

https://dspace.org/

Deep Multi-Instance Learning for automated pathology screening in frontal chest radiographs

Gerrand, Jonathan David

2018

https://hdl.handle.net/10539/26625

Gerrand, Jonathan David (2018) Deep multi-instance learning for automated pathology screening in frontal chest radiographs, University of the Witwatersrand, Johannesburg, <http://hdl.handle.net/10539/26625>

Downloaded from a DSpace repository. https://dspace.org/

Deep Multi-Instance Learning for Automated Pathology Screening in Frontal Chest Radiographs

Jonathan David Gerrand

A dissertation submitted to the Faculty of Engineering and the Built Environment, University of the Witwatersrand, Johannesburg, in fulfilment of the requirements for the degree of Master of Science in Engineering.

Johannesburg, September 2018

Declaration

I declare that this dissertation is my own, unaided work, except where otherwise acknowledged. It is being submitted for the degree of Master of Science in Engineering to the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination to any other university.

Signed this ____ day of _____ 20__

_____

Jonathan David Gerrand

Abstract

In the face of an ever increasing patient burden and limited medical expertise, developing nations require effective screening strategies to deal with the high mortality rates associated with infectious diseases such as pulmonary TB and pneumonia. The recent use of deep convolutional neural network (DCNN) models to perform automated chest X-ray screening presents a viable solution to this need. However, in order to operate with high accuracy, these deep models are often required to detect subtle cases of pathology under weak training supervision.

This dissertation addresses the challenge of improving DCNN-based pathology classification under a weakly-supervised setting, by developing a multi-instance learning (MIL) method to detect local discriminative information within weakly-labelled frontal chest X-ray images.

The developed MIL method extends from previous works in the literature, and consists of two training stages which can be fully automated to enable end-to-end learning. The first stage sensitises a model to local sources of discriminative information, while the second stage boosts this model to recognise non-discriminative sources.

Validation of the method is initially performed on two synthetic datasets, after which it is experimentally tested across four chest X-ray datasets containing pathological findings for TB and pneumonia. The overall results show that the MIL method enables the detection of small and subtle findings for pathology, outperforming conventional weakly-supervised classification for pulmonary TB detection, while producing poorer performances in detecting pathologies with larger spatial extents. Consideration of these findings motivates for the combined use of the MIL method along with other weakly supervised techniques, such holistic classification, in order to improve overall pathology classification.

To the teachers in my life who cared,

because you understood the true power of your profession.

Acknowledgements

I would like to recognise and my express deep gratitude to my supervisors, Adam Pantanowitz, Dr. Quentin Williams, and Dr. Dalton Lunga, for the mentorship and direction they have offered throughout the stages of my research journey. Their belief in this work and continued optimism have been of great worth.

I gratefully acknowledge the Council for Scientific and Industrial Research (CSIR) for its funding of this research.

I further wish to acknowledge and thank Dr. Nasreen Mahomed for her instruction and assistance surrounding the medical aspects of this work. The time Nasreen has taken to assist with the labeling of radiographs and their interpretation is greatly appreciated.

To my parents, Priscilla and Daniel, I express my sincere appreciation for the value they have always placed on learning and its unending pursuit.

Finally, I would like to give my heartfelt thanks to my dear wife, Kirsten, who has patiently loved, supported and inspired me throughout my studies.

Contents

Declaration i

Abstract ii

Dedication iii

Acknowledgements iv

Contents v

List of Figures ix

List of Tables xi

Nomenclature xii

1 Introduction 1

1.1 Background 1

1.2 Problem Statement 3

1.3 Research Objectives 4

1.4 Thesis Statement 5

1.5 Dissertation Structure 5

2 Supporting Theory 7

2.1 Introduction 7

2.2 Chest Radiography 7

2.2.1 Why Chest X-rays? 8

2.2.2 Examination Procedure 8

2.2.3 PTB and Childhood Pneumonia Screening 12
2.3 Computer Aided Diagnosis (CAD) 13
2.3.1 Rule-based and Machine Learning Approaches 14
2.3.2 CAD and Computer Automated Diagnosis 15
2.3.3 Application to PTB and Pneumonia Screening 16
2.4 Deep Convolutional Neural Networks 19
2.4.1 Traditional Artificial Neural Networks 19
2.4.2 Convolutional Neural Networks 20
2.4.3 Enabling Deeper Learning 22
2.4.4 Common Building Blocks 23
2.4.5 Training 28
2.4.6 Regularisation 32
2.4.7 Cited Models 33
2.5 Deep Convolutional Neural Networks in CAD 36
2.5.1 Enabling End-to-End Learning 36
2.5.2 Dataset Limitations 37
2.5.3 Transfer Learning and Fine-tuning 39
2.5.4 Patch-based versus Holistic Classification 41
2.5.5 Application to Chest Radiography 43
2.5.6 Multi-instance Learning Approaches 46
2.6 Summary 50

3 Deep Multi-instance Learning for Radiographic Chest Screening 51

3.1 Introduction 51
3.2 Existing Techniques 51
3.3 Multi-instance Deep Learning Method 54
3.3.1 Stage 1 - Discriminative Patch Classification 55
3.3.2 Stage 2 - Non-Discriminative Instance Boosting 57
3.3.3 Testing , Validation and Inference 61
3.3.4 Computational Cost of MIL 62
3.4 Summary 63

4 Model Selection, Training and Evaluation 64

4.1 Introduction . . . 64

4.2 Model Architecture Selection . . . 65

4.2.1 GoogLeNet . . . 65

4.2.2 Network-in-Network . . . 66

4.3 Training . . . 67

4.3.1 Cross-Validation . . . 67

4.3.2 Data Augmentation . . . 68

4.3.3 Model Initialisation and Fine-tuning . . . 69

4.3.4 Regularisation . . . 70

4.3.5 Hyperparameter Selection . . . 72

4.4 Testing and Analysis . . . 74

4.4.1 Performance Metrics . . . 74

4.4.2 Confidence Intervals . . . 78

4.4.3 Saliency Maps . . . 79

4.5 Experimental Setup . . . 80

4.6 Summary . . . 81

5 Validation of Deep Multi-instance Learning for Radiographic Screening 82

5.1 Introduction . . . 82

5.2 Datasets and Data Preparation . . . 83

5.3 Experimental Setup . . . 84

5.4 Results and Discussion . . . 85

5.4.1 CXR - Small . . . 86

5.4.2 CXR - Large . . . 89

5.5 Efficacy of MIL Method Training Stages . . . 92

5.6 Summary . . . 94

6 Experimental Comparison of MIL and Holistic Classification 95

6.1 Introduction . . . 95

6.2 Datasets and Data Preparation . . . 95

6.3 Model Selection, Training and Evaluation . . . 97

6.4 Indiana OpenI Opacity Dataset . . . 98

6.4.1 Description . . . 98

6.4.2 Results . . . 99

6.5 PCV-13 Paediatric Pneumonia Dataset . . . 103

6.5.1 Description . . . 103

6.5.2 Results . . . 104

6.6 ChestX-ray14 Infiltration Dataset . . . 108

6.6.1 Description . . . 108

6.6.2 Results . . . 109

6.7 Montgomery-Shenzhen Pulmonary Tuberculosis Dataset . . . 113

6.7.1 Description . . . 113

6.7.2 Results . . . 115

6.8 Implementation Considerations . . . 118

6.9 Discussion . . . 121

6.9.1 Holistic versus Multi-Instance Learning . . . 121

6.9.2 Comparison to Human Performance . . . 122

6.10 Summary . . . 124

7 Conclusion 126

7.1 Overview . . . 126

7.2 Summary of Findings . . . 127

7.3 Research Conclusions . . . 128

7.4 Suggestions for Future Work . . . 128

A MHREC - Clearance Certificate 146

List of Figures

2.1 Common chest X-ray orientations. 9
2.2 Chest radiograph of an adult male. 10
2.3 Traditional CAD Pipeline for chest radiography. 14
2.4 Feed-forward neural network. 20
2.5 Visual representation of 2-D convolution. 23
2.6 Depth within a convolutional layer. 24
2.7 Max and average pooling operations. 26
2.8 Individual neuron. 27
2.9 Common non-linear activation functions. 27
2.10 Comparison of traditional and DCNN-based CAD systems. 37
2.11 DCNN fine-tuning. 39
2.12 Patch-based versus holistic classification. 42
2.13 Depiction of multi-instance learning. 47
3.1 Attention model accuracy vs lung nodule size. 53
3.2 High-level overview of the multi-instance deep learning method. . . . 54
3.3 Stage-1 of the multi-instance deep learning method. 56
3.4 Stage-2 image patch sampling. 60
3.5 Stage-2 non-discriminative instance boosting. 60
4.1 The GoogLeNet architecture. 65
4.2 The Network-in-Network architecture. 66
4.3 Data augmentation operations. 68
4.4 Learning rate grid search. 72
4.5 Possible outcomes for binary classification. 74

4.6 Example ROC curves. 77
5.1 Examples from the ‘Small’ and ‘Large’ synthetic CXR datasets. . . . 83
5.2 MIL validation patch settings. 85
5.3 Accuracy results for the‘CXR - Small’ dataset. 86
5.4 ROC curves for the ‘CXR - Small’ dataset. 87
5.5 Saliency maps generated for the ‘CXR - Small’ dataset. 88
5.6 Accuracy results for the ‘CXR - Large’ dataset. 89
5.7 ROC curves for the ‘CXR - Large’ dataset. 90
5.8 Saliency maps generated for the CXR-Large dataset. 91
5.9 Synthetic CXR dataset patches. 93
6.1 MIL evaluation patch settings. 97
6.2 Example CXRs from the Indiana OpenI dataset. 99
6.3 Model accuracies for the Indiana OpenI dataset. 100
6.4 ROC curves for the Indiana OpenI dataset. 101
6.5 Saliency maps for the Indiana OpenI dataset. 102
6.6 Example CXRs from the PCV-13 dataset. 103
6.7 Model accuracies for the PCV-13 dataset. 104
6.8 ROC curves for the PCV-13 dataset. 106
6.9 Saliency maps for the PCV-13 dataset. 107
6.10 Example CXRs from the ChestX-ray14 dataset. 108
6.11 Accuracy results for the ChestX-ray14 Infiltration dataset. 109
6.12 ROC curves for the ChestX-ray14 infiltration dataset. 111
6.13 Saliency maps for the ChestX-ray14 infiltration dataset. 112
6.14 Example CXRs from the Montgomery-Shenzhen dataset. 113
6.15 Accuracies obtained for the Montgomery-Shenzhen dataset. 114
6.16 ROC curves for the Montgomery-Shenzhen dataset. 116
6.17 Saliency maps for the Montgomery-Shenzhen dataset. 117
6.18 Model training and validation curves. 119
6.19 Overall model performance. 121
6.20 Comparison to human reader performance. 123

List of Tables

2.1 Summary of popular DCNN architecture characteristics. 34

4.1 Selected experimental hyperparameters 71

4.2 Hardware and software specifications for experimentation. 80

5.1 'Small' and 'Large' CXR dataset statistics 84

5.2 Stage-1 vs Stage-2 performance for the synthetic CXR datasets. . . . 92

6.1 Utilised CXR datasets for evaluation. 96

6.2 Performance results for the Indiana OpenI dataset. 99

6.3 Performance results for the PCV-13 dataset. 105

6.4 Performance results for the ChestX-ray14 Infiltration dataset. 110

6.5 Performance results for the Montgomery-Shenzhen dataset. 115

6.6 Average model training times. 119

6.7 Overall performance of the holistic and MIL classification methods. . 120

6.8 Comparison of CXR reading performance for human readers. 124

Nomenclature

AP	Anterior-Posterior
AUC	Area Under Curve
CAD	Computer-Aided Diagnosis
CI	Confidence Interval
CNN	Convolutional Neural Network
CT	Computed Tomography
CV	Cross-Validation
CXR	Chest X-Ray
DCNN	Deep Convolutional Neural Network
FN	False Negative
FP	False Positive
ILSVRC	ImageNet Large Scale Visual Recognition Competition
LOO	Leave-One-Out
MIL	Multi-instance Learning
PA	Posterior-Anterior
PEP	Primary End-point Pneumonia
PTB	Pulmonary Tuberculosis
ROC	Receiver Operating Characteristic
ROI	Region of Interest
SGD	Stochastic Gradient Descent

SVM Support Vector Machine

TN True Negative

TP True Positive

WHO World Health Organisation

Chapter 1

Introduction

This dissertation presents a multi-instance, deep learning method for the automated screening of frontal chest radiographs in adult and paediatric populations. The goal of the multi-instance method is to enable deep convolutional neural networks, trained under a weakly-supervised setting, to perform this automated screening with increased accuracy compared to conventional holistic classification.

1.1 Background

Samora Machel, the late former president of Mozambique, once lamented that "[t]he rich man's dog gets more in the way of vaccination, medication and medical care than do the workers upon whom the rich man's wealth is built". While this statement may be considered extreme, when viewed in the light of a recent report from the WHO [1] which states that individuals infected with pulmonary tuberculosis (PTB) in developing African countries are three-times more likely to die solely because of inadequate health-care, one is given pause for consideration. Another report from the World Economic Forum offers little consolation to this reality, indicating that by the year 2030, twelve times the current volume of medical physicians will be required in developing countries to meet an ever-growing patient burden [2]. Indeed, in the face of such challenges, solutions are required which will equip modern clinicians with the means to provide the quality of health care that patients in developing parts of the world (such as Sub-Saharan Africa) deserve. To this end, intelligent technologies such as Computer-Aided Diagnosis (CAD) can be considered.

CAD refers to a technology within the medical imaging domain which sees the combination of medical imaging, machine learning and computer vision algorithms to enable a computing device to infer information from a medical image. This information can assist medical workers and clinicians to interpret the often complex medical images examined and assist in activities such as image comprehension, automated patient

screening or pathology diagnosis. The use of CAD is further seen to enable the rapid processing of images within these activities, which are typically time-consuming in nature.

Within resource-scarce nations, currently one possible high-impact area for the application of CAD is its utilisation as a radiographic screening tool. Recent estimates place pneumonia and PTB as leading causes of mortality in child and adult populations within developing countries [1, 3], with an estimated 1.3 million African children dying from pneumonia in 2011, and 1.8 million individuals in developing nations dying from PTB in 2016. These lower respiratory infections can be detected through the use of chest radiography, and as such avail themselves to be detected through the use of CAD.

With the above in mind, recent advances in a field of machine learning known as Deep Learning have lead to marked improvements in the accuracy of CAD tools. Here the use of deep convolutional neural networks (DCNN) have refined the state-of-the-art across multiple imaging modalities such as seen within [4], [5] and [6].

One drawback surrounding the usage of deep learning techniques, however, are their heavy reliance on large amounts of data. Unlike the computer vision field where sizeable datasets such as ImageNet [7] (which contains over 12 million labelled images) are used to train deep learning models, datasets within the medical imaging field are considered large if they contain several hundred images. This scarcity of data negatively affects the performance of trained DCNN models [8].

It might seem logical then to simply create larger medical image datasets in response to this predicament. Unfortunately several factors, including the time and cost required to prepare substantial amounts of medical data on a specific modality or pathology, hinder progression. To make matters worse, within the medical imaging domain most state-of-the-art techniques for classification require Region-of-Interest (ROI) labelled data in order to be trained [5]. This labelling, which sees specific pathological areas being outlined from the background of a medical image, further adds to the complexity required to make a medical image dataset and greatly limits their size.

The two challenges highlighted above can be tackled using different approaches. To compensate for the small quantities of data found within medical imaging datasets a technique known as deep fine-tuning [9] has been investigated, with recent results showing its suitability for several medical imaging tasks [10, 11]. Fine-tuning does not, however, remove the need for ROI labelling during training. For this to occur the actual setting under which DCNN training occurs needs to be relaxed from a supervised to a weakly-supervised one in which a model only receives image-level (instead of region-level) information.

Recently fine-tuned DCNN models have been explored in a weakly-supervised setting, known as holistic image classification, for CAD-based frontal radiographic screening [12, 13, 14]. Under this setting, a DCNN model is trained using only image-level pathology labels and entire radiograph images as its inputs to perform classification.

Holistic image classification eases the ROI labelling constraint on radiograph datasets at the cost of providing a far more difficult classification task for models; one which requires them to both detect and classify local pathological regions while ignoring noisy global information present within the input. It imposes a further constraint when one considers the typical fixed-size inputs of DCNN architectures. Here the input sizes of these architectures are far smaller in comparison to the radiograph images they are required to process. Thus the classification of pathology is required to occur in the absence of certain local information within an input image which may be lost during its spatial compression.

As an alternative to this holistic approach, Yan et al. [15] have investigated weak-supervision for medical image classification through the creation of a multi-instance, deep learning framework. Within their approach, which is tailored for body-part recognition from slice-level computerised tomography (CT) scan images, the authors detail the ability of their framework to outperform holistic image classification. This performance increase is explained as their method is capable of detecting local regions, which are discriminative for classification, within the CT slice images while ignoring non-discriminative global information. However, the authors perform their research using an ad-hoc DCNN which is shallower then those used to define the state-of-the-art within other CAD applications. Furthermore, they exclude the use of fine-tuning which has been shown to increase the performance of the holistic image classification setting [10].

1.2 Problem Statement

The above introduction allows for the consideration of a research problem in which the merits of a multi-instance, deep learning based approach for weakly-supervised frontal chest X-ray (CXR) screening is studied. In this sense one could ask which technique, whether holistic or multi-instance based, is more appropriate for training DCNN models for pathology screening.

This research seeks to investigate the merits of a MIL-based method used to train a DCNN for pathology screening within image-level labelled frontal CXRs. Indeed, it seeks to evaluate how the use of local information (obtained from the multi-instance

method) affects the performance of a fine-tuned DCNN relative to current holistic image classification methods.

As of yet, a deep learning based MIL method for frontal CXR pathology classification has not been investigated. The development of such a method could improve the performance of CAD-based chest screening interventions motivated within this text. The exploration of the trade-offs, both in terms of performance and implementation for this method also holds merit in understanding its real feasibility for use.

It is further noted that very little work prior to this point has been done to investigate the efficacy of any weakly-supervised deep learning methods (holistic or MIL-based) for frontal chest radiograph screening on a paediatric population [16]. This age group, which is often considered most vulnerable within disease burdened countries, deserves the improved interventions which could be offered by these methods.

1.3 Research Objectives

In order to investigate the research problem defined, the following research objectives are set forth:

- **To evaluate the performance of holistic chest radiograph classification using DCNN architectures fine-tuned on image-level labelled CXRs.** This objective seeks to provide a baseline to which experimental results of the developed multi-instance method can be compared. This objective will further explore the efficacy of the holistic image classification method for paediatric populations.
- **To develop an appropriate multi-instance deep learning method capable of performing frontal chest radiograph screening using these same deep CNN architectures and image data.** The work of Yan et al. [15] has provided evidence that a multi-instance deep learning method for body-part recognition can be applied to the CT modality. The successful adaption of this method for pathology screening with chest radiographic images will provide a novel technique of weakly-supervised classification within this field.
- **To gauge the relative performance of both the holistic and multi-instance classification methods.** Obtaining a performance comparison between these two methods will allow for a pronouncement to be made surrounding the suitability of the respective methods for frontal CXR screening,

as well as a critical discussion surrounding the merits of the latter proposed method.

- **To establish trade-offs inherent in the selection between these two methods.** The achievement of this objective holds practical significance for the implementation of the multi-instance deep learning method and can advise the research community surrounding the feasibility of such a method within the physical world.

1.4 Thesis Statement

In consideration of the above research problem and objectives, the thesis of this work is:

A multi-instance learning method, developed to extract local discriminative information from image-level labelled CXR images, will increase the performance of a DCNN fine-tuned for pathology screening within frontal CXR images, relative to a holistic image classification method.

Within the thesis statement, the term performance is restricted to detection classification metrics including accuracy, sensitivity, specificity and receiver operating characteristics which the method demonstrates. Metrics pertaining to the localisation performance of the method are not considered.

This work also only concerns itself with the detection of pathological findings related to adult and paediatric PTB, as well as paediatric pneumonia. The developed multi-instance deep learning method will be tested in a binary manner (abnormal vs normal) using datasets which are representative of these settings. The selected target datasets are detailed in Sections 6.4, 6.5, 6.6 and 6.7 respectively.

Furthermore, this work only considers two DCNN architectures presented within Sections 4.2.1 and 4.2.2 respectively, with fine-tuning being initiated from the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2012 dataset.

1.5 Dissertation Structure

The dissertation will take the following structure:

Chapter 2 reviews ***background*** information pertaining to the research problem. Initially a broad review of the relevant theory, including radiographic screening, computer aided diagnosis and machine learning techniques is given. The chapter

then narrows in scope, detailing firstly the theory surrounding deep CNN models and secondly their use within modern medical image processing.

Chapter 3 is responsible for introducing the main contribution of this work; a ***multi-instance deep learning method*** for pathology detection in frontal CXR's. A review of similar multi-instance and deep learning methods is first conducted. The formalisation and motivation of the method is then presented in detail.

Chapter 4 describes aspects of the work's ***methodology*** relating to model selection, training and evaluation. The choice of two DCNN models commonly studied within the medical imaging community is motivated. The procedure used to train these models, including fine-tuning and data augmentation is then discussed. Finally, the metrics and techniques used to assess the performance of these trained models are provided.

Chapter 5 ***validates the proposed method experimentally***. Two synthetic datasets are formed which allow the operating characteristics of the method to be gauged. Using these datasets, the advantage of the multi-instance method over holistic image classification for the detection of local discriminative information is established.

Chapter 6 details the ***results and analysis*** for the comparison of the multi-instance deep learning method to holistic classification. Four CXR datasets containing pathological findings relating to pulmonary TB and paediatric pneumonia are initially introduced and are used to evaluate the methods across several metrics. The results from these evaluations are then used as a basis to discuss the merits of the proposed method, as well as its computational cost relative to holistic classification.

Chapter 7 provides a ***summary*** of the findings of the work. These findings, which are related to the research objectives presented within this work, allow for ***future work*** to be considered and presented.

Chapter 2

Supporting Theory

2.1 Introduction

This chapter presents a review of the supporting theory and literature relevant to the research considered within this dissertation.

Initially the use of chest radiography and its automated reading via computer-aided diagnosis (CAD) is motivated to perform screening for pulmonary tuberculosis (PTB) and paediatric pneumonia. The development, structure and training of deep convolutional neural networks (DCNNs) models are then presented, which enables a critical discussion surrounding the improvements and limitations associated with the use of these models in recent CAD research. At the conclusion of this discussion, a basis is formed upon which the use of multi-instance learning (MIL) can be considered as an alternative to holistic classification for DCNN-based automated screening.

2.2 Chest Radiography

In November of 1895, German-born physicist Wilhelm Conrad Rontgen stumbled across the first medical application for X-rays. Within his laboratory at the University of Wurzburg, Rontgen discovered that when the hand of his wife was placed between a cathode ray tube and a photographic plate, a silhouette of the bones in her hand appeared when the plate was developed. This was the first ever 'Rontgenogram', or radiograph, ever produced [17].

Since their discovery, the medical use of x-rays, known as radiography, have seen tremendous growth and have paved the way for numerous fields of study, such as mammography, fluoroscopy, and angiography [18]. These imaging techniques have been of great benefit within the practice of medicine.

This section describes the field of chest radiography, providing motivation for its continued use both within general medical contexts and especially as a screening tool for infectious lung diseases within developing nations.

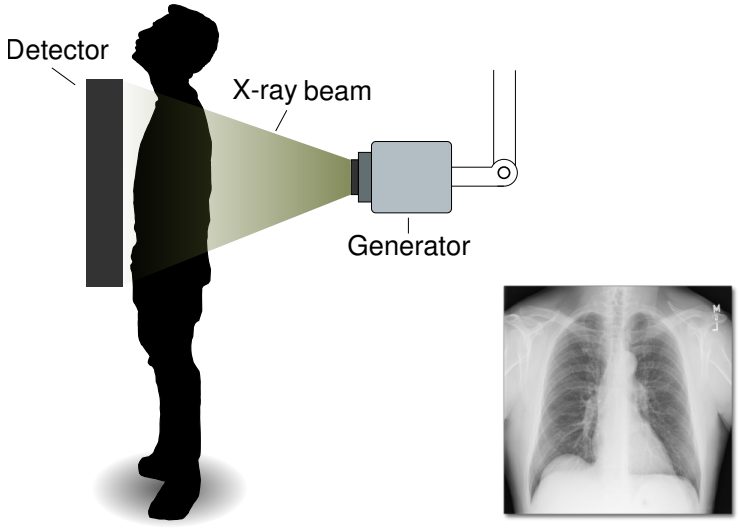
2.2.1 Why Chest X-rays?

The chest has been described as the 'mirror of health and disease' [18]. Indeed, even though more sophisticated radiological procedures have been discovered within the field of medical imaging, the chest X-ray (CXR) is one of the most common examinations which patients undergo - constituting around a third of all the procedures within radiological departments [18, 19]. From a diagnostic perspective, the extensive use of the CXR is due to the amount of information which can be gauged surrounding the health and condition of a patient based on the interpretation of their chest radiograph. Obtaining a radiograph and its interpretation can also be performed at a greater speed compared to other medical imaging techniques [18]. Other additional factors which make the use of CXRs favourable include their lower operational cost, the low amounts of radiation they expose patients to and their versatility; with portable radiograph scanners being used extensively at the bedside of patients within intensive care units [19].

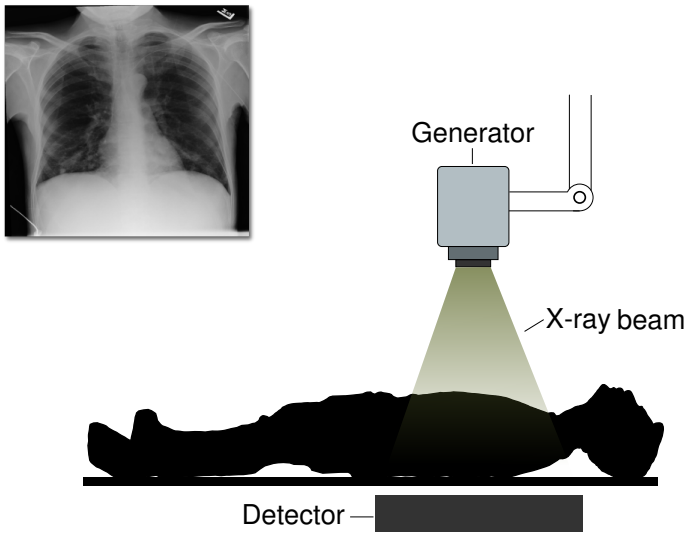
In the context of Africa, several technological advances in digital radiography have increased access levels to basic radiology services within its developing nations. Within these nations, which are often typified by their low to middle income status, multiple factors traditionally limit the use of chest radiography. These include intermittent electricity and clean water supply (required for system power and film development), the prohibitive cost of older radiography systems and the inaccessibility of certain medical facilities in order to perform regular system maintenance [20]. However, units such as the World Health Imaging System for Radiography (WHIS-RAD) [21] have been designed specifically with these limitations in mind; operating in remote areas with minimal maintenance, being powered via battery or solar energy sources and remaining less expensive than previous radiographic systems. As such, the ability of individuals to receive a CXR for diagnostic and screening purposes has continued to increase [20].

2.2.2 Examination Procedure

While the difficulty of obtaining a CXR has been greatly simplified by recent technological developments, correct interpretation thereof remains a challenging task. This complexity stems from the nature of radiographic images; with multiple anatomical



(a) Posterior-anterior (PA) position

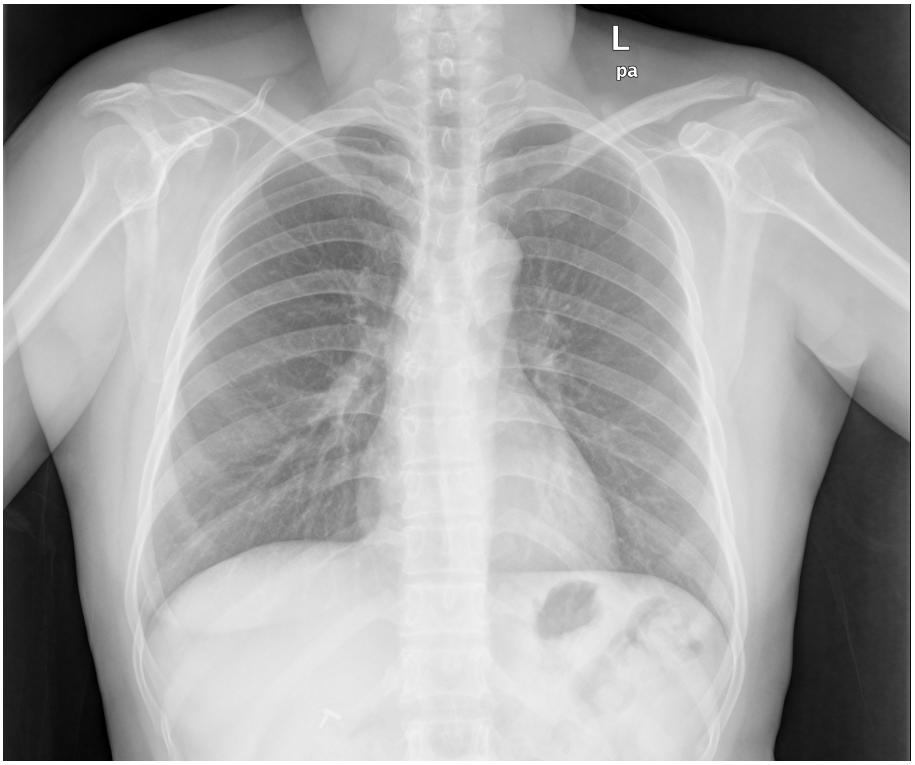


(b) Anterior-posterior (AP)

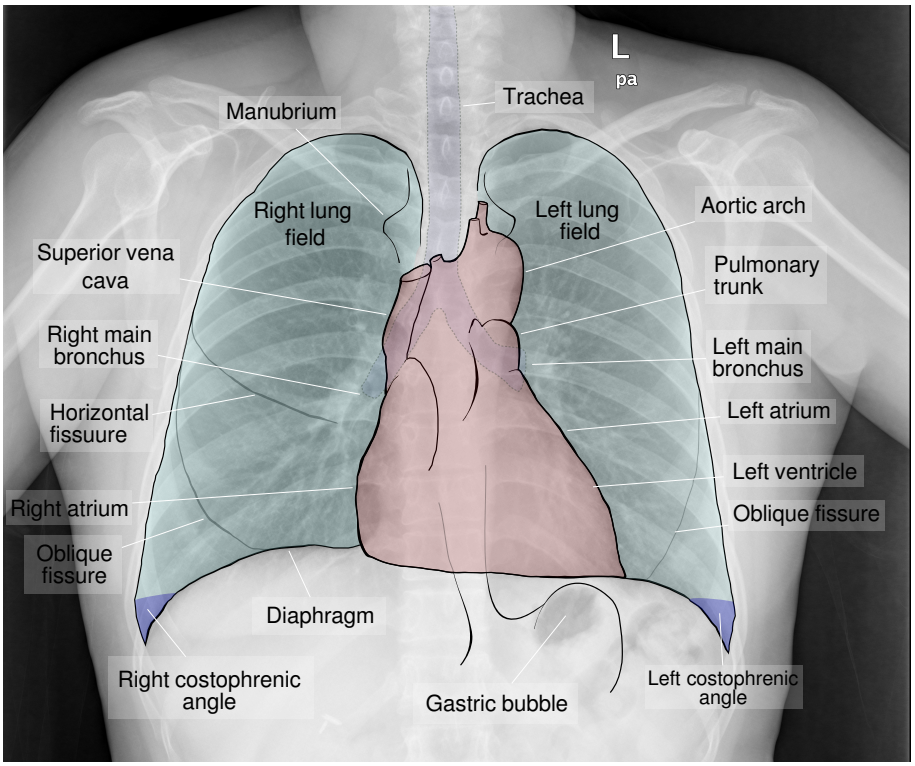
Figure 2.1: Common patient orientations during chest X-ray capturing. (Source: Personal collection. CXR images obtained from the ChestX-ray14 dataset [14], CC0 1.0)

structures within the thoracic cavity being superimposed upon one another. An understanding of how these images are obtained allows for this complexity to be appreciated.

For the typical posterior-anterior (PA) CXR projection, shown in Figure 2.1, an individual is stood upright, with their chest facing a detector plate which is sensitive to excitation by X-rays [18]. A source is positioned behind the individual and is responsible for the production of these X-rays which enter the individual's back and pass out through his/her chest onto the detector. X-rays are attenuated by



(a) Standard radiograph.



(b) With anatomical regions indicated.

Figure 2.2: Chest radiograph of an adult male in the posterior-anterior (PA) projection. (CXR image obtained from the ChestX-ray14 dataset [14], CC0 1.0, labelling assistance provided by Dr. Simon King)

materials of different densities, with air causing the least attenuation and bone the most. As such, the resultant radiographic image contains a silhouette-like impression of all the anatomical structures of the chest including the heart, lungs, vertebrae, thoracic cage and trachea [18]. While these structures are distinct anatomically, the nature of a radiographic image makes it appear as though they are overlapping, as represented in Figure 2.2. This makes the inspection of the lung fields for abnormalities especially complex. Beyond this complexity, several factors additionally deteriorate the quality of a CXR during the capturing process. These include: under/overexposure to X-rays (mainly applicable to non-digital systems); poor inspiration (the lungs being inadequately filled with air); and rotational artefacts (the source not being perpendicular to the detector) [18].

CXRs are taken in a variety of projections depending on clinical needs and what is feasible for a given patient. Typically, the upright PA projection is preferred with a left lateral CXR (x-rays penetrate the patient from his/her right side) sometimes being used as a follow up to confirm prior radiological findings [18]. In certain cases however, such as when the patient is unable to stand due to trauma, unconsciousness or non-cooperation (typically associated with paediatric patients), an anterior-posterior (AP) radiograph is taken [18]. As depicted in Figure 2.1, this projection sees a patient lying down with the detector placed behind his/her back and X-rays penetrating the front of the chest. While AP CXRs can be taken for a wider range of patients, they are generally less preferred due to effects such as cardiac and mediastinum magnification which they cause [18].

Once a CXR has been obtained, a systematic procedure is often followed for its interpretation. This procedure includes the inspection of the airways, bones, cardiac and costophrenic regions, diaphragm, extra-thoracic region and lung fields [18]. While such approaches are useful, the subtle nature of chest pathologies along with the inherent complexity of a CXR often requires that a clinician receive extensive training before they become competent at CXR interpretation. The findings of Melbye and Dale [22] support this claim, wherein it was shown that medical residents performed poorer at diagnosing pneumonia within PA and lateral CXRs as compared to experienced radiologists within the same hospital. Similar findings have been shown for paediatric patients [23, 24] as well as for other infectious diseases such as PTB [25]. It is further noted that even when experts are present, the interpretation of a CXR can require multiple individuals to reach a consensus reading [24, 26]. These important results emphasise the difficulty of CXR interpretation and the need for medical expertise in order for CXR-based examinations to be clinically meaningful.

2.2.3 PTB and Childhood Pneumonia Screening

Recently the advances surrounding radiographic systems have lead to greater interest in the use of CXRs for the diagnosis of prevalent lower respiratory tract infections such as pneumonia and PTB within developing countries. These infections are synonymous with high morbidity and mortality rates, with childhood pneumonia being responsible for an estimated 896 000 deaths within 2015 - 99% of which occurred in low to middle income countries [27] - and PTB killing an estimated 1.8 million individuals in 2016 [1]. Unfortunately, these numbers are projected to increase, with over 10.4 million new cases of PTB being reported in 2016 compared to the 9.6 million for 2015. To make matters worse, while the burden of these diseases continues to grow, reports such as that from the World Economic Forum predict that developing nations will soon face a severe deficit in the number of skilled professionals within their health care systems [2]. In response to this looming crisis, several diagnostic procedures have been considered to widely screen for these infections and to allow for the reliable application of medical treatment.

The current gold-standard for PTB detection is the identification of the bacteria *Mycobacterium tuberculosis* within sputum or pus samples obtained from an individual suspected of having an infection [28]. However, culturing this organism is a slow process which can take several months of growth before identification can occur [29]. Another approach for consideration is sputum smear microscopy, which sees sputum samples taken over one to several days and their observation under a microscope. While this technique is simple and inexpensive, its efficacy is seen to decrease significantly for patients with a low bacterial load, who have paediatric PTB, or are co-infected with HIV [30]. In contrast to these shortcomings, Xpert MTB/RIF® (GeneXpert, Cepheid, Sunnyvale, CA, USA) presents a newer molecular diagnostic test which produces quicker and more accurate results when testing for PTB. Unfortunately, Xpert is also seen to have a substantial drawback in that its high cost makes it currently unfeasible for widespread usage in poorer nations [30, 31].

In the context of screening interventions for childhood pneumonia, both clinical signs and microbiological tests are seen to perform poorly [32]. These results were recently affirmed for the use of induced sputum microscopy, wherein the study conducted by [33] highlighted the low sensitivity of the technique under multiple settings. Furthermore, Van Vugt et al. [34] studied the usefulness of clinical signs (which are difficult to obtain for paediatric patients) for diagnosing adult patients infected with pneumonia, and determined that these signs provided little diagnostic value due to their general nature. For this reason, the current diagnostic gold-standard for childhood pneumonia is considered to be the evaluation of a CXR according to

the World Health Organisation (WHO) guidelines [24]. However, it is acknowledged even by this regulatory body that in the presence of inadequate training and reader experience, high rates of inter-observer disagreement can be experienced, leading to possible misdiagnosis [35, 36].

Considering the above-stated limitations, several studies have shown the effectiveness of a radiographic examination, namely a CXR, as a diagnostic screening tool for PTB [37, 38]. Furthermore the diagnosis of childhood pneumonia via the same screening techniques has shown considerable efficacy when performed by expert radiologists [20, 39, 40]. In light of these findings, advances within digital radiography systems have enabled high quality CXRs to be produced in a rapid and inexpensive manner within developing nations, and the identification of lung diseases within these images poses an accurate screening tool which is more time and cost effective compared to other alternatives for both PTB and childhood pneumonia. Thus the widespread usage of CXR screening could prove valuable in the treatment of these prevalent diseases.

At this point, one might stop and question how this approach presents a viable solution. After all, the interpretation of a chest radiograph is a complex task requiring expert-level training, and it is this very expertise which many developing nations are seen to severely lack. In answer to this predicament, research has turned towards the use of artificial intelligence techniques to compensate for this lack of expertise, namely computer aided diagnosis.

2.3 Computer Aided Diagnosis (CAD)

CAD refers to a technique in which computer algorithms are used to assist clinicians in diagnosing radiological images [41]. The study of CAD forms an interdisciplinary field which sees the combination of computer vision, machine learning and medical image processing techniques in order to provide this diagnostic assistance [42]. The extent to which a CAD tool assists a clinician varies widely based on its application and ranges from the detection and marking of conspicuous tissue areas, referred to as computer aided detection (CADe), to the analysis of these areas and a confidence-interval diagnosis being produced (sometimes referred to as CADx) [42].

Within the following subsections, a review of CAD and its application to chest radiography is provided. A summary of the historical approaches to CAD is initially given which provides a basis to review current CAD techniques, especially those applied to PTB and pneumonia screening.

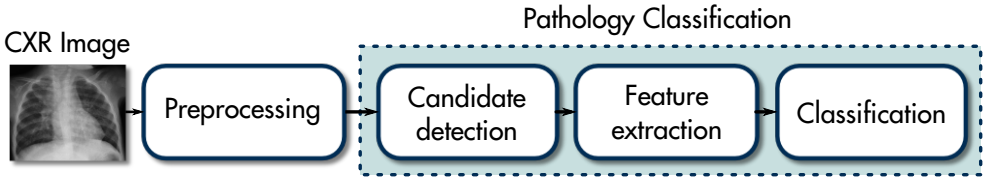


Figure 2.3: Traditional CAD Pipeline for chest radiography. (Source: Personal collection)

2.3.1 Rule-based and Machine Learning Approaches

The earliest form of CAD was introduced in 1963 by Lodwick et al. [43]. Lodwick presented a process in which features captured from the chest radiographs of patients with lung cancer could be given to a computer to predict the likelihood of their survival. The features were manually gathered, using an extensive series of visual measurements such the size and boarder definition of a tumour, and were provided in a numerical format to a computer which ran a Bayesian classifier in order to perform the prediction.

