can introduce significant variability in the dataset, which might make it difficult for the model to generalise and the model may be learning to distinguish between the PA and AP views rather than focusing on the features of fibrosis.

While none of the models in experiment three demonstrated standout performance across all metrics, run 1 showed the least tendency to overfit, with a more consistent performance between training and validation data, except for a notable overfitting on specificity. There were some prominent differences in the optimal model configuration for this combined dataset, compared with the PA and AP views, namely the smaller ViT-B/16 model, ExponentialLR scheduler and earlier stopping (15 epochs for combined v 30 for PA and AP).

- Model: ViT-B/16
- Weights: IMAGENET1K_SWAG_E2E_V1
- Batch Size: 128
- Optimizer: AdamW
- Scheduler: ExponentialLR
- Loss Function: Cross entropy loss
- Learning Rate: 1e-4
- Data Augmentation: None
- Device: Cuda

This model is represented in the Final Results Table 7.5 as "ViT Combined".

7.4 Final Results

After selecting the best models based on training and validation performance, the final evaluation was conducted on a holdout test dataset. This dataset was completely separate from the training and validation datasets and was only used once, after the model was fully trained and finalised, thus providing an unbiased evaluation of the model's

generalisation ability to new, unseen data, simulating real-world scenarios where the model would encounter data it has never seen before.

As ML models can exhibit variability in their performance across different runs, where evaluation metric scores can differ slightly across different runs of the same model, random seeds for all sources of randomness were set to help mitigate this variability. In addition, the evaluation process was run a minimum of five times with results averaged to provide a more reliable estimate of model performance. Standard deviations are shown in brackets.

Model	**Acc**	**Prec**	**Sen**	**Spec**	**F1**	**AUROC**
ViT PA	43% (±8%)	47% (±5%)	53.5% (±9.5%)	31.5% (±7.5%)	50% (±7%)	42.5% (±8.5%)
ViT AP	65.9% (±10.9%)	71.8% (±23.8%)	45.8% (±17.8%)	85.9% (±9.9%)	56.2% (±20.2%)	65.8% (±13.8%)
ViT Combined	40.5% (±11.5%)	45% (±12%)	64.5% (±2.5%)	19.5% (±10%)	51.5% (±7.5%)	41.5% (±8.5%)

TABLE 7.5: Final Results

Figure 7.1 illustrates the comparative average performance metrics, expressed as percentages, across different models developed and evaluated in this project.