It is striking that this first approach, in which features were extracted from a radiograph and then evaluated by a computer-based learning model, was not pursued for nearly two decades thereafter [41]. During this time technological advances allowed for the digitisation of radiograph films and automatic feature extraction, however, the classification of these features was seen to occur through the use of 'rule-based' systems using 'if-then-else' type logic [41]. The work of Toriwaki et al. [44] demonstrates this principle well. Here the authors detailed a system which methodically recognised anatomical regions of the chest before it was able to make a diagnosis from this information using a series of logical threshold 'tests'. Generally, these rule-based systems had very limited success in practice, and were often seen to generalise poorly to new data presented to them due to the brittle logic they possessed [41]. In essence it was ultimately discovered that humans performed poorly at defining appropriate decision boundaries for the large quantities of data which could be extracted from medical images [41]. For this reason CAD research within chest radiography and the greater CAD community at large turned towards the use of machine learning techniques to advance performance.

Lodwick's approach to CAD made the assumption that a computer could use the features extracted from a radiograph in a more sophisticated manner than could a human. Beginning in the 1980s, the work of many scholars such as that of Doi et al. [42] validated this assumption and demonstrated the effectiveness of machine learning algorithms within CAD. Under this new approach, research efforts shifted away from defining manual classification rules to instead the discovery of more

sophisticated feature descriptors and powerful machine learning models to classify these features [41]. The latter of these efforts saw many improvements due to advances within the machine learning community, with powerful learning models such as Artificial Neural Network (ANN) and Support Vector Machine (SVM) models being introduced. However, the development and selection of appropriate feature descriptors has remained a constant challenge.

As motivated previously, the reading of a chest radiograph is a highly complex process. Numerous anatomical structures are present and superimposed upon one another, which makes the task of pathology detection susceptible to this 'anatomical noise'. This complexity has lead to the formation of many feature descriptors which are based upon concepts such as co-occurrence matrices, Fourier spectra, morphological gradients, fractal dimensions and wavelet theory [45]. A salient problem, however, surrounds the inability of these features descriptors to generalise to different pathologies or diagnostic tasks. Here the use of one or several features may be very appropriate for pathology detection under a certain setting, while being poorly suited under another. For example, [46] use several advanced segmentation and delineation techniques to detect pleural effusion within frontal chest radiographs with high accuracy, and Xu et al. [47] use a coarse to fine set of features for PTB cavity detection. Yet these approaches are of little value for their respective counterparts or any other pathology detection method. While the use of more robust feature descriptors from the computer vision community has found popularity, such as Intensity Local Patches [48], Local Binary Patterns [49] or Shift Invariant Feature Transforms [50], these gains are still limited. The work of Li et al. [51] highlighted these limitations, wherein the use of several general feature descriptors were shown to produce appreciable performance in detecting specific classes of lung disease, yet no single feature descriptor was shown to generalise across all disease classes.

These limitations have lead to the belief that manually selecting a set of optimal features for a range of tasks is generally impossible. Indeed, as observed by Van Ginneken, traditional feature selection within CAD is believed to be 'more art than science' [41].

2.3.2 CAD and Computer Automated Diagnosis

Since its inception, the goal of CAD has always been to enable computer-based systems to fully automate the process of a radiological assessment [52]. This lead to the term 'automated diagnosis' being used by the early research community to define its goal of creating computational systems which could replace radiologists entirely [42]. It was soon observed, however, that such approaches were not realisable

due to the complexities of interpreting radiological data and a lack of sophisticated algorithms which, as motivated within the above section, could not generalise well. Thus the search for an approach which could completely automate a radiological examination was diverted to one in which a system's output could be used to 'aid' a radiologist within specific tasks instead [42].

Today CAD tools are used to assist clinicians in a wide range of activities including detection [49, 53], segmentation [54, 55, 56] and suggestive evaluation [57] during the examination of medical media. In all of these assisted activities, it is the clinician who ultimately provides a diagnosis while the tool takes the form of a second opinion, providing additional information which may, or may not, have been captured by the clinician [42]. This means that it is acceptable for a CAD tool to perform with less accuracy at a given diagnostic task compared to its human counter-part, as it may indicate missed or supportive diagnostic information to them [58, 59].

Within recent years however, promising accuracy gains surrounding various CAD techniques have reignited the belief that automated diagnosis is possible for certain radiographic examinations [41]. Under this new setting, a computer's performance is required to be extremely high - equalling or surpassing that of a clinical expert - in order for its operation to be accepted [42]. For most diagnostic tasks the current performance levels of such tools are not compelling enough for their use in first-world countries, in which several experts are often available to perform radiographic interpretation and diagnosis. However, in resource constrained environments where there is a shortage of specialist expertise (such as those seen within developing nations), these performance levels have been seen to approach those expert radiologists and outperform regular clinicians [60]. It is within these contexts that the merits of high accuracy CAD systems can be evaluated for their use in performing fully-automated diagnostic tasks, such as that of a radiographic screening tool.

2.3.3 Application to PTB and Pneumonia Screening

Several authors have studied the efficacy of CAD approaches in performing automatic CXR screening for lung diseases [29, 60, 61]. In particular, Muyoyeta et al. [60] investigated the automatic scoring of chest-radiographs for patients suspected of pulmonary PTB infection within a Zambian clinic. Here the authors used an advanced CAD tool, *CAD4TB* (version 1.08) [62], which was able to assign a score out of 100 for each patient based on the likelihood of the individual having PTB. The tool assigned these scores using a number of inter-connected subsystems which performed functions such as pixel-level and shape abnormality detection, along with clavicle detection. Individuals who possessed a score above a set threshold (in this case 60)

were seen as likely carriers of the infection and were required to go for additional medical testing to provide a ground-truth diagnosis. The study showed that the tool possessed 100% sensitivity, detecting every case of PTB it was shown. In contrast to this, the specificity of the tool was only 33.1% - with 195 out of 291 patients being incorrectly labelled as likely carriers of the infection.

A similar result was obtained by Jaeger et al. [29] in which the authors compared automatic scoring obtained from their CAD system against scoring from radiologists. The system followed the traditional CAD pipeline represented within Figure 2.3, with processing stages consisting of lung segmentation via template-fit optimisation, object detection inspired feature descriptors and SVM classification. Results were derived from the Montgomery CXR dataset [26] in which there are a total of 58 positive and 80 negative cases labelled for PTB classification (further described in Section 6.7). Here the CAD system achieved an overall accuracy of 78.3%, which was only 3.7% lower than that of the radiologists (82%). However, when required to function with high sensitivity (95%), the tool obtained a specificity of 39% compared to the radiologist's 69.9%. This meant that the majority of healthy patients (48 individuals) were labelled as disease carriers when seen by the CAD system.

Philipsen et al. [31] investigated the cost and throughput efficiencies associated with automated CXR reading in a PTB screening clinic. To do this, the authors simulated the effect of pre-screening patients by capturing and reading their CXRs, before deciding whether to administer PTB testing using Xpert. Here CXR reading was performed using either an expert radiologist or an updated form of the *CAD4TB* tool (version 3.07), which facilitated automated reading. The results showed that automated pre-screening could reduce the cost per screened subject by as much as 50% compared to only using Xpert, and the cost per detected PTB case by up to 60%. Furthermore, using automated reading the throughput of the screening clinic could be increased by 250%. Despite these gains in efficiency, however, the automated approach was also seen to reduce the sensitivity of screening by 12.3% compared to only using Xpert, performing marginally poorer than a human reader. In this regard the authors argue that the use of a more accurate CAD tool could overcome this limited performance.

In regard to childhood pneumonia, there currently is an absence of work in which the performance between radiologists and CAD is directly compared. This is seen to be general for the application of CAD to paediatric populations, with few works existing to establish the effectiveness of this automated screening approach for younger patients. Two reasons may be attributed to this. First, although the use of a CXR evaluation is considered as the gold-standard for diagnosing childhood pneumonia [24], obtaining a chest radiograph from a paediatric patient can be a difficult procedure due

to their age [63]. This often leads to CXR images being obtained which are of a low quality (AP projection, with inadequate inspiration) and thus increases the complexity of CAD-based CXR interpretation [63]. Second, the burden of childhood pneumonia is highly correlated to low income countries, with no need for wide-spread screening in first-world nations. Van Ginneken observes that this absence of need within developed countries provides little financial impetus for researchers (who typically hail from wealthier nations) to invest resources into creating such technologies [64]. Instead, a need is present for researchers within developing countries to provide these solutions.

In answer to such a need, the work of Olivera et al. [65] provides a feasible basis for the study of CAD in relation to childhood pneumonia. Here the authors present a CAD system, known as *Pneumo-CAD*, which uses a combination of wavelet-based textural features and content-based image retrieval techniques to predict the likelihood of a paediatric CXR as being abnormal. While limitations exist within this study, such testing which is performed on only 20 images and the use of manually digitised CXR films (photographed in low detail), the results show the merits of such an approach with the system obtaining an accuracy of 90%. More recently, the use of this same system on a larger dataset of 158 CXR images along with various machine learning models for feature classification has produced a top accuracy of 77% [66]. While no direct comparison can be made between these CAD results and other studies which gauge the performance of human readers for paediatric CXR screening, it is important to note that the CAD results are compelling and fair similarly to human readers [24].

In consideration of the above presented findings, three salient aspects emerge:

- **The viability of automated CXR screening within high disease burden areas:** Within the works presented, the detection of chest pathologies through the use of automated CXR screening was seen to occur at high sensitivity levels. The work of Philipsen et al. [31] further motivates how these automated approaches can significantly reduce the cost and throughput of current screening strategies for diseases such as PTB.

- **The need for increased study surrounding CAD for paediatric CXR settings:** The paucity of results surrounding CAD-based approaches to paediatric populations is of great concern. This is especially true for prevalent diseases, such as childhood pneumonia, in which a CXR reading constitutes the current gold-standard for diagnosis.

- **Limitations in the efficacy of traditional CAD tools for PTB screening:** For the works of [60], [29], and [31], it was seen that the performance of other diagnostic tests such as Xpert or that of human readers attained higher

levels of accuracy. This emphasises the limitations of these CAD techniques for their use as effective screening tools, and the need for more accurate approaches to be developed.

From these points, it is seen that there is both a need and potential for CAD systems to perform CXR screening in an automated manner. However, if these systems are to be of practical use, greater levels of accuracy are required of them. In seeking for this performance improvement, it is recognised that each of the above presented CAD systems rely on traditional machine learning approaches using 'handcrafted' features. As Van Ginneken [41] and Litjens et al. [67] both observe, a natural progression from these traditional approaches is the use of machine learning techniques which are able learn the most appropriate features for a given task automatically. Such a class of machine learning techniques do exist, however, and are embodied within many of the algorithms in the field of deep learning. These algorithms, which take the form of neural network models, are created from multiple layers which are able to transform input data to output responses using a series of self-learnt feature representations [68]. Of particular importance to the medical imaging community are a set of these algorithms, known as deep convolutional neural networks (DCNNs), which have redefined the state-of-the-art for many medical imaging tasks [67]. In light of these advances, the use of DCNN models to improve the performance of CAD systems for automated CXR screening can be considered.

2.4 Deep Convolutional Neural Networks

Deep Convolutional Neural Networks (DCNNs) present as a set of machine learning models which are specifically suited for image classification tasks [67]. Several technological advances have allowed these models to be studied and trained in a feasible manner, resulting in state-of-the-art performances across several medical imaging modalities [67]. In consideration of these achievements, the goal of this section is to provide general reviews of the development, structure and training methods surrounding modern DCNN models, as well as of popular DCNN architectures studied within the literature.

2.4.1 Traditional Artificial Neural Networks

Artificial Neural Networks (ANNs) are a class of supervised learning models which were studied as early as the 1940's [70]. As depicted in Figure 2.4, the typical architecture of a feed-forward neural network consists of an input, several hidden

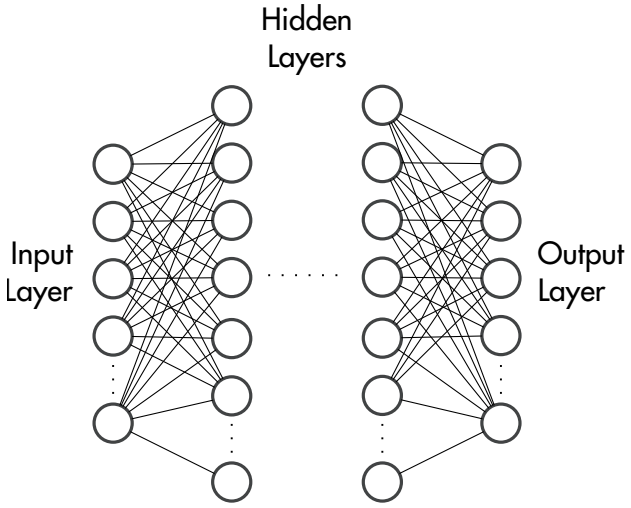


Figure 2.4: A representation of a traditional feed-forward artificial neural network model. (Adapted from: Murphy [69], 2012)

layers and an output layer which each contain a varying number of 'neural' units known as neurons. The connections within these networks are dense and acyclic, causing information to be propagated in a single direction from the input to the output layer [70]. During this forward pass through the network, each unit computes a sum based on information passed via weighted 'synapses' from other units in lower layers, to which a non-linear activation function is applied [70]. This information flow ultimately causes units within the output layer to activate, and constitutes the network's output. Learning can then occur as a correction signal is passed backwards through the network from its output to input, adjusting the activation bias of each unit and its corresponding synapse weights [70].

2.4.2 Convolutional Neural Networks

During the 1960's, research within the field of neurophysiology began to uncover mechanisms of visual pattern recognition within the portion of the mammalian brain responsible for sight, known as the visual cortex. Hubel and Wiesel [71] discovered that neurons located within the early part of a cat's visual system acted as 'detectors', responding strongly to specific patterns of light (such as horizontal bars) within a certain portion of a display, yet produced no response whatsoever for other patterns presented [72]. In addition to this, more complex neurons connected to these simple detectors were seen to be invariant to small translations within the patterns of light, producing the same response when a pattern was moved in a given direction. From these insights, general structural and processing concepts for visual perception were established [72]. These concepts included ideas such as local neural activity,

translational invariance from lower to higher neural layers, and the composition of visual scenes as two dimensional spatial maps [72].

Inspired by these findings, in 1980 Fukushima [73] proposed a neural network model known as the 'Neocognitron'. The network consisted of three layers in which both 'simple cell' and 'complex cell' sub-layers were present, and sought to replicate the mechanism of visual pattern recognition within the brain by connecting these layers in a hierarchical manner - with each succeeding layer sampling a region of the previous layer's output. Once trained, the network yielded astonishing results; learning feature representations within its various layers which could differentiate between inputs, as well as becoming mostly invariant to slight translational changes. As a drawback however, training the network itself (which occurred in an unsupervised, layer-wise manner) proved to be a computationally expensive task, and limited the size and applicability of the Neocognitron [72].

In answer to this, LeCun [74] presented a visual perception inspired model which could be trained using the more suitable backpropagation algorithm. The model, known as 'LeNet', was referred to as a Convolutional Neural Network (CNN) architecture as it replaced the simple cells of the Neocognitron with neural layers which performed 2-D convolution on their inputs. The network further replaced the complex cells of the former model with 'pooling' layers, which allowed it to still exhibit translational invariance. These changes produced an architecture which was well suited for handling image data, was more efficient to train, and yielded compelling performance in visual pattern recognition challenges [75].

Spurred on by initial promising results, researchers began investigating the suitability of CNN models for their use within CAD systems. These applications mainly involved the detection of lung nodules [76] or the discovery of other lung structures [77]. Within each of these works, the use of CNNs saw them provided with small image patches which needed to be classified as normal or abnormal. Surprisingly however, the usage of these models became less popular, with other traditional machine learning approaches maintaining a dominant position [41, 67]. The principle reason for this lack of interest stemmed from the computational effort and training difficulties associated with more complex variants of these models [41], with deeper models suffering from effects such as gradient vanishing and overfitting during training [68]. These were problems which were general to the machine learning community, with other deep network structures, such as recurrent neural networks, presenting similar difficulties during training [72].

2.4.3 Enabling Deeper Learning

In December 2012, Krizhevsky et al. [8] entered a CNN named 'AlexNet' into the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) whose depth was an unprecedented 8 layers with over 60 million parameters. Despite its size (1000x more parameters than LeNet), the network was trained successfully - claiming first position in the contest and beating the runner-up entry by over 10% in top-5 test error rate. Three important factors were seen to contribute to this milestone achievement within the computer vision community [8]:

1. **The development of more sophisticated regularisation techniques and appropriate activation functions for the training of deep models:** Previously the training of deeper neural network models suffered from vanishing/exploding gradients during backpropagation, with these models also being prone to overfitting [68]. The use of non-saturating activation functions as well as new regularisation techniques such as dropout [78] greatly reduced these undesirable properties.

2. **The availability of enormous amounts of training data:** Within the natural image domain, a need for more training data gave rise to datasets, such as ImageNet [7] and CIFAR-100 [79], whose sizes are over two orders of magnitude larger than previous ones such as MNIST [75]. This increase in data further reduced the extent of overfitting within deeper networks.

3. **Advances in the use of GPU hardware to enable efficient network training:** The availability of GPU-enabled programming languages such Nvidia's CUDA allowed for many of the computationally intensive operations required to train deep neural network models to be performed in parallel. With these enhancements, the amount of time required to train deep models was significantly reduced.

Since the entrance of AlexNet, deeper and more sophisticated DCNN models have been produced. These models have been extensively studied within the computer vision community, and have been applied within various domains, such as that of CAD. In order to understand the operation of DCNN models within these applications, a brief review of the common structural aspects, training methods and regularisation techniques associated with these models is given within the following subsections.

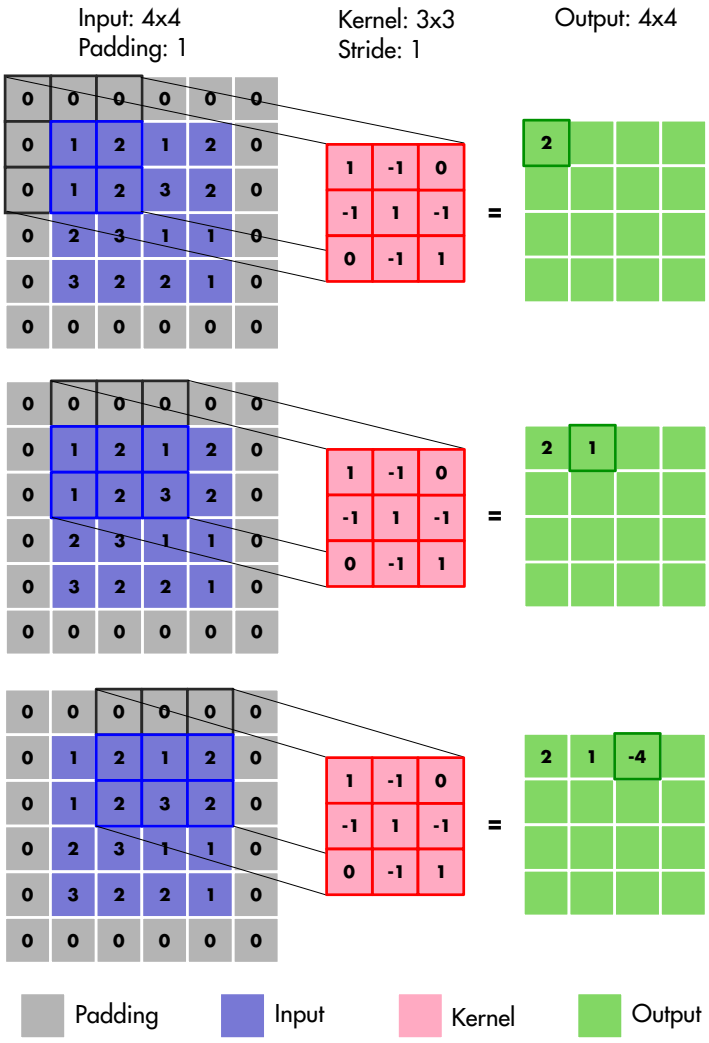


Figure 2.5: A visual representation of 2-D convolution. Here a 4x4 input is convolved with a 3x3 kernel, using a stride of 1. To ensure that the input and output maintain the same dimensions, padding information is added to the input. (Adapted from Karpathy [80], 2017)

2.4.4 Common Building Blocks

A simple view of DCNN models typically sees them comprised of several layer-types and activations which are repeated with various permutations to their parameters [72]. In this sense, a given DCNN architecture can be seen as a combination of these layers and activations in a modular, building block manner. Included within these building blocks are convolutional, pooling, fully-connected, detector and softmax layers. A brief description and mathematical formulation of these is given below.

Convolutional Layers As described by their name, convolutional layers are responsible for performing linear convolution upon their inputs [72]. These inputs,

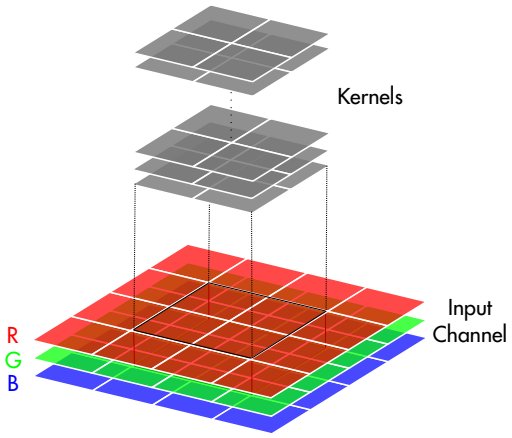


Figure 2.6: A representation of depth within a convolutional layer. Here an input image can be considered as three distinct colour channels, each representing an input receptive field, to which each convolutional kernel is applied. (Adapted from Karpathy [80], 2017)

known as receptive fields $\boldsymbol{X}$, typically consist of one or more grid-like channels of data to which several convolution kernels $\boldsymbol{K}$ are applied in order to produce further grid-like outputs known as activation maps $\boldsymbol{Z}$.

To see how these activation maps are formed, let $\mathsf{X}_{i,j,k}$ and $\mathsf{Z}_{i,j,k}$ represent units located at channel i, row j, and column k of the input and output respectively. Furthermore, let $\mathsf{K}_{i,j,k,l}$ represent a weight positioned at row k and column l of a kernel which maps from the input channel j to the output channel i. Using this notation, the output of each unit of a convolutional layer can be given as [72]:

$$\mathsf{Z}_{i,j,k} = \sum_{l=0}^{D} \sum_{m=0}^{H} \sum_{n=0}^{W} \mathsf{X}_{l,j+m,k+n} \mathsf{K}_{i,l,j,k}, \tag{2.1}$$

where D represents the depth or number of channels within the input, with H and W representing the height and width of the each kernel within $\boldsymbol{K}$.

Three important hyper-parameters govern the operation of convolutional layers [72]. First, the *spatial extent* of a kernel $H \times W$, determines its height H and width W dimensions. Second, *stride* s determines the number of times a kernel is applied to the layer input. If $s = 1$, then the kernel will be applied at every possible spatial location of the input. If $s > 1$ however, the kernel will skip $s - 1$ positions of the input every time it is applied. This can be formally expressed as:

$$\mathsf{Z}_{i,j,k} = \sum_{l=0}^{D} \sum_{m=0}^{H} \sum_{n=0}^{W} \mathsf{X}_{l,j\times s+m,k\times s+n} \mathsf{K}_{i,l,j,k}. \tag{2.2}$$

Based on the spatial extent of a kernel, the height and width of the input channels may be reduced during convolution. In order to prevent this, the third *padding* hyperparameter is specified which appends extra rows and columns of a predetermined value to the edges of the input. To assist understanding, a visual representation of a convolutional layer being applied to an input with a depth of 1 is given in Figure 2.5. To further illustrate this principle, Figure 2.6 shows how several of these kernels would each be applied to the entire depth of the layer's input.

Unlike fully connected layers, in which each input value is used to determine every output response within the layer, convolutional layers are sparsely connected and exhibit parameter sharing [72]. These properties also allow the effective depth of the model to be increased, as their presence decreases the computational burden and memory requirements of forward and backpropagation used during model training [72].

Convolutional layers are responsible for input feature detection and output feature generation. A remarkable finding in studies surrounding the convolutional kernels of sufficiently trained DCNN models is their ability to approximate the statistical structure of natural images in a hierarchical manner [9]. Here earlier kernels, especially those within the first layer of a network, have been seen to form edge and blob detectors highly similar to Gabor functions. Later kernels present feature descriptors with greater complexity which are able to detect structures such as an eye, gun or even an entire face.

Pooling Layers Like convolutional layers, pooling layers act in a sparsely connected manner, mapping sub-regions of dimension $H \times W$ in an input channel to single unit in the output [8]. These pooling operations are performed independently in a per-channel manner. They are used to increase the translational invariance of a network, while gathering local statistics of each sub-region within the input channels [72]. This gathering is akin to aggregation of the input data, and reduces the spatial extent of the pooling layers output.

The two most commonly used forms of pooling operators include *max* and *average* pooling [72]. Max pooling has the effect of propagating the highest value found within a sub-region of the input to the output, while average pooling will propagate the average of all values found within the sub-region. These two pooling operations are given by 2.3 and 2.4 respectively.

Max pooling:

$$\mathsf{Z}_{i,j,k} = \underset{0 \leq l < H, 0 \leq m < W}{\arg\max} \mathsf{X}_{i,j \times s+l,k \times s+m}. \tag{2.3}$$

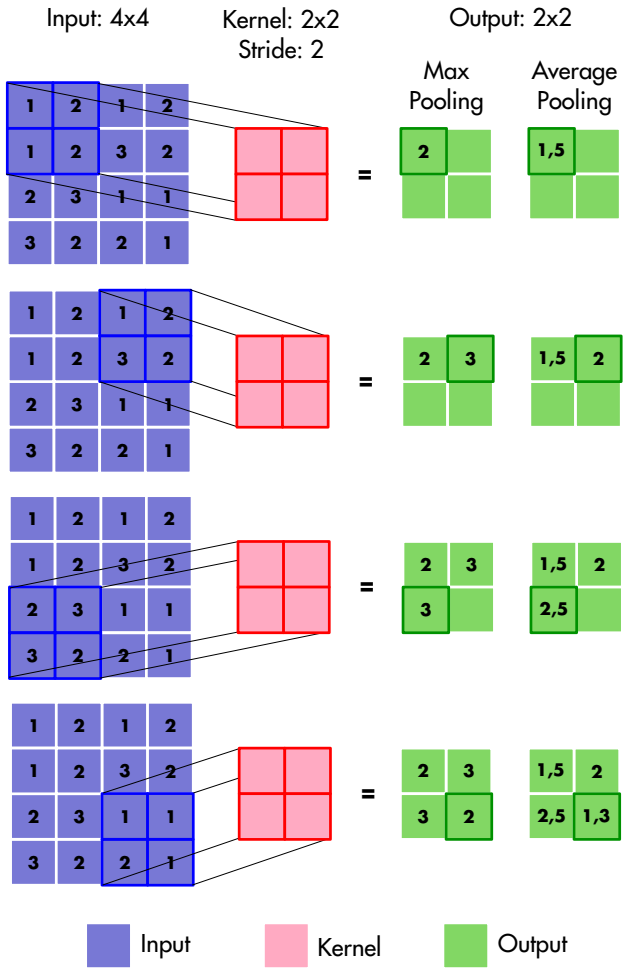


Figure 2.7: A representation of the 'max' and 'average' pooling operations. (Adapted from Karpathy [80], 2017)

Average pooling:

$$\mathsf{Z}_{i,j,k} = \frac{1}{HW} \sum_{0 \leq l < H, 0 \leq m < W} \mathsf{X}_{i,j \times s+l, k \times s+m}. \tag{2.4}$$

As with convolutional layers, pooling layers have a spatial extent ($H \times W$) and a stride (s) associated with them. These hyperparameters are often chosen such that there is no overlap between the various sub-regions of pooling ($H = W = s$) [72]. A visual demonstration of the pooling operation is provided within Figure 2.7.

Fully-connected Layers While the above-mentioned layers perform feature generation and spatial aggregation, they are not often responsible for the translation of these features into class predictions. This is the role of fully-connected layers within modern DCNN models, whose weights are trained to approximate a classifier which will map generated input features to an output vector corresponding to the correct class label/s of the input [72].

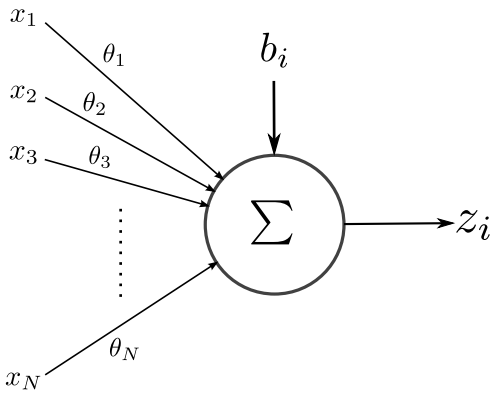


Figure 2.8: An individual artificial neuron within a neural network. (Adapted from Karpathy [80], 2017)

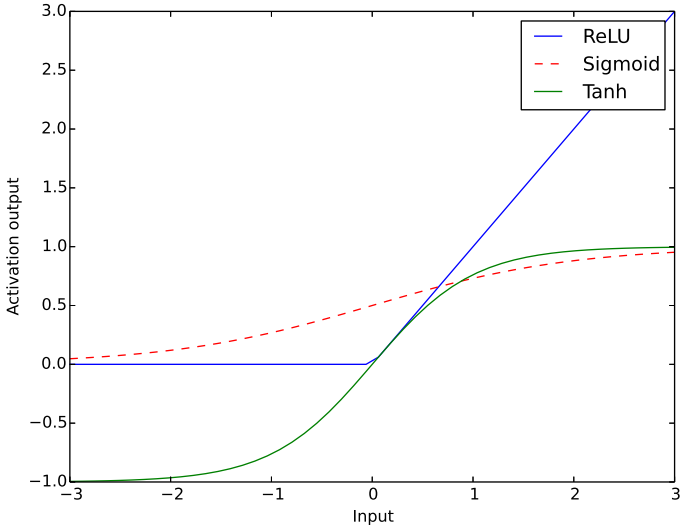


Figure 2.9: Typical activation functions used to add non-linearity to the functional approximations of neural networks. (Source: Personal collection)

Given a fully connected layer with N inputs, the output of a single unit z_i within the layer can be given as:

$$z_i = \sum_{j=0}^{N} \theta_{i,j} x_j + b_i, \tag{2.5}$$

where $\theta_{i,j}$ represents a weight mapping input x_j to z_i, and b_i is the output bias. This weighted sum can be visually seen in Figure 2.8.

Detector Layers (Non-linear Activations) Following the linear output of a convolutional or fully-connected layer, a 'detector layer' is often used to introduce

non-linearity within the functional approximations produced by a DCNN. This layer operates in an element-wise manner, mapping each of its inputs x_i to an output z_i using a non-linear activation function. Originally, activations such as the Sigmoid (2.6) and Tanh (2.7) functions were used within the research community. Due to unsatisfactory properties of these functions however, such as gradient saturation and longer training times [68], the Rectified Linear Unit (ReLU) activation (2.8) and its variants are now commonly used as activation functions [70]. A graphical comparison of these functions can be seen in Figure 2.9

$$z_i = sig(x_i) = \frac{1}{1 + e^{-x_i}} \tag{2.6}$$

$$z_i = tanh(x_i) = \frac{e^{x_i} - e^{-x_i}}{e^{x_i} + e^{-x_i}} \tag{2.7}$$

$$z_i = ReLU(x_i) = max(x_i, 0) \tag{2.8}$$

Softmax Layers In many applications it is desirable to represent the output of a network as a discrete probability distribution over the K classes it seeks to classify. To do this, the responses of the K output units of a network are fed into a final layer which applies the softmax function across its inputs x_i. The output z_i of each unit within this softmax layer can be given by:

$$z_i = \frac{e^{x_i}}{\sum_{k=1}^{K} e^{x_k}}. \tag{2.9}$$

Within most classification tasks, the output of the softmax layer represents the elements of a multinoulli or categorical distribution [69, Chapter 4]. Thus for a given input $\boldsymbol{x}_i$ to the entire network, the predicted probability for the jth class can be given as [70]:

$$P(y_j = 1|x; \boldsymbol{\theta}) = \frac{e^{\boldsymbol{z}_j}}{\sum_{k=1}^{K} e^{\boldsymbol{z}_k}}, \tag{2.10}$$

where $\boldsymbol{z}$ is a vector representing the network's output unit values, which in turn are functions of $\boldsymbol{x}$ and the network parameters $\boldsymbol{\theta}$.

2.4.5 Training

Training a neural network can be thought of as finding values for the network's parameters $\boldsymbol{\theta}$ which minimise a cost function $J(\boldsymbol{\theta})$ applied to the entire network [70].

This cost function typically comprises two components. The first component consists of a performance measure which is gauged over the complete training set, and takes the form of a loss function $L(\hat{y}, y)$ computed pair-wise between each predicted class label $\hat{y}$ and actual class label y produced from this set. The second component consists of regularisation terms $\mathfrak{R}(\boldsymbol{\theta})$ which can be used to help the network generalise to unseen sources of data [81].

Using $f(\boldsymbol{x}_i, \boldsymbol{\theta})$ to represent the model prediction $\hat{y}_i$ for a given input $\boldsymbol{x}_i$ with corresponding true label y_i, the generic cost function applied to a network trained on N datapoints can be given as:

$$J(\boldsymbol{\theta}) = \frac{1}{N} \sum_{i=1}^{N} L(f(\boldsymbol{x}_i, \boldsymbol{\theta}), y_i) + \mathfrak{R}(\boldsymbol{\theta}). \tag{2.11}$$

Having obtained the cost function, an optimisation strategy can be chosen to minimise its value by updating $\boldsymbol{\theta}$. It is noted here that this optimisation is non-convex in nature due to the non-linearity associated with neural networks [82]. As such, a widely followed optimisation strategy involves a variant of gradient descent being applied to iteratively update a network's parameters [83]. Using the cost function represented in 2.11, a single update step of general gradient descent is performed by first computing the gradient $\boldsymbol{g}$ of the cost function with respect to $\boldsymbol{\theta}$:

$$\boldsymbol{g} = \frac{1}{N} \nabla_{\boldsymbol{\theta}} \sum_{i=1}^{N} L(f(\boldsymbol{x}_i, \boldsymbol{\theta}), y_i) + \nabla_{\boldsymbol{\theta}} \mathfrak{R}(\boldsymbol{\theta}), \tag{2.12}$$

and then applying this gradient to update the parameters of the network:

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \epsilon \boldsymbol{g}. \tag{2.13}$$

Here ϵ is a scalar hyperparameter known as the learning rate, which determines the size of the update step [82]. The calculation of the gradient terms themselves, given in 2.12, is achieved using the backpropagation algorithm [70].

Several well-defined optimisation algorithms exist for the training of DCNN models. These algorithms are discussed below, with brief mentions of their variants provided for further reference.

Stochastic Gradient Descent Stochastic Gradient Decent (SGD) is a popular learning algorithm which can reduce the computational burden of training a DCNN model [83]. As shown in Equation 2.11, cost functions are generally calculated over the entire input-prediction results of a training dataset. When this calculation is

considered along with the typical sizes of the sets used to train DCNNs (which range in size from thousands to millions of images), the complete calculation of the cost function becomes prohibitive. To alleviate this problem, SGD sees the value of the cost function being calculated for random examples drawn in batches of k items from the training set [83]. These values are then used directly for updating network parameters by computing the cost gradient via backpropagation [83]. The cost function as approximated by SGD is given in Equation 2.14 (omitting the regularisation term), with the gradient estimate $\hat{\boldsymbol{g}}$ produced by this function given in Equation 2.15

$$J(\boldsymbol{\theta}) = \frac{1}{k}\sum_{i=1}^{k} L(f(\boldsymbol{x}_i, \boldsymbol{\theta}), y_i). \tag{2.14}$$

$$\hat{\boldsymbol{g}} = \frac{1}{k}\nabla_\theta \sum_{i=1}^{k} L(f(\boldsymbol{x}_i; \boldsymbol{\theta}), y_i). \tag{2.15}$$

In SGD, because the entire training set is never sampled, noise is continuously added to the parameter updates. This noise is seen to be of little concern during the initial training steps of the network, however as the training begins to converge this noise can prevent a satisfactory minimum being reached [83]. To offset this problem, the notion of learning-rate decay is introduced. Here the value of the learning-rate ϵ, decreases every iteration until a given step τ. This decay is described in Equation 2.16, where $\alpha = \frac{k}{\tau}$, ϵ_0 is the initial learning-rate and ϵ_τ is the final learning-rate (often about 1% of ϵ_0) [83].

$$\epsilon_k = (1-\alpha)\epsilon_0 + \alpha\epsilon_\tau. \tag{2.16}$$

Algorithm 2.1 describes how learning-rate decay is used alongside SGD in the weight update step within a network:

Momentum While SGD decreases the computational burden of the learning process, it can still be considerably slow [83]. For this purpose, the concept of momentum can be used to increase the speed of the learning process for a network. Momentum for SGD uses the analogy of its physical counterpart in that it accumulates an exponentially decaying moving average of past gradients, denoted as velocity $\boldsymbol{v}$, and continues to update the weights of the network based off the direction of $\boldsymbol{v}$ [83]. Using momentum, the parameters of a network can be updated according to:

$$\boldsymbol{v} \leftarrow \phi\boldsymbol{v} - \epsilon\nabla_\theta(\frac{1}{k}\sum_{i=1}^{k} L(f(\boldsymbol{x}_i, \boldsymbol{\theta}), y_i)), \tag{2.17}$$

Algorithm 2.1 Stochastic Gradient Descent (SGD) with learning rate decay. (Adapted from Goodfellow et al. [83], 2015).

Require: Initial learning rate ϵ_0, Final learning rate ϵ_τ
Require: Inintial parameters $\boldsymbol{\theta}$
Require: Number of training steps M, Final decay iteration τ
Ensure: Training set X_N
1: **for** $m \leftarrow 1, ..., M$ **do**
2: Sample k items from X_N: $\{\boldsymbol{x}_1, ..., \boldsymbol{x}_k\}, k \leqslant N$ with targets $\{y_1, ..., y_k\}$
3: Compute gradient estimate: $\hat{g} \leftarrow +\frac{1}{k}\nabla_\theta \sum_i L(f(\boldsymbol{x}_i; \boldsymbol{\theta}), y_i)$
4: Update decay factor: $\alpha \leftarrow \frac{m}{\tau}$
5: Update learning rate: $\epsilon_m \leftarrow (1-\alpha)\epsilon_0 + \alpha\epsilon_\tau$
6: Apply gradient update: $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \epsilon_m \hat{g}$
7: **end for**

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \boldsymbol{v}. \tag{2.18}$$

Here ϕ is a 'memory' hyper-parameter which determines how much impact previous velocities have on the current velocity value.

Adaptive Learning Rate Algorithms Another learning approach sees each parameter within a network being updated with an independent learning rate throughout the course of training. This describes a class of algorithms which are considered to have adaptive learning rates.

Several adaptive learning rate algorithms have been proposed for the training of deep neural networks, each showing differing levels effectiveness. One of the first of these algorithms, *AdaGrad,* was proposed by Duchi et al. [84] and adapts model parameters inversely based upon their past accumulated values. Thus parameters with large partial derivatives have significant decreases in their learning rates, while small partial derivatives have minimal reductions applied to them. While AdaGrad has been shown to work well for convex optimisation settings, several limitations including premature reductions in the learning rates of models have been identified for non-convex problems [83]. In order to rectify these limitations, Hinton et al. [85] proposed the *RMSProp* algorithm which replaced the summed accumulation of past gradients performed by AdaGrad, with an exponentially weighted moving average. The weighted average allows the learning rates to 'forget' previous non-convex structures passed through during training and is controlled by an additional memory hyper-parameter. The effectiveness of the RMSProp algorithm has made it a common choice for deep-learning practitioners.

A further alteration to RMSProp is the *Adam* or 'Adaptive Moments' algorithm [86], which combines momentum to the adaptive learning rate setting. Like RMSProp, Adam is considered to be a robust algorithm within the research community although its initial input learning rate sometimes needs alteration for specific training instances [83].

2.4.6 Regularisation

Regularisation refers to any technique applied during training which seeks to reduce the validation and test-error of a model [81]. In other words, a regularisation technique which is applied to a machine learning model seeks to increase the ability of that model to generalise. Within the context of deep learning, common regularisation techniques include weight decay, dropout, data augmentation and early-stopping [81]. These techniques are described in brief within the subsections below.

Weight Decay Weight decay, or the L^2 parameter norm penalty, is a regularisation technique used widely in the field of machine learning [81]. To apply weight decay during training, the regularisation term within 2.11 is set such that $\mathfrak{R}(\boldsymbol{\theta}) = \frac{\eta}{2}\|\boldsymbol{w}\|_2^2$. Here $\boldsymbol{w}$ represents the weight parameters of a model, and η is a hyperparameter which governs the contribution of the regularisation term to the overall cost function [81].

The intuition behind weight decay is that its use moves the weights of a model closer to zero. By doing this, the capacity of the model is reduced and its weights are prevented from overfitting to examples only present within the model's training set.

Dropout Dropout is a regularisation technique introduced by Hinton et al. [78] which effectively sees a single network being trained as an ensemble of networks. This is achieved by introducing dropout layers which cause various network branches to be 'dropped' or excluded from a training step based upon a specified dropout probability hyperparameter. This process forces the network to learn multiple representations for a given class during training and reduces the overfitting which the network may experience [81]. During network testing, these representations can be combined by setting the dropout probability to 0.

Data Augmentation Data augmentation refers to the process of training a model on additional data which is generated from the original training set [81]. The generation process involves applying predefined affine transformations, such as rotation, scaling, translation and reflections to individual images. These generated images become new training pairs in the form of (x', y), where y (image label) is

obtained from the original unaltered image. Through this process, although no new original training data is obtained for the model, the effective data volumes seen during training are amplified.

Care must be taken when performing data augmentation as certain transformations can violate class labels [81]. Examples of this violation could include reflecting the letter 'b' to 'd' or scaling the number '7' to appear as a '1' [81]. Furthermore, it is imperative that data augmentation is applied under controlled experimental conditions as its inconsistent usage can heavily influence the performance of models being compared [81].

Although data augmentation has existed since the early 1990s [87], Krizhevsky et al. [8] first showed its effectiveness in training DCNN's. Within this work it was also demonstrated that the process of augmentation could be performed on a CPU while training of the model occurred on a GPU concurrently. This essentially allowed the gains of data augmentation to come without increasing the time complexity of training [8].

Early Stopping When training models with the representational capacity sufficient to overfit on a specific task, care must be taken to prevent prolonged exposure to the same data samples during training. Overfitting can be observed in a model when after a number of iterations the model's train-error decreases while the validation-error begins to increase steadily. At this point it is favourable to stop the training process and to utilise the model parameters which produced the lowest validation-error. This concept has been formalised within the machine learning research community and is called 'early stopping' [81].

Early stopping observes the validation-error of a model during training and causes the training procedure to halt after a specified number of iterations if this error value has not decreased [81]. The amount of iterations used to wait before terminating training is referred to as 'patience' p, and can be seen to act within early stopping as given by Algorithm 2.2.

2.4.7 Cited Models

Having detailed the typical layers found within a DCNN and prominent methods used within training, a brief review of several frequently studied and successful DCNN models is provided in the paragraphs below. Key details of these models are summarised in Table 2.1:

Table 2.1: Summary of popular DCNN architecture characteristics.

Name	**Year**	**Defining feature/s**	**# Params** ($\times 10^6$)	**# Layers**	**ILSVRC-2012 error rate*** (%)
AlexNet [8]	2012	Introduced Dropout, ReLU activations, GPU-based training.	60	8	15.3
NIN [88]	2013	Multi-layer-Perceptron convolution layers using 1x1 convolution kernels.	-	9	-
VGG-19 [89]	2014	Very deep CNN structure, smaller convolution kernels (3x3).	144	19	7.3
GoogLeNet [90]	2015	Hebbian-style inception layers, computational efficiency.	5	22	6.7
Inception-V3 [91]	2016	Convolution layer factorisation.	24.5	42	5.6
ResNet-152 [92]	2016	Residual layer connections, Extremely deep architecture.	**1.9**	**152**	**3.4**

* Top-5 error rate

Algorithm 2.2 The early stopping algorithm. This technique acts as a meta-algorithm during training and ensures that a model does not overfit due to over-exposure to data. (Adapted from Goodfellow et al. [81], 2015)

Require: Number of steps between evaluations n, Maximum training steps M
Require: Patience p, Patience counter p_c
Require: Initial model parameters θ_o

$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta}_o$
$v \leftarrow 0$
$\boldsymbol{\theta}^* \leftarrow \boldsymbol{\theta}$
for $j \leftarrow 1, ..., M$ **do**
    Update model parameters: $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta}_j$
    if $j \mod n = 0$ **then**
        $v' \leftarrow \text{ValidationSetPerformance}(\boldsymbol{\theta})$
        if $v' > v$ **then**
            $p_c \leftarrow 0$
            $\boldsymbol{\theta}^* \leftarrow \boldsymbol{\theta}$
            $v \leftarrow v'$
        else if $p_c > p$ **then**
            return $\boldsymbol{\theta}^*$ ▷ Early stopping achieved
        else
            $p_c \leftarrow p_c + 1$
        end if
    end if
end for
return $\boldsymbol{\theta}^*$ ▷ Early stopping not acheived, consider increasing M

AlexNet Created by Krizhevsky et al., AlexNet constituted the first popular DCNN of the current research era and significantly outperformed all its competitors in the 2012 ILSVRC competition [8, 67]. The model pioneered the way for new research in the field by combining several techniques which overcame previous training difficulties associated with DCNNs. These techniques included the ReLU activation unit, Dropout and training the network model over several GPUs.

Network-In-Network Regarded as the philosophical inspiration behind Google's later inception-themed DCNN models [93, 67], the Network-in-Network (NIN) model consisted of three multi-layer-perceptron-convolution (mlpconv) layers which were claimed to act as 'micro neural networks' within the overall network structure [88]. Central to these mlpconv layers was the novel use of convolutional kernels with only

a 1x1 spatial extent, which increased the ability of each layer to abstract information without imposing a heavy computational burden during training [88].

VGG VGG introduced the notion of 'very deep convolutional networks' by tripling the number of layers used within a DCNN. With this increase in depth, a price was paid in the form of computational cost and parameter increase. The resulting network set a new state-of-the-art performance for the ILSVRC-2012 test dataset for the task of object detection.

GoogLeNet Szergdy et al. [91] used principals of Hebbian theory and the structure of the visual cortex to inspire the creation of *inception* layers within their DCNN. Inception layers can be decomposed into multiple parallel convolutional layers of different sizes which are layered upon one-another and whose outputs are all fed into the same pooling layer. By layering multiple inception layers together, the network was able to achieve state-of-the-art results on the ILSVRC-2012 test dataset at a fraction of the computational cost required by VGG. A visual depiction of the GoogLeNet network can be seen in Figure 4.1.

Inception-V3 Inveption-V3 was the successor of the GoogLeNet model, with larger convolution operations in the model being replaced by several layered smaller ones. This design choice allowed for Inception-V3 to contain four times the number of parameters compared to GoogLeNet, while only increasing its computational cost to train by a factor of 2.5 [91].

ResNet ResNet-152 showed the possibility for extremely deep CNNs to produce state-of-the-art results by winning the 2015 ILSVRC competition [92]. The model implemented over 150 layers (three times greater than Inveption-V3) and used residual mapping functions to help in its training. In spite of the number of layers, it is noted that this model contains 30 times fewer parameters than AlexNet and possesses the least amount of parameters compared to any other successful deep network architecture.

2.5 Deep Convolutional Neural Networks in CAD

2.5.1 Enabling End-to-End Learning

The application of DCNN models to CAD systems has advanced the state-of-the-art in many medical imaging modalities. These results have been shown to be comparable

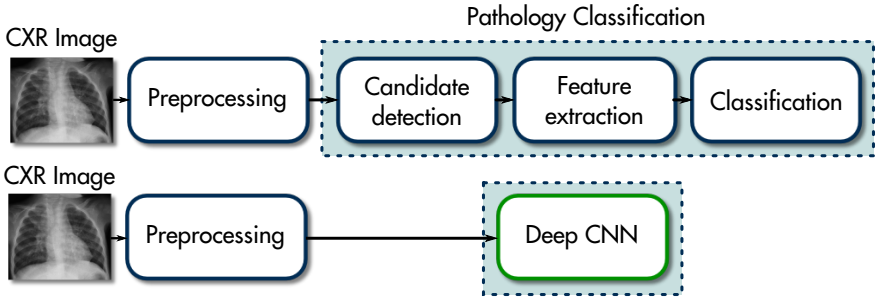


Figure 2.10: Comparison of traditional and DCNN-based CAD systems. (Source: Personal collection)

to the performance of clinicians [94], and in some cases, such as meiosis detection, have even exceeded them [95]. In [96], Anthimopoulos et al. created a DCNN model for Interstitial Lung Disease detection which outperformed handcrafted state-of-the-art methods by 14%, while in [6] the *Overfeat* CNN was used to produce feature descriptors which enabled a CAD system to detect pulmonary peri-fissural nodules with accuracy comparable to a human expert. Similar results have been obtained for brain segmentation [97], EEG pattern detection [98], and breast cancer detection [99].

A primary reason for the success of DCNN models in CAD systems is their ability to enable end-to-end learning. Here, as depicted in Figure 2.10, the complex tasks of feature selection and classification occur in a joint manner during training of a DCNN-based system. This fundamental change, reminiscent of the CAD community's abandonment of manual rule-based systems, removes the need for human experts to define the most appropriate feature descriptors for a given task, but instead allows these to be learnt from the data instead [41].

2.5.2 Dataset Limitations

The ability of a DCNN to learn features in a data-driven manner requires an inconvenient trade-off within the medical imaging domain; the need for large amounts of training data. Here 'large' can be defined as a dataset containing anywhere from hundreds of thousands to millions of data points. This is less of problem within the computer vision community where massive public datasets such as ImageNet [7] exist. Indeed, it is in this natural image setting in which the labelling of images can occur using large-scale crowd sourcing solutions such as Amazon Mechanical Turk [100]. Within the medical field, however, the acquisition of large open datasets of images and their relevant labelling poses a far greater problem.

The first constraining factor surrounding the size of medical image datasets is the need for ground-truth labelling. Most CAD systems are trained in a supervised

manner. Here each input sample provided to the system x_i is provided an explicit label y_i which describes the input's class (such as being *normal* or *abnormal*). For natural images, obtaining this labelling is normally straight-forward as little expertise is required to generate input-labels such as 'rose', 'elephant' or 'office chair'. In the case of medical data, however, ground-truth labels require far more stringent procedures for their assignment. This is due to their reliance on techniques which are considered the gold-standard for establishing a diagnosis, such as tissue biopsies, computed tomography and sputum cultures [41]. For the creation of most chest radiography datasets, the gold-standard takes the form of a panel of experienced radiologists whose consensus reading, based on current radiographic findings as well as prior medical history, sets the ground-truth label for each image [26, 101]. While this approach is quicker and less expensive than other diagnostic means, it is noted that a considerable amount of time and financial resources are still required to produce relatively small datasets [14].

Once the appropriate ground-truth labelling has been obtained for a given image, a second limiting factor is the frequent need to explicitly annotate certain areas within an abnormal image. Once again, unlike problems posed in the general computer vision community in which class-related objects are often centred within images or can easily be labelled by an unskilled individual, pathological information within medical images is often small and localised. Due to these characteristics, Region-of-Interest (ROI) labelling performed by medical experts is often required for medical images in order to train CAD systems. This labelling is a resource intensive process, with large amounts of time and financial expenditure required to label relatively small amounts of data which can additionally suffer from inter-reader disagreement [5].

In response to this labelling limitation, Albarqouni et al. [100] recently investigated the influence of using non-medical participants in the labelling process of a large open-source histology dataset. To do this, the authors created a framework known as *Annot8* which combined large amounts of crowdsourced ROI labels along with a small number of clinician-based labels in order to train a CNN model for meiosis detection. While the approach greatly amplified the volume of data available for training, the results ultimately showed that the crowdsourced labels were of an unsatisfactory nature with high inter-reader disagreement being observed [100].

Another approach to boost the effective volume of medical datasets was explored by Shin et al. [102] in which topic modelling techniques such as latent Dirichlet allocation were used to extract high level labels for radiographic cases stored in their hospital Picture Archiving and Communication System (PACS). This study allowed for an unprecedented amount of medical images to receive labels ($> 216k$), however, in the

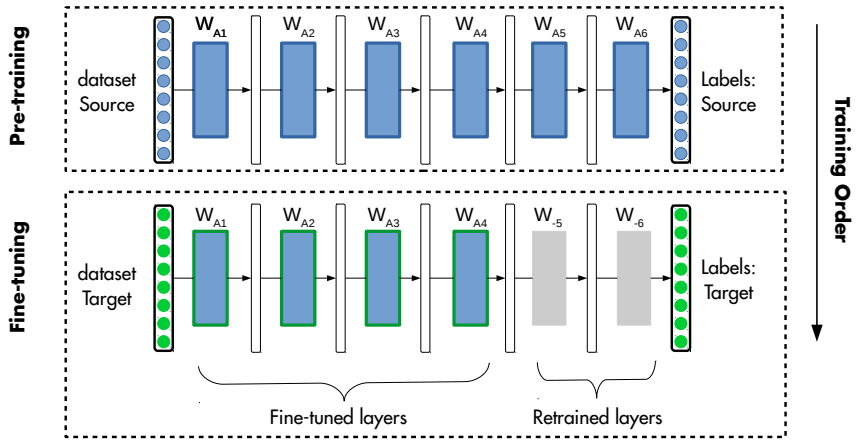


Figure 2.11: A depiction of fine-tuning a DCNN network. Typically, fine-tuning involves pre-training a network on a large set of data (dataset 'Source'), after which it is trained on a smaller dataset (dataset 'Target') which represents the real task to be performed. During this second training stage only a selected number of network layers are chosen to be re-trained, while the remainder of the layers are fine-tuned using a slower learning rate. (Adapted from Yosinski et al. [9])

absence of ROI labelling these mined labels were deemed to be too general in practice for effective learning to occur.

It must also be noted at this point that unlike the computer vision community, larger medical image datasets do not always present balanced class sizes. This is especially true for chest radiography collections in which the size of a pathology class is tightly correlated to its occurrence within a given area, population or age group. For example, within the recently formed *ChestX-ray14* dataset [14] the pneumonia class comprises less than 1% of the total image classes. Other large datasets such as the Indiana OpenI collection [101] also present similar class imbalances where additionally there are no paediatric patients represented. These imbalances in essence cause a severe class shortage in datasets which appear to be substantial in size. For these reasons, research has turned towards techniques such as transfer learning to increase the performance of DCNN approaches to CAD under these data constrained settings.

2.5.3 Transfer Learning and Fine-tuning

In transfer learning, layers from a DCNN trained on a source dataset are re-purposed or transferred to assist in a new target task [9]. Re-purposing sees these trained layers used to create feature descriptors, generating inputs for other classifiers. When transferred, these layers are used as part of a new DCNN network which is retrained

on a target dataset. This practice is seen to work effectively as long as the source dataset produces features which are general and can represent aspects of the target dataset as well [103].

During the process of retraining, the decision can be made to either leave the transferred layers frozen, or to allow their weights to be updated. This later choice is known as fine-tuning and is visually depicted in Figure 2.11. In the work of Yosinski et al. [9], it was observed that fine-tuning improves the generality of a network, and in the worst case this improvement is as good as training the network on the target dataset from scratch. Another important finding surrounded the concept of source and target dataset 'distance'. Here the further apart the purposed tasks for the source and target data were, the worse the performance of the fine-tuned network would become. This result was demonstrated within [103], where CNN models trained on the generic ImageNet and Places [104] datasets performed poorly when fine-tuned and tested on task specific datasets such as SUN [105] (scene recognition). From these results one might assume that the application of transfer learning and fine-tuning from the natural to the medical image domain has little efficacy. However, the work of several authors has disproved this intuition and has shown the remarkable success of these techniques for use within CAD [67, 94].

Initial attempts to utilise transfer learning for CAD systems saw the output feature descriptors of DCNN models such as Overfeat and AlexNet (trained on natural image datasets such as ImageNet) being fed into linear classifiers. The use of these 'off-the-shelf' CNN features produced encouraging results, as seen within [106] and [6], and demonstrated how these features were comparable to state-of-the-art approaches. Following these successes the efficacy of fine-tuning for medical imaging was investigated within the works of Tajbakhsh [11] and Shin et al. [10] respectively, with both studies yielding complementary results.

In the presence of small medical datasets for four distinct CAD tasks, Tajbakhsh [11] initially posed the question of whether fine-tuning a CNN or training it from scratch would increase its performance within the given tasks. In this regard, fine-tuning was consistently seen to improve performance relative to training from scratch. The same work also introduced the notion of 'deep' fine-tuning for medical contexts and showed how updating all the layers across a fine-tuned network further improved its ability to generalise on smaller datasets.

Similarly, Shin et al. [10] explored the use of fine-tuning CNN models as opposed to previous transfer learning approaches for two popular CAD tasks, namely: lung nodule detection and interstitial lung disease classification. Within these tasks it was found that not only did fine-tuning surpass other transfer learning techniques in performance, but that its use provided new state-of-the-art results as well.

Interestingly, in the presence of these compelling results arising from fine-tuning, both the work of Shin et al. and Tajbakhsh discuss the question of whether continued performance improvements will arise either from more sophisticated classifier models or the use of larger and more comprehensive image datasets. To this point both sets of authors are in agreement and affirm the greater need for larger medical datasets. A limitation to this belief, however, is that neither the work of Shin et al. or Tajbakhsh make use of more advanced DCNN architectures, such as Inception-V3 or ResNet, to reach this conclusion. Here the more recent work of both Esteva et al. [94] and Gulshan et al. [107] could be considered, wherein both sets of authors achieved human expert performance using the Inception-V3 network fine-tuned on large medical datasets. Thus, while the use of more sophisticated DCNN architectures can be seen as being advantageous, their reliance on large datasets is still observed.

2.5.4 Patch-based versus Holistic Classification

An additional approach to overcome the limited size of medical datasets is to reduce the amount of supervision required during training; adapting learning from a supervised to a weakly-supervised setting and thereby allowing larger quantities of weakly-labelled data to be gathered. As described in Section 2.5.2, most CAD tasks rely on the presence of ROI labels during the process of training. The need for these labels stems from the tendency of state-of-the-art approaches to implement patch-based classification, which sees discriminative pathology-class patches being sampled from a ROI in order to provide a strong supervision signal during training. Several factors make the use of these ROI labels undesirable in a medical context, however, and effectively constrain the sizes of medical datasets [108].

In response to this limitation, the efficacy of using image-level labels for medical image datasets has been investigated. Under this new label setting, the complexity required to label a dataset is eased at the expense of increasing the difficulty of the learning task. This difficulty is represented in Figure 2.12 and arises as the strong supervision signals provided by well-defined ROI labels in patch-based classification are replaced by weaker labels in which global noise is present.

One of the first studies of weakly-supervised image classification within a CAD setting was undertaken by Gao et al. [5]. Within this work, the authors propose the holistic classification of slice-level-labelled computerised tomography (CT) scans for interstitial lung disease detection, and compare the performance of this learning setting against that of patch-based classification. Several preprocessing steps were initially applied to the ILD dataset used, which contains a total of 907 slice-level CT images taken from 120 patient examinations. Following this preprocessing, each

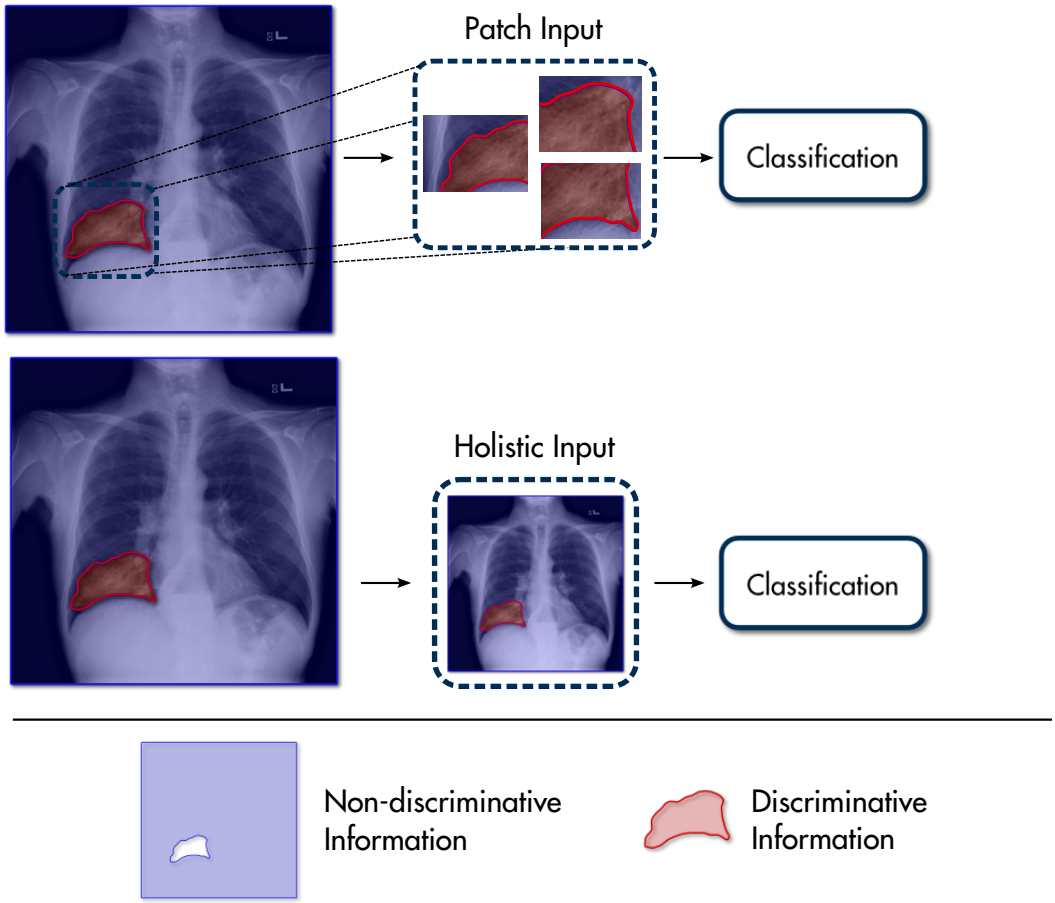


Figure 2.12: A comparison between patch-based and weakly-supervised holistic classification. While the holistic setting relaxes the need for region-of-interest labelling, it increases the complexity of classification as a smaller proportion of the input contains discriminative information. (Source: Personal collection. CXR obtained from the ChestX-ray14 dataset [14], CC0 1.0)

slice-level image was down-sampled and fed directly into a pre-trained DCNN model for fine-tuning or testing. The results of this study showed promise, with holistic classification obtaining an overall accuracy of 68.8% across 6 pathology classes, yet these were inferior to the 87.9% achieved by patch-based classification following the same methodology.

Shin et al. [10] argued that this gap in performance was due to the low volume of data points provided by the holistic setting (907 images) compared to the patch-based one (4600 image patches). The authors repeated the holistic classification task presented in [5], however they did so with pre-trained models of varying depths compared under the more appropriate evaluation settings of 5-fold cross-validation (CV) and leave-one-out (LOO) at a patient level. The obtained results showed an appreciable contrast in performance, with the best accuracies obtained under the 5-fold CV and

LOO being 57% and 90.2% respectively. This latter result was far more compelling when compared to the accuracy of 91.1% obtained for patch-based classification under the same LOO protocol.

Lastly, the strong performances produced within the works of Esteva et al. [94] and Gulshan et al. [107] can once again be considered for their relevance to weakly-supervised medical image classification. Here Esteva et al. [94] used over 130 000 weakly-labelled images of skin cancer to fine-tune their DCNN model performing holistic classification. The same weakly-supervised approach was also used by Gulshan et al. [107], wherein their DCNN model was trained under the holistic classification setting using 128 175 diabetic retinopathy images with only image-level labels.

These findings affirm that through the use of DCNN models, the trade-off between relaxed labelling and increased learning complexity associated with weakly-supervised image classification can be afforded, thus enabling potential advances for modern CAD systems and novel research associated with it.

2.5.5 Application to Chest Radiography

The combined use of holistic image classification along with weakly-labelled image datasets has recently lead to large improvements within DCNN-based CAD for chest radiography. Several of these advances are discussed below, highlighting both the contributions and shortcomings of these studies.

One of the initial results which highlighted the performance enabled by DCNN techniques was presented by Rajkomar et al. [109] in which the authors created a pipeline for radiograph orientation classification. Within this pipeline, nine variants of the GoogLeNet model [90] were fine-tuned with varying quantities of data in both colour and grey-scale format from the ImageNet dataset. These models were then used to classify input CXR images to determine if they were either in a lateral or frontal (AP or PA) projection. The results obtained from this study contain several salient points:

- The **importance of performing complete fine-tuning** for a DCNN pre-trained on a generic colour image dataset such as ImageNet. Two of the models used had only their final classification layers fine-tuned during training. These models performed substantially worse than those for which all their layers were updated during training.
- The **suitability of data augmentation for radiographic datasets**. Artificially increasing the size of the image datasets studied through the use of affine

transformations and cropping was seen to improve the accuracy of the trained classifier models.

- The **ability to obtain high accuracies through the use of DCNN models for chest radiography**. The authors report a top accuracy of 99.73% for their classification pipeline. This result is comparable to prior state-of-the-art techniques for radiograph orientation classification in which multiple expert feature descriptors were required [109].

An apparent limitation of this study is the relative ease surrounding the task of radiograph orientation classification. Unlike other screening and diagnostic tasks in chest radiography, which need clinical expertise for the detection of subtle pathologies, projection classification is based upon a radiograph's global structure and can be performed using non-medical personnel [109]. As such, this study does not reflect upon the ability of deep leaning approaches to perform pathology classification.

Shin et al. [110]. tackled the challenge of DCNN-based pathology classification as part of their work to automatically annotate chest radiographs. The authors observed the ability of deep learning techniques to generate image captions which encoded contextual information from natural images [111], and as such desired to create a framework which could detail disease context within chest radiographs in the form of medical annotations. As a first stage to achieve this outcome, the authors trained and compared two DCNN models (NIN and GoogLeNet) to classify 17 pathology classes comprising of 3000 adult radiographs with image-level labels selected from the Indiana *OpenI* dataset. Several regularisation techniques were used to accommodate data imbalances present within these classes and included data augmentation, batch normalisation and data dropout. It is noted that fine-tuning was not used, and both models were trained using raw parameter initialisation. Under these settings, the resulting overall accuracies for the NIN and GoogLeNet models were 62.43% and 69.84% respectively. While these performance levels do not approach those of radiologists, they indicate the ability of DCNN models to detect a wide range of pathologies within chest radiographs. This was seen to occur even in the presence of heavily unbalanced and under-represented classes.

In the presence of a sizeable radiographic dataset, however, Cicero et al. [13] report compelling performance metrics. Here the authors gathered data from their institution's radiology information system comprising of 35038 image-level labelled frontal adult chest radiographs which were used to train and validate a GoogLeNet model from scratch. Each radiograph contained one or multiple pathology labels and as such the model was trained to perform multi-class classification. Following this training, an overall accuracy of 82.83% (AUC = 0.897) was achieved during testing -

which is considerably higher than that of [110]. While a direct comparison between these two studies may not be appropriate due to the lack of a common test dataset, the influence of larger amounts of data within this latter study can be appreciably observed.

At this point one may wonder if high accuracies are obtainable using chest radiographic datasets of a smaller size. In this regard, recent work by Lakhani [12] demonstrated the ability of a fine-tuned DCNN network (GoogLeNet) to achieve an overall accuracy of 95.00% for the task of PTB screening within a combined CXR collection including the *Montgomery* and *Shenzhen* datasets. These results outperformed previous hand-crafted approaches such as those used by [29] discussed in Section 2.3.3.

In contrast to these findings, however, Wang et al. [14] obtained an overall accuracy of 69.62% for multi-class chest disease classification even though their study consisted of a large dataset, several sophisticated DCNN models (including VGG and ResNet) and fine-tuning. As an explanation, the authors discussed the inter-class performance of each of their models. Here large variances were seen between individual class accuracies, which were uncorrelated to the number of examples within these classes of the dataset. With these differences it was suggested that classes such as 'Mass' (accuracy = 56.44%) and 'Infiltration' (61.28%) were more difficult to classify as they are defined by finer local features. This was not a unique finding. Indeed, even though Cicero et al. [13] achieved higher overall accuracies within their study compared to Wang et al. [14], they identified large inter-class variances for their models - obtaining lower performances for the 'Consolidation' (75%) and 'Pneumothorax' (78%) classes. These differences were motivated to arise from subtle local information (which describes these classes) being lost during the down-sampling required to process large radiographic images through DCNN models. If this is the case, then both [14] and [13] identify the inability of DCNN models to fully use local discriminative information when trained under a holistic classification setting.

An additional limitation inherent within each of the works presented thus far is a severe under-representation of paediatric patients for which chest pathologies are considered. As motivated within Section 2.3.3, several factors make the interpretation of paediatric radiographs highly complex, such as high anatomical variance and the technical difficulty of conducting a radiological examination for younger patients. Considering this complexity, there exists a need to investigate the efficacy of deep learning techniques surrounding the classification of chest pathologies not only in adults, but in paediatric populations as well.

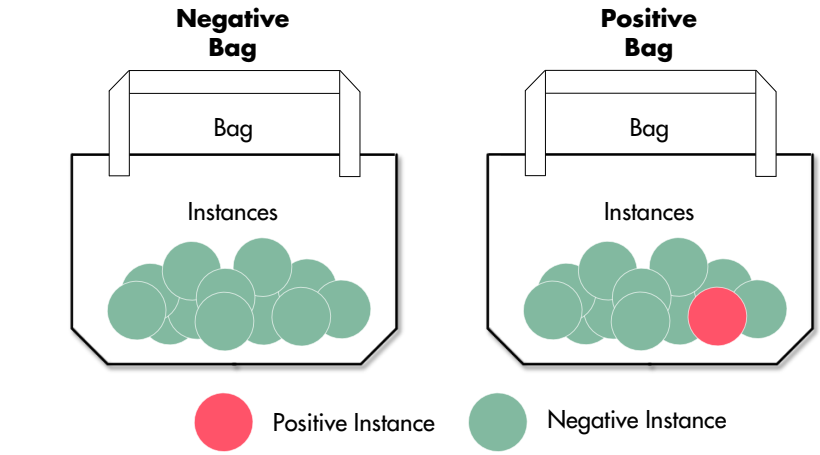
2.5.6 Multi-instance Learning Approaches

As previously discussed, an apparent limitation of training DCNN models using holistic image classification is the inability of this setting to fully enable the use of local discriminative information. This problem can be thought of from two perspectives. The first is that holistic classification forces information loss. Here, as suggested by [13], local information present within an original radiograph is lost during the down-sampling process required to feed it as input into a DCNN model. The second perspective is that holistic classification attenuates local information through the introduction of global noise. This noise, as motivated by [14], causes smaller local regions within the input to be ignored in the presence of larger global structures. Drawing from these perspectives, a weakly-supervised technique which is able to both limit information loss during image sub-sampling and reduce the amount of global noise presented during classification is desired. For this reason, the use of multi-instance learning (MIL) has recently been proposed.

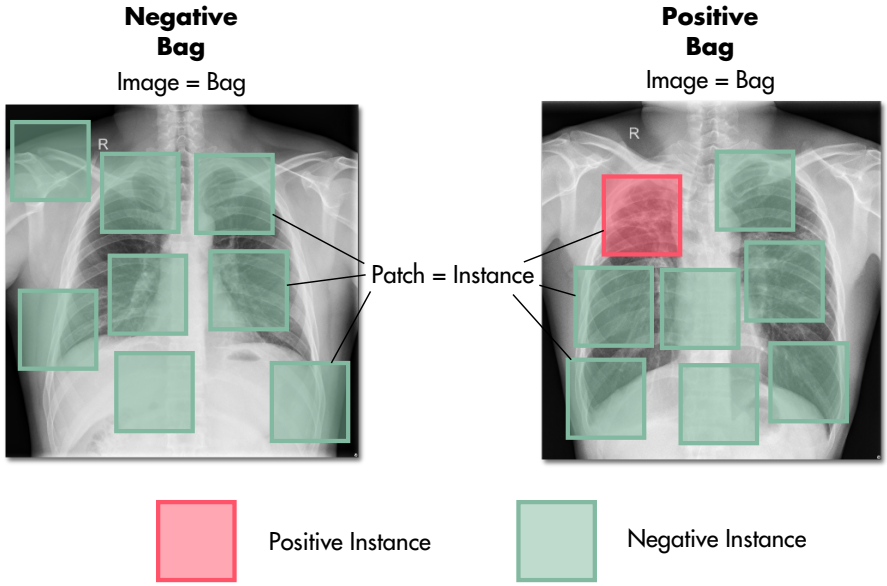
Multi-instance learning is a generalised form of supervised learning introduced by [112] in which a label is no longer associated with a single training instance, but instead is assigned to a collection of training instances often referred to as a *bag*. A common relationship exists between these individual instances and the bag of which they form part, and the task of a learner is often to infer this relationship during training and testing [113]. For classification problems, a bag's label typically originates from one or more instances within it. Here, in the case of binary classification for example, a bag can receive a positive label if one or more instances within it are themselves positive, while the bag receives a negative label if all of the instances are negative [113]. This example is depicted in Figure 2.13a.

Using these concepts, one can naturally extend the task of weakly-supervised chest radiograph screening to a multi-instance learning problem [108]. In this sense, an entire chest radiograph image can be considered as a bag while sub-regions or patches from this image can be considered as instances [108]. Following on from this, the image-level label provided for the chest radiograph (bag) can be used to indicate the presence of image patches (instances) which are either normal or abnormal within it. During training, the relationship between these image patches and their associated bag label can be learnt [108]. Thus, under the multi-instance learning setting, the classification of an entire radiograph can occur through the examination of its individual image patches. For clarity of explanation, this application to chest radiograph screening is represented in Figure 2.13b

The contributions of several authors have shown the applicability of multi-instance learning for the task of weakly-supervised chest radiography screening. Within the



(a) Conceptual



(b) Actual MIL

Figure 2.13: A depiction of multi-instance learning. (Source: Personal collection. CXR images obtained from the Shenzhen dataset [26], CC0 1.0)

work of [108], Melendez et al. adapted the maximum pattern margin support vector machine (miSVM) algorithm [114] to enable a traditional CAD system to classify PTB within chest radiographs with only image level labels. Using this MIL-based system, the authors obtained an AUC performance of 0.86 compared to the 0.88 produced by their fully supervised CAD system. This is a considerable outcome as the former method only required image level labels during training while the latter, supervised method, required highly detailed ROI labelling.

Lopes et al. [115] addressed this same task of MIL-based PTB classification using a combination of DCNN features and unsupervised methods. Here the authors first used transfer learning to extract feature descriptors generated from a DCNN model

whose input was overlapping sub-regions extracted from a chest radiograph. These features were aggregated into dictionaries of visual characteristics using k-means clustering which were then classified using conventional supervised learning. The authors compared this patch-based MIL approach against a holistic classification one in which the feature descriptors were obtained using an entire radiograph as input. From this comparison, two important observations can be made. First, the MIL approach consistently outperformed the holistic one, with average AUC scores of 0.915 and 0.875 obtained respectively. Second, this performance increase was attributed to less information being lost when using multiple image patches instead of the entire image as input into the learning system. It is noted that this approach does not reduce the amount of global noise exposed to the classifier during inference, however, as the feature vectors of each image patch were considered equally within the dictionaries formed.

An argument could be made at this point surrounding the work of [115] in that it still relied on traditional machine learning methods. As such, their pipeline imposes a large amount of complexity and inhibits end-to-end learning. Indeed, the form of their proposed system did not allow for advanced regularisation techniques such as fine-tuning, and the weights of their models were never adapted to the datasets upon which they are applied. These techniques would be desirable within a MIL-based approach as they characterise current state-of-the-art approaches [12].

In relevance to these needs, Yan et al. [15] presented a DCNN-based CAD system which was able to perform end-to-end learning under a multi-instance learning setting. The authors applied this system to the task of body-region identification in which CT slices with only image-level labels were available. Body-region recognition is inherently complex due to small local changes within the slice-level anatomy of a given body-part, and it is for this reason that a multi-instance approach was taken. Similar to the work of [115], the input to the learning pipeline consisted of a series of sub-region instances extracted from a given CT slice. However, instead of using a DCNN model to merely extract features from these instances, the authors trained this model using a probability aggregation function to generate bag label predictions. It was motivated that this altered training allowed the DCNN model to become sensitive to local discriminative information within each slice while learning to ignore other global structures which add noise during classification [15]. Furthermore, this training setting enabled end-to-end learning to be performed. Using the same custom DCNN model as a basis for comparison, the overall MIL-based system was tested against one which used holistic classification. The results affirmed the suitability of the MIL-based approach, with the former system achieving an F1-score (defined in Section 4.4.1) of 92.23% compared to that of 87.73% for the holistic classification

setting.

An evident limitation in considering the approach of [15] for CXR screening, however, surrounds the context in which it was applied. Here the task of bodypart recognition, the authors made an assumption that various regions were easily identifiable via anatomical structures at a local level. This assumption was seen to be true for the most part, with the MIL-based system performing well in distinguishing between well-defined bodypart regions such as 'nose', 'kidney ' and 'cardiac'. Unfortunately this was not the case for regions which exhibited low inter-class variance such as 'liver middle' and 'liver upper', in which F1-score performance degraded by up to 10%. This degradation poses a challenge for CXR screening, in which tasks such as pneumonia detection are typified by classes with low inter-class and high intra-class variances respectively [36]. In the face of these complexities for CXR screening, it remains uncertain as to whether the above MIL approach holds merit.

Based on the works discussed above, findings exist which indicate the suitability of using a multi-instance learning approach for weakly-supervised CXR pathology screening. However, several questions still remain open regarding the adaption of such an approach using deep learning techniques:

1. **Can a MIL method using DCNN techniques, based upon the work of [15], be applied to CXR screening?** The work of Yan et al. outperforms holistic classification for a task in which the classes are easily distinguishable at a local level. This is not always true for tasks within CXR screening where subtle differences at a local level define classes.

2. **How would this MIL method compare to fine-tuned holistic image classification?** Each of the works presented thus far have neglected to consider fine-tuning in their comparison. This is in contrast to results which show that when used, fine-tuning outperforms raw model training and other forms of transfer learning which have previously been used for comparison [9].

3. **Would such a method generalise under multiple pathology and age settings?** Within CAD, deep learning based approaches have shown a remarkable ability to generalise over multiple tasks. Each MIL method proposed thus far has only been applied to PTB detection - with emphasis on adult populations. In this sense it is still yet to be investigated whether a MIL technique could be applied across multiple chest pathologies.

The work of dissertation seeks to provide answers to these questions.

2.6 Summary

In developing nations characterised by high patient burdens and limited medical resources, the use of CAD to perform automated CXR reading presents as a viable means to screen for infectious diseases such as pulmonary PTB and paediatric pneumonia. To perform this screening with increased efficacy compared to traditional 'hand-crafted' machine learning systems, a class of deep learning algorithms known as DCNNs have recently been studied within the literature.

DCNN models are composed of multiple layer-types, including convolutional, pooling, fully-connected, detector, and softmax layers, with specific arrangements of these layers forming popular architectures which are widely studied. The three factors which have enabled the successful training of these deep architectures are the combined use of several regularisation techniques, the adoption of GPU-based computation, and the presence of large sets of training data. Unfortunately this last factor posses a limitation to the medical imaging community, as the cost and effort required to gather large image datasets with detailed labelling is substantial.