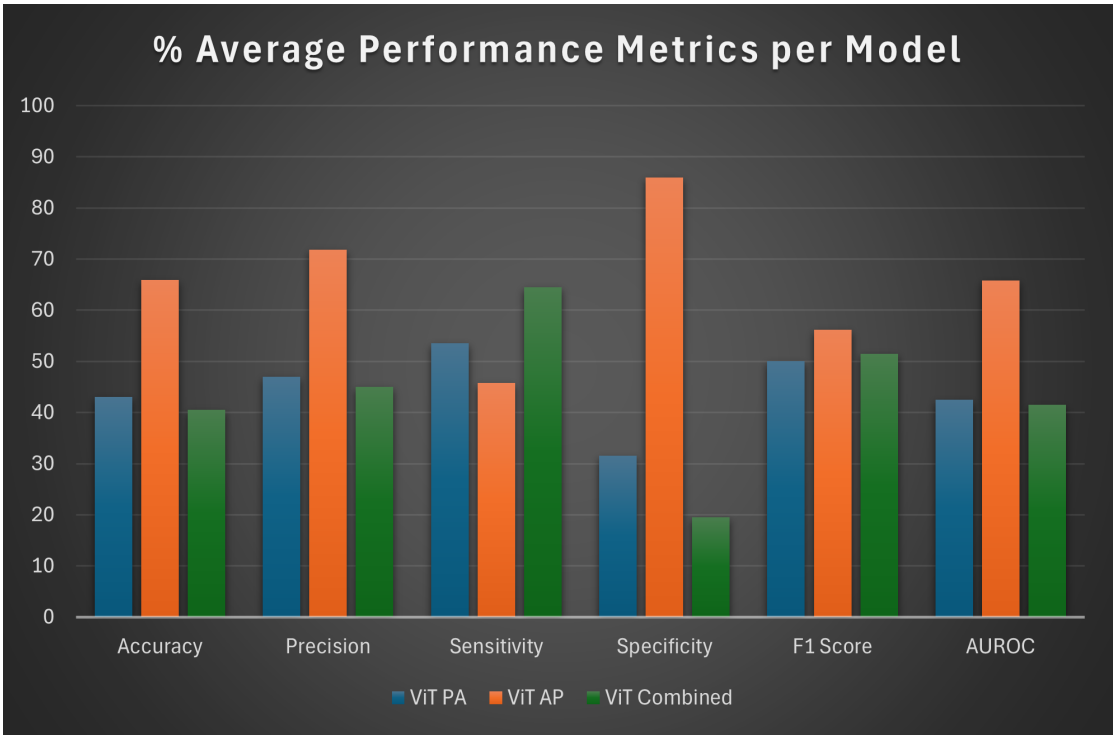

FIGURE 7.1: Final Results Graph

The ViT PA model's performance on the test dataset is relatively low. The accuracy, precision, sensitivity, and F1 score hover around or below 50%, suggesting that the model is not much better than random guessing. The AUROC being below 50% is particularly concerning, as it implies the model is systematically worse than random guessing at distinguishing between classes. The standard deviations, while moderate, indicate some variability in performance across the test runs.

There is a notable drop in all performance metrics from the training/validation set to the test set, which is indicative of overfitting. The model appears to perform well on the data it was trained on (and validated against) but fails to generalise to new, unseen data. Further development work on the PA view model would be required to increase the performance across all metrics, potentially achieved through stronger regularisation techniques, data augmentation and further hyperparameter tuning.

The ViT AP model's performance on the test dataset has high specificity, consistent with the train/validation performance, suggesting the model is more reliable at identifying non-fibrosis images. For accuracy, precision, and AUROC it performs moderately, in the 66-72% range. The low sensitivity (46%) is a significant weakness, indicating the model struggles to correctly identify fibrosis images.

The relatively lower standard deviation in specificity suggests the model's ability to identify non-fibrosis images is more consistent across different experiment runs. The significant variability in key metrics like precision, sensitivity, and F1 score implies that the model's performance is not stable, which would be problematic in medical diagnosis, where consistent results are crucial.

Additional tuning of the AP model is required to increase performance and provide more consistency across different runs. This could be achieved by incorporating advanced regularisation techniques, such as introducing dropout or further experimentation with weight decay, to prevent overfitting and stabilise the model's learning process. Additionally, applying more advanced data augmentation strategies, such as MixUp or CutMix, could enhance the diversity of the training data, making the model more robust. Further, implementing cross-validation with a high number of folds may help ensure the model generalises well to unseen data. Fine-tuning hyperparameters and their values through techniques like grid search could also optimise the model's performance and reduce variability across runs.

For the combined data view, the averages across most metrics suggest that the model is underperforming, exemplified in the accuracy, precision, and AUROC metrics. Sensitivity shows relatively better performance and stability, suggesting the combined view

model is more reliable in detecting fibrosis, though this is not sufficient to balance out the poor performance in other areas.

The standard deviations indicate the combined model's performance is inconsistent across different runs, particularly in precision and specificity. This inconsistency could be due to factors like data variability, model instability, or sensitivity to training configuration. The model requires significant improvement, particularly in enhancing specificity, AUROC, and overall consistency. Techniques like better regularisation, hyperparameter tuning, or even re-evaluating the model architecture might be necessary.

Chapter 8

Conclusions

8.1 Conclusions

The ViT PA model demonstrates significant challenges in generalising from the training and validation data to the test dataset. The substantial drop in performance across all key metrics, including accuracy, precision, sensitivity, specificity, F1 score, and AUROC indicates that the model suffers from overfitting. This overfitting has resulted in the model performing only marginally better than random guessing on the test data, with the AUROC score falling below 50%, highlighting a critical issue in its ability to distinguish between fibrosis and non-fibrosis images.

The observed variability in test performance, as indicated by the standard deviations, further underscores the model's lack of robustness. This variability suggests that the model's predictions are inconsistent across different runs, making it unreliable for clinical application in its current state.

The evaluation of the ViT AP model on the test dataset revealed both strengths and weaknesses that highlight the need for further refinement to improve its real-world applicability. The model demonstrated high specificity, consistent with its performance during the training and validation phases, indicating a reliable ability to correctly identify non-fibrosis cases. This suggests the model is effective at ruling out fibrosis in patients who do not have the condition, which is important for reducing false positives in clinical settings.

However, the model's moderate performance in terms of accuracy, precision, and AUROC, coupled with low sensitivity underscores a critical limitation. The model's inability to reliably detect fibrosis cases poses a significant risk in clinical applications,

where missing positive cases could delay necessary treatment and worsen patient outcomes. The low sensitivity, paired with the high specificity, suggests the model may be overly conservative, potentially at the cost of failing to identify patients with fibrosis. Given these findings, it is evident additional tuning of the ViT AP model is necessary to enhance its overall performance and stability.

The AP ViT model performing better than the PA model across a number of metrics is a somewhat curious result because traditionally, the PA view is considered to provide a clearer image of the chest compared to the AP view, and a more accurate anatomical representation due to the way the X-rays pass through the body. There is also less overlap of the lungs by the shoulder blades in the PA view, critical for the models in this project.

Given these factors, it was expected the PA view would produce better or at least comparable results to the AP view. Therefore, the AP ViT model outperforming the PA model across several metrics raises interesting questions about the underlying reasons for this outcome. This may be because a ViT might be inherently more suited to processing the type of images generated by the AP view, with the AP images used in training potentially having clearer markers of fibrosis.

This unexpected result suggests further investigation is necessary to understand the reasons behind the AP model's superior performance, which could provide valuable insights into both the model's behaviour and the characteristics of the CXR data.

The results from the model using a combined view test dataset showed better sensitivity performance, compared with separate PA and AP models, but a reduction in performance across all other metrics. This highlights the challenge in balancing sensitivity with overall model accuracy and specificity, particularly when combining the CXR datasets that contain inherent variability between the PA and AP views.

Having a high performing and consistent model that can distinguish between fibrosis and non-fibrosis CXRs, from either a PA and AP view, would provide clinicians with a flexible solution applicable in various clinical settings. A model that consistently performs well across different views would enhance clinicians' confidence in the results, enabling more informed decision-making.

While the models developed during this project show promise in some areas, to take them into clinical practice, further work is required to define clinical goals and patient outcomes to improve, to drive the work to tune the models and improve their performance and consistency across runs.

To assess the robustness of the models, working with an expert radiologist or clinician to verify whether a model is correctly identifying fibrosis and non-fibrosis in the images, rather than spurious correlations, would be beneficial. Saliency maps could help in this regard, to show where the model has identified fibrotic lesions, or the lack thereof, and ensure they match the assessment of a human expert.

The lack of diversity metadata in the NIH CXR dataset is restrictive, as it means there is no understanding of how the model performs across ethnically diverse datasets. This would need to be addressed in future work to facilitate deployment in a real-world setting.

Reflecting on how the requirements outlined in the functional 3.2 and non-functional 3.3 requirements sections have been addressed, the project successfully met the primary objective of developing binary classification models to detect fibrosis, establishing their effectiveness. In doing so, all healthcare data privacy and security regulations were met. Mandatory criteria, such preventing data leaks, ensuring data was distributed across data splits, and using a holdout training dataset were also achieved. However, some of the recommended requirements, such as segmenting lung regions and using saliency maps to visualise and interpret which parts of the CXR most influence the classification of the ML models, remain opportunities for future work. While this reflection highlights some of the achievements of the project, the following section provides a clear direction for future work, ensuring the foundation laid here can be built upon to further advance the field of fibrosis detection.

Chapter 9

Future Work

9.1 Future Work

Future work in the domain of fibrosis detection in CXRs holds substantial promise for enhancing the accuracy and applicability of ML models, ultimately achieving better outcomes for patients. Real-world deployment and validation of models in clinical settings will be crucial for assessing its practical utility and reliability, ensuring the technology can be effectively integrated into routine medical practice.