To alleviate the labelling constraint associated with medical image classification, current studies see DCNN models being trained on weakly-labelled images under a holistic image classification setting. Holistic classification may not be the most appropriate setting for CXR screening, however, as its use can increase the difficulty of detecting small and subtle pathological features which are spatially local within an image. In this regard the use of MIL, a technique which is able to extract local discriminative information in a weakly-supervised manner, is motivated for its application to DCNN-based CXR screening.

Chapter 3

Deep Multi-instance Learning for Radiographic Chest Screening

3.1 Introduction

This chapter presents a multi-instance learning (MIL) algorithm for a deep convolutional neural network (DCNN) performing radiographic chest screening. The algorithm seeks to enable the detection of local discriminative information within an image, thereby increasing classification performance.

Initially a review of similar works within the literature is undertaken to provide context around the proposed algorithm. The main text then proceeds to present the proposed MIL algorithm which comprises two stages: namely bag classification and model boosting respectively. Finally, a summary of the salient points discussed within this chapter is provided.

3.2 Existing Techniques

Multi-intstance learning is a generalised form of supervised learning in which labels are no longer directly associated with single instances, but instead are assigned to collections of instances known as bags [112]. Under this new setting, the objective of a learner is to determine which instances within a bag are responsible for its given label. In Section 2.5.6, the use of multi-instance learning as a technique to perform pathology screening within CXRs is motivated due to its ability to use local discriminative information during image classification. This section provides a review of existing techniques within the literature, which similarly use local information to improve the performance of image classification.

In the natural image domain, region-based proposal methods [116, 117], component attributes [118] and object bank representations [119] form techniques which outperform those which only utilise global information to perform image classification. In each of these cited approaches, the task of classification is seen to occur by extracting salient objects or features which are representative of a given image category. The disadvantage of each, however, is the need for detailed labelling in the form of bounding-box or region-of-interest (ROI) annotations during training. As motivated in section 2.5.2, the use of such labelling becomes unfeasible for larger medical datasets.

To circumvent this dependence, several methods have been investigated which use local discriminative information within images without the presence of a strong supervision signal to guide learning. These methods include unsupervised approaches such as that of Singh et al. [120], in which image patches are selected for model training via steps of iterative discriminative clustering, and weakly-supervised approaches such as deformable-part models introduced by Felzenszwalb et al. [121].

Under a weakly-supervised setting, visual attention models form a class of methods which have shown particular success in exploiting discriminative information in images which are cluttered or only distinct at a local level. The works of Cao et al. [122] and Xiao et al. [123] demonstrate this success within the natural image domain, with techniques such as feedback DCNN connections [122] and multi-level attention [123] enabling performance comparable to fully-supervised approaches.

Visual attention models have also been applied to the task of frontal CXR classification. Pesce et al. [124] introduce the CONAF model (convolutional network with attention feedback) which is trained using both localisation and classification error to more accurately detect lung nodules within a large CXR dataset. The localisation error is computed by comparing an activation map of the model's output layer against the position of ROI labelling, and is considered as a soft attention mechanism which corrects where the model looks in an image. While the CONAF model produces superior performance over previous state-of-the-art techniques for nodule detection, a notable limitation is its dependence on ROI labelling. Guan et al. [125] remove this need for expert labelling and extend the use of soft attention by reclassifying a portion of their model's input which is considered salient. Here the authors first use a 'global' DCNN model to generate an attention map of an input CXR. The area of the input which receives the greatest attention for a given pathology label is then cropped and classified by a second 'local' DCNN model. An overall class prediction is then formed by jointly considering the local and global DCNN responses. The resulting attention-guided framework is seen to outperform approaches which only consider global information during classification.

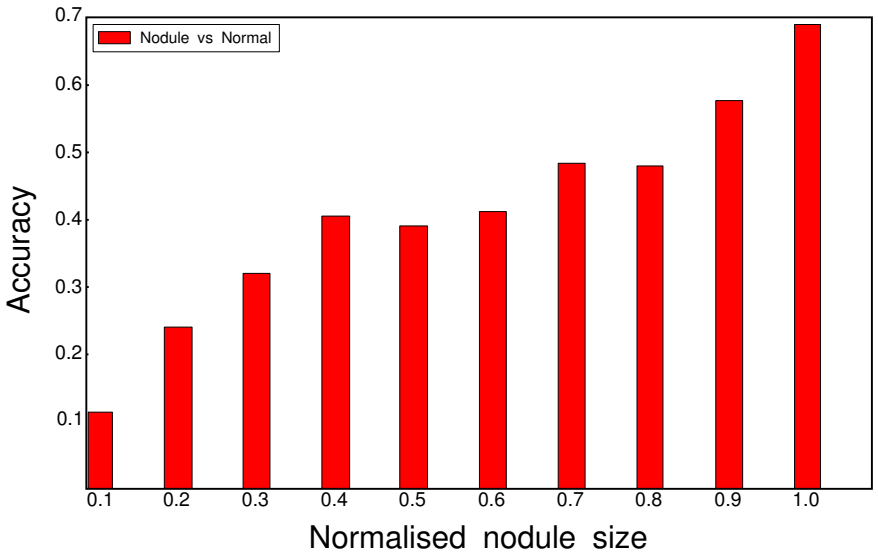


Figure 3.1: Results from the work of Pesce' et al. in which the performance of their attention guided model diminishes with decreasing lung nodule size. (Amended from Pesce' et al. [124], 2017).

Despite the gains offered by each of the presented visual attention techniques, a collective limitation is their reliance on global information to drive attention to local regions of their input. Thus small or subtle discriminative findings which can easily be missed at a global scale may never receive the additional attention which these techniques afford. Figure 3.1 illustrates this limitation, wherein Pesce et al. [124] acknowledge a large drop in accuracy when classifying smaller sized lung nodules with the CONAF model. This decrease in performance is indicative of the smaller discriminative features associated with these lung nodules being incorrectly classified or missed altogether.

In light of this limitation posed by visual attention methods, DCNN models have been applied along with MIL techniques to extract local discriminative information in a robust manner. Xu et al. [126] demonstrate such an approach in order to classify weakly-labelled histopathology images as being cancerous or normal. Initially the authors consider each image as a bag containing several image patch instances. A DCNN model pretrained for cancer detection is then applied to extract features from each of these instances, with the resulting outputs being used to produce instance level predictions. Using the multi-instance learning criterion (given in Section 2.5.6), these instance level predictions are then aggregated to form a bag label which corresponds to the image's predicted class. While the performance this MIL-based approach is comparable with that of fully-supervised methods implemented by the same authors, a major limitation surrounds its dependence on a DCNN model which is trained on the target dataset in a fully-supervised manner.

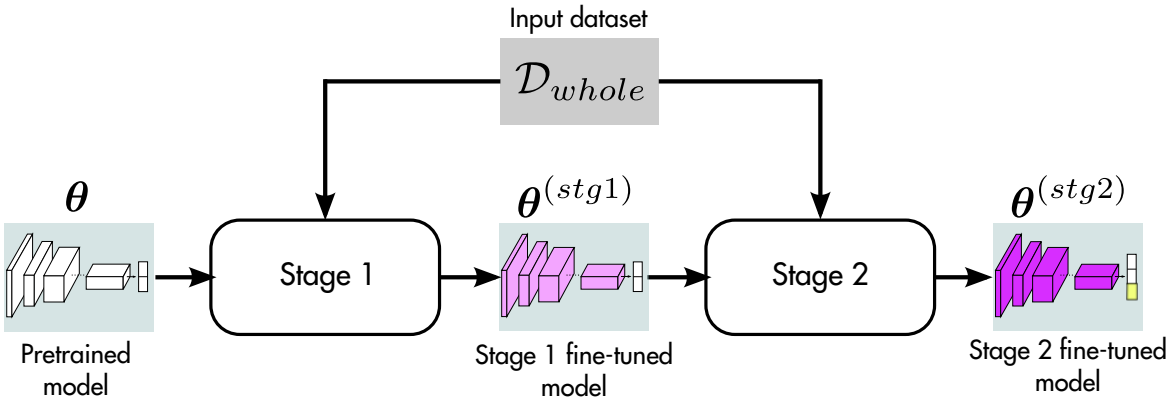


Figure 3.2: High-level overview of the multi-instance deep learning method. Within the two represented stages, the goal of this method is to train a DCNN model with input parameters $\boldsymbol{\theta}$ to detect local information which is both discriminative and non-discriminative for pathology. (Source: Personal collection)

Tang et al. [127] relax this requirement for fully-supervised training by using a DCNN pretrained on a large dataset to perform transfer learning - allowing all training to occur in a weakly-supervised manner. Lopes et al. [115] follow a similar approach, using features extracted from a pre-trained GoogLeNet model along with k-means clustering to form bag-level feature vectors which can be classified using a support vector machine. Within both of these works, however, the use of weakly-supervised classifiers, which are separate from DCNN models, prevents end-to-end learning from occurring. Thus the gains offered by techniques such as fine-tuning are not available.

Wu et al. [128] overcome this limitation by implementing a bag-label prediction function whose form is differentiable. Applying this function, the authors are able to predict patch instance labels using the AlexNet DCNN model pre-trained on ImageNet, while jointly fine-tuning this model through successive steps of backpropagation. Yan et al. [15] extend this work into the medical image domain wherein they also add an additional boosting stage after training their unique DCNN model from scratch. This boosting refines the predicted image patch labels, reducing the number of false positives at an instance level.

3.3 Multi-instance Deep Learning Method

The MIL method proposed within this dissertation follows on from the successful work of Yan et al. [15] by adapting the body-part recognition algorithm presented by the authors to enable the detection of pathology within frontal chest X-rays.

The proposed method consists of two learning stages which are represented in Figure 3.2, and can be automated to perform end-to-end learning. The first stage of this method is responsible for discovering CXR image patches which are discriminative for pathology by fine-tuning a DCNN using only image-level labels. The second stage is used to boost this DCNN model, reducing the number of false positives it produces and allowing it to detect patches which are non-discriminative for pathology. These two stages are presented within this section.

3.3.1 Stage 1 - Discriminative Patch Classification

Figure 3.3 presents an overview of stage-1 training for the MIL algorithm.

Initially a training set of CXR images $\mathcal{D}_{whole} = \{\boldsymbol{X}_i, i = 1, ..., m\}$ with corresponding class labels $Y_i \in \{0, 1\}$ is given, with a label of 0 indicating a negative/normal image and 1 representing a positive/abnormal image. Using the multi-instance criterion, each image within the training set can be considered as a bag containing n instances or patches. Here the ith image can be given as $\rho(\boldsymbol{X}_i) = \{\boldsymbol{x}_{ij}, j = 1, ..., n\}$ with $\boldsymbol{x}_{ij}$ representing the jth image patch of the image, and $\rho(\cdot)$ being a function to determine the spatial extent of each patch.

Under this assumption, the bag-level label for each image Y_i is known, while the instance-level labels for its patches y_{ij} may be unknown. In the face of this uncertainty, the goal of this learning stage is to assign high class probabilities to instances which are correctly discriminative for a given class.

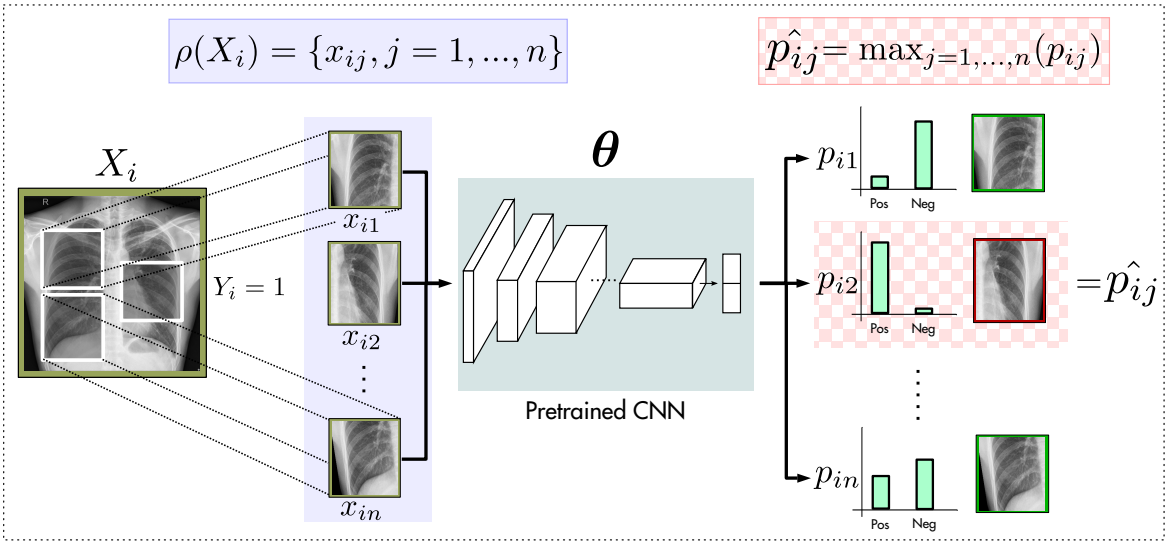
For a given image, each of its patches form the input to a DCNN model to produce a positive instance probability p_{ij} such that:

$$p_{ij} = P(y_{ij} = 1|\boldsymbol{x}_{ij}; \boldsymbol{\theta}), \tag{3.1}$$

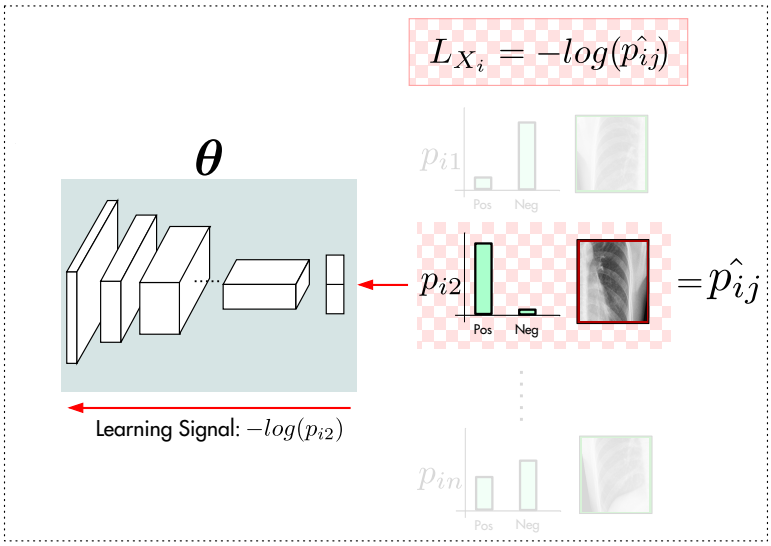
where $\boldsymbol{\theta}$ represents the fine-tuned parameters of the DCNN model and the probability P is obtained by applying the softmax function (2.9) to the output layer of the model. At this point a function f is required which will map the set of instance probabilities at a patch-level to a bag level probability p'_i:

$$p'_i = f(p_{i1}, p_{i2}, ...p_{in}). \tag{3.2}$$

Several such forms of f have been studied within the literature (See [113] for a review of these). Under the CXR screening setting, it is assumed that a normal CXR will contain no discriminative instances for pathology, while an abnormal CXR will contain one or more discriminative instances. Using this reasoning, a function can be chosen which is suitable for single positive-instance recognition. Comparing several



(a) Patch sampling and classification



(b) Learning from discriminative image patches

Figure 3.3: A representation of stage-1 training for the multi-instance learning method. (Source: Personal collection)

variants of f, Wu et al. [128] assert that for this condition, the most appropriate form is the max function:

$$p'_i = \hat{p_{ij}} = \max_{j=1,...,n}(p_{ij}). \tag{3.3}$$

Yan et al. [15] have considered the use of the max function before and argue that an intermediate function is required to map individual instance probabilities to image regions, which in turn are used to form bag-level probabilities. This intermediate step is motivated to reduce spurious predictions induced by local noise within single image patches, and effectively aggregates the predictions of neighbouring instances. However, this process requires that $\rho(\cdot)$ be chosen such that the number of patch

instances produced for an image is large and are densely overlapping. This sampling strategy is suitable for classification with the smaller custom DCNN model which the authors use within their work, yet it proves computationally prohibitive for deeper DCNN models such as those within Section 2.4.7. Instead, an alternative approach involves setting $\rho(\cdot)$ such that fewer patch instances are produced which are already large enough to become resilient against local sources of noise. When experimenting with this alternative patch sampling strategy, Yan et al. [15] report only a minor drop in accuracy ($< 0.7\%$) compared to theirs involving region-based aggregation. As such, the use of larger image patches is selected for this work, with further exploration surrounding the effect of the varying the patch size being given in chapters 5 and 6.

The resulting process of image patch sampling and bag-level probability assignment is represented in Figure 3.3a. Having obtained this bag-level probability, negative log-loss can be applied to calculate the loss between this bag estimate and the real label the during training:

$$L(\boldsymbol{X}_i, \boldsymbol{\theta}) = L_{X_i} = -\{Y_i log(p'_i) + (1 - Y_i) log(1 - p'_i)\}. \tag{3.4}$$

Substituting $p'_i = \hat{p_{ij}}$, for the case of a positive bag $Y_i = 1$, (3.4) becomes:

$$L_{X_i} = -log(\hat{p_{ij}}), \tag{3.5}$$

while for a negative bag $Y_i = 0$, (3.4) simplifies to:

$$L_{X_i} = -log(1 - \hat{p_{ij}}). \tag{3.6}$$

From 3.5 and 3.6, and as represented in Figure 3.3b, it is observed that the correction signal propagated back through the network during training will be based solely on the patch which is most discriminative for pathology in a given image. For positive bags, 3.5 is seen to produce a weak correction signal for high positive-class probability estimates - rewarding the detection of instances which may contain pathology. Conversely, for negative bags, the presence of these same high probability estimates in 3.6 produces a strong correction signal - penalising instances which are falsely predicted as being discriminative for pathology. Over successive training iterations, these correction signals lead the network to become sensitive to discriminative pathology patches, while equally recognising instances which are discriminative of the normal class.

3.3.2 Stage 2 - Non-Discriminative Instance Boosting

During the first stage of the MIL method, the model is trained to become sensitive to instances which are discriminative for pathology. However, this may lead to false

positive predictions being generated due to the low class variance typically seen between healthy and abnormal CXRs even at an instance level. For this reason, a second training stage inspired by the work of Yan et al. [15] is motivated to enable the DCNN model to detect instances which are in fact non-discriminative for pathology classification. In this manner, the model produced from stage-1 training is boosted [129] to become a stronger learner which is aware of both discriminative and non-discriminative information. This stage is similar to the instance-label re-assignment step followed within the mi-SVM algorithm [114], where bag level labels are used to iteratively refine the instance-level predictions of the algorithm.

The second training stage begins by forming three new training subsets which are generated from the original training set $\mathcal{D}_{whole}$. These are subsets are $\mathcal{D}_{dis+}$, $\mathcal{D}_{dis-}$ and $\mathcal{D}_{!dis}$. To achieve this formation, each image $\boldsymbol{X}_i \in \mathcal{D}_{whole}$ is decomposed into patches using $\rho(\cdot)$, with an instance-level class probability estimate being assigned each patch using the DCNN fine-tuned for instance detection within stage-1. The first two subsets, D_{dis-} and D_{dis+}, contain image patches which are considered discriminative for the negative and positive classes respectively. As such, (3.7) and (3.8) are used to allocate image patches to these subsets:

$$\mathcal{D}_{dis-} = \{\boldsymbol{x}_{ij} \in \rho(\boldsymbol{X}_i) | P(y_{ij} = 0 | \boldsymbol{x}_{ij}; \boldsymbol{\theta}^{(stg1)}) \geq \alpha_0, Y_i = 0\}, \tag{3.7}$$

$$\mathcal{D}_{dis+} = \{\boldsymbol{x}_{ij} \in \rho(\boldsymbol{X}_i) | P(y_{ij} = 1 | \boldsymbol{x}_{ij}; \boldsymbol{\theta}^{(stg1)}) \geq \alpha_1, Y_i = 1\}. \tag{3.8}$$

Here $P(y_{ij} = 1|\boldsymbol{x}_{ij}; \boldsymbol{\theta}^{(stg1)})$ represents the probability estimate of a given patch being positive using fine-tuned parameters $\boldsymbol{\theta}^{(stg1)}$, and $P(y_{ij} = 0|\boldsymbol{x}_{ij}; \boldsymbol{\theta}^{(stg1)})$ represents the negative class probability estimate. The terms α_0 and α_1 are used as a scalar-valued confidence thresholds,$\alpha_0 \in (0.5, 1), \alpha_1 \in (0, 1)$, which allow for multiple discriminative patches to be extracted from each training image. The need for these thresholds arises as during stage-1 training only the maximum instance-level positive probability estimate is used to predict the label of a given input image. However, during the patch extraction step within stage-2, several instances within one image may produce high responses for either the positive or negative class and thus can to be collected for model boosting. Both α_0 and α_1 are fixed for all processed images and as such form hyperparameters for the MIL algorithm.

According to the MIL criterion, each instance within a negative bag automatically inherits a negative label (for visual clarity, see Figure 2.13a). Thus, each patch from a negative image which produces a higher positive class probability estimate is in fact a false-positive response and is non-discriminative for pathology. A similar

false-negative assignment is not true, however, for patches from a positive image which are assigned higher negative-class probabilities. Here, for positive images, only one patch needs to be considered positive for an entire image to receive a positive label. Thus the presence of multiple negative patches within a positive bag is valid. To this end Mendelez et al. [108] state that only with additional expert labelling can the true patch-level labels within positive images be correctly given. The authors further caution that due to the diffuse nature of chest pathologies, even patches from positive images which receive higher negative-class probabilities may still contain subtle disease features.

Following on from the above reasoning, (3.9) is used to allocate patch instances extracted from $\mathcal{D}_{whole}$ into the $\mathcal{D}_{!dis}$ subset:

$$\begin{aligned}\mathcal{D}_{!dis} = \{\boldsymbol{x}_{ij} \in \lambda(\boldsymbol{X}_i) | P(y_{ij} = 1 | \boldsymbol{x}_{ij}; \boldsymbol{\theta}^{(stg1)}) \geq 0.5, Y_i = 0\} \\ \cup \{\boldsymbol{x}_{ij} \in \lambda(\boldsymbol{X}_i) | P(y_{ij} = 1 | \boldsymbol{x}_{ij}; \boldsymbol{\theta}^{(stg1)}) < \alpha_1, Y_i = 1\}.\end{aligned} \tag{3.9}$$

Once extracted, the three subsets are combined to form a new training set:

$$\mathcal{D}_{patch} = \mathcal{D}_{dis-} \cup \mathcal{D}_{dis+} \cup \mathcal{D}_{!dis},$$

such that $\mathcal{D}_{patch} = \{\boldsymbol{x}_i | i = 1, ..., l\}$ with l patches and corresponding class labels $y_i \in \{0, 1, 2\}$. These labels are assigned to each patch according to which subset it originated from. In this sense, all patches from $\mathcal{D}_{dis-}$ inherit the negative/normal class 0, those within $\mathcal{D}_{dis+}$ are considered positive/abnormal 1, and patches from $\mathcal{D}_{!dis}$ are assigned as being non-discriminative 2. The overall formation of this new training set is represented in Figure (3.4)

Following the formulation of the new training set, the DCNN model with accompanying parameters $\boldsymbol{\theta}^{(stg1)}$ undergoes an additional fine-tuning training step as represented within Figure 3.5. Within this step, the final classification layer of the model is reinitialised and an additional output unit is included. This additional unit allows the model to recognise and classify patches into the non-discriminative class. Similar to stage-1, the input to the model during this training takes the form of image patches, however, these are sourced from $\mathcal{D}_{patch}$. With stronger input-label pairs (x_i, y_i) being available to train the model, the probability estimate for each input becomes:

$$p'_i = P(y_i | \boldsymbol{x}_i; \boldsymbol{\theta}^{(stg1)}), \tag{3.10}$$

where $P(y_i | \boldsymbol{x}_i; \boldsymbol{\theta}^{(stg1)})$ represents the estimate of the input $\boldsymbol{x}_i$ belonging to class y_i. Having obtained this estimate, learning is performed by applying negative log-loss

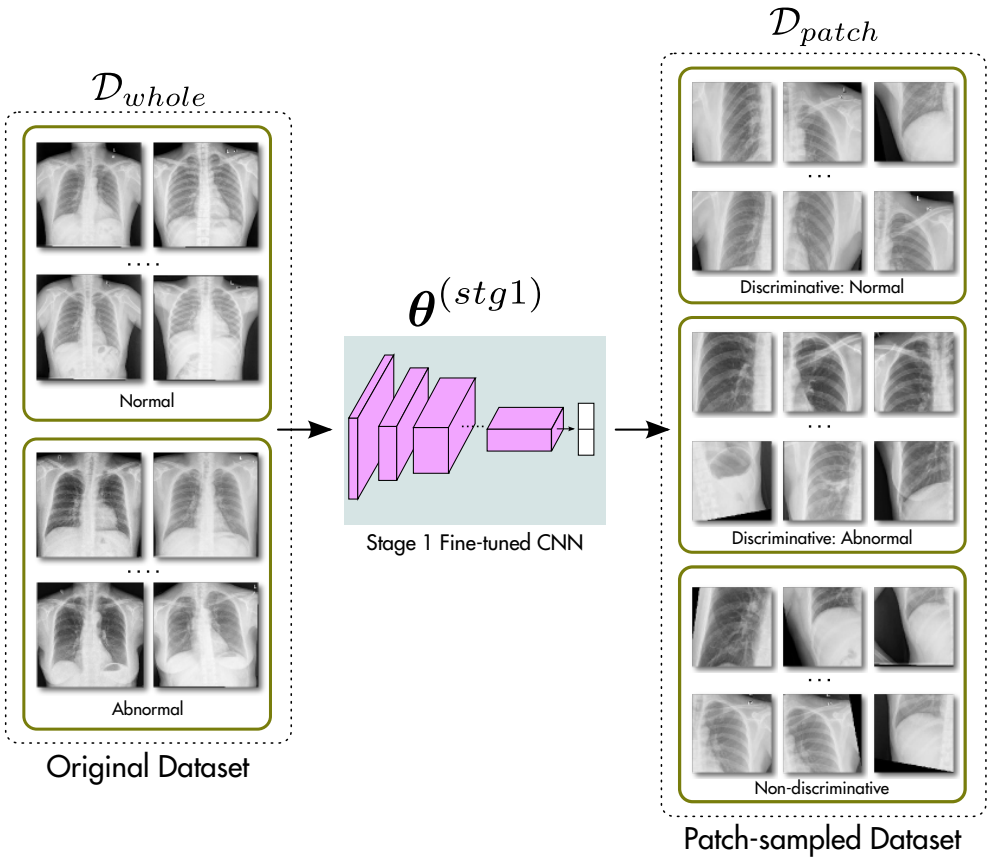


Figure 3.4: Generation of the patch dataset $\mathcal{D}_{patch}$ using the DCNN model fine-tuned during the first stage of training. The generated dataset contains image patches sampled from the original dataset which are either discriminative for normal lung tissue, discriminative for pathology or are non-discriminative respectively. (Source: Personal collection).

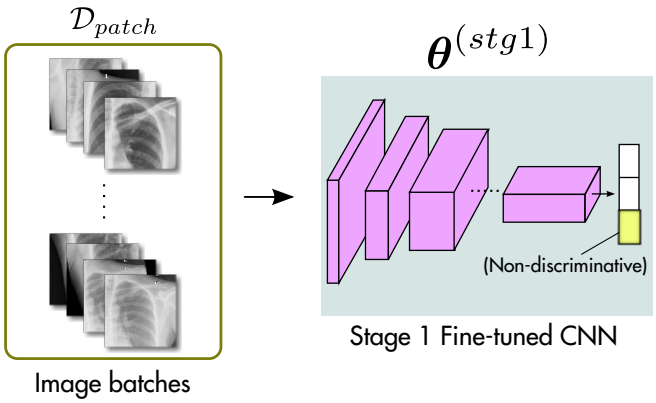


Figure 3.5: Non-discriminative instance boosting. This process involves further fine-tuning the DCNN model obtained from stage-1 using batches drawn from the newly created training set $\mathcal{D}_{patch}$. During this fine-tuning, an additional output unit is added to the model to enable it to produce responses for inputs belonging to the 'non-discriminative' class. (Source: Personal collection)

to compute the model loss for the current input, after which parameter updates are performed using backpropagation and a chosen optimisation strategy.

At the conclusion of this second training stage, the model (with updated parameters $\boldsymbol{\theta}^{(stg2)}$) is able to identify both discriminative and non-discriminative instances for pathology within an image-level labelled CXR.

3.3.3 Testing , Validation and Inference

Once trained using the MIL method, a model can be used to detect discriminative pathology patches from holistic CXR images through the procedure described in Algorithm 3.1.

Algorithm 3.1 Multi-instance deep learning algorithm for CXR class inference during implementation. Given a CXR image X and DCNN model with parameters $\boldsymbol{\theta}^{(stg2)}$ obtained using the steps defined in sections 3.3.1 and 3.3.2, the following procedure can be used to obtain a class prediction $\hat{Y}$ for the image.

Require: Patch extraction function, $\rho(\cdot)$
Require: $\boldsymbol{\theta}^{(stg2)}$
Require: Discriminative threshold, $\mathcal{T}$
$\boldsymbol{x} \leftarrow \rho(X)$
for $i \leftarrow 1, ..., len(\boldsymbol{x})$ **do**
  $\hat{p_i} \leftarrow P(y_i = 1 | x_i; \boldsymbol{\theta}^{(stg2)})$
  if $\hat{p_i} \geq \mathcal{T}$ **then**
    return $\hat{Y} \leftarrow 1$ ▷ Abnormal patch discovered
  end if
end for
return $\hat{Y} \leftarrow 0$ ▷ No abnormal patches found

The preprocessing stage involves the clipping of patches from the CXR using the $\rho(\cdot)$ setting implemented during training. These patches are then fed into the model for inference, with the positive probability estimate for each patch being considered. Using the MIL criterion, if one or more patches has been assigned a probability estimate for being abnormal above a confidence threshold $\mathcal{T}$, the entire image receives an abnormal classification prediction. Conversely, if all of the patches receive low estimates for being abnormal (high normal or non-discriminative class probabilities), the entire image is assigned a normal classification prediction.

The exact choice of $\mathcal{T}$ implemented for inference can be informed by metrics obtained during model validation. In this regard one approach is to choose $\mathcal{T}$ based upon

the optimal value of Youden's index [130] for a receiver operating characteristic curve (introduced in Section 4.4.1) plotted during evaluation on the validation dataset [12, 109]. For this work, $\mathcal{T} = 0.5$ is chosen as the threshold value applied across all models to standardise the comparison between the MIL and holistic classification methods.

3.3.4 Computational Cost of MIL

Having presented the MIL method used within this dissertation, an assessment is made surrounding its computational cost versus that of standard holistic classification. It is noted that due to varying network architecture design, dataset selection and non-convex optimisation associated with the training of a DCNN, the time complexity of a method is motivated to be analysed empirically [131]. As such, this section briefly discusses the space complexity associated with the MIL method, while Section 6.8 presents measurements surrounding its training and inference times during experimentation.

Training During the training of a DCNN model, several factors are seen to make up its total memory footprint. While some of these factors remain static throughout the training procedure, such as the number of model parameters and a given framework's memory overhead, others are seen to scale in an approximately linear manner with the batch-size used to train the model [132, 93]. These dynamic factors include parameter gradients, optimiser parameters and layer outputs for both forward and backward propagation. In consideration of this, if the combined memory required by these dynamic factors for a single image is represented by M, then the space required for training with a batch-size of N using holistic image classification can be approximated as $M \times N$ [133]. In comparison, in implementing the MIL method it is observed that during stage-1 training the effective size of a given batch of images is increased by a factor P, which represents the number of patches extracted using $\rho(\cdot)$. These patches are all required to be processed collectively in order to determine the maximum positive probability for a single given image. This same requirement does not hold during stage-2 of training, where images from $\mathcal{D}_{patch}$ can be processed independently of one another. As such, the space required for training using the MIL method is dominated by stage-1 training and can be given as $M \times N \times P$.

The impact of the increased space complexity associated with the MIL algorithm either requires that additional computing hardware or smaller batch sizes be utilised during model training.

Inference During inference, the MIL method is required to process a maximum of P instances in order to classify a single image. As these instances can be processed independently of one another, however, the space required for inference with the MIL method can be reduced such that it is identical to that of holistic classification.

3.4 Summary

The proposed MIL method is based on the work of Yan et al. [15] and is adapted to overcome the limited use of local information associated with holistic classification and current visual attention methods. It consists of two training stages, which enable a DCNN model to detect and classify local regions of discriminative pathology using only image-level labelled CXRs.

Within stage-1 of the MIL method, the max function is used to aggregate the positive class predictions of multiple patches, which can be sampled from an image under different bag settings. Over multiple training iterations, this aggregation allows a model to become sensitive to local information which is discriminative for pathology or normal tissue. Stage-2 boosts the performance of the stage-1 trained model by fine-tuning it to recognise non-discriminative instances for pathology. This process involves initially re-sampling the original training set into a new one containing patches which are discriminative for each class, as well as those which are considered non-discriminative for pathology. This new dataset provides stronger input-label pairs to assist in model training. During inference, input images are broken up into patches associated with the bag setting used during training, and are classified with the stage-2 trained using the multi-instance criterion.

A brief consideration of the computational cost of the MIL sees its space complexity increase linearly with the number of patches sampled per image during model training.

Chapter 4

Model Selection, Training and Evaluation

4.1 Introduction

The field of machine learning has long been considered an experimental science [134]. This designation is given as most of the algorithms studied within the field are too complex to be evaluated in a purely theoretical manner. It is especially true for deep learning methods - where models often contain in excess of 1 million parameters. Research surrounding machine learning therefore requires experimentation with the aims of discovering which conditions allow for improved learning to occur for a given task, and to specify possible reasons for this improvement. In light of this requirement, the intent of this dissertation is to establish empirically whether the Multi-Instance Learning (MIL) method introduced in Chapter 3 produces superior performance over holistic classification for the given task of chest X-ray (CXR) classification. It also seeks to provide reasoning for findings surrounding the presence or absence of this improvement. In order to accomplish these tasks in a reliable and repeatable manner, both the conditions of learning for the respective methods, as well as the manner in which they are systematically compared need to be provided. This chapter seeks to provide these details.

Initially a review of the model architectures selected for experimentation is provided. Following this, the choice of training procedure and model hyperparameters used during experimentation are discussed. Metrics used to evaluate the trained models are then introduced. Finally, the chapter concludes by specifying the experimental framework in which model training and evaluation occurs.

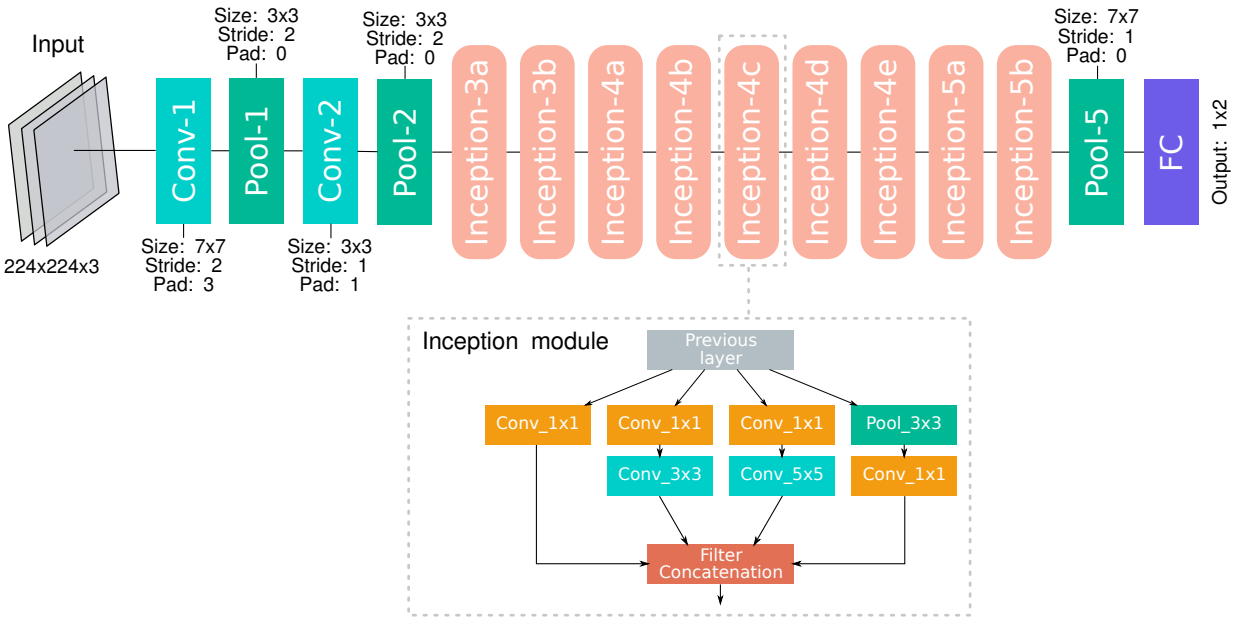


Figure 4.1: The GoogLeNet architecture. (Amended from Shin et al. [10], 2016)

4.2 Model Architecture Selection

The *GoogLeNet* and *Network-in-Network* (NIN) Deep Convolutional Neural Network (DCNN) architectures are selected to compare the MIL and holistic classification methods studied within this dissertation. The choice of two model architectures is motivated as it allows for both the inter-model and intra-model effects of the proposed method to be observed. Furthermore, while deeper architectures have consistently shown improved performance within the computer vision community [89, 90, 91, 92, 135], the medical imaging community has tended to study smaller, ad-hoc network architectures [15, 51, 67, 95, 136]. Tajbakhsh et al. [11] contend that this preference for smaller architectures is driven by the reduced size of medical datasets which are often available to perform learning on, with Shin et al. [10] similarly arguing the need for deeper architectures to be trained on larger medical datasets. As such, due to the relatively small size of the datasets used within this work, both GoogLeNet and NIN are seen to be adequate for study. The structure of these networks as well as their relevance to automated CXR reading are provided within the subsections below.

4.2.1 GoogLeNet

Originally described in section 2.4.7, GoogLeNet is a 22-layer DCNN architecture created by Szegedy et al. [90] which won the 2014 ImageNet Large Scale Visual Recognition Competition (ILSVC). The distinguishing characteristic of the architecture

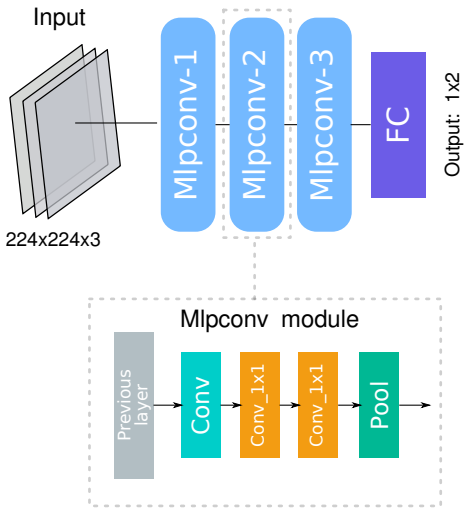


Figure 4.2: The Network-in-Network (NIN) architecture. (Amended from Lin et al. [88], 2013)

is its several 'inception' modules, depicted in Figure 4.1, which allow for a deeper and wider structure to be achieved by the network without increasing the number of parameters which it possesses. These parameter savings enable GoogLeNet to achieve greater accuracies compared to other networks which are similar to its size, while still maintaining a smaller memory footprint [133].

GoogLeNet has been utilised in several works which study forms of automated CXR reading. These uses include single-class and multi-class pathology classification within frontal CXR images [12, 13, 14], CXR orientation detection [109], as well as automated radiological chest report generation [110].

4.2.2 Network-in-Network

Presented in Figure 4.2, the NIN architecture consists of three 'MLPConv' (Multi-Layer-Perceptron-Convolution) layers which its authors purport to act as individual sub-networks within the overall network architecture [88]. This structure, which is considered to have been the philosophical inspiration for the inception modules of GoogLeNet [93], allowed NIN to outperform other state-of-the-art models of its time on the CIFAR-100 dataset [88]. As such the model has been widely cited and studied within the computer vision community [67, 93].

Shin et al. [110] motivated the use of the NIN network for their study involving automated CXR classification and report generation due to its small size, which both prevents overfitting and allows for faster training times, as well as its superior performance over other similar sized networks. For these same reasons the NIN architecture is seen as an adequate choice for the work within this dissertation.

4.3 Training

4.3.1 Cross-Validation

A common procedure for training and evaluating a machine learning algorithm is to divide randomly the target dataset upon which it is to be assessed into three subsets which are used for training, validation and testing respectively. The training subset typically contains the largest proportion of data (between 60 - 80%), with the remaining data often being shared equally by the other two subsets [69, Chapter 1]. In cases where the dataset is of a sufficiently large size, this procedure works well as the test subset contains enough data points to be representative of the entire dataset - allowing for an accurate estimation of the algorithm's performance when tested on this set. On the other hand, in situations in which only a smaller dataset is available for evaluation - as is common within medical imaging - the opposite is true and the representative ability of the test subset cannot be guaranteed. This in turn can lead to large variances in algorithmic performance being recorded depending on which samples are present within the training and test subsets, thus providing a poor estimate of the algorithm's true performance. In light of this challenge, various sampling strategies have been proposed within the literature in an attempt to provide more accurate estimates of performance for smaller datasets.

Two commonly seen sampling strategies used for the evaluation of computer aided diagnosis (CAD) systems involve either performing Leave-One-Out (LOO) or k-fold cross-validation (CV). Within this latter strategy, the target dataset is initially divided into k distinct folds after which k evaluation runs are performed. During each of these runs, a new fold is designated as the test set, while the remaining $k-1$ folds are used for training or validation purposes. At the conclusion of the evaluation runs, the individual scores are combined to provide an overall estimate for algorithmic performance [82]. For *stratified* k-fold CV, each fold is sampled such that it contains an equal number of data points from the respective classes within the dataset. LOO cross-validation is seen to occur as a specific case of k-fold CV where k is set to equal the number of data points within the target dataset.

In regard to these two strategies, Shin et al. [10] assert 5-fold cross validation as being more challenging and clinically relevant compared to LOO, as it is likely that a CAD tool will see a large amount of variance in the cases it is exposed to during in-field use. Kohavi [137] also argues for moderate values of k to be utilised along with stratification during CV, with experimentation showing that LOO produces optimistically-biased performance results. Several authors have also utilised k-fold cross validation in the analysis of their CAD systems recently, including Roth et al. [4],

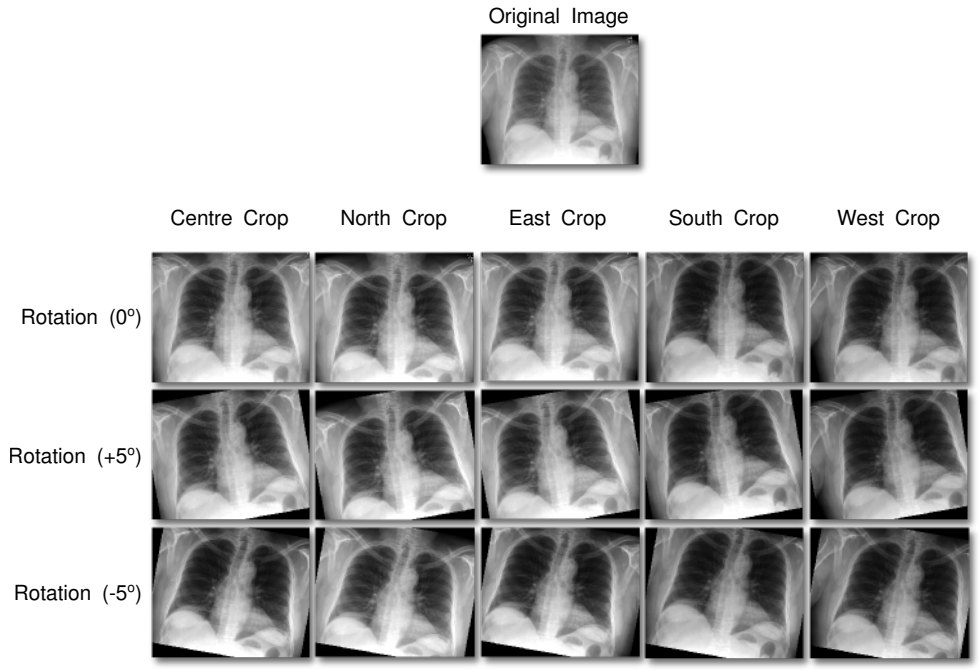


Figure 4.3: Class-preserving data augmentation operations applied during image pre-processing. (Source: Personal collection)

Melendez et al. [138], Lopes et al. [115] and Ciompi et al. [6]. While various values of k are studied within these works, $k = 10$ is motivated to reduce the pessimistic-bias associated with 5-fold CV [137]. However, it is noted that increasing the value of k comes with an increased computational cost as additional training/evaluation runs are required to compute a model's overall performance. This computation is especially expensive for DCNN models which can still take several hours to train on a powerful machine.

In consideration of the above stated factors, the resulting program selected for model training and evaluation is stratified 5-fold cross-validation, with a final split of 60% training, 20% validation and 20% testing data per cross-validation run. Furthermore, it is noted that due to the non-convex nature of training DCNN models, evaluating two identical model settings on the same data fold may produce slight differences in their final performance. As such, five runs for each fold of cross validation are performed to produce a more stable overall result for each classifier.

4.3.2 Data Augmentation

Within the medical imaging community, data augmentation (introduced in section 2.4.6) has emerged as a standard practice when applying deep learning techniques to a given problem [67]. Here a wide range of augmentation operations are present; ranging from simple geometric transformations such as scaling, rotation,

translation and reflection [96, 109, 110], to more specialised forms such as elastic deformation [136] and Adaptive Histogram Equalisation [12]. While the use of these more advanced techniques have been shown to improve performance, their implementation is considered to be outside the scope of this work's objectives. As such, Figure 4.3 shows the simple augmentation class-preserving operations which are selected. These are: cropping (centre, upper, lower, left and right anchored), and rotation ($\pm 10^o$), with combinations of these operations also being utilised. The resulting augmentations increase the effective size of a target dataset by a factor of 15. During image pre-processing, each of the above operations is implemented using ImageMagick v.6.7.7-10.

4.3.3 Model Initialisation and Fine-tuning

Section 2.5.3 introduced the concept of fine-tuning a DCNN as a means to improve its performance when evaluated on a smaller dataset. The author's own work [16], along with that of [12] and [109], shows how this practice is similarly beneficial when applied to frontal CXR images. In accordance with these findings, each GoogLeNet and NIN model presented within this dissertation is trained using fine-tuning. This process is described below:

For both of the selected network architectures, fine-tuning commences by loading a version of the corresponding network model which has been pre-trained on the ILSVRC2012 natural image dataset [7]. The last fully-connected layer of this imported model is then discarded, and is replaced with one possessing an output unit for each class within the target dataset. For both holistic classification and stage-1 of the MIL method, this corresponds to two output units, while for stage-2 of the MIL method three output units are created. The raw weights of this new layer are set using Xavier initialisation [139], which allows for faster training convergence compared to random initialisation via a Gaussian distribution. Following this initialisation process, training proceeds with the parameters of both the reinitialised and transferred being updated in successive mini-batch steps. Motivated by the findings of Tajbakhsh et al. [11], these updates are guided by two separate learning rates which are applied to the reinitialised and transferred layers respectively. Here as in the works of [109] and [10], the reinitialised layer is assigned an initial learning rate which is 10x greater than that of the transferred layers. To aid convergence during training, both of these rates are decayed by a factor of 10 after every 20 epochs as is done similarly in the works of [140] and [10].

To guide the overall computation and application of parameter updates, the Adam optimiser [86] (introduced within section 2.4.5) is selected. While many state of the

art network architectures continue to use Stochastic Gradient Descent (SGD) with momentum to perform their training [89, 141, 142], Adam has recently been used within several works studying automated CXR reading [96, 124, 140, 143]. Wilson et al. [144] argue that this preference is due to the faster convergence times produced by Adam under default settings. However, the authors further caution that the use of such adaptive learning rate optimisers may produce increased generalisation error compared to SGD with well tuned hyperparameters for a given task. Thus while Adam is utilised in this work to enable faster training times, it is acknowledged that its use may contribute to conservative model performances.

As a last consideration of the above-defined fine-tuning process, one may question the suitability of using natural images to pre-train a model for its use in a medical application. To this end the work of [109] shows that fine-tuning from natural images which are adapted to resemble medical data outperforms fine-tuning from standard natural image sources alone. However, in the absence of such data-sources and the computational means to perform extensive pre-training, as is the case in this context as described in Chapter 2, the performance gains seen using standard fine-tuning processes are still seen to be desirable. Thus while it is conceded that gains in model performance may be lost due to the choice of pretraining data within this work, the use of standard fine-tuning is still seen to be beneficial as shown in [16].

4.3.4 Regularisation

Three regularisation techniques which are common within the literature, namely; early stopping, dropout and weight decay, are further implemented to assist training and model performance. These techniques are considered individually below:

Early Stopping Within their text on deep learning, Goodfellow et al. [82] cite the use of early stopping during model training as being an extremely useful practice. This is due to the fact that early stopping acts as both a regulariser - reducing generalisation error during testing - as well as a means to shorten model training times. The use of this practice within this work also enables a fairer comparison between the holistic and MIL classification methods; with models trained under each method being exposed to an appropriate amount of data to achieve their best performance. This fact weakens the argument that the second stage of MIL training unfairly exposes a model to additional training time, as a model trained under the holistic setting also has the opportunity to continue training to improve its performance.

Table 4.1: Hyperparameters selected for use during final experimentation.

Attribute	Value
Learning rate - reinitialised layer	1.5×10^{-3}
Learning rate - transferred layers	1.5×10^{-4}
Weight decay - η	1×10^{-4}
Dropout probability*	0.4
MIL - α_0	0.6
MIL - α_1	0.4
Adam - β_1	0.9
Adam - β_2	0.999
Adam - ϵ	1×10^{-8}

* Dropout only applied to GoogLeNet.

During training, the process of early stopping is implemented as represented within algorithm 2.2. Here the validation set error is determined using the F1-score (introduced in section 4.4.1 below), with n being set such that this error is measured four times per epoch. The patience is set such that $p = 20$, and effectively stops training after five epochs if no further improvement on the validation set is recorded. A maximum number of 150 epochs are reserved for training in this manner, however this limit is never meaningfully reached.

Dropout Similar to the works of [15], [51] and [110], dropout [78] is applied to the inputs of the fully-connected layer within each GoogLeNet model. During training the network's prescribed dropout probability of 0.4 is used, while during testing and validation this probability is set to 0. By default the NIN architecture does not include dropout, and as such the technique is not utilised for its models training and evaluation.

Weight Decay As discussed within section 2.4.6, the use of weight decay or the L^2 parameter norm penalty is a common practice in the training of DCNN models. This is also seen to be true for the medical imaging community, with both [109] and [125] using the technique for frontal CXR reading. Thus weight decay is implemented during training for each model under the holistic and MIL settings.

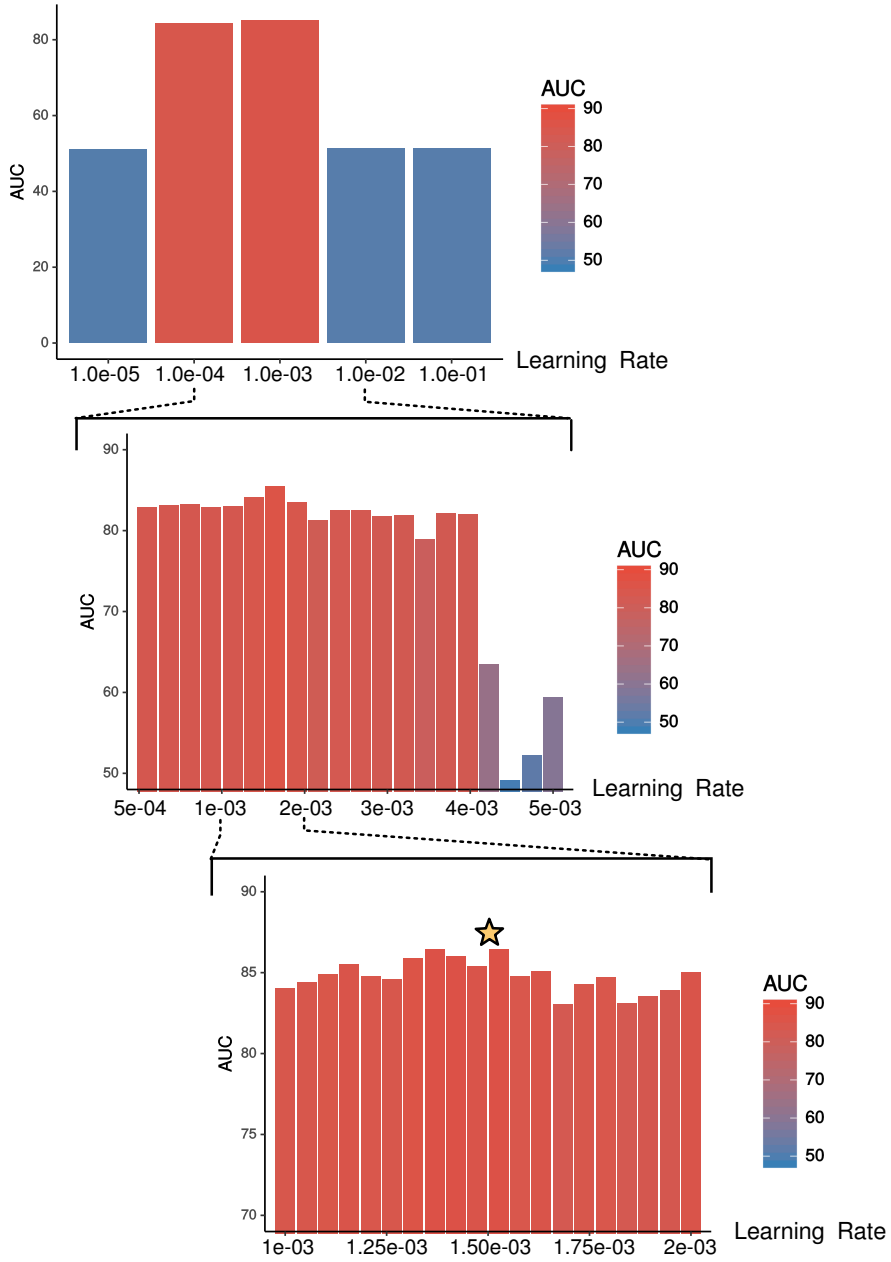


Figure 4.4: Learning rate selection via a coarse-to-fine grid search.

4.3.5 Hyperparameter Selection

Prior to final experimentation, various hyperparameter configurations are investigated to guide model training. As poignantly noted within their recent review of over 300 medical imaging papers using deep learning, Litjens et al. [67] here observe that 'no clear recipe' exists for discovering the optimal hyperparameters for a given task. Instead the authors observe that intuition often guides a systematic search for these values. With this in mind, table 4.1 presents the final hyperparameters selected for use in training both the GoogLeNet and NIN models.

Several approaches were used to obtain the values presented in Table 4.1. For the global learning rate, a coarse-to-fine grid search using the GoogLeNet network was performed over the PCV-13 dataset[1] introduced in section 6.5. This search, which is depicted in Figure 4.4, was based on the initial values used within [10, 12, 110] who all use GoogLeNet for medical image analysis. A similar search was also performed for the weight decay coefficient, with initial values sourced from [109] and [10]. Other hyperparameters are used according to default values provided. These consist of the dropout probability used by [78], as well as the β_1, β_2 and ϵ values recommended for Adam within [86].

Similar to the work of [10], the maximum effective batch size is constrained to 196 during training to limit the requirement of computing resources.

To determine appropriate α_0 and α_1 values to be used within the MIL method, initially several GoogLeNet models underwent stage-1 training on the PCV-13 dataset, with each model using one of the patch settings defined in section 6.3. At the conclusion of this training, a grid search was performed for each resulting model in which the stage-2 performance was recorded for $\alpha_0 \in \{0.5, 0.6, 0.7\}$ and $\alpha_1 \in \{0.3, 0.4, 0.5, 0.6, 0.7\}$. During this process it was observed that the performances produced under these settings were largely insensitive to changes in these investigated hyperparameters. This indicates that already after stage-1 training, a model trained under the MIL method produces high confidence estimates for discriminative image patches.

In consideration of the above presented hyperparameter values and the means whereby they were derived, two prominent limitations are noted. Firstly, an assumption is made that the values obtained for the GoogLeNet network will be equally suitable for the NIN network. While this same assumption is made within the literature by [10] and [110], other values may more suitable for the NIN architecture. Secondly, the hyperparameters presented are only obtained on a single dataset, even though they are used across various other datasets presented in chapters 5 and 6. In defence of both limitations, however, it is additionally noted that the raised limitations inherently disadvantage model performance and do not advantage it. In this sense the supplied hyperparameters provide a pessimistic estimate of performance as opposed to an overly optimistic one. Furthermore, it is recognised that the objectives within this work seek to comparatively evaluate the MIL and holistic classification methods. As such, the use of values which are shared in common allows for a more fair comparison to occur. Lastly, the viewpoint of Litjens et al. [67] is once again provided wherein

[1] Performance based-off the Area Under Curve scores from a single held-out test set consisting of 20% of the data points in each class. The remaining 80% of the data is used for model training and validation.

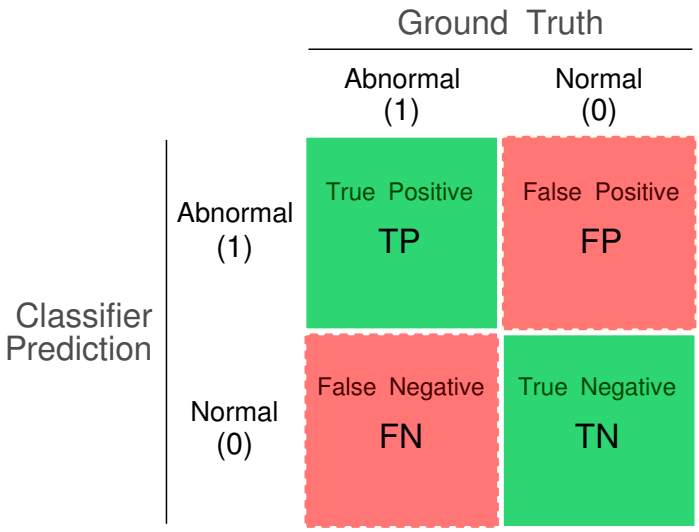


Figure 4.5: Possible outcomes for binary classification. (Source: Personal collection).

the authors believe the selection of optimal hyperparameters to be of secondary importance to other factors such as the quality of data and their correct utilisation.

4.4 Testing and Analysis

Within the field of machine learning, the ability of an algorithm to learn is often characterised by its improved performance on a given task [134]. Thus, in order to effectively study and compare various machine learning algorithms, the ability to obtain reliable measurements surrounding performance is essential. In light of this need, this section provides various metrics and techniques which are used within this work to evaluate and compare the holistic and MIL classification techniques.

4.4.1 Performance Metrics

During the process of binary inference or classification, a model will typically produce a decision surrounding the class y of a given input x according to the decision rule [69, Chapter 5]:

$$y' = \mathbb{I}(f(x) > \mathcal{T}), \tag{4.1}$$

Where:

y'**:** The predicted class label.

$\mathbb{I}(\alpha)$**:** The indicator function. $\mathbb{I}(\alpha) = 1$ if α is true. Otherwise $\mathbb{I}(\alpha) = 0$.

$f(x)$**:** The model's prediction confidence that $y = 1$ based off input x.

$\mathcal{T}$**:** Confidence threshold.

Here the model is seen to predict $y' = 1$ if its confidence $f(x)$ is greater than the set threshold $\mathcal{T}$, otherwise it predicts $y' = 0$. Based off this decision, four classification outcomes are possible and are represented within Figure 4.5. These outcomes consist of either being a true positive (TP): the model correctly predicts abnormality $y' = 1$ which is actually present $y = 1$, false positive (FP): the model predicts abnormality $y' = 1$ when normality is present $y = 0$, false negative (FN): the model predicts normality $y' = 0$ when abnormality is actually present $y = 1$, or a true negative (TN): the model predicts normality $y' = 0$ when normality is present $y = 0$. Using these classification outcomes as a basis, several metrics can be formulated to measure the performance of a classifier and are presented below:

Accuracy For the task of classification, accuracy is a common metric used to assess a model [82]. Accuracy simply measures the fraction of cases correctly classified within an evaluated dataset, and can be given by:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}. \tag{4.2}$$

Several authors, including Lopes et al. [115], Lakhani et al. [12], Cicero et al.[13] and Pesce et al. [124], use accuracy to gauge the performance of their CAD systems for frontal CXR screening. Accuracy provides a useful aggregate of the overall results produced by such systems, however, it does not provide insight into how well they can actually detect pathology or recognise its absence. For this reason, additional metrics such as sensitivity and specificity are also considered in model assessment.

Sensitivity Often referred to as *recall* within the greater machine learning community, sensitivity measures the fraction of abnormal cases which are correctly detected within an evaluated set. It is given by:

$$Sensitivity = \frac{TP}{TP + FN}. \tag{4.3}$$

Sensitivity is a metric which is commonly used within medical image analysis and the study of CAD systems [67]. As observed by Doi [42], such systems are required to have very high sensitivity levels if they are to be justifiably used, as any sensitivity lower than that of a human reader would greatly reduce a system's efficacy as a screening tool. Works using sensitivity in the evaluation of automated frontal CXR reading include Jaeger et al. [29], Maduskar et al. [61] and Van Ginneken et al. [60].

Specificity Specificity, also known as the *true negative rate*, measures the fraction of normal cases within an evaluated set which are correctly classified as being normal. It is given by:

$$Specificity = \frac{TN}{TN + FP}. \tag{4.4}$$

Specificity is seen to go hand-in-hand with sensitivity in its wide usage [67], and forms a complementary measure of a classifier's suitability for CAD [42]. Here Doi once again places emphasis on a CAD system possessing high specificity levels, as low levels would cause many false positives predictions to be generated. This in turn would require an individual with medical expertise to constantly validate the results of the system, ensuring that predicted abnormal cases are in fact abnormal, and thus defeats the notion of automated reading.

F1-Score Based-off the merits of both sensitivity and specificity mentioned above, within a medical context it is often desirable for a screening tool to possess both high levels of sensitivity and specificity in order for it to have efficacy. As such, the F1-score is often used as a measure to jointly observe these two metrics:

$$F1\,Score = \frac{2 \times (Sensitivity \times Specificity)}{Sensitivity + Specificity} \tag{4.5}$$

As described by Murphy [69, Chapter 5], the F1-score forms a harmonic mean of the sensitivity and specificity. This harmonic form is advantageous in situations where the sensitivity and specificity are seen to differ greatly for a given task and provides a more suitable indication of performance compared to an arithmetic mean. Several authors make use of the F1-score in assessing the performance of their CAD systems, including Yan et al. [15], Rajpurkar et al. [140] and Shin et al. [10].

Receiver Operating Characteristic (ROC) As mentioned within this section's introduction, the classification outcomes presented within Figure 4.5 (and subsequently the sensitivity and specificity) are determined at a given threshold $\mathcal{T}$. However, these same outcomes can be measured across a range of threshold values - producing a relationship between sensitivity and specificity as $\mathcal{T}$ increases or decreases. In this same manner, when sensitivity is plotted against 1-specificity (known as the false-positive rate) under a range of $\mathcal{T}$ values, the resulting chart is referred to as a Receiver Operating Characteristic (ROC) curve.

Figure 4.6 presents an example of two ROC curves plotted on the same set of axes as a means to form a comparison. In considering these curves several observations

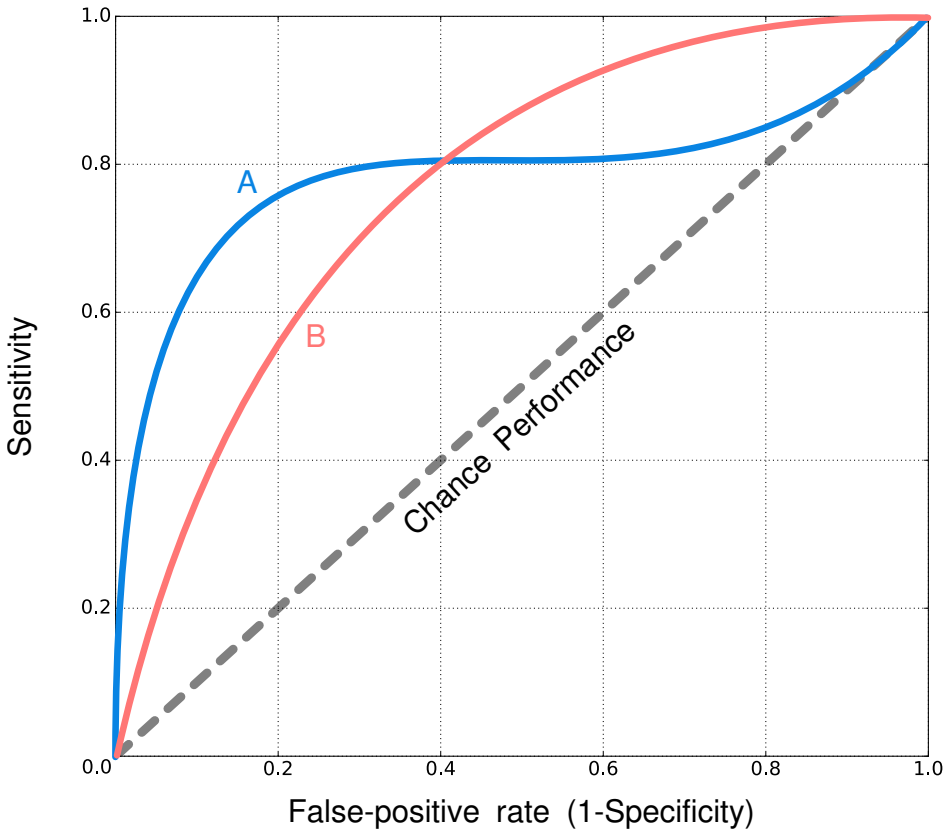


Figure 4.6: Example Receiver Operating Characteristic (ROC) curves. Even though curves A and B have roughly the same area, A is still more clinically desirable due to its higher sensitivity at lower false-positive rates. (Amended from Obuchawski [145], 2003).

are made. First, it is seen that each curve can start at $(0, 0)$ when $\mathcal{T} = 100$ and can end at $(1, 1)$ when $\mathcal{T} = 0$. Second, when operating at chance level, a classifier's performance can be mapped to any point on the diagonal line. Third, as noted by Obuchawski [145], the clinically relevant region of an ROC curve typically sits at a false-positive rate between $0 - 0.2$. Thus while both curves A and B have roughly the same area, curve A is preferable as a diagnostic tool. Lastly, the curve of a perfect classifier touches the upper-left corner of the chart, with good performing classifiers closely approaching it.

The use of an ROC curve is seen as an important means to gauge a classifier's performance [145], with numerous works citing its usage within the literature [12, 13, 14, 115, 143, 125]. This importance is especially true when considering multiple classifiers, as their curves allow them to be compared at the same specificity or sensitivity levels regardless of the associated $\mathcal{T}$ values.

Area Under Curve (AUC) While an ROC curve enables several sensitivity/specificity readings to be inferred surrounding a classifier's performance, typically it is desirable to express a classifier's characteristic as a single numerical value. To this end the Area Under Curve (AUC) metric is introduced which, as its name indicates, computes the area under a given ROC curve. Values for the AUC metric range from 1.0, representing a perfect classifier, to 0.5 which represents a chance classifier. The AUC metric is typically reported alongside ROC curves. As such its use within the literature is similarly prevalent.

4.4.2 Confidence Intervals

Having defined various measures of performance which can be applied to competing classifiers, an important final task lies in determining which of the compared classifiers is the best. In light of this need, this section details the use of confidence intervals within this work to enable a fairer decision surrounding the relative merit between the holistic and MIL classification methods.

As described within section 4.3.1,each evaluated model undergoes training and testing over five runs of 5-fold cross validation. From these runs, the computed average sensitivity, specificity, F1-score and AUC metrics are reported. Demsar [146] here argues that the use of these averages alone provides a poor indication of relative performance between classifiers. This is motivated as averages are susceptible to situations in which a few excellent results compensate for a classifier's predominantly poor performance. In this regard, the use of a confidence interval (CI) as a means to further quantify average performance is motivated.

For a given set of n recorded measurement samples with mean value $\bar{x}$ and unknown standard deviation, the upper and lower bounds on the confidence interval surrounding the mean can be given as:

$$(\bar{x} - t_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}, \bar{x} + t_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}). \tag{4.6}$$

Here $t_{\frac{\alpha}{2}}$ is the critical value for a t-distribution with $n-1$ degrees of freedom measured at a significance level of α, while s is the sample standard deviation calculated from n. Based off empirical evidence offered by Bouckaert [147], a decision is made to use each of the 25 model runs ($n = 25$) to compute confidence intervals, as this approach is seen to yield greater repeatability compared to simply averaging over individual folds. The significance level is set to $\alpha = 0.05$ to derive the 95% confidence intervals commonly reported within the literature [13, 96, 140].

One limitation that may be brought into attention at this point surrounds the creation of confidence intervals using measurements obtained from cross validation. Here, as cautioned by Salzberg [148], even though measurements are derived from independent test sets, the use of common training data during cross-validation may lead to slight biases in the resulting intervals. In response, while Goodfellow et al. [82] acknowledge this concern, they contend that these differences are still small enough to competently compare two classifiers by considering cases in which their resulting confidence intervals do not overlap.

All confidence intervals reported are calculated using standard packages within the R statistical computing framework (v3.4.1).

4.4.3 Saliency Maps

A common criticism of neural network methods is that they function as 'black box' classifiers [67]. In this sense, once a neural network has performed classification on a given input, the size of the network's internal parameter space often prohibits any insight to be gained as to why the specific classification outcome was actually achieved. Litjens et al. [67] points out that this lack of accountability is especially harmful in a medical context where poor predictions can have legal consequences. Thus a CAD system cannot be based on predictive performance alone, but requires the ability to explain its decisions in a manner that they can be adequately interrogated.

In light of these challenges, several techniques have been proposed which seek to reduce the lack of articulation present within deep learning approaches. These include the visualisation of intermediate layers within a network [149], generating text-based descriptions of input images [110, 111] and quantifying the true level of uncertainty surrounding a prediction [150].

One technique of particular interest surrounds the generation of a saliency map, which provides a visual means to inspect where a DCNN places its attention within an image when making a prediction surrounding its content [151]. Saliency maps are created by exploiting the typical forward and backward propagation of information which occurs during DCNN model training. Initially a given image is passed through the various layers of the network, with its respective layer responses and final layer activation (corresponding to a given class prediction) being recorded. This activation can then be passed back through the network with its gradient being determined in respect to the original input image. By doing this, a mapping onto the input can be produced which indicates the most discriminative components of its content which lead to the final activation.

Table 4.2: Operating system and hardware specifications used during training and testing.

Specification	Description/Quantity
Operating system	Ubuntu 14.04 LTS
CPU	28 cores @2.4 GHz
GPU	Nvidia 1080GTX (8 Gb)
Primary memory	128 Gb
Secondary memory	5 Tb

While several methods exist to produce saliency maps, these either require the use of a specialised network architecture [152], modification to an existing architecture [153], or dilute the true attention of a model [149]. As such, the *guided backpropagation* method of Springenberg et al. [151] is used to produce the saliency maps generated within this work.

4.5 Experimental Setup

All algorithmic development, model training, evaluation and saliency map generation is implemented[2] within the **Tensorflow** framework [154] (version 1.2.1) on a machine of which specifications are provided within Table 4.2.

In motivation of the use of Tensorflow, it is noted that multiple libraries and frameworks support the creation, training and evaluation of DCNN models. Popular choices amongst these include Torch [155], Tensorflow [154], MXNet [156], and Caffe [157], with all possessing model zoos which act as an open-access repositories for current and prior state-of-the-art DCNN models. Tensorflow has support from both academia and industry, with a large community following which includes prominent AI research institutions such as Google's Deep Mind [158]. The framework is open-sourced, offers full GPU acceleration, and provides regular updates ensuring that it supports recent academic advances.

[2]Implementation code available at: **https://github.com/JonGerrand/Deep-Multi-Instance-Learning**

4.6 Summary

The empirical nature of machine learning requires that the comparison of the MIL and holistic classification methods be performed using experimentation. To do this in a reliable and repeatable manner, the models, training procedure, analysis techniques and experimental setup used within this work are defined.

The GoogLeNet and NIN networks are selected to each act as a model architecture during training, allowing for both the inter-model and intra-model effects of the studied methods to be assessed. Training itself occurs using five runs of 5-fold CV for each model to reduce optimistic bias in performance measurements. During this training, fine-tuning from the ILSVRC-2012 dataset, along with train set data augmentation, dropout, L^2 norm weight decay and early stopping are all used to increase model generalisation. Hyperparameter values for the learning-rate and weight decay are set using a coarse-to-fine grid search with GoogLeNet models, while the α_0 and α_1 hyperparameters for the MIL method are set using a standard grid search. Other hyperparameter values are adopted from the literature.

Once trained, the accuracy, sensitivity, specificity, F1-score, ROC curve and AUC metric of each model is obtained to measure its performance. The use of confidence intervals allows these respective metrics to be aggregated across CV model runs for comparison of the classification methods in a more reliable manner. To qualitatively assess the performance of each model, saliency maps are generated using guided backpropagation.

All experimentation and evaluation is performed using the Tensorflow framework (version 1.2.1), with code used for the implementation of the MIL method available at **https://github.com/JonGerrand/Deep-Multi-Instance-Learning**.

Chapter 5

Validation of Deep Multi-instance Learning for Radiographic Screening

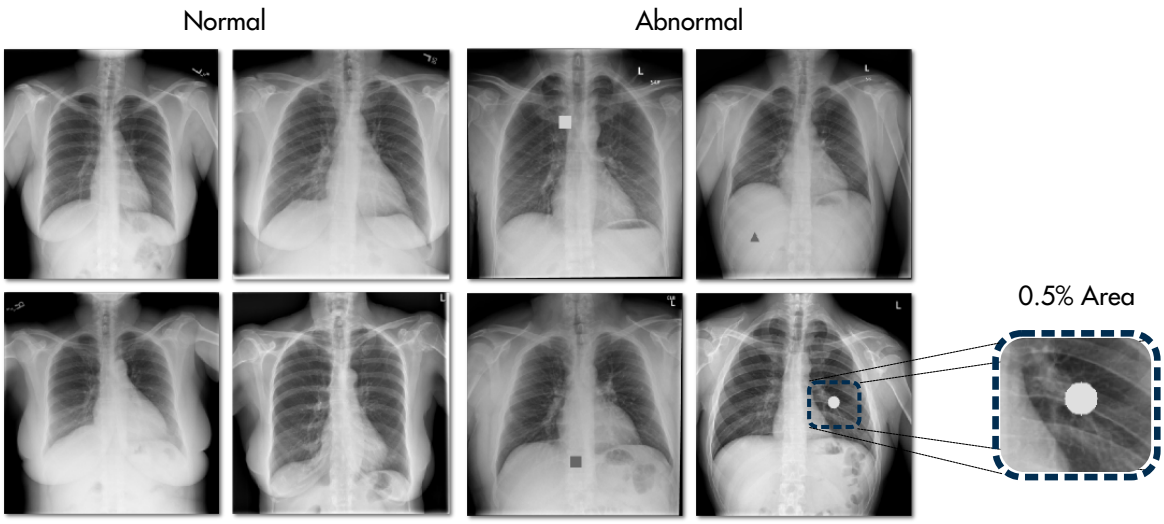
5.1 Introduction

Within Chapter 3 a multi-instance learning (MIL) method for training deep convolutional neural networks to perform frontal radiograph screening was presented. This chapter validates the functioning of the MIL framework and discusses its characteristics. As a means to facilitate this validation, an artificial classification problem is constructed using two synthetic datasets. This problem seeks to:

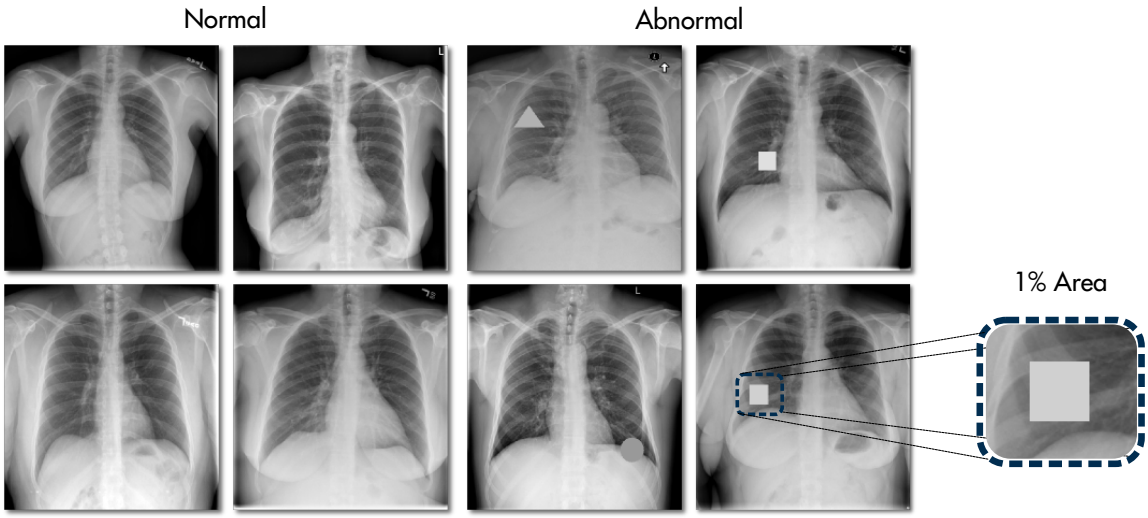
1. Verify the ability of the proposed MIL method to detect and classify local discriminative information within an input image.

2. Establish the relative performance of the MIL and holistic methods in classifying images with this local discriminative information.

Initially, a review of the two synthetic datasets used within the validation process is provided. Following this, the experimental setup for the validation procedure is described, introducing the various model settings used during training and testing. Experimental results are then provided for the respective datasets, allowing for a comparison between the MIL and holistic classification methods. A review of additional data collected during model training is then made to further verify the proposed MIL method. Finally, the outcomes of the validation procedure are summarised.



(a) "Small" CXR



(b) "Large" CXR

Figure 5.1: Samples drawn from the '*Small*' and 'Large' synthetic CXR datasets used during the validation of the multi-instance deep learning method.

5.2 Datasets and Data Preparation

Two synthetic datasets each containing 2782 images are used to validate the functioning of the MIL method. Figure 5.1a and 5.1b presents examples drawn from these datasets which are referred to as the 'Small' and 'Large' chest X-ray (CXR) datasets respectively. This section describes the formation and characteristics of these two datasets.

Frontal radiograph images from the *Normal* subset of the Indiana OpenI CXR collection (described in detail within Section 6.4) are partitioned evenly into two new classes. Within the first class, each image is left unaltered and maintains its label from the original collection. For the second class, each image is assigned an *abnormal*

Table 5.1: 'Small' and 'Large' CXR dataset statistics

Dataset	**# Datapoints**		
	Train	**Validation**	**Test**
CXR - Small	1670	554	558
CXR - Large	1670	554	558

label and is modified by placing one of three geometric shapes (circle, square or triangle) of a varying greyscale intensity within its foreground. The intensity, position and choice of shape for each abnormal image is chosen randomly; with the intensity values lying in the range of $[50, 240]$ and position coordinates between 10% to 90% of the image width and height pixel values. For the 'Small' dataset, the size of each shape is fixed such that it occupies 0.5% of the image area, while for the 'Large' dataset each shape is fixed to occupy 1% of the area.

The resulting datasets can each be used to train a deep convolutional neural network (DCNN) model to perform binary classification. During this training, although each image shown to a model contains variability due to patient-dependent anatomical differences, the only true discriminative information between classes is the presence or absence of a local geometric shape. Thus, in order to perform well at the binary classification task, the model is required to detect and utilise this local information correctly. The use of different shape sizes within the two databases is motivated to allow for the difficulty of shape detection to be varied. Furthermore, this variation in size is representative of the different spatial extents observed between various chest pathologies.