9.1.1 Enhancing the Model's Diagnostic Effectiveness

One avenue for future research involves enhancing the model's diagnostic effectiveness and generalisability. Additional testing of this could be achieved by implementing custom code for cross-validation that keeps the same patient within the same fold to prevent data leaks, while also stratifying age and gender across the folds.

Exploring the impact of image preprocessing on the model's effectiveness is another key area. Specifically, investigating lung segmentation before splitting the image into patches (with positional embedding) and inputting these into a ViT could yield significant insights. Comparing this approach with CXRs that have no segmentation fed into the same model, and using saliency maps to verify the segmentation, would help determine the preprocessing's impact.

Additionally, seeking out other datasets with fibrosis labelled data, such as PadChest, could increase the number of fibrosis images available. The PadChest CXR dataset from Spain has 751 CXR images labelled "pulmonary fibrosis" with 991 CXRs with child labels, which include not only the primary label but also related sub-conditions

derived from the primary condition. Further work would be required to understand whether these images could be used to increase the fibrosis image class and lead to more effective and robust models. The Open Access Biomedical Image Search Engine collates image data from multiple sources and appears to have fibrosis labelled images. Further investigation would be required to determine the licence of the images and custom API queries developed to obtain the relevant fibrosis images and metadata.

Ensuring future datasets include ethnicity metadata would allow for the development of more inclusive and representative models, addressing potential biases and improving diagnostic accuracy across diverse populations. This would help build more reliable and accurate models for diagnosing fibrosis. Different ViT architectures, and alternatives such as CNNs, could also be explored to assess whether they could provide better performance and more robust models.

9.1.2 Facilitating the Model's Use in Clinical Decision-Making

Implementing an ML model to diagnose fibrosis in clinical decision-making requires rigorous validation to ensure its reliability, accuracy, and safety. A human expert would need to conduct several validation steps before the model could be used in practice.

- Data source - to ensure the data used for training the model comes from reliable, high-quality sources, ensuring the quality and consistency of the images and associated annotations.
- Image preprocessing - to check these steps do not introduce biases or artefacts that could impact the model's performance.
- Expert comparison - to compare the model's performance against the diagnostic performance of human experts, both in terms of accuracy and consistency.
- Clinical usefulness - to evaluate whether the model's outputs are clinically useful. This includes checking if the model can identify fibrosis at different stages, if it can be integrated into the clinical workflow, and if it adds value to the existing diagnostic process.
- Interpretability - to ensure the model's decisions are interpretable by experts, understanding why the model made specific diagnoses, especially in borderline or complex cases.
- Robustness testing - to test the model's robustness against variations in imaging conditions, such as different hospitals, consulting rooms, scanners and image resolutions.

- Bias testing - to examine the model for potential biases, ensuring it performs well across different subgroups (e.g. age, gender, ethnicity).
- Regulatory approval - to obtain necessary regulatory approvals from relevant bodies by demonstrating the model's safety and efficacy.
- Ethical and legal review - to ensure the model's deployment adheres to ethical and legal requirements, including patient privacy, informed consent, and transparency.
- Performance monitoring - to continuously monitor the model's performance post-deployment to ensure it remains accurate and reliable.
- Model updates - to liaise with the model maintenance team to update the model as new data becomes available and as clinical, ethical and legal guidelines evolve.

9.1.3 Related Models Which Could Provide Benefit to CF Patients

While detection of fibrosis in CXRs is valuable for clinicians to facilitate treatment, fibrosis is not the only clinical observation identifiable via CXRs, which is relevant to CF patients. Working with CF specialists to define multiple anatomical and pathological variations that impact CF patients, and training a multi classification model to provide a comprehensive set of diagnoses for CF patients could enhance the utility of the same image.

Individuals with CF experience progressive lung damage over time. Predicting this damage would enable clinicians to implement proactive treatments and measures to slow the progression of lung deterioration. Developing an ML model to predict lung condition over time could be extremely beneficial for clinicians. This model could be further enhanced by incorporating additional lung function biomarkers, such as the FEV1 score, creating an even more effective tool for proactive care.