For model evaluation, the data is split into 60/20/20 % sets of training, validation and testing data respectively. Due to the larger quantity of datapoints present within these datasets, the use of both cross-fold validation and data-augmentation as outlined within Section 4.3 is omitted. Thus five evaluation runs on the single data split are performed, with the metric averages across these runs being presented. The resulting dataset statistics are provided in Table 5.1

5.3 Experimental Setup

The experimental setup implements the NIN and GoogLeNet DCNN architectures detailed within sections 4.2.2 and 4.2.1 respectively. For both networks, four model variants are trained - each using one of the bag settings depicted within Figure 5.2.

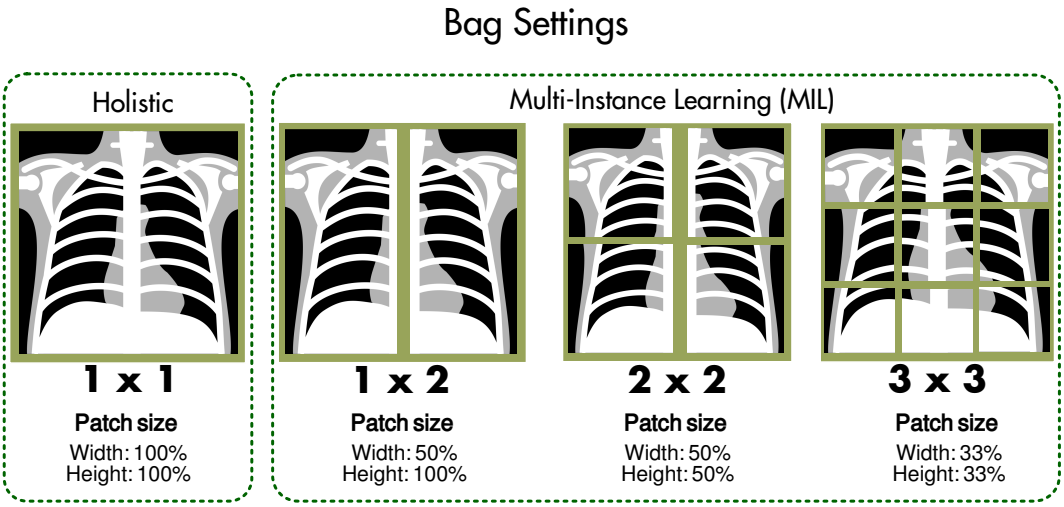


Figure 5.2: Patch settings utilised by the MIL method during training and testing. Note that the 1x1 setting corresponds to holistic image classification for which a second training stage is not implemented. (CXR image attributed to: Jmarchn, 2016, cc BY-SA 3.0)

The choice of the patch extents associated to these bags shows a staggered trade-off between global and local context within an image. Here the 1x1 bag setting provides the most global context and least local context, while the opposite is true for the 3x3 setting which provides the least global context and the most local context. To provide patches under these settings, crops corresponding to the patch dimensions given in Figure 5.2 are extracted from each full-sized image and are subsequently resized to 224x224 pixels using bilinear interpolation.

In order to form the boosting datasets for each of these models, the patch extraction thresholds (discussed in Section 3.3.2) are set as $\alpha_0 = 0.6$ and $\alpha_1 = 0.4$ respectively. During training, parameter initialisation, fine-tuning and hyper-parameter selection are followed as outlined within Section 4.3. The resulting models are referred to as *GoogLeNet-1*x*1*, *GoogLeNet-1*x*2, GoogLeNet-2*x*2, GoogLeNet-3*x*3* as well as *NIN-1*x*1*, *NIN-1*x*2, NIN-2*x*2* and *NIN-3*x*3* respectively. Under the '1x1' patch setting, the MIL method corresponds to standard holistic image classification. As such, both the *GoogLeNet-1*x*1* and *NIN-1*x*1* models provide a baseline comparison for the MIL method against this alternative classification technique.

5.4 Results and Discussion

The proposed MIL classification method is tested on both synthetic CXR datasets. The resulting performance metrics and evaluation data for each model are provided within the subsections which follow.

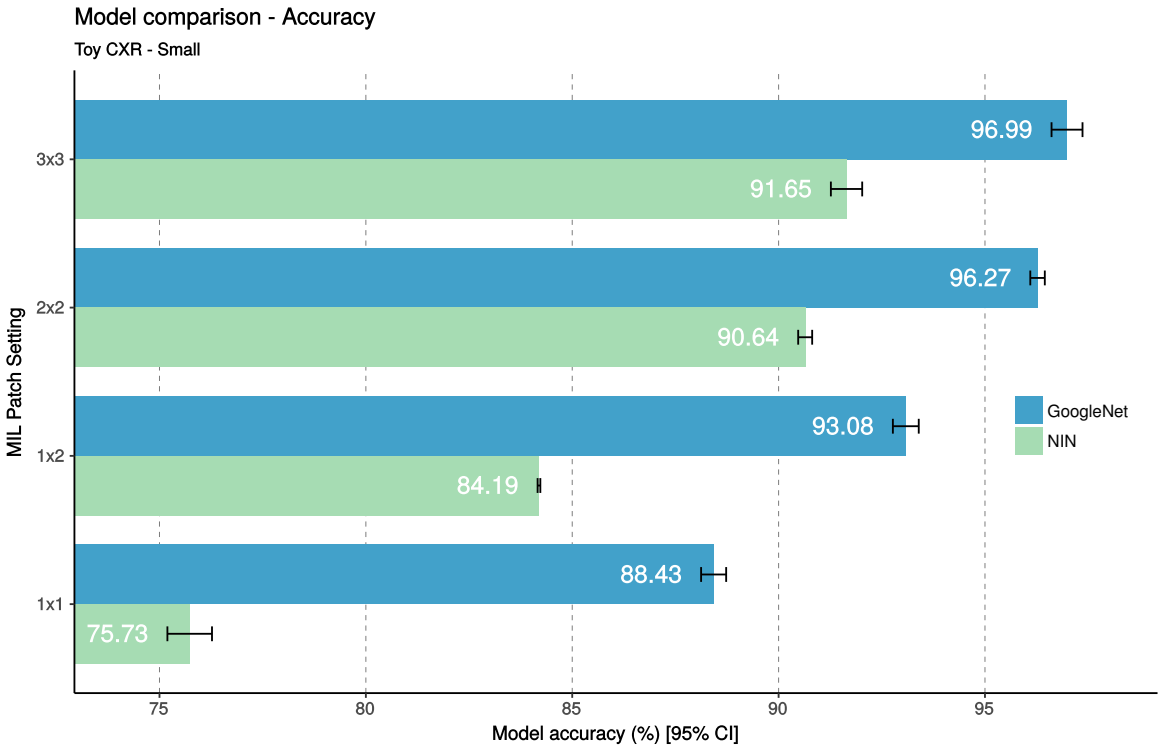


Figure 5.3: Accuracy results for the synthetic CXR-Small dataset.

5.4.1 CXR - Small

Figure 5.3 shows the model accuracies obtained for the CXR-Small dataset. Here the highest performances are both for the 3x3 patch setting, with the GoogLeNet and NIN models achieving 96.99% and 91.65% respectively. The lowest performances are obtained for the holistic classification setting, with accuracies of 88.43% and 75.73% for the respective networks being recorded. In this regard, the use of the MIL method is seen to boost performance, with the 1x2 patch setting providing increases of approximately 4% and 8% over the 1x1 (holistic) setting. This improvement is seen to continue as smaller patch settings (2x2 and 3x3) are used.

The effect of the MIL method on both the sensitivity and specificity levels for the trained models can be observed from the receiver operating characteristic (ROC) curves given in Figure 5.4. For the GoogLeNet models, the 3x3 patch setting yields the highest AUC score of 0.988, while the 1x1 setting achieves the lowest of 0.908. Furthermore, the 3x3 setting achieves a sensitivity of 98% versus the 85% obtained by the 1x1 setting when compared at a false-positive rate of 0.2. Similar findings are also produced for the NIN models, with the 3x3 patch setting recording the highest AUC and sensitivity scores of 0.954 and 94% respectively, while the lowest AUC and sensitivity scores of 0.811 and 70% arise from the 1x1 setting.

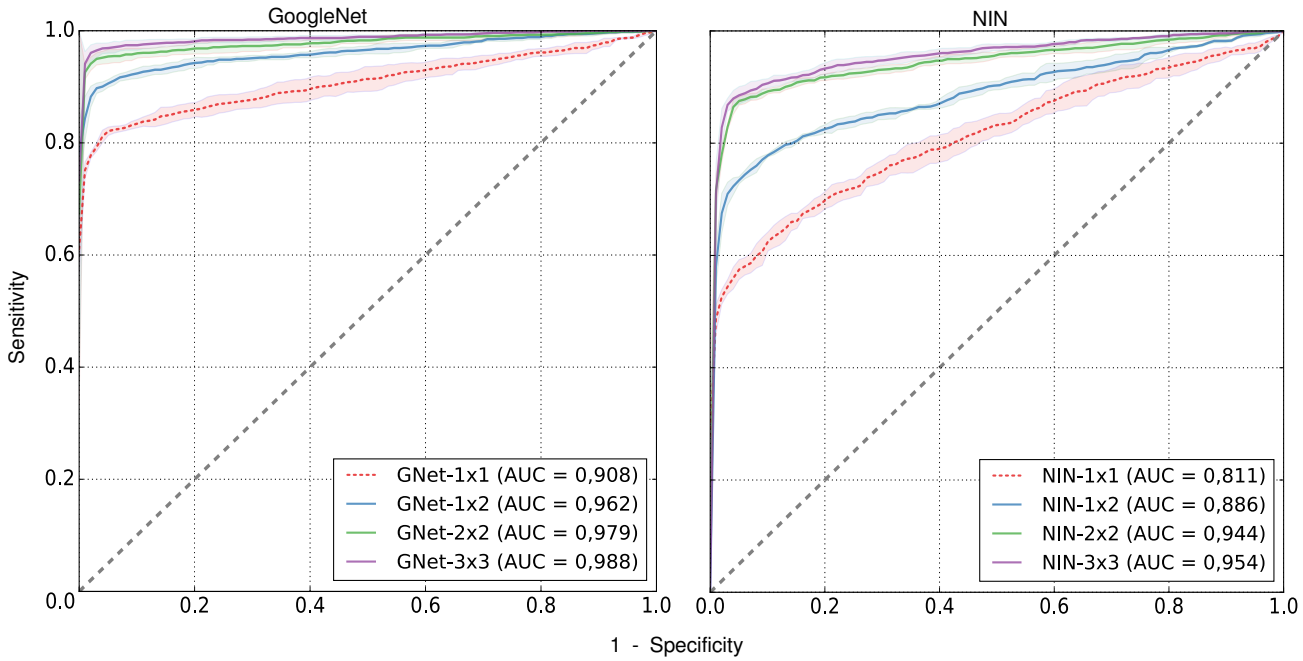


Figure 5.4: ROC curves and AUC metrics for the 'CXR - Small' dataset.

The presented results indicate that the MIL method allows a DCNN model to become more sensitive to local discriminative information than does holistic classification. This is further illustrated by observing that the GoogLeNet-1x1 model is required to operate at 8% specificity to achieve the same sensitivity as the 3x3 setting operating with 96% specificity. Relating to this point, it is noted that for both the GoogLeNet and NIN models, the large difference in sensitivity diminishes for the smaller patch sizes. This suggests that after reaching a sufficient scale of representation for local information, the use of smaller patch settings within the MIL method has a diminished effect on improving model performance.

In terms of inter-model performance for the given classification task, the GoogLeNet models are seen to produce superior results over the NIN models. However, under the MIL setting, both the 3x3 and 2x2 NIN models are seen to outperform the GoogLeNet-1x1 model trained using holistic image classification. This highlights how the MIL method offers a viable alternative to increasing model complexity as a means to enhance the detection of local information.

In order to qualitatively review the classification results obtained, Figure 5.5 provides saliency maps for false-positive, true-positive and false-negative findings produced by the best (3x3) and worst (1x1) performing GoogLeNet models during testing.

For the false-positive examples given in Figure 5.5 (a) - (d), it is noted that both the 1x1 and 3x3 models incorrectly classify certain imaging artefacts (jewellery or ECG electrodes) as being discriminative for the *abnormal* class. However, as shown

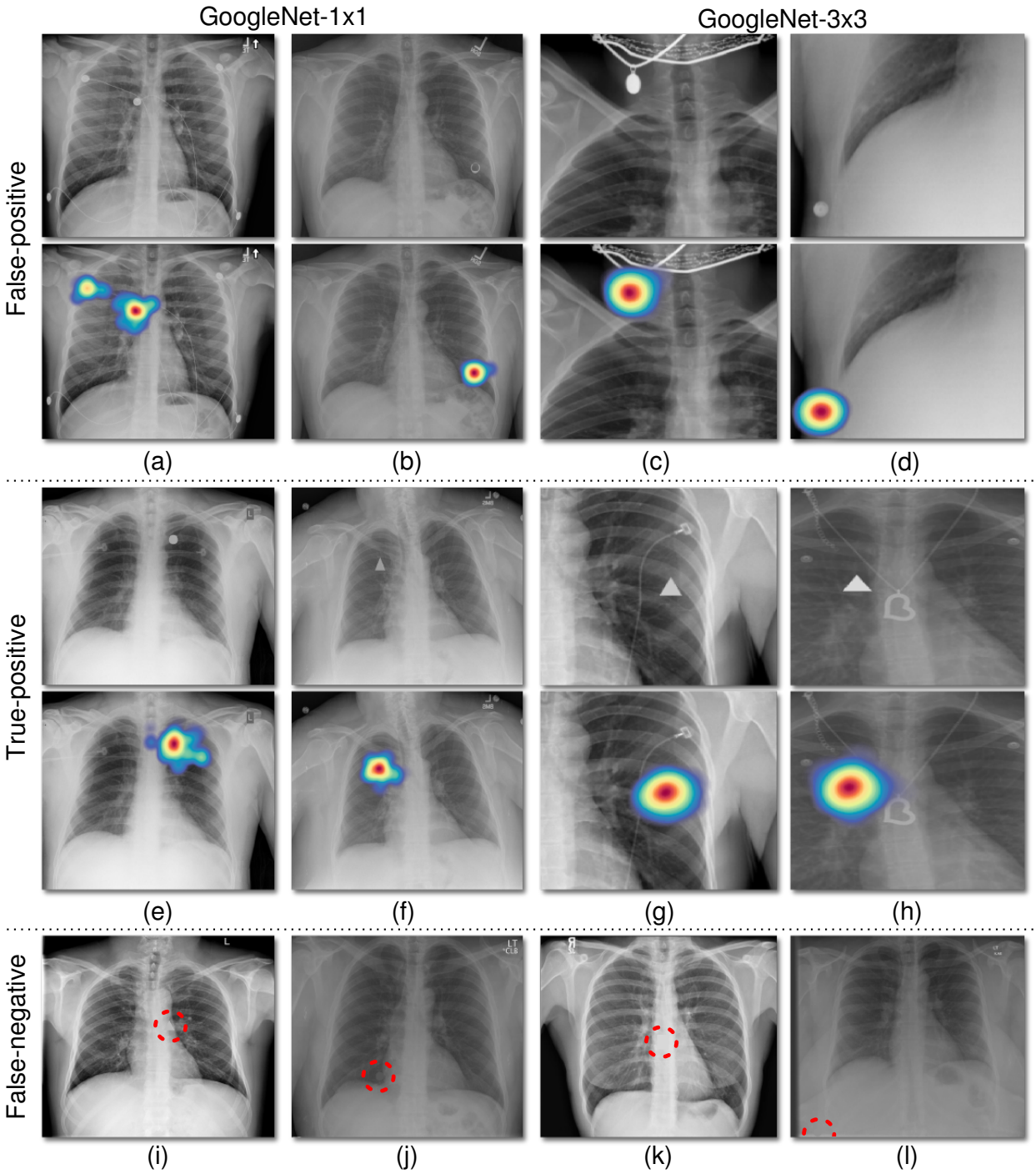


Figure 5.5: Saliency maps generated for the best (3x3) and worst (1x1) GoogLeNet models on the CXR-Small dataset. (a) - (d) present false positive examples wherein both the 3x3 and 1x1 settings are seen to incorrectly recognise imaging artefacts as discriminative shapes. (e) - (h) provide examples of true positive findings in which the class-aligning shapes are correctly detected. Within (i) - (l) false-negative findings are presented with missed shapes indicated with a red boundary. Those shapes missed by the 3x3 setting are particularly difficult for even a human observer to find.

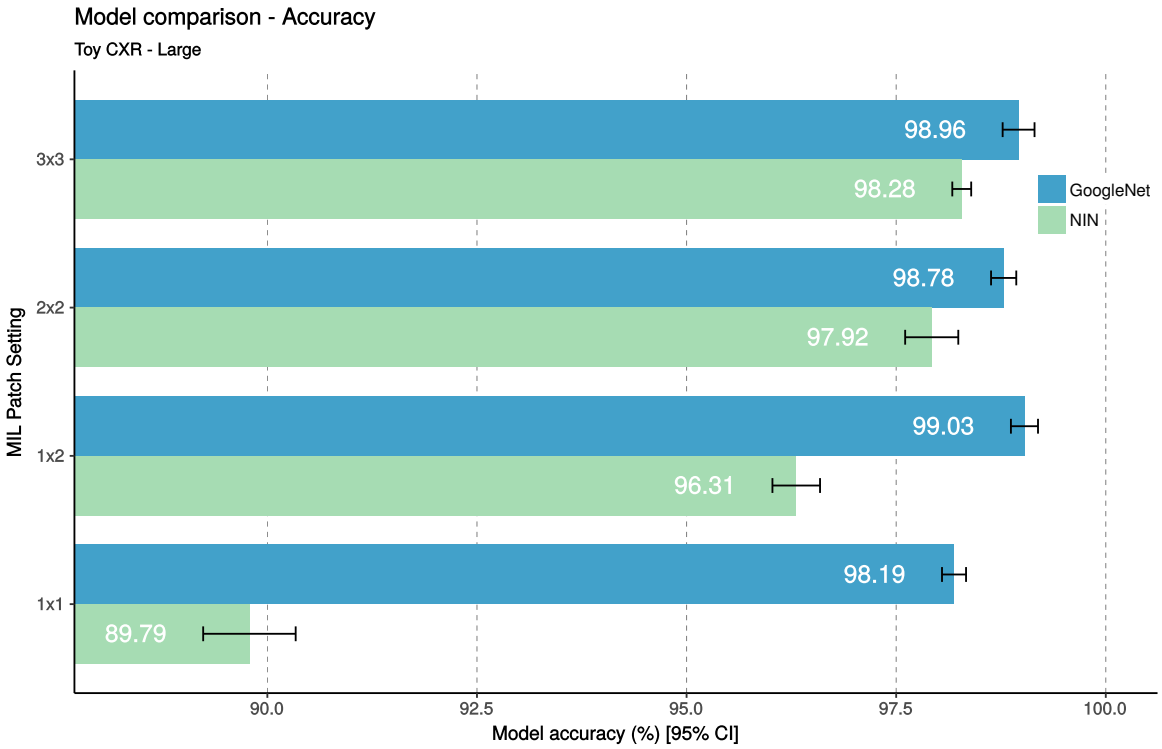


Figure 5.6: Accuracy results for the synthetic 'CXR-Large' dataset.

within Figure 5.5 (g) - (h), the use of MIL allows the trained models to become more resilient to these sources of imaging noise. It is further observed that while both models produce false-negative results, as given in Figure 5.5 (i) - (l), those for the 3x3 setting pose a challenging detection problem even for human readers - emphasising the ability of the MIL method to robustly detect local discriminative information.

5.4.2 CXR - Large

Figure 5.6 presents the accuracies obtained for the GoogLeNet and NIN models on the CXR-Large dataset. The holistic setting produces the worst performances, with the GoogLeNet-1x1 and NIN-1x1 models achieving scores of 98.19% and 89.79% respectively. However, while the 3x3 patch setting still yields the top accuracy of 98.28% for the NIN models, the 1x2 patch setting is seen to produce the best accuracy of 99.03% for the GoogLeNet models. Smaller performance improvements over holistic classification are also observed compared to the CXR-Small dataset. This declining improvement is expected and indicates that the larger shape sizes within each *abnormal* image are of a sufficient scale to be detected and classified appropriately by the holistic setting.

ROC curves obtained from the various models trained on the large CXR dataset are

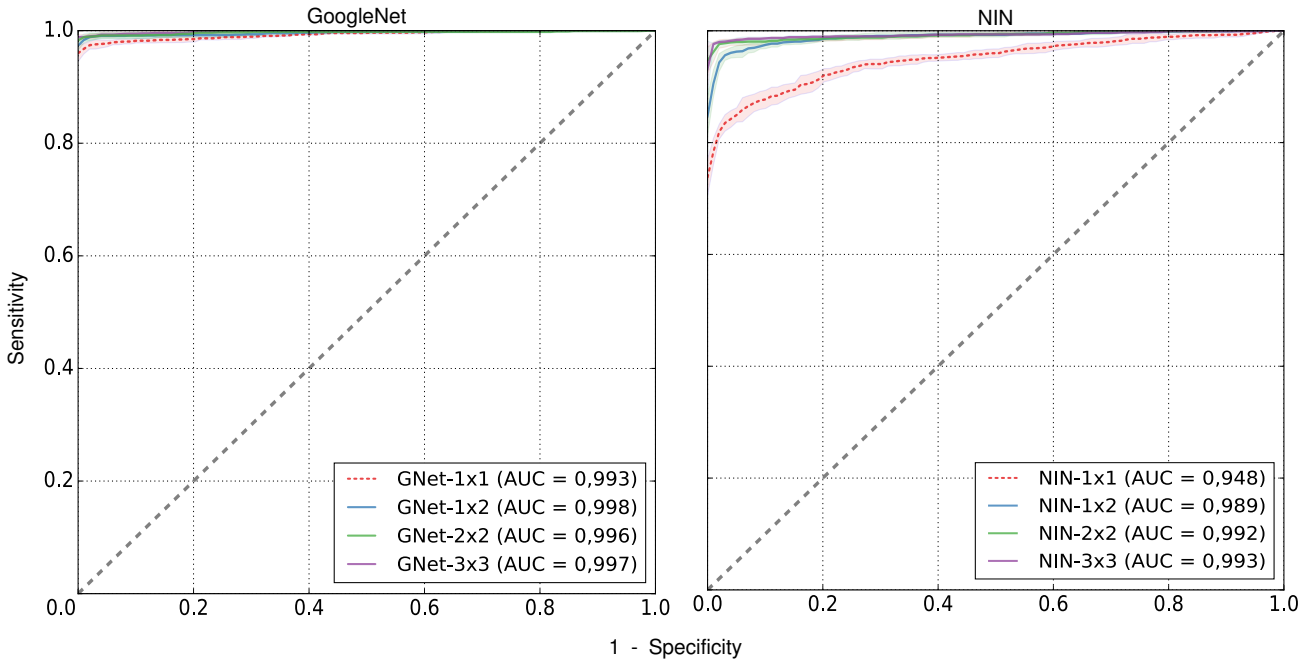


Figure 5.7: ROC curves and AUC metrics for the 'CXR-Large' dataset.

given in Figure 5.7. For the GoogLeNet models, these curves present near-perfect classifier performances with AUC scores ranging between 0.993 (1x1 setting) and 0.998 (1x2 setting) being recorded. This range is larger for the NIN models, with the lowest AUC score of 0.948 being achieved by the 1x1 setting and the highest of 0.993 being achieved for the 3x3 setting. Furthermore, the 1x1 setting is seen to achieve a sensitivity of 0.9 compared to the 0.989 achieved under the 3x3 setting at a false-positive rate of 0.2. In this regard, the MIL method is still seen to be beneficial in training the NIN models even though larger sources of discriminative information are present.

To further explore the results obtained, Figure 5.8 presents sample saliency maps generated by the 1x1 and 3x3 NIN models on examples drawn from the synthetic CXR-Large test set. Here (a) - (d) present false-positive findings, while (e) - (h) and (i) - (l) present true-positive and false-negative findings respectively. Interestingly, while both models produce several false-positive findings, the 1x1 setting is seen to be incorrectly sensitive to anatomical noise such as the aortic arch of the heart (upper region of Figure 5.8 (a) - (b)). In contrast to this limitation by holistic classification, (g) - (h) illustrate how the MIL method allows the 3x3 model to become insensitive to this anatomical noise and classify discriminative shape information correctly.

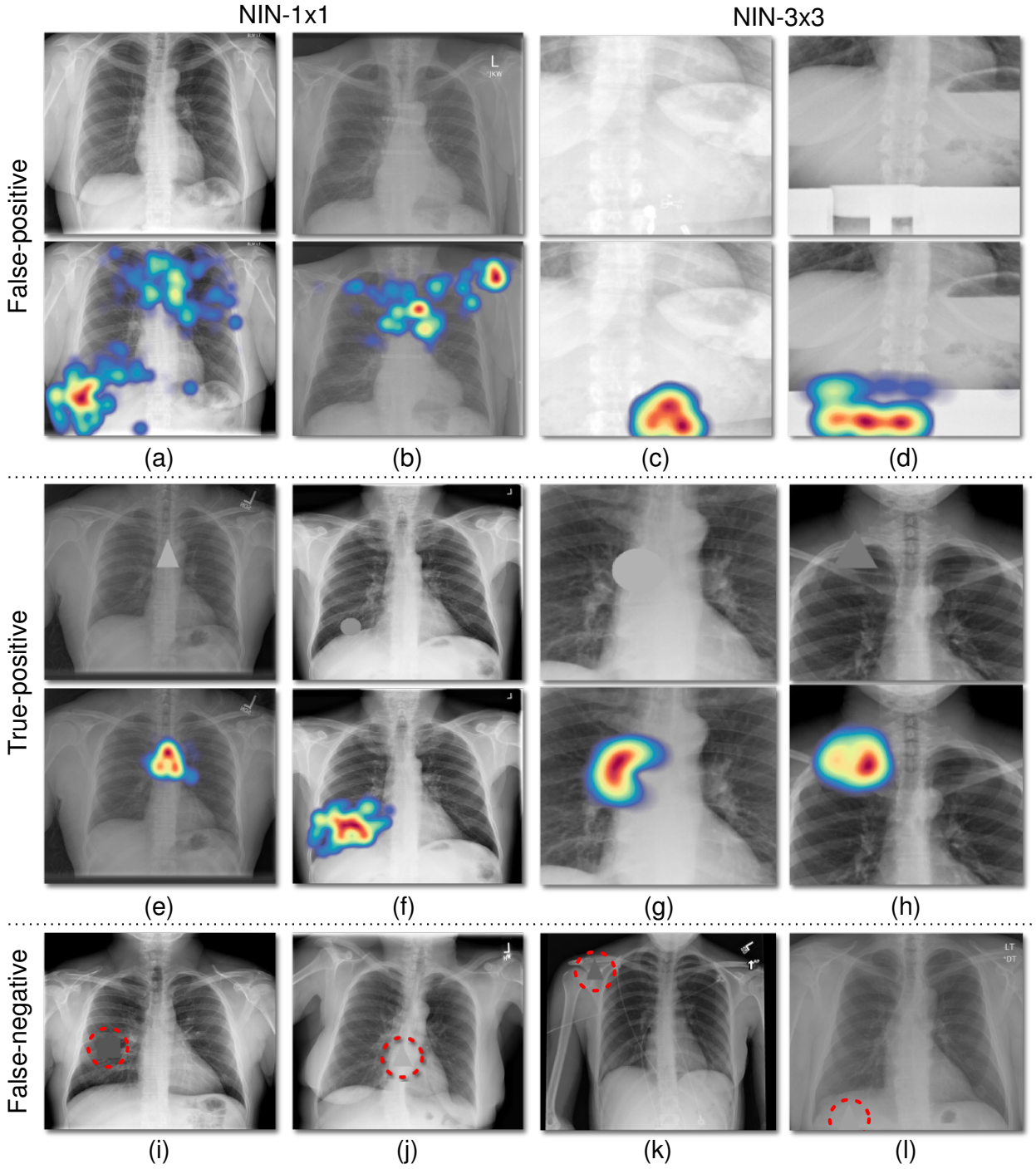


Figure 5.8: Saliency maps generated for the best (3x3) and worst (1x1) NIN models on the CXR-Large dataset. (a) - (d) provide false-positive findings in which the models are seen to be sensitive to both anatomical and imaging artefact noise. (e) - (h) give true-positive examples from both models. Within (i) - (l), a red boundary is used to indicate missed shapes for false-negative examples.

Table 5.2: Stage-1 vs Stage-2 performance for the synthetic CXR datasets.

		AUC (%)			
	Patch Setting	**1x2**	**2x2**	**3x3**	**Avg**
Small	GoogLeNet: Stage-1	95.09	97.07	98.52	96.89
	GoogLeNet: Stage-2	**96.16***	**97.86***	**98.76**	**97.59**
	NIN: Stage-1	87.42	92.72	94.90	91.68
	NIN: Stage-2	**88.60***	**94.29***	**95.43***	**92.77**
Large	GoogLeNet: Stage-1	99.63	**99.69**	99.47	99.59
	GoogLeNet: Stage-2	**99.75**	99.62	**99.67**	**99.61**
	NIN: Stage-1	98.56	98.92	99.15	98.87
	NIN: Stage-2	**98.92***	**99.20***	**99.26**	**99.13**

* Indicates a result for which 95% confidence intervals don't overlap.

Note: Bold figures represent best performance.

5.5 Efficacy of MIL Method Training Stages

Chapter 3 introduced two training stages of the MIL method which enable a model to become sensitive to both discriminative and non-discriminative information within local regions of an image. To validate the functioning of these training stages, this section presents and reviews data collected during model training and testing on the synthetic CXR datasets.

To observe the correct functioning of stage-1 training, Figure 5.9 presents randomly-selected examples of the discriminative and non-discriminative patch classes used for model boosting. Here patches belonging to the positively discriminative class all correctly contain shapes indicating *abnormal* images. Furthermore, the negatively discriminative class patches have an absence of these shapes to correctly indicate *normal* images. Several of the non-discriminative class patches contain either anatomical or imaging noise, highlighting the susceptibility of these examples to be incorrectly classified.

The efficacy of the second, boosting stage of the MIL method is motivated by the results given in Table 5.2. This table presents the average AUC scores achieved under the various patch settings for the synthetic CXR test sets after stage-1 and stage-2 training. As can be seen within these results, the use of boosting causes a general increase in performance within these models, with these improvements being especially pronounced for the CXR-Small dataset and the NIN models. In this regard,

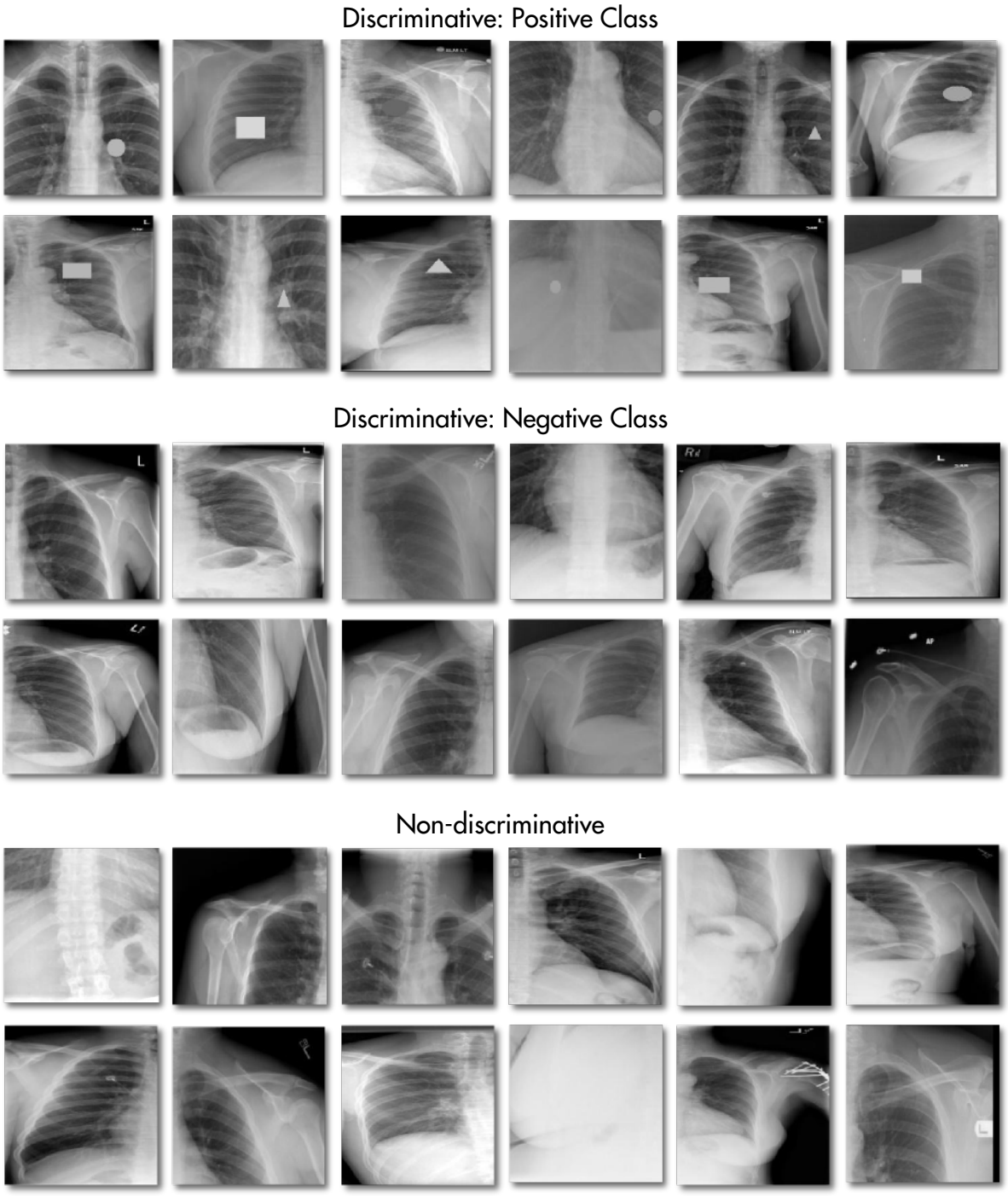


Figure 5.9: Example patches extracted from the CXR-Small and CXR-Large datasets by various models during stage-2 of the MIL method.

the second stage of training is seen to be beneficial to the performance of models trained using the MIL method.

In consideration of the reviewed two-stage approach, the reader may possibly question whether additional MIL training iterations - afforded by the second stage - pose an unfair advantage over the single training stage used by holistic classification. To this point, it is noted that all training is monitored by the early stopping algorithm (described in Section 2.4.6). As such, each model (whether MIL or holistic) is exposed to an appropriate amount of data before training ceases, thereby limiting its susceptibility to overfit if additional training iterations were to continue.

5.6 Summary

The MIL method is successfully implemented and outperforms holistic classification on the 'Small' and 'Large' synthetic CXR datasets. These results see smaller bag settings (3x3, 2x2) producing the best performances over each metric, while holistic classification consistently produces the worst performances. For the CXR-Small and CXR-Large datasets the differences in accuracy between the GoogLeNet-3x3 and GoogLeNet-1x1 models are 8.55% and 0.77% respectively, while for the NIN-3x3 and NIN-1x1 models these differences are 15.92% and 8.49%. From these results it is concluded that:

1. The MIL method enables a model to become more sensitive to local discriminative information compared to holistic classification.

2. The efficacy of the MIL method diminishes when larger sources of discriminative information are present within an image.

The training stages of the MIL method are also validated during evaluation on the synthetic CXR datasets. Here a review of image patches extracted for model boosting shows that stage-1 training allows a model to discover local discriminative information. An inspection of averaged model performances obtained following stage-1 and stage-2 training further demonstrates that the latter stage enables model boosting, increasing model performance following stage-1.

Chapter 6

Experimental Comparison of MIL and Holistic Classification

6.1 Introduction

In Chapter 5, the MIL method proposed within this dissertation was empirically tested on two synthetic datasets, with the results showing superior performance in classifying local discriminative information using a DCNN as compared to holistic classification. This chapter extends the evaluation of the proposed MIL method by assessing its performance on four digital CXR datasets containing pathology cases for paediatric pneumonia, pulmonary tuberculosis (PTB), airspace opacity and lung infiltrates respectively. In conjunction to this evaluation, the performance of holistic classification is also gauged across these datasets to provide a comparative standard against which the MIL method's results can be measured.

Initially the experimental setup is defined, providing the data and model settings used to compare the MIL and holistic classification methods. The description and results associated with each dataset are then given, followed by a review of training and inference times for various model settings used during experimentation. The chapter then concludes by first presenting a critical discussion surrounding the overall performances yielded by the compared classification methods, and then considering these performances against those of human readers recorded in the literature.

6.2 Datasets and Data Preparation

Four medical CXR datasets are selected for the testing of the MIL method and its comparison with holistic image classification. These datasets, which are described in detail within sections 6.4-6.6 and summarised in Table 6.1, present radiological cases

Table 6.1: Summary of datasets used during evaluation.

Dataset	**Pathology type**	**Origin**	**Orientation**	**Population age**	**Number of cases**		**Format**	**Resolution** ($w \times h$)
					Positive	**Negative**		
Indiana OpenI: Opacity subset [101]	Airspace opacity	Indiana Network for Patient Care, USA	PA	Adult	425	425	PNG	$512 \times 419 - 624$
PCV-13 [159]	Pneumonia	PCV case control study, South Africa	AP	Paediatric	586	586	JPEG	$698-3893 \times 504-4158$
ChestX-ray14: Infiltration [14]	Lung infiltrate	National Institutes of Health Centre, USA	PA/AP	Paediatric/ Adult	5619	5619	PNG	1024×1024
Montgomery-Shenzhen [26]	Pulmonary TB	Montgomery county, USA. Shenzhen province, China	PA/AP	Paediatric/ Adult	394	406	PNG	$948 - 3001 \times 4020 - 4892$

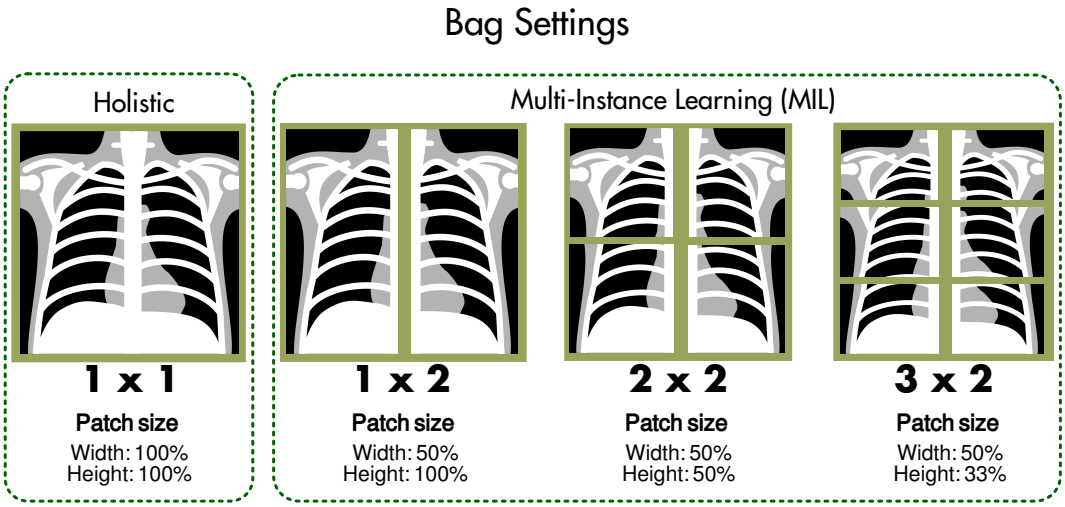


Figure 6.1: Patch settings utilised during the evaluation of the MIL method on the CXR datasets. (CXR image attributed to: Jmarchn, 2016, CC BY-SA 3.0)

of PTB, pneumonia or chest pathologies associated with this latter disease [24, 36] (airspace opacity and lung infiltrate). The use of several datasets for evaluation is motivated in order to determine the efficacy of the proposed method across multiple patient populations - allowing for variations in age, study location, and CXR orientation to be considered when assessing the ability of the method to generalise.