The primary challenge for such a model is gathering a comprehensive, longitudinal dataset of repeated CXRs from the same CF patients. While such data is routinely collected during annual reviews at CF clinics, including CXRs, there currently appears to be no centralised repository for this information.

The NIH CXR dataset does have some longitudinal patient images available. Extending this dataset would facilitate the development or fine-tuning of models to help predict the progression of fibrosis. This could be used alongside other CF patient longitudinal data, such as infection history, comorbidities and complications, weight, and intravenous antibiotic and medication usage, to provide a predictive capability that enables clinicians

to adopt a more preventative and personalised approach to CF patient care. Ouis and A. Akhloufi [2024] detail a number of CXR datasets that could be used in future work.

As the project concludes, it is worth reflecting on a journey which began as a deeply personal motivation and has grown into a significant piece of ML engineering. The models developed here represent a step forward in the application of ML for fibrosis detection in CXRs, but the work is far from complete. Future advancements could focus on refining the algorithms, expanding the dataset to include a more diverse population, and integrating the models into real-world clinical workflows. By continuing this line of research, we can move closer to the goal that inspired this project — to make a meaningful contribution to CAD, ultimately improving the lives of individuals with CF, like my son, Isaac.

Bibliography

Alaa, A. M. and van der Schaar, M. (2018). Prognostication and risk factors for cystic fibrosis via automated machine learning. *Scientific reports*, 8(1):11242.

Alaa, A. M. and van der Schaar, M. (2019). Attentive state-space modeling of disease progression. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Assayag, D., Morisset, J., Johannson, K., Wells, A., and Walsh, S. (2020). Patient gender bias on the diagnosis of idiopathic pulmonary fibrosis. *Thorax*, 75:thoraxjnl–2019.

Bak, S. H., Park, H. Y., Nam, J. H., Lee, H. Y., Lee, J. H., Sohn, I., and Chung, M. P. (2019). Predicting clinical outcome with phenotypic clusters using quantitative ct fibrosis and emphysema features in patients with idiopathic pulmonary fibrosis. *PloS one*, 14(4):e0215303.

Baratella, E., Fiorese, I., Minelli, P., Veiluva, A., Marrocchio, C., Ruaro, B., and Cova, M. A. (2023). Aging-related findings of the respiratory system in chest imaging: pearls and pitfalls. *Current Radiology Reports*, 11(1):1–11.

Belaid, O. N. and Loudini, M. (2020). Classification of brain tumor by combination of pre-trained vgg16 cnn. *Journal of Information Technology Management*, 12(2):13–25.

Bradley, A. P. (1997). The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7):1145–1159.

Bressem, K. K., Adams, L. C., Erxleben, C., Hamm, B., Niehues, S. M., and Vahldiek, J. L. (2020). Comparing different deep learning architectures for classification of chest radiographs. *Scientific reports*, 10(1):13590.

Candemir, S., Jaeger, S., Palaniappan, K., Antani, S., and Thoma, G. (2012). Graph cut based automatic lung boundary detection in chest radiographs.

Candemir, S., Rajaraman, S., Thoma, G., and Antani, S. (2018). Deep learning for grading cardiomegaly severity in chest x-rays: An investigation. In *2018 IEEE Life Sciences Conference (LSC)*, pages 109–113.

Chan, H.-P., Hadjiiski, L. M., and Samala, R. K. (2020a). Computer-aided diagnosis in the era of deep learning. *Medical physics*, 47(5):e218–e227.

Chan, H.-P., Samala, R. K., Hadjiiski, L. M., and Zhou, C. (2020b). *Deep Learning in Medical Image Analysis*, pages 3–21. Springer International Publishing, Cham.

Chen, L., Li, S., Bai, Q., Yang, J., Jiang, S., and Miao, Y. (2021). Review of image classification algorithms based on convolutional neural networks. *Remote Sensing*, 13(22).

Chen, S., GE, C., Tong, Z., Wang, J., Song, Y., Wang, J., and Luo, P. (2022). Adaptformer: Adapting vision transformers for scalable visual recognition. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A., editors, *Advances in Neural Information Processing Systems*, volume 35, pages 16664–16678. Curran Associates, Inc.