During evaluation, each dataset is split into five distinct folds using 5-fold cross validation as described in Section 4.3.1. Each resulting fold contains training, validation and testing sets of 60%:20%:20% of the fold data respectively. Data augmentation is then applied to each fold's training set using the translation, rotation and reflection operations detailed within Section 4.3.2 to increase the effective volumes of data exposed to each model during training. To provide patches under the various bag settings, crops corresponding to the patch dimensions given in Figure 6.1 are extracted from each full-sized image and are subsequently resized to 224x224 pixels using bilinear interpolation.

6.3 Model Selection, Training and Evaluation

As motivated in Section 4.2, the NIN and GoogLeNet DCNN architectures are selected to experimentally compare the MIL and holistic classification methods. For each network architecture, four model variants are trained corresponding to the respective patch settings represented in Figure 6.1. At this point, it is noted that a difference exists between the 3x3 patch setting used during the MIL method's validation (described in Section 5.3), and the 3x2 patch setting presented above. This difference

arises as the latter 3x2 setting requires fewer patches overall during evaluation, decreasing effective model training times, while still providing an increased number of patches sampled from the lung fields of each CXR. In the same manner as described within the MIL method validation, it is further noted that for the 1x1 setting model training and evaluation correspond to holistic classification. Thus both the GoogLeNet and NIN models trained under this setting provide a baseline for comparison between holistic and MIL-based classification. Overall, the resulting models are referred to as *GoogLeNet-1*x*1*, *GoogLeNet-1*x*2, GoogLeNet-2*x*2, GoogLeNet-3*x*2* as well as *NIN-1*x*1*, *NIN-1*x*2, NIN-2*x*2* and *NIN-3*x*2* respectively.

Training is performed as explained within Section 4.3. Initially, the above models undergo pre-training and parameter initialisation as described in Section 4.3.3. Following this, each model is trained upon the target dataset for which it will be evaluated. For the holistic models (GoogLeNet-1x1 and NIN-1x1), this training step takes the form of standard fine-tuning. For the MIL models, this training involves both the stage-1 and stage-2 MIL training steps defined within Chapter 3. In order to form the boosting datasets for each of these models during stage-2, the patch extraction thresholds (discussed in Section 3.3.2) are set as $\alpha_0 = 0.6$ and $\alpha_1 = 0.4$ respectively.

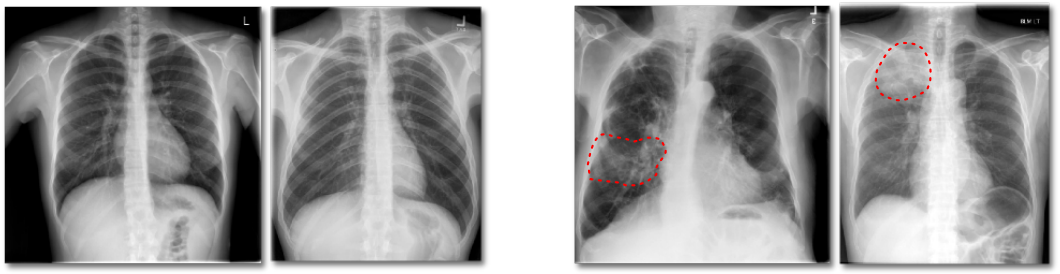
The process for model testing and validation is presented in Section 4.4. In this regard, five training and evaluation runs are performed for each model setting using 5-fold cross-validation at an image level - providing a total of 25 models for each patch setting per dataset. The overall average accuracy, sensitivity, specificity, F1-score and ROC curve metrics obtained from these models are presented as results.

6.4 Indiana OpenI Opacity Dataset

6.4.1 Description

The Indiana University chest X-ray dataset comprises of 7470 fully anonymised posterior-anterior (PA) and lateral CXRs with accompanying radiology reports obtained from the Indiana Network for Patient Care image archiving system [101]. The digital radiographs, which are stored at $512 \times 419 - 624$px in PNG (8-bit grey level) format, are freely available as a subset of the OpenI biological image and medical literature collection (https://openi.nlm.nih.gov/).

Associated with each radiograph is a set of Medical Subject Headings (MeSH) which provide a basis for pathology assignment to each image. In total, 101 MeSH terms are present within the dataset, several of which can be overlapping and simultaneously



(a) Normal (b) Opacity

Figure 6.2: Example CXRs drawn from the Indiana OpenI opacity dataset.

Table 6.2: Performance results for the Indiana OpenI dataset.

Setting	**Sensitivity**	**Specificity**	**F1-score** (95%CI)	**AUC** (95%CI)
GoogLeNet-1x1	**76.00**	80.04	**77.68** (±0.34)	**85.71** (±0.41)
GoogLeNet-1x2	71.36	**83.44**	76.55 (±1.37)	85.26 (±0.63)
GoogLeNet-2x2	74.20	78.60	75.98 (±0.89)	83.70 (±0.77)
GoogLeNet-3x2	73.96	77.28	75.14 (±0.66)	83.99 (±1.01)
NIN-1x1	**77.40**	75.68	**76.15** (±0.92)	**84.62** (±0.98)
NIN-1x2	72.20	**79.36**	75.37 (±1.06)	82.70 (±1.04)
NIN-2x2	75.36	76.12	75.26 (±1.94)	83.18 (±1.38)
NIN-3x2	76.48	74.76	75.15 (±0.67)	83.19 (±0.93)

Note - All values are expressed as percentages. Bold indicates best results.
AUC = Area Under Curve

present within any one report. The most frequent MeSH term, *normal,* covers 2969 radiographs and has no overlap with other terms. While both PTB and pneumonia are under-represented within the dataset, a large number of airspace opacity cases - whose radiological findings are similar to that of pneumonia - are present. Due to this higher availability, a balanced subset of 425 *normal*, and 425 *opacity* frontal cases are formed into two distinct classes and are chosen for evaluation. Figures 6.2a and 6.2b present visual examples of these two classes respectively. Following data augmentation, each fold within the dataset is seen to contain 8256 train, 170 validation and 170 testing images respectively.

6.4.2 Results

Figure 6.3 presents the model accuracies achieved for the Indiana OpenI opacity dataset, while the overall sensitivity, specificity, F1-score and AUC metrics obtained

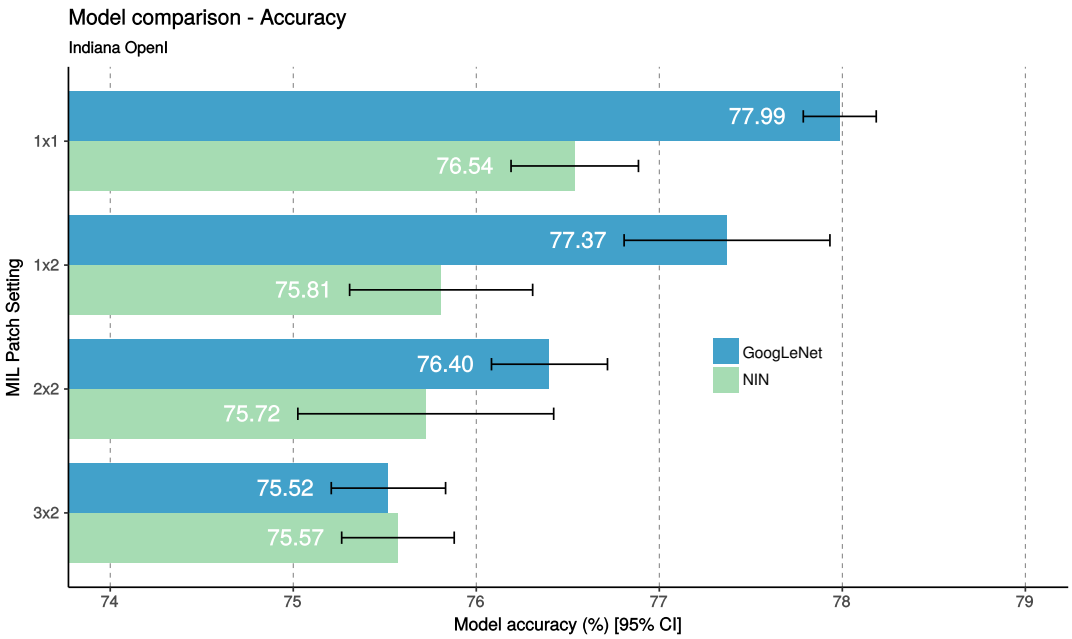


Figure 6.3: Model accuracies obtained for the Indiana OpenI dataset. Error bars indicate 95% confidence intervals.

by these same models are summarised in Table 6.2. For both network architectures, the holistic setting produces the best performance, with the GoogLeNet-1x1 and NIN-1x1 models achieving accuracies of 77.99% (95% CI: ±0.2) and 76.54% (95% CI: ±0.35) respectively. In comparison, the models trained under the MIL setting show decreasing performance as the patch size is reduced - with the lowest accuracies of 75.52% (95% CI: ±0.31) and 75.57% (95% CI: ±0.31) being yielded for the GoogLeNet-3x2 and NIN-3x2 models.

The ROC curves for the GoogLeNet-1x1, -1x2, -2x2 and -3x2 models, along with the corresponding curves obtained for the NIN models are shown in Figure 6.4. Here both 1x1 model settings are again seen to produce the best characteristics, with the GoogLeNet-1x1 model possessing a sensitivity of 84% at 70% specificity, while the NIN-1x1 model operates with 82% sensitivity at 70% specificity. At these same specificity levels, the worst performing GoogLeNet model (-2x2) possesses a sensitivity of 80% and the worst performing NIN model (-3x2) sensitivity of 80%.

As a means to visually inspect the results yielded by the compared methods on the OpenI opacity dataset, Figure 6.5 presents several saliency maps of false-positive (a - d), true-positive (e - h), and false-negative (i - l) examples drawn from the OpenI test set during evaluation with the NIN-1x1 and NIN-1x2 models. These findings are discussed briefly below:

Figure 6.5 **(a - d):** Within the false positive images shown in (a) - (d) both models

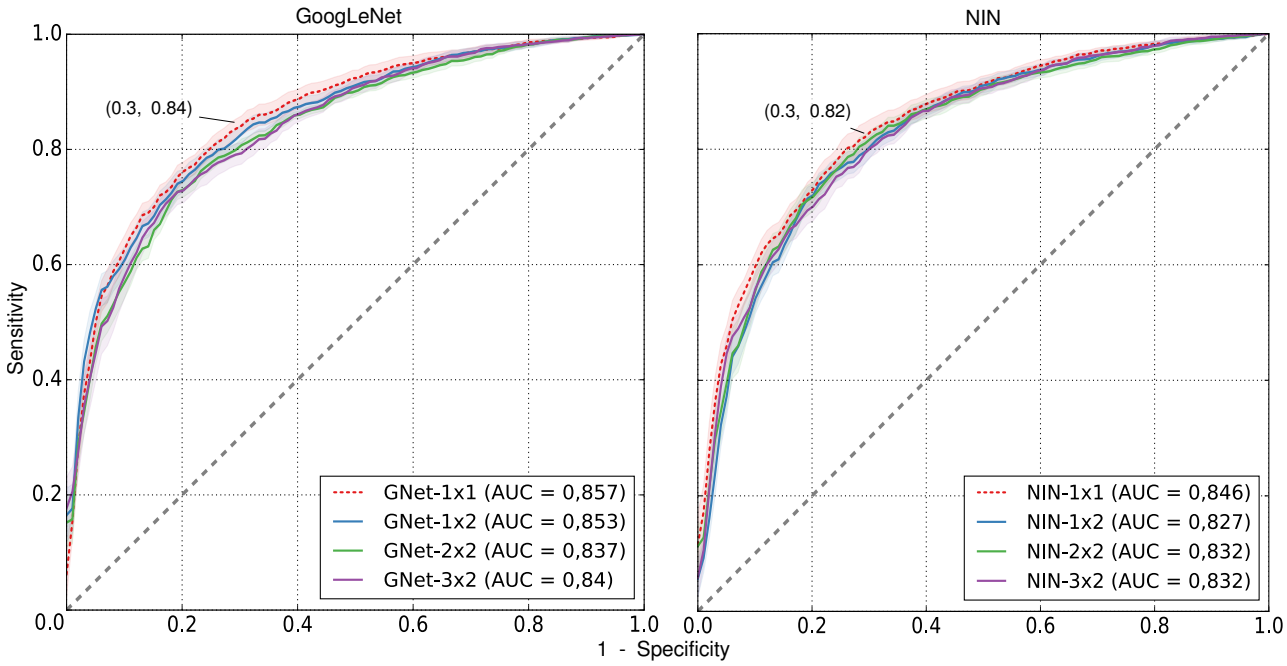


Figure 6.4: ROC curves generated for the various models evaluated on the Indiana OpenI dataset. Shaded areas represent a 95% CI for each curve.

are seen to be erroneously sensitive to normal anatomical features surrounding the hilar and basal regions of the lungs. This sensitivity could be attributed to the fact that subtle changes within these areas are often indicative of pathology (see examples below).

Figure 6.5 **(e - f):** In image (e) the NIN-1x1 model accurately detects opacity within the right midlung, while in image (f) the model indicates the presence of opacities which are bilaterally present in both lung bases. Interestingly, the attention shown by the model to other regions within these images is not unwarranted as additional pathological findings are present within them. In this sense, a strong response is seen at the left lung apex within (e) where pulmonary fibrosis is present, while the calcified granuloma at the right base of (f) also receives attention. This suggests that the model is sensitive to additional abnormal lung findings.

Figure 6.5 **(g - h):** Both the opacities within the peripheral aspect of the left lung in image (g), as well as the left lower lobe consolidation in image (h) are correctly detected by the NIN-1x2 model.

Figure 6.5 **(i - l):** The pathological findings missed by the NIN-1x1 and NIN-1x2 models are defined by features which are small and subtle in nature. Images (i) and (j) both present examples of localised circular opacities which are present within the right upper lung for (i), and bilaterally within the upper lungs for (j) respectively. Even at their original resolution of 512x512 pixels, these findings are difficult to

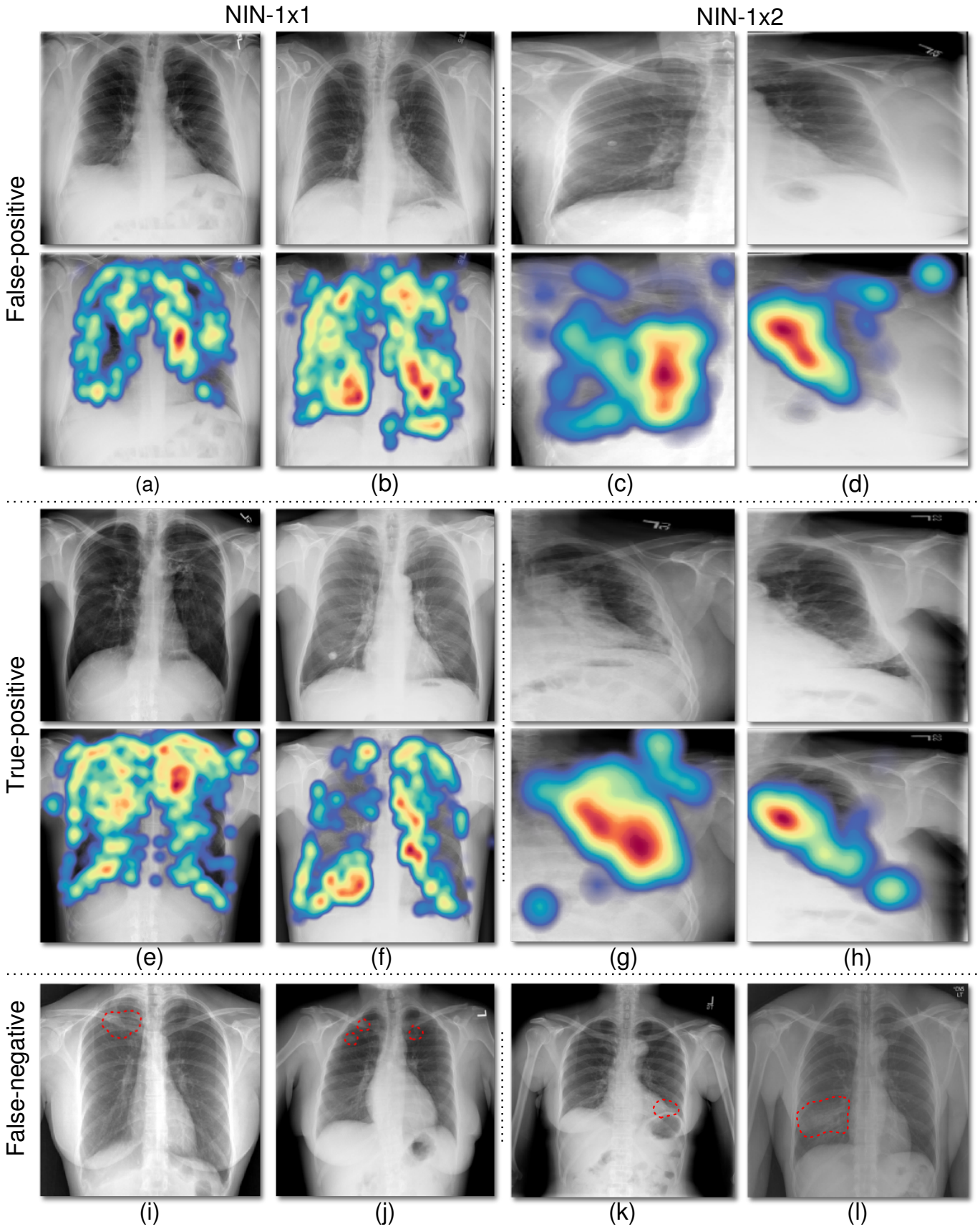
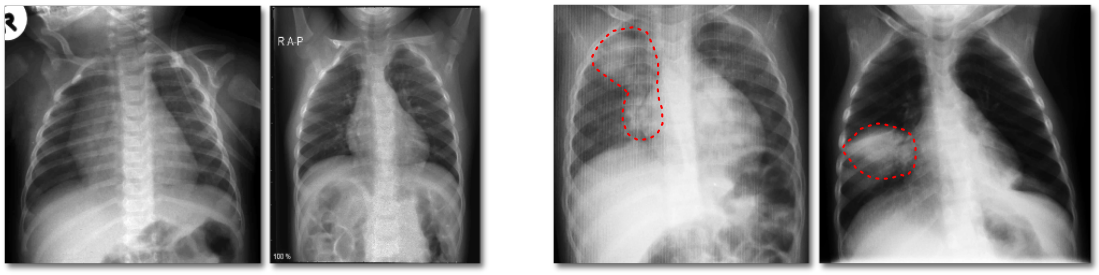


Figure 6.5: Saliency maps generated for various examples drawn from the Indiana OpenI opacity dataset using the NIN-1x1 and NIN-1x2 models.



(a) No infiltration/consolidation/effusion (Non-PEP) (b) Primary end-point pneumonia (PEP)

Figure 6.6: Example CXRs drawn from the PCV-13 paediatric pneumonia dataset.

detect. Images (k) and (l) present further subtle findings; with a small opacity at the left lung lingula within (k) appearing as a continuation of the cardiac border, and a diffuse airspace opacity within the right basal region of (l) being missed.

6.5 PCV-13 Paediatric Pneumonia Dataset

6.5.1 Description

The PCV-13 chest X-ray dataset forms part of a larger case control study evaluating the effectiveness of pneumococcal conjugate vaccine (PCV) against presumed bacterial pneumonia infections for hospitalised HIV-uninfected children [159]. The study took place in South Africa, with study sites situated in Cape Town (Red Cross Children's Hospital), KwaZulu-Natal (Ngwelezane Hospital) and Soweto (Chris Hani-Baragwanath Hospital), and produced a total of 3905 anterior-posterior (AP) and 2374 lateral CXRs (digital) of children from 1 - 59 months of age. The radiographs were stored in JPEG format (8-bit grey level), with the broad spread of paediatric ages accounting for the variation in image dimensions ($698 \times 504 - 3893 \times 4158$ px). To ensure confidentiality, complete anonymisation of the radiographs was performed at a patient level.

Two paediatricians trained in the WHO standardised chest X-ray interpretation methodology for paediatric pneumonia [24] formed a panel and were used to independently classify the frontal radiographs into one of three classes at an image level. In the event of a discord reading, a third senior radiologist was used to provide the reference standard. The final consensus readings were obtained as: *Primary end-point pneumonia* (PEP) - 1394 cases; *Other infiltrates* (OI) - 1751 cases; and *No infiltration/consolidation/effusion* (Non-PEP) - 586 cases. The remaining 174 radiographs were deemed to be uninterpretable by each reader.

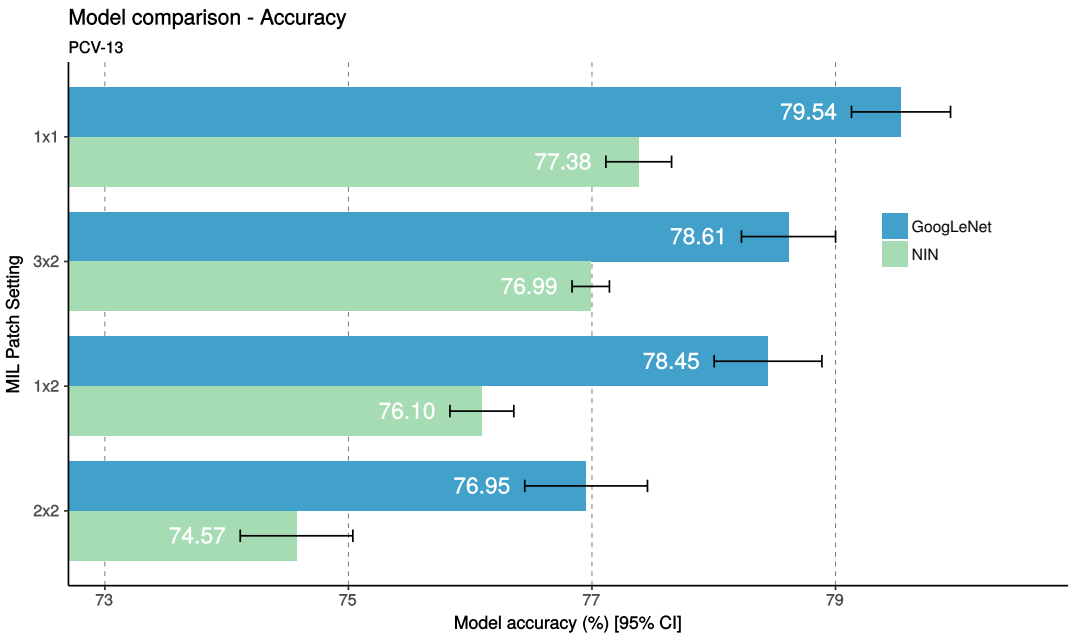


Figure 6.7: Model accuracies obtained for the PCV-13 dataset. Error bars indicate 95% confidence intervals.

For the diagnosis of paediatric pneumonia, the use of the OI class has recently yielded poor inter-observer agreement between clinicians, while the PEP class has shown higher levels of agreement and disease identification [20]. As such, a balanced subset of the PCV-13 dataset is chosen for evaluation consisting of two classes of 586 Non-PEP and 586 PEP cases respectively. Following the process of data augmentation, these two classes result in 11264 training, 234 validation and 234 testing data points per dataset fold respectively. Examples of the PEP and Non-PEP classes are provided in Figure 6.6a and 6.6b.

The PCV-13 dataset is not currently accessible to the public. As such, ethical clearance (certificate no. M161133 - Appendix A) from the medical human research ethics committee (MHREC) of the University of the Witwatersrand, Johannesburg has been obtained in order for its use.

6.5.2 Results

The model accuracies obtained for the PCV-13 dataset are provided in Figure 6.7. In a similar manner to the OpenI opacity subset, the GoogLeNet-1x1 and NIN-1x1 models are seen to produce the highest accuracies of 79.54% (95% CI: ±0.41) and 77.38% (95% CI: ±0.26) respectively, while the GoogLeNet-2x2 and NIN-2x2 produce the lowest accuracies of 76.95% (95% CI: ±0.5) and 74.57% (95% CI: ±0.46). The

Table 6.3: Performance results for the PCV-13 dataset.

Setting	**Sensitivity**	**Specificity**	**F1-score** (95%CI)	**AUC** (95%CI)
GoogLeNet-1x1	**79.52**	80.40	**79.79** (±1.04)	**87.26** (±0.67)
GoogLeNet-1x2	74.71	**82.75**	78.18 (±1.27)	86.82 (±0.67)
GoogLeNet-2x2	75.84	78.00	76.54 (±1.41)	85.16 (±1.23)
GoogLeNet-3x2	78.36	78.80	78.03 (±1.17)	86.71 (±0.43)
NIN-1x1	**76.96**	77.80	**76.97** (±0.7)	84.90 (±0.84)
NIN-1x2	74.40	77.92	75.77 (±0.61)	81.84 (±0.81)
NIN-2x2	73.04	76.16	74.18 (±1.17)	82.56 (±1.02)
NIN-3x2	75.76	**78.16**	76.66 (±0.22)	**85.16** (±0.87)

Note - All values are expressed as percentages. Bold indicates best results.
AUC = Area Under Curve

sensitivity, specificity, f1-score and AUC metrics additionally obtained during model evaluation are presented in Table 6.3.

Figure 6.8 depicts ROC curves computed for the various GoogLeNet and NIN models during evaluation on the PCV-13 dataset. While the 1x1 setting achieves the best characteristic for the GoogLeNet models, attaining a sensitivity of 81% at 80% specificity, the 3x2 setting is seen to slightly improve over the 1x1 setting for the NIN models, with a sensitivity of 83% at 70% specificity. At these same specificity levels, the worst performing models (GoogLeNet-2x2 and NIN-2x2) achieve sensitivities of 74% and 78% respectively.

To qualitatively assess the models trained on the PCV-13 dataset, Figure 6.9 presents several saliency map/example pairs representing false positive (a - d), and true positive (e - h) results obtained from the GoogLeNet-1x1 and GoogLeNet-3x2 models during evaluation. In addition to these results, false negative outcomes (i - l) are also provided within this figure. These findings are discussed briefly below:

Figure 6.9 **(a - d)**: Each of the images within (a) - (d) present false positive results which arise due to the close similarity of typical anatomical structures and pathological findings associated with paediatric patients. For images (a), (b), and (d), these similarities are seen within the right hemithorax where normal structures around the hilum appear as diffuse opacities. For image (c), the softening of the right lower heart border - which is typically indicative of the silhouette sign [36] - causes a false positive prediction to occur.

Figure 6.9 **(e - f)**: Within images (e) and (f), the GoogLeNet-1x1 model is seen to correctly detect the signs of end-point paediatric pneumonia. Here (e) shows correct

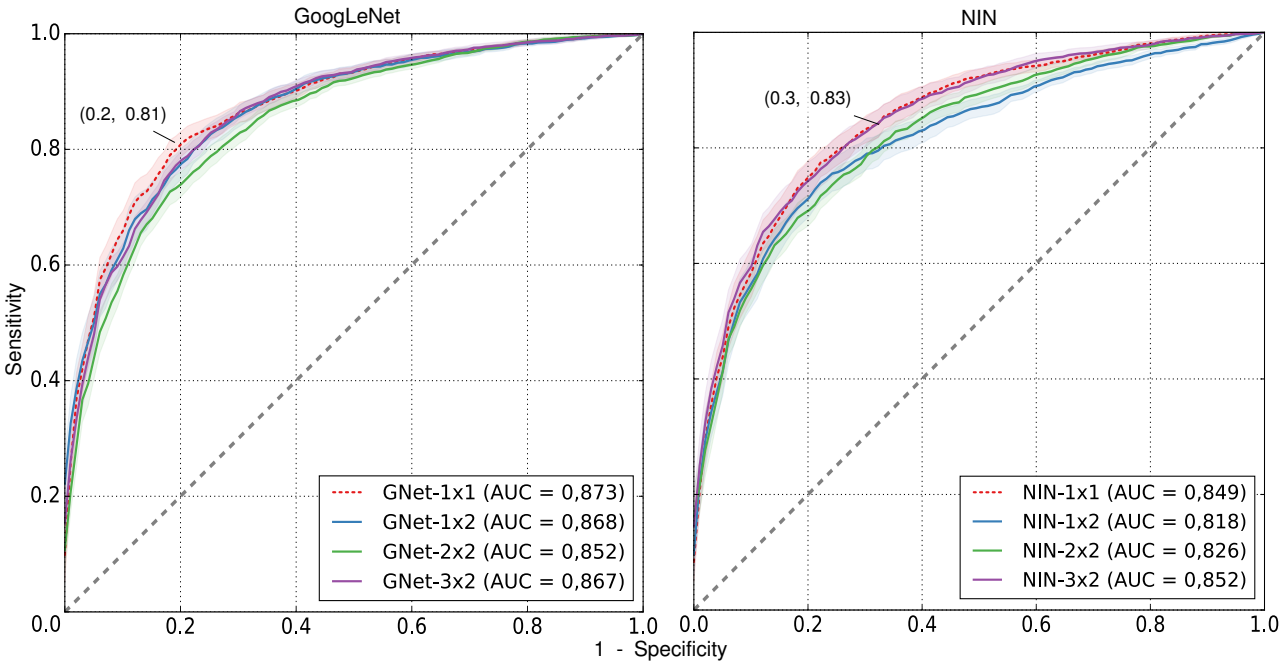


Figure 6.8: ROC curves generated for the various models tested on the PCV-13 dataset. Note that shaded areas represent a 95% confidence interval for each model.

attention being given to a dense wedge-shaped opacity within the right lung hilar region, which indicates right upper-lobe consolidation. Additionally, in image (f) the 1x1 model is also seen to be sensitive to diffuse opacification which is present throughout the entire left lung field

Figure 6.9 **(g - j)**: Images (g) - (j) provide examples of subtle pathology which is correctly detected using the MIL method, but is missed during holistic classification. In the first example, diffuse consolidation present within the right hilar and basal lung regions of (g) is given correct attention by the GoogLeNet-3x2 model. In contrast, these findings are seen to be missed under the holistic setting as shown in image (i). The second example motivates the need to detect subtle, localised information in order to make a correct classification of pathology. Within image (j) the dense opacity within the right upper-lung field may immediately appear as lobar consolidation. However, it is noted that in neonatal patients this opacity is typically associated with the thymus - a normal anatomic structure which is not indicative of pathology [63]. Thus the false negative finding produced by the GoogLeNet-1x1 model within (j) is understandable. Upon closer inspection of this region however, as provided by image (h), it can be seen that there is in fact consolidation which overlaps with the thymus. Under the MIL setting, this subtle discriminative information is correctly identified.

Figure 6.9 **(k - l)**: Pathological findings missed by the GoogLeNet-3x2 model are presented in images (k) and (l). Here subtle dulling at the right and left cardiac

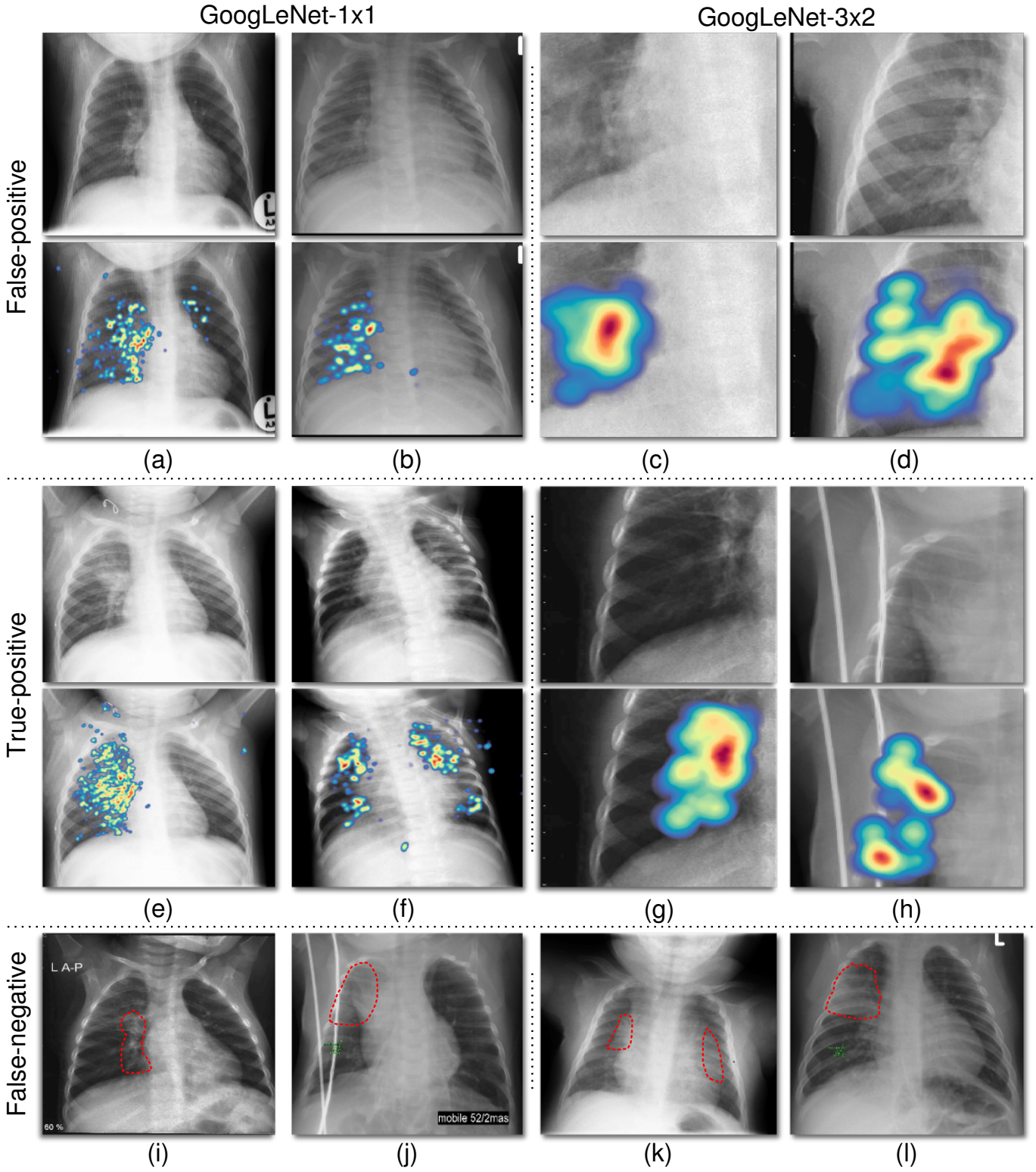
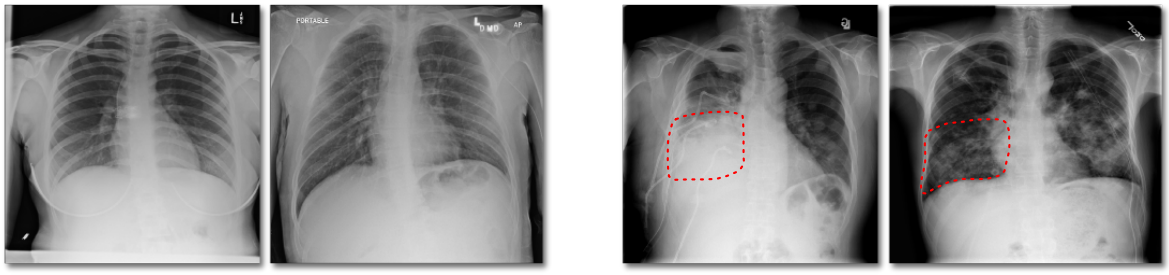


Figure 6.9: Saliency maps generated for various examples drawn from the PCV-13 dataset.



(a) Normal (b) Infiltration

Figure 6.10: Example CXRs drawn from the ChestX-ray14 infiltration dataset.

borders, which is indicative of the silhouette sign, does not receive appropriate attention in order to be correctly classified within (k). This may be due to the suboptimal exposure of the lung fields within this CXR, which causes uncertainty between pathological and normal anatomical findings when viewing only a portion of the lung fields under the MIL setting. Image (l) gives a clearer example of pathology being missed, wherein dense lobar consolidation in the right upper lung is left undetected.

6.6 ChestX-ray14 Infiltration Dataset

6.6.1 Description

ChestX-ray14 is a 'hospital-scale' radiographic dataset released as a means to benchmark thoracic disease classification and localisation systems built using deep learning methods [14]. The collection, which was sourced from the National Institutes of Health's picture archiving and communication system, contains 112 120 frontal CXRs (AP and PA) which cover 14 different chest pathologies in addition to normal/healthy findings. In order to provide labelling for a medical dataset of such magnitude, a two-pass text mining process was followed to extract pathology information from the radiological reports associated with each image. This automated approach gave rise to each CXR receiving one or more pathology labels at an image level, with 'normal' label being used as a background label in the event of no pathologies being detected. To make the original CXR data more usable (originally stored in DICOM format), each CXR image within the collection was re-sized to 1024×1024 px and is stored in PNG (8-bit grey level) format.

From the original ChestX-ray14 dataset, the *infiltration* label is used for experimental comparison between the holistic and MIL classification settings. Here, all images which

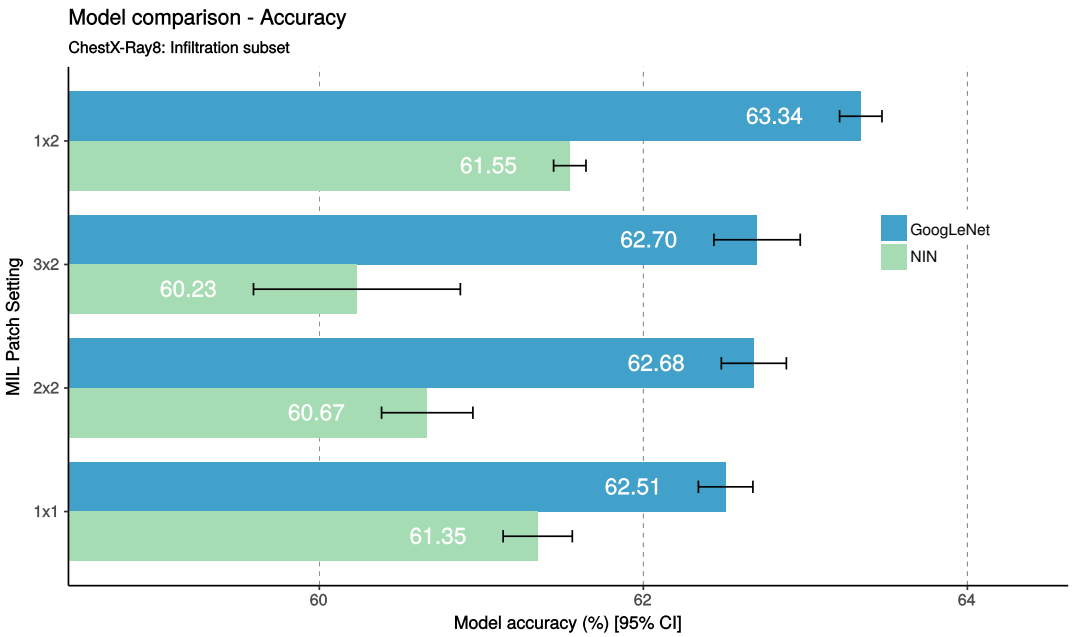


Figure 6.11: Accuracy results for the various GoogLeNet and NIN models evaluated on the ChestX-ray14 Infiltration subset. Error bars indicate 95% confidence intervals.

are exclusive to this label are formed into a class, with an equal number of *normal* images being randomly selected to form an opposing control class. The resulting *infiltration* dataset contains 5619 images per class respectively, with Figure 6.10 providing examples drawn from this subset. During data pre-processing, a conservative number of image augmentation operations are used due to the already well represented size of each pathology class. These augmentations, which consist of $\pm 10^o$ rotations and central cropping, give rise to a total of 11462 training, 3820 validation and 3820 testing data points per fold respectively.

6.6.2 Results

Figure 6.11 presents accuracies obtained by the various GoogLeNet and NIN models during evaluation on the ChestX-ray14 infiltration dataset. Unlike previous results, the highest values are produced under the 1x2 setting, with the GoogLeNet-1x2 and NIN-1x2 models achieving accuracies of 63.34% (95% CI: ±0.13) and 61.55% (95% CI: ±0.1) respectively. In comparison, the lowest values are obtained by the GoogLeNet-1x1 and NIN-3x2 models with accuracies of 62.51% (95% CI: ±0.17) and 60.23% (95% CI: ±0.64) respectively.