Chetoui, M. and Akhloufi, M. A. (2022). Explainable vision transformers and radiomics for covid-19 detection in chest x-rays. *Journal of Clinical Medicine*, 11(11).

Chowdhury, M. E. H., Rahman, T., Khandakar, A., Mazhar, R., Kadir, M. A., Mahbub, Z. B., Islam, K. R., Khan, M. S., Iqbal, A., Emadi, N. A., Reaz, M. B. I., and Islam, M. T. (2020). Can ai help in screening viral and covid-19 pneumonia? *IEEE Access*, 8:132665–132676.

Cichosz, P. (2011). Assessing the quality of classification models: Performance measures and evaluation procedures. *Central European Journal of Engineering*, 1:132–158.

Cohen, J. P., Morrison, P., and Dao, L. (2020). Covid-19 image data collection.

Cossio, M. (2023). Augmenting medical imaging: a comprehensive catalogue of 65 techniques for enhanced data analysis. *arXiv preprint arXiv:2303.01178*.

de Bruijne, M. (2016). Machine learning approaches in medical image analysis: From detection to diagnosis. *Medical Image Analysis*, 33:94–97. 20th anniversary of the Medical Image Analysis journal (MedIA).

Drabiak, K., Kyzer, S., Nemov, V., and El Naqa, I. (2023). Ai and machine learning ethics, law, diversity, and global impact. *The British journal of radiology*, 96(1150):20220934.

Elgendi, M., Nasir, M. U., Tang, Q., Smith, D., Grenier, J.-P., Batte, C., Spieler, B., Leslie, W. D., Menon, C., Fletcher, R. R., et al. (2021). The effectiveness of image augmentation in deep learning networks for detecting covid-19: A geometric transformation perspective. *Frontiers in Medicine*, 8:629134.

Ge, Y., Wang, Q., Wang, L., Wu, H., Peng, C., Wang, J., Xu, Y., Xiong, G., Zhang, Y., and Yi, Y. (2019). Predicting post-stroke pneumonia using deep neural network approaches. *International Journal of Medical Informatics*, 132:103986.

Godbole, V., Dahl, G. E., Gilmer, J., Shallue, C. J., and Nado, Z. (2023). Deep learning tuning playbook. Version 1.0.

Han, Z., Gao, C., Liu, J., Zhang, S. Q., et al. (2024). Parameter-efficient fine-tuning for large models: A comprehensive survey. *arXiv preprint arXiv:2403.14608*.

He, H., Cai, J., Zhang, J., Tao, D., and Zhuang, B. (2023). Sensitivity-aware visual parameter-efficient fine-tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11825–11835.

Heidari, M., Mirniaharikandehei, S., Khuzani, A. Z., Danala, G., Qiu, Y., and Zheng, B. (2020). Improving the performance of cnn to predict the likelihood of covid-19 using chest x-ray images with preprocessing algorithms. *International Journal of Medical Informatics*, 144:104284.

Heyen, N. B. and Salloch, S. (2021). The ethics of machine learning-based clinical decision support: an analysis through the lens of professionalisation theory. *BMC Medical Ethics*, 22:1–9.

Horry, M. J., Chakraborty, S., Paul, M., Ulhaq, A., Pradhan, B., Saha, M., and Shukla, N. (2020). X-ray image based covid-19 detection using pre-trained deep learning models.

Ibrahim, A. U., Ozsoz, M., Serte, S., Al-Turjman, F., and Yakoi, P. S. (2021). Pneumonia classification using deep learning from chest x-ray images during covid-19. *Cognitive Computation*, pages 1–13.

Jin, Y., Lu, H., Zhu, W., Yan, K., Gao, Z., and Li, Z. (2021). Ctfc: A convolution and visual transformer based classifier for few-shot chest x-ray images. In *2021 2nd International Conference on Artificial Intelligence and Computer Engineering (ICAICE)*, pages 616–622.

JM, B. et al. (2018). Data wrangling and data leakage in machine learning for healthcare.

Kalafatis, D., Gao, J., Pesonen, I., Carlson, L., Sköld, C. M., and Ferrara, G. (2019). Gender differences at presentation of idiopathic pulmonary fibrosis in sweden. *BMC Pulmonary Medicine*, 19:1–8.

Kaur, T. and Gandhi, T. K. (2019a). Automated brain image classification based on vgg-16 and transfer learning. In *2019 International Conference on Information Technology (ICIT)*, pages 94–98.

Kaur, T. and Gandhi, T. K. (2019b). Automated brain image classification based on vgg-16 and transfer learning. In *2019 international conference on information technology (ICIT)*, pages 94–98. IEEE.

Kermany, D. S., Zhang, K., and Goldbaum, M. H. (2018). Labeled optical coherence tomography (oct) and chest x-ray images for classification.

Kravchenko, T., Bogdanova, T., and Shevgunov, T. (2022). Ranking requirements using moscow methodology in practice. In Silhavy, R., editor, *Cybernetics Perspectives in Systems*, pages 188–199, Cham. Springer International Publishing.

Krishnan, K. S. and Krishnan, K. S. (2021). Vision transformer based covid-19 detection using chest x-rays. In *2021 6th International Conference on Signal Processing, Computing and Control (ISPCC)*, pages 644–648.

Lee, C., Yoon, J., and Schaar, M. v. d. (2020). Dynamic-deephit: A deep learning approach for dynamic survival analysis with competing risks based on longitudinal data. *IEEE Transactions on Biomedical Engineering*, 67(1):122–133.

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.

Lones, M. A. (2021). How to avoid machine learning pitfalls: a guide for academic researchers. *arXiv preprint arXiv:2108.02497*.

Lynch, D. A., Godwin, J. D., Safrin, S., Starko, K. M., Hormel, P., Brown, K. K., Raghu, G., King Jr, T. E., Bradford, W. Z., Schwartz, D. A., et al. (2005). High-resolution computed tomography in idiopathic pulmonary fibrosis: diagnosis and prognosis. *American journal of respiratory and critical care medicine*, 172(4):488–493.

Matsoukas, C., Haslum, J. F., Sorkhei, M., Söderberg, M., and Smith, K. (2022). What makes transfer learning work for medical images: Feature reuse & other factors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9225–9234.

Matsoukas, C., Haslum, J. F., Söderberg, M., and Smith, K. (2021). Is it time to replace cnns with transformers for medical images?

Minssen, T., Gerke, S., Aboy, M., Price, N., and Cohen, G. (2020). Regulatory responses to medical machine learning. *Journal of Law and the Biosciences*, 7(1):lsaa002.

Mondal, A. K., Bhattacharjee, A., Singla, P., and Prathosh, A. P. (2022). xvitcos: Explainable vision transformer based covid-19 screening using radiography. *IEEE Journal of Translational Engineering in Health and Medicine*, 10:1–10.

Murphy, Z. R., Venkatesh, K., Sulam, J., and Yi, P. H. (2022). Visual transformers and convolutional neural networks for disease classification on radiographs: A comparison of performance, sample efficiency, and hidden stratification. *Radiology: Artificial Intelligence*, 4(6):e220012.

Naralasetti, V., Shaik, R. K., Katepalli, G., and Bodapati, J. D. (2021). Deep learning models for pneumonia identification and classification based on x-ray images. *Traitement du Signal*, 38(3).

Okolo, G. I., Katsigiannis, S., and Ramzan, N. (2022). Ievit: An enhanced vision transformer architecture for chest x-ray image classification. *Computer Methods and Programs in Biomedicine*, 226:107141.

Ouis, M. Y. and A. Akhloufi, M. (2024). Deep learning for report generation on chest x-ray images. *Computerized Medical Imaging and Graphics*, 111:102320.

Pal, K. K. and Sudeep, K. S. (2016). Preprocessing for image classification by convolutional neural networks. In *2016 IEEE International Conference on Recent Trends in Electronics, Information Communication Technology (RTEICT)*, pages 1778–1781.

Rahman, T., Khandakar, A., Kadir, M. A., Islam, K. R., Islam, K. F., Mazhar, R., Hamid, T., Islam, M. T., Kashem, S., Mahbub, Z. B., Ayari, M. A., and Chowdhury, M. E. H. (2020). Reliable tuberculosis detection using chest x-ray with deep learning, segmentation and visualization. *IEEE Access*, 8:191586–191601.

Rahman, T., Khandakar, A., Qiblawey, Y., Tahir, A., Kiranyaz, S., Kashem, S. B. A., Islam, M. T., Al Maadeed, S., Zughaier, S. M., Khan, M. S., et al. (2021). Exploring the effect of image enhancement techniques on covid-19 detection using chest x-ray images. *Computers in biology and medicine*, 132:104319.

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. International Conference on Machine Learning and Data Engineering.

Singh, V., Sharma, S., Goel, S., Lamba, S., and Garg, N. (2021). Brain tumor prediction by binary classification using vgg-16. *Smart and sustainable intelligent systems*, pages 127–138.

Somayaji, R. and Chalmers, J. D. (2022). Just breathe: a review of sex and gender in chronic lung disease. *European Respiratory Review*, 31(163).

Sunanthini, V., Deny, J., Kumar, E. G., Vairaprakash, S., Govindan, P., Sudha, S., Muneeswaran, V., and Thilagaraj, M. (2022). Comparison of cnn algorithms for feature extraction on fundus images to detect glaucoma. *Journal of Healthcare Engineering*, 2022.

Tammina, S. (2019). Transfer learning using vgg-16 with deep convolutional neural network for classifying images. *International Journal of Scientific and Research Publications (IJSRP)*, 9(10):143–150.

Taylor, P., Alberdi, E., and Lee, R. (2001). Incorporating radiological knowledge in a cad system. *International Congress Series*, 1230:593–598. Computer Assisted Radiology and Surgery.

Tomoto, M., Mineharu, Y., Sato, N., Tamada, Y., Nogami-Itoh, M., Kuroda, M., Adachi, J., Takeda, Y., Mizuguchi, K., Kumanogoh, A., et al. (2024). Idiopathic pulmonary fibrosis-specific bayesian network integrating extracellular vesicle proteome and clinical information. *Scientific Reports*, 14(1):1315.

Ukwuoma, C. C., Qin, Z., Heyat, M. B. B., Akhtar, F., Smahi, A., Jackson, J. K., Furqan Qadri, S., Muaad, A. Y., Monday, H. N., and Nneji, G. U. (2022). Automated lung-related pneumonia and covid-19 detection based on novel feature extraction framework and vision transformer approaches using chest x-ray images. *Bioengineering*, 9(11).

Uparkar, O., Bharti, J., Pateriya, R., Gupta, R. K., and Sharma, A. (2023). Vision transformer outperforms deep convolutional neural network-based model in classifying x-ray images. *Procedia Computer Science*, 218:2338–2349. International Conference on Machine Learning and Data Engineering.

Vollmer, S., Mateen, B. A., Bohner, G., Király, F. J., Ghani, R., Jonsson, P., Cumbers, S., Jonas, A., McAllister, K. S., Myles, P., et al. (2020). Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. *bmj*, 368.

Walsh, S. L., Calandriello, L., Silva, M., and Sverzellati, N. (2018). Deep learning for classifying fibrotic lung disease on high-resolution computed tomography: a case-cohort study. *The Lancet Respiratory Medicine*, 6(11):837–845.

Wang, L., Lin, Z. Q., and Wong, A. (2020). Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images. *Scientific reports*, 10(1):19549.

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.

Watadani, T., Sakai, F., Johkoh, T., Noma, S., Akira, M., Fujimoto, K., Bankier, A. A., Lee, K. S., Müller, N. L., Song, J.-W., et al. (2013). Interobserver variability in the ct assessment of honeycombing in the lungs. *Radiology*, 266(3):936–944.

Weidman, S. (2019). *Deep learning from scratch: Building with python from first principles*. O'Reilly Media.

Yadav, S. S. and Jadhav, S. M. (2019). Deep convolutional neural network based medical image classification for disease diagnosis. *Journal of Big data*, 6(1):1–18.

Yanase, J. and Triantaphyllou, E. (2019). The seven key challenges for the future of computer-aided diagnosis in medicine. *International journal of medical informatics*, 129:413–422.