As a means to draw additional insight surrounding the compared models, Table 6.4 provides sensitivity, specificity, F1-score and AUC metrics obtained during evaluation.

Table 6.4: Performance results for the ChestX-ray14 Infiltration dataset.

Setting	**Sensitivity**	**Specificity**	**F1-score** (95%CI)	**AUC** (95%CI)
GoogLeNet-1x1	62.00	**63.12**	62.35 (±0.40)	67.60 (±0.59)
GoogLeNet-1x2	64.96	61.68	**63.22**(±0.47)	**68.29** (±0.53)
GoogLeNet-2x2	67.58	57.44	61.99 (±0.29)	67.09 (±0.50)
GoogLeNet-3x2	**69.07**	56.32	61.78 (±1.17)	67.54 (±0.55)
NIN-1x1	62.12	60.56	61.12 (±0.54)	65.85 (±0.43)
NIN-1x2	61.80	**61.29**	**61.46** (±0.23)	**66.05** (±0.34)
NIN-2x2	62.50	58.83	60.45 (±0.69)	64.74 (±0.83)
NIN-3x2	**62.84**	57.63	59.97 (±1.52)	64.17 (±2.14)

Note - All values are expressed as percentages. Bold indicates best results.
AUC = Area Under Curve

Here it can be seen that under the MIL setting, decreasing the patch scale results in an increase in sensitivity for both the GoogLeNet and NIN models, while at the same time decreasing the specificity possessed by these models. This trade-off in sensitivity and specificity indicates that by using a smaller patch size, the models can more confidently detect pathological instances within an image at the expense of being able to identify healthy ones. Further attention surrounding this phenomenon is given in Section 6.9.1.

In order to determine the operating characteristics for each model, Figure 6.12 presents ROC curves obtained during evaluation. Within this figure, the 1x2 patch setting is once again seen to produce the best characteristics for both network architectures, with the GoogLeNet-1x2 and NIN-1x2 models achieving sensitivities of 67% and 63% respectively at 60% specificity. At this same specificity level, a sensitivity of 65% is obtained for the GoogLeNet-1x1 model while the NIN-1x1 model achieves a sensitivity of approximately 63%.

It may be of concern at this point that the performances recorded for the ChestX-ray14 dataset appear to be substantially lower when considered against the results obtained on the previous datasets. However, when compared to the work of Wang et al. [14] through which the dataset was introduced, it can be seen that the performances achieved under the holistic and MIL settings for the infiltration class are in fact similar to those of the authors own GoogLeNet models. In explanation, Wang et al. cite the presence of large amounts of noise present within the dataset as the source for this poor performance.

Once again, as a final means to visually inspect the performance of the trained

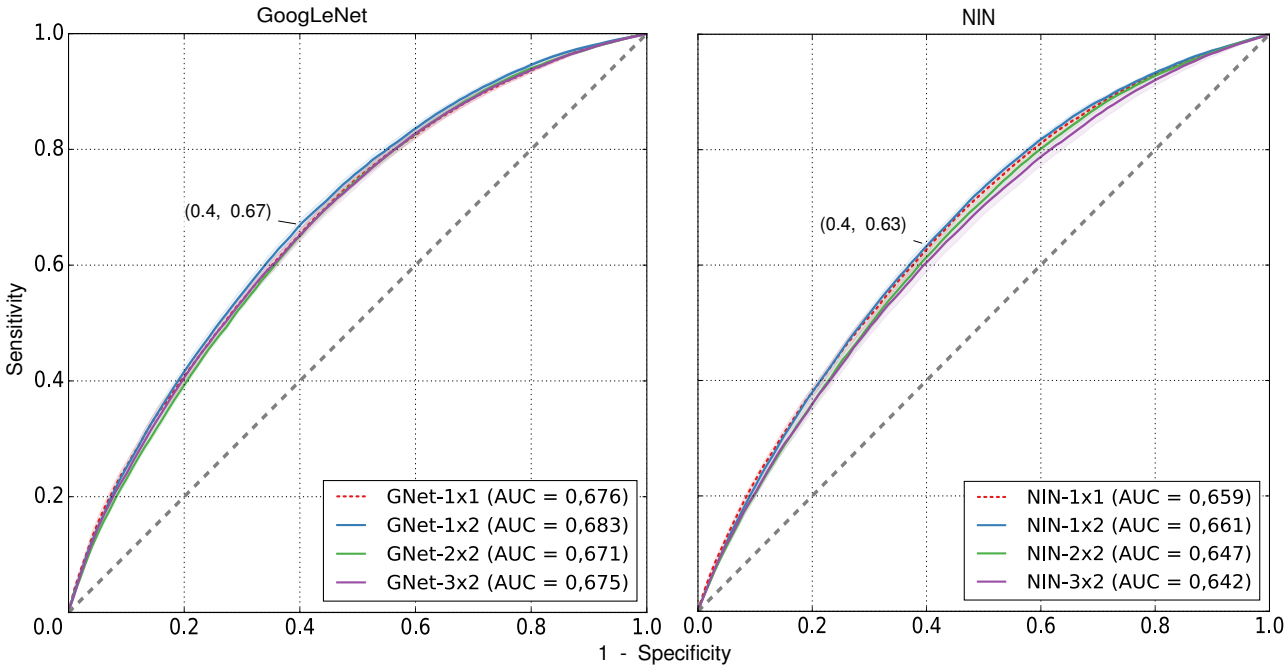


Figure 6.12: ROC curves generated for various models tested on the ChestX-ray14 infiltration dataset.

models, Figure 6.13 presents several false-positive (a - d), true-positive (e - h) and false-negative (i - l) classification results drawn from the ChestX-ray14 infiltration dataset during evaluation. Saliency maps are provided for the false-positive and true-positive examples in order to visualise each model's attention during classification, while markings are given in the false-negative examples as a means to indicate missed pathology. These findings are discussed below:

Figure 6.13 (a - b): Images (a) and (b) both present examples of a false-positive result arising from medical artefacts included within a CXR. Within (a), an inserted portacath over the right upper-lung is incorrectly detected as being pathogenic, while the same is true in image (b) for several drainage lines and surgical clips situated over the right lung area.

Figure 6.13 (c - d): Several false-positive findings are seen to arise for the GoogLeNet-1x2 model due to high amounts of anatomical or radiographic noise. In image (c), this radiographic noise is manifest by poor inspiration which is seen to mimic the presence of infiltrate within the right lower lung. Image (d) presents an example of anatomical noise causing misclassification wherein diffuse patterns around the right hilar region are incorrectly detected.

Figure 6.13 (e - f): The images within (e) and (f) highlight examples of infiltrate being correctly detected and classified by the GoogLeNet-1x1 model. Within image (e) denser infiltrate within the right lower and mid lung regions, indicating lobar

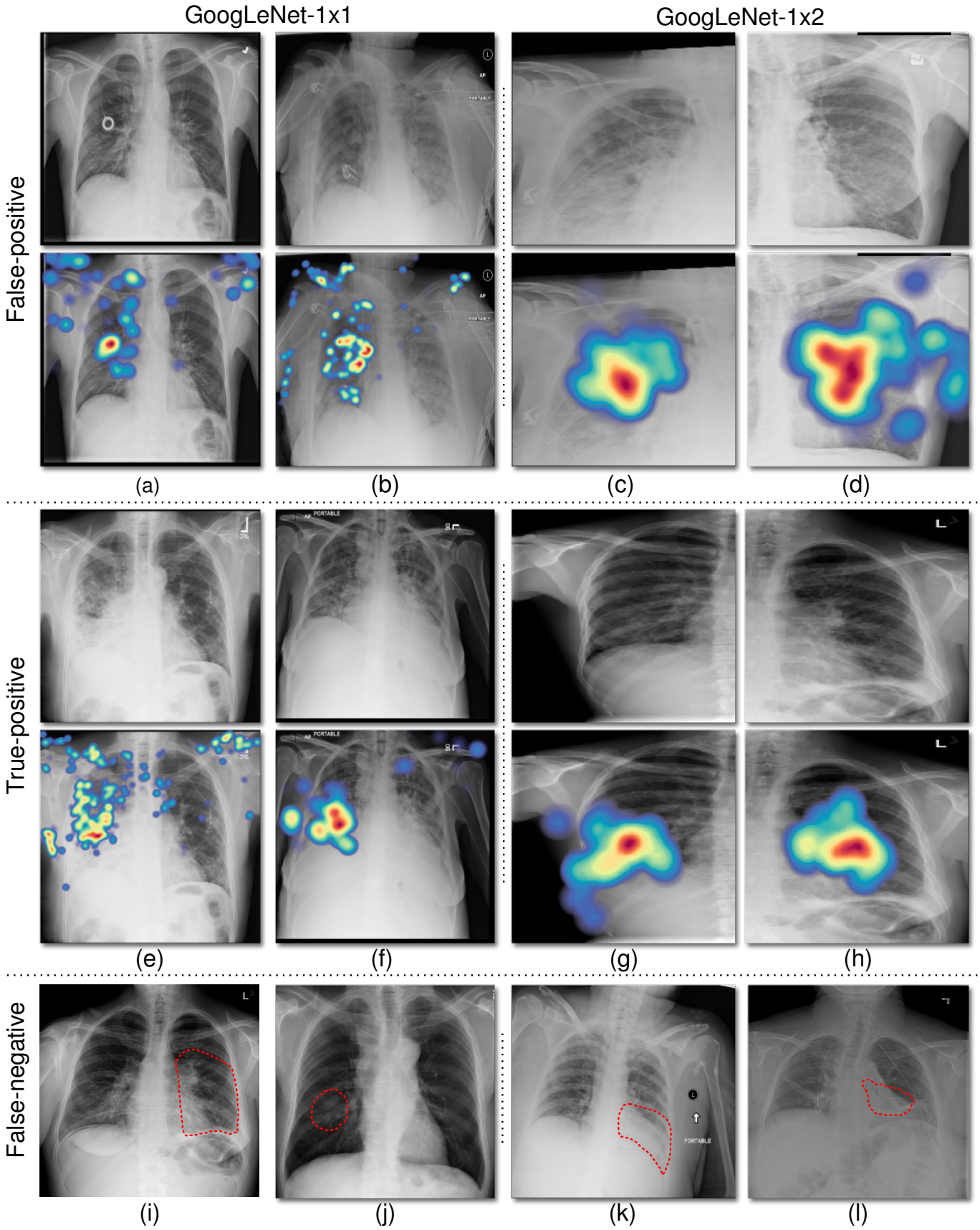
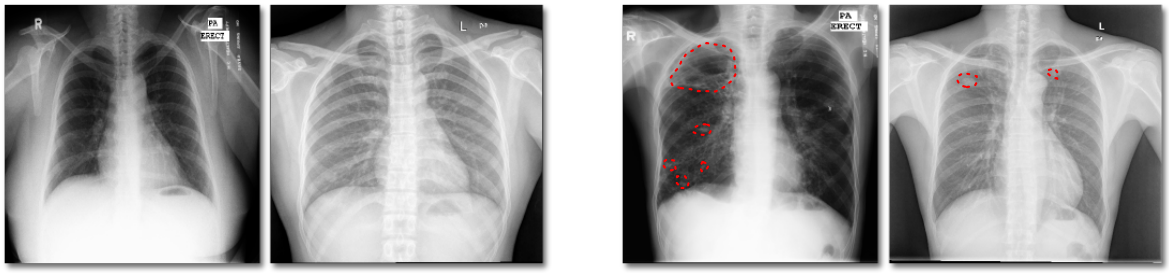


Figure 6.13: Saliency maps generated for various examples drawn from the ChestX-ray14 infiltration dataset using the GoogLeNet-1x1 and GoogLeNet-3x2 models.



(a) Normal

(b) Pulmonary Tuberculosis (PTB).

Figure 6.14: Example CXRs drawn from the Montgomery-Shenzhen TB dataset

consolidation, is given greatest attention. Image (f) shows a more patchy infiltrate within the right lower lung also being correctly detected.

Figure 6.13 (g - j): Examples (g) to (j) present subtle manifestations of infiltrate which are either correctly detected or missed by the 1x2 and 1x1 settings of the GoogLeNet models. For (g) and (h), this infiltrate appears diffusely within the right lower and left lower lobes of the lung respectively and is detected by the GoogLeNet-1x2 model. In contrast, the infiltrate within (h) is missed when the same CXR is processed by the GoogLeNet-1x1 model as indicated in image (i). A further false-negative result is seen under the holistic setting within (j) where several small nodular opacities within the left lower lobe avoid detection.

Figure 6.13 (k - l): Images (k) and (l) present findings missed under the GoogLeNet-1x2 setting. Within image (k) a dense opacity surrounding the left lingula and left base of the lung is seen to completely obfuscate the heart silhouette. While this abnormality is visibly apparent to a human reader, to the model the similarity of this infiltrate to surrounding tissue may cause it to be ignored. A less apparent case of infiltration is seen in (l), where an air bronchogram over the left ventricle, indicating consolidation, is missed.

6.7 Montgomery-Shenzhen Pulmonary Tuberculosis Dataset

6.7.1 Description

The Montgomery-Shenzhen dataset is a radiographic collection consisting of both normal and PTB frontal CXR findings which have been publicly released by the U.S National Library of Medicine in an attempt to advance the automated diagnosis of

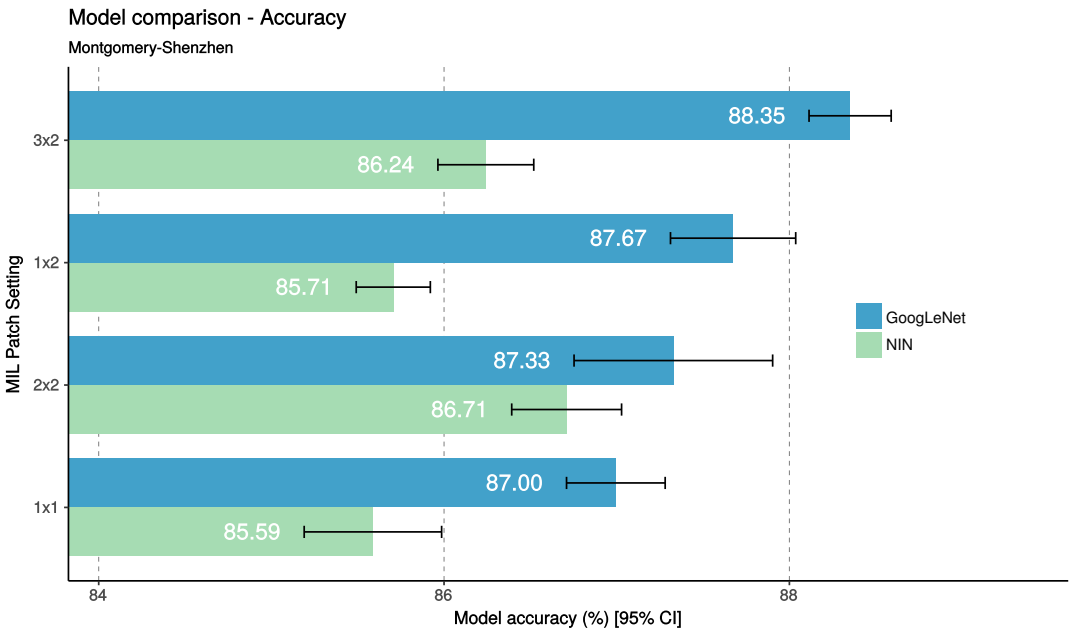


Figure 6.15: Model accuracies obtained for the Montgomery-Shenzhen dataset. Error bars indicate 95% confidence intervals.

chest pathologies [26]. Within the collection, data was obtained from two sites, namely the Montgomery County Tuberculosis Screening Program (USA) and the Shenzhen No.3 People's Hospital (China), with full anonymisation being performed for each case. For the Montgomery site, 80 normal and 58 PTB-positive cases (PA orientation) were captured using computed radiography and stored as 4020×4892 px or 4892×4020 px PNG (12-bit grey level) images. The Shenzhen site produced 326 normal and 336 PTB-positive cases (PA orientation) captured using digital radiography which were stored as 3001×3001 px PNG (8-bit grey level) images, with 21 of these cases being stored at a lower resolution of 948×1130 px and belonging to paediatric patients (AP orientation). Each radiograph within the dataset is accompanied by a summarised clinical report from an experienced radiologist, from which its image-level label is derived.

For evaluation, the combined total of 394 *PTB-positive* and 406 *normal* radiographs are formed into two distinct classes which are left unbalanced due to their similar sizes, with Figure 6.14a and 6.14b presenting examples of these classes respectively. Following pre-processing via data augmentation, the number of images within each fold's training, validation and testing sets are 7776, 156, and 156 respectively.

Table 6.5: Performance results for the Montgomery-Shenzhen dataset.

Setting	**Sensitivity**	**Specificity**	**F1-score** (95%CI)	**AUC** (95%CI)
GoogLeNet-1x1	**86.00**	87.53	86.49 (±0.20)	93.27 (±0.35)
GoogLeNet-1x2	81.36	92.84	86.61 (±0.23)	94.08 (±0.23)
GoogLeNet-2x2	81.60	93.00	86.87 (±0.62)	93.94 (±0.42)
GoogLeNet-3x2	83.43	**93.33**	**88.00**(±0.28)	**94.56**(±0.45)
NIN-1x1	**82.76**	88.44	85.39 (±0.37)	92.88 (±0.50)
NIN-1x2	81.56	89.83	85.33 (±0.27)	92.35 (±0.69)
NIN-2x2	82.00	**91.48**	**86.36**(±0.32)	**93.45** (±0.44)
NIN-3x2	81.92	90.56	85.87 (±0.31)	92.91 (±0.72)

Note - All values are expressed as percentages. Bold indicates best results.
AUC = Area Under Curve

6.7.2 Results

Figure 6.15 provides the holistic and MIL model accuracies obtained on the Montgomery-Shenzhen dataset. Unlike previous results obtained for the OpenI and PCV-13 datasets, each GoogLeNet model trained under the MIL setting is seen to produce higher accuracies compared to the holistic setting, with the highest accuracy of 88.35% (95% CI: ±0.24) achieved by the GoogLeNet-3x2 model. In comparison, the GoogLeNet-1x1 model achieves an accuracy of 87% (95% CI: ±0.29). A similar trend is also followed for the NIN models, with both the 2x2 and 3x2 models producing accuracies of 86.71% (95% CI: ±0.32) and 86.24% (95% CI: ±0.28) respectively, surpassing the NIN-1x1 accuracy of 85.59% (95% CI: ±0.40).

Table 6.5 presents sensitivity, specificity, F1-score and AUC metrics for the GoogLeNet and NIN models which further support the superior results obtained under the MIL setting. Within this table an apparent contradiction exists in which the holistic setting is seen to produce higher sensitivity levels in comparison to the MIL setting for each network - despite the fact that the latter setting is implemented to detect smaller local pathologies and thereby improve overall sensitivity. In explanation, it is noted that these sensitivity results are determined with a 0.5 probability estimate threshold for class prediction across all models. As such, they do not account for the lower confidences which can be tolerated under the MIL setting when each possible class prediction for a given input is similarly weighted.

In relation to the preceding discussion surrounding model sensitivities, Figure 6.16 provides ROC curves generated for the GoogLeNet and NIN models on the Montgomery-Shenzhen dataset. The best characteristic is achieved by the GoogLeNet-3x2 model,

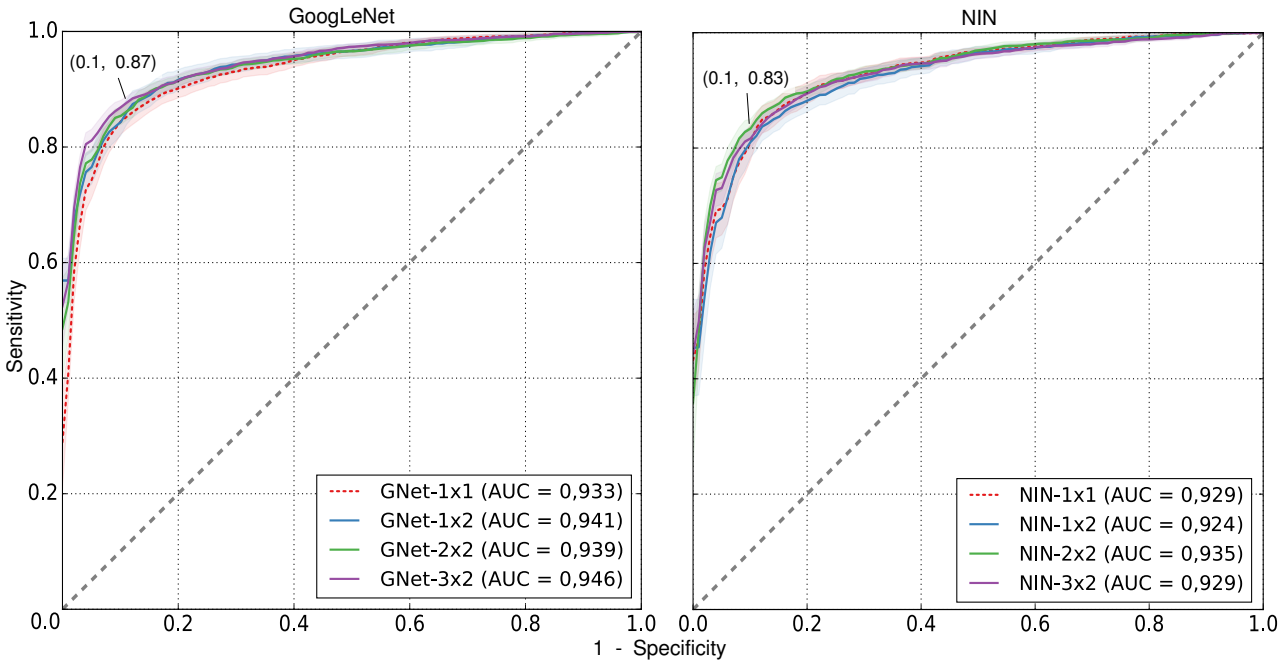


Figure 6.16: ROC curves generated for various models tested on the Montgomery-Shenzhen dataset. Shaded areas represent 95% confidence intervals for the respective models.

which possesses a sensitivity of 87% at 90% specificity. In comparison, the GoogLeNet-1x1 model achieves a sensitivity of 84% at this same specificity. This difference in performance is even more distinct at higher specificity levels, with the 3x2 model possessing a sensitivity of 81% versus the 74% produced by the 1x1 model when compared at a specificity of 95%. Similar results are seen for the NIN models, with the NIN-2x2 model possessing a sensitivity of 83% at 90% specificity, while the NIN-1x1 model achieves 81% sensitivity at 90% specificity.

In order to visually assess the efficacy of the trained models at classifying PTB, Figure 6.17 presents several false-positive (a - d) and true-positive (e - h) saliency maps, along with false-negative (i - l) examples drawn from the Montgomery-Shenzhen dataset during evaluation with the GoogLeNet-1x1 and GoogLeNet-3x2 models. These instances are discussed briefly below:

Figure 6.17 (a - b): In both image (a) and (b) the 1x1 model is seen to be overly sensitive to normal anatomical noise present within the lung fields. For image (a), incorrect attention is given to the right upper region of the lung which is clear, while in image (b) this attention is given to the left hilar region and left lung base. It is also noted within these images that attention is given outside of the lung fields, indicating misappropriate features which may have been learnt under the 1x1 setting.

Figure 6.17 (c - d): These examples demonstrate false-positive responses due

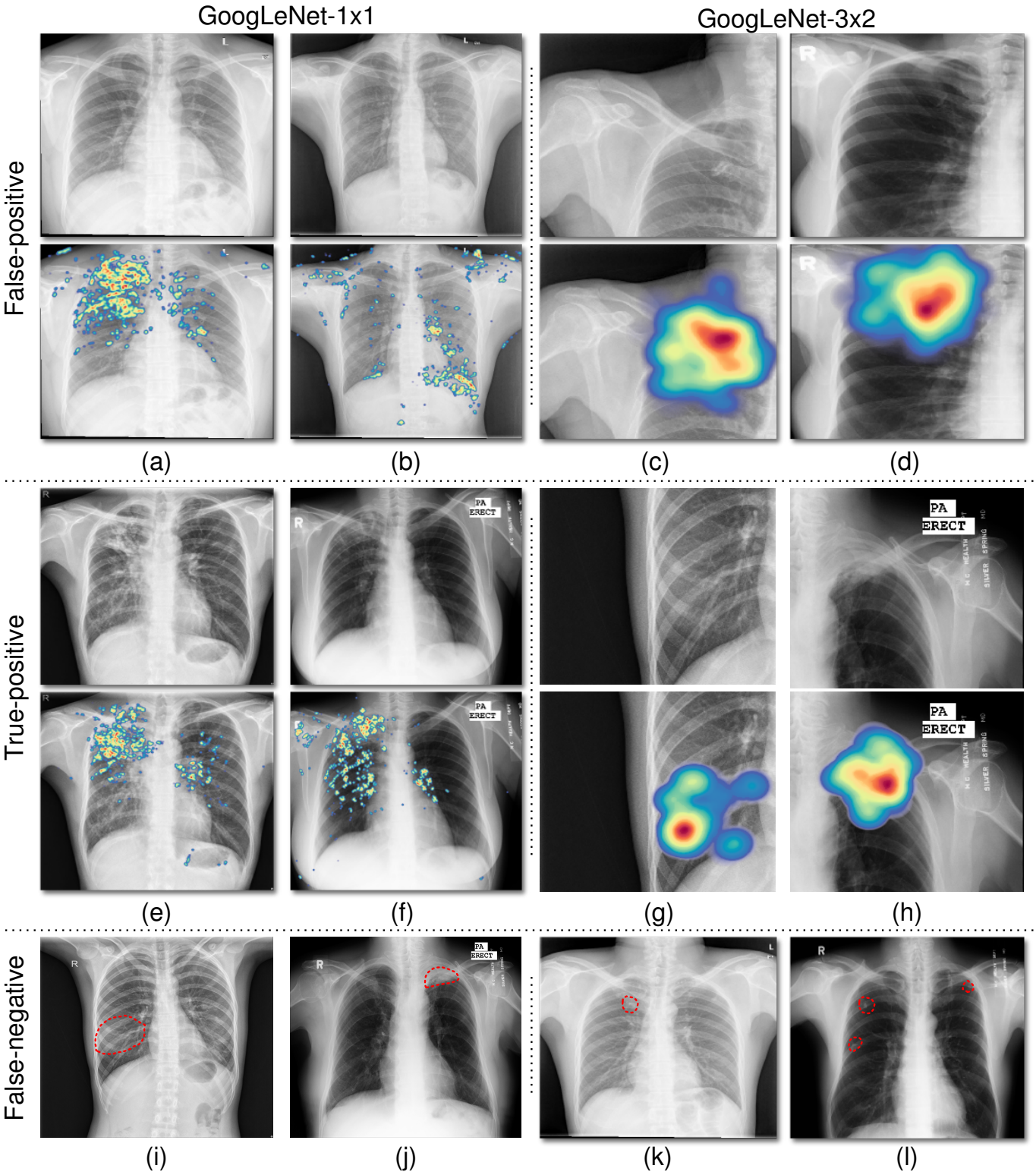


Figure 6.17: Saliency maps generated for various examples drawn from the Montgomery-Shenzhen PTB dataset using the GoogLeNet-1x1 and GoogLeNet-3x2 models.

to seemingly pathological features found within each CXR. In image (c), the 3x2 model is seen to incorrectly recognise a fracture within the third rib and surrounding anatomical noise as pathology for PTB. A similar error arises within image (d) wherein a subtle lucency in the right upper lung is falsely detected (possibly as cavitation).

Figure 6.17 (e - f): Images (e) and (f) present true-positive responses produced by the GoogLeNet-1x1 model. Within image (e), the model correctly detects the presence of bilateral PTB, with strong attention being produced surrounding cavitation within the right upper lung. A more subtle finding is shown in image (f) wherein fibronodular infiltrate within the right upper lung is correctly detected and classified.

Figure 6.17 (g - j): The images contained within (g) to (j) help illustrate the increased sensitivity produced by the MIL setting compared to holistic classification. In image (g), the 3x2 GoogLeNet model is correctly able to detect subtle patterns of cavitation within the right mid-lung which are indicative of a secondary PTB infection. In contrast, the GoogLeNet-1x1 model misses these patterns and produces a false-negative prediction for this same CXR as illustrated in image (i). This same scenario is seen to occur for images (h) and (j) as well, wherein localised infiltrate within the left upper-lung which missed under the 1x1 setting in image (j) is seen to be accurately detected by the GoogLeNet-3x2 model in image (h).

Figure 6.17 (k - l): The false-negative findings presented in images (k) and (l) are characterised by subtle pathological features for PTB. Within image (k), the GoogLeNet-3x2 models misses several small granulomas which are bilaterally present within the lung apices. In a similar manner image (l) depicts small calcified granulomas indicative of inactive PTB which are also missed.

6.8 Implementation Considerations

Within section 3.3.4, a discussion was given surrounding the additional memory constraints of using the MIL method as opposed to holistic classification. As a means to further consolidate this discussion, Table 6.6 presents the average training time for each GoogLeNet model setting evaluated on the Montgomery-Shenzhen and ChestX-ray14 datasets. Here a general increase in the training time for each model setting is seen to occur as the number of patches used under the MIL method increases. For the Montgomery-Shenzhen dataset, which represents the smallest collection of data, the shortest average training time of 21 minutes is possessed by the holistic setting, while the longest average training time of 80 minutes occurs under the 3x2 MIL setting. A similar pattern is observed for the ChestX-ray14 dataset, the largest

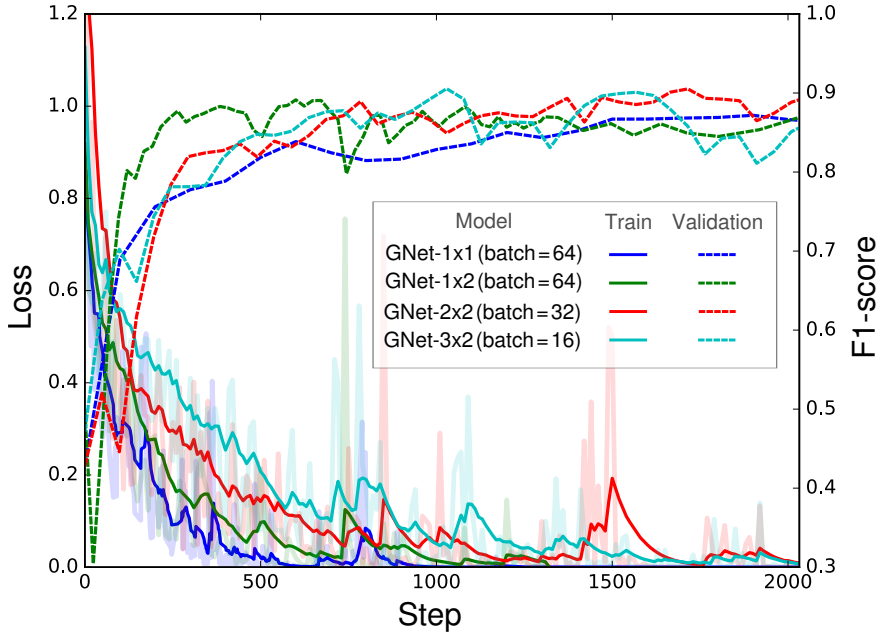


Figure 6.18: Sample curves showing model loss and validation accuracy for GoogLeNet models during stage-1 training on the Montgomery-Shenzhen dataset. Here the holistic setting (GNet-1x1) can be seen to lead to faster training convergence as opposed to the MIL settings.

Table 6.6: Average training times for the GoogLeNet models on the Montgomery-Shenzhen and ChestX-ray14 datasets.

Dataset	**Setting**	**Batch Size**	**Avg train time** (min.)			**Increase**
			Stage		**Total**	
			1	**2**		
Montgomery-Shenzhen	1x1	64	21	-	21	1×
	1x2	64	23	7	30	1.4×
	2x2	32	37	19	56	2.7×
	3x2	16	54	26	80	3.8×
ChestX-ray14	1x1	64	56	-	56	1×
	1x2	64	59	26	85	1.5×
	2x2	32	113	51	164	2.9×
	3x2	16	167	72	239	4.3×

Table 6.7: Overall performance of the holistic and MIL classification methods.

Setting	Accuracy	F1-score	AUC
GoogLeNet-1x1 (Holistic)	**76.76**	**76.58**	83.46
GoogLeNet-1x2	76.71	75.39	**83.62**
GoogLeNet-2x2	75.84	75.35	82.47
GoogLeNet-3x2	76.3	75.74	83.2
NIN-1x1 (Holistic)	**75.22**	**74.85**	**82.06**
NIN-1x2	74.42	74.09	80.63
NIN-2x2	74.42	74.06	80.98
NIN-3x2	74.76	74.41	81.36

Note - All values are expressed as percentages. Bold indicates best result.

collection of data, wherein the 3x2 setting is seen to take 4.3 times longer to train compared to the holistic setting.

Three main factors are seen to contribute to the increased training times recorded. The first factor is the reduction in batch sizes associated with greater numbers of patch-samples, with an increased number of steps being required for both learning and validation during training. The second factor is the slower convergence rates achieved under the MIL setting for training, which can be observed within the sample stage-1 training curves provided in Figure 6.18. The third factor is the requirement of an additional training stage for model boosting.

With regard to the computational time of the MIL method for inference, during evaluation the time taken by each model to classify an input image is seen to almost increase in a linear relation to the number of patches used by its setting. Here the average inference time achieved by the holistic setting is $2.78\,ms$ per image, while the 1x2, 2x2 and 3x2 settings achieve times of $6.04\,ms$, $11.70\,ms$ and $17.73\,ms$ respectively. These findings indicate that once trained, a model operating under the MIL setting could be used for automated reading in the same manner that one implementing holistic classification could, without any delays perceptible to a human operator being incurred.

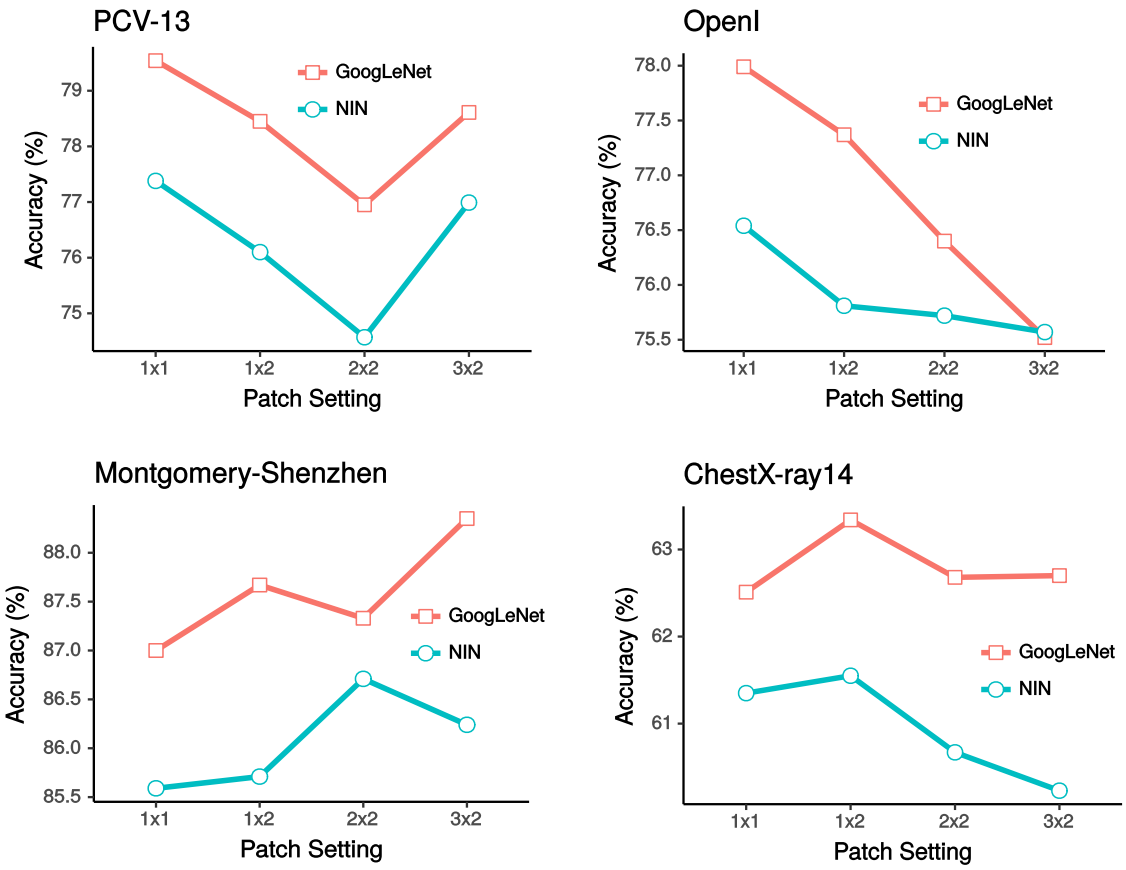


Figure 6.19: Overall model performance on each of the evaluated datasets.

6.9 Discussion

6.9.1 Holistic versus Multi-Instance Learning

Table 6.7 presents the overall performance results of the holistic and MIL classification methods. For the GoogLeNet models, the holistic setting is seen to obtain the highest accuracy and F1-score metrics of 76.76% and 76.58% respectively, while the GoogLeNet-1x2 setting achieves the highest AUC of 83.62%. In comparison, the holistic setting is seen to produce the highest result for each metric surrounding the NIN models, obtaining 75.22%, 74.85% and 82.06% for accuracy, F1-score and AUC respectively. It is important to note, however, that for either network architecture these superior results for the holistic setting are not present uniformly across each evaluated dataset. In this sense, both the MIL and holistic classification methods are seen to excel relative to one another on different sets of data.

Figure 6.19 helps illustrate the relative performance between the classification methods by providing a comparative view of the model accuracies obtained for each of the evaluated datasets. Within the figure it can be seen that for the Indiana OpenI and PCV-13 datasets, the holistic classification method produces the best performance,

while for the ChestX-ray14 Infiltration and Montgomery-Shenzhen datasets the best results are obtained by the MIL method. In consideration of these findings it is observed that for the first three datasets, whose pathological findings relate to pneumonia and are spatially larger than those of pulmonary tuberculosis, patch sizes with larger spatial contexts (1x2 setting) tend to produce better results. Conversely, for the smaller pathological findings of the Montgomery-Shenzhen dataset, smaller sized patches which allow for local attention (2x2 and 3x2 settings) are seen to increase model performance.

The experimental results given in Chapter 5 provide understanding as to why this latter finding occurs, with the MIL method improving the detection of local discriminative information and thereby increasing model sensitivity to small and subtle pathological findings. However, they do not provide an explanation for the deterioration in performance seen when using smaller sized patches for spatially larger pathologies. Here the recent work of Guan et al. [125] becomes relevant wherein the authors discovered that cropping sub-regions of pneumonia within CXRs reduced the performance of their attention-guided DCNN model, compared to the use of image crops which encompassed the pneumonia's entire pathological area. The authors subsequently reasoned that a loss of information and context caused by the smaller crops were responsible for the drop in performance. While it is noted that these findings were obtained on a separate dataset using a different DCNN model architecture, they support the notion that for spatially larger pathological findings the use of smaller sized patches with the MIL method may lead to context loss and thereby are less appropriate.

Finally, even in the light of the above discussion, it is noted from examples provided within Figures 6.9 and 6.13 that the MIL method enables a model to detect smaller localised findings within pathologies of a generally larger spatial extent. It is these findings which are perhaps the most important to detect, as larger manifestations of pathology can be easily recognised by human readers. As such, the use of both holistic and MIL classification methods are seen as being valuable in the automated reading of frontal CXRs.

6.9.2 Comparison to Human Performance

In the same manner as the works of [29] and [61], Figure 6.20 presents a comparison of the MIL and holistic model settings against the performance of human readers reported within the literature. Table 6.8 additionally provides details surrounding the sample size, reader experience level, sensitivity and specificity associated with each of these compared studies. While it is noted that these sources do not allow for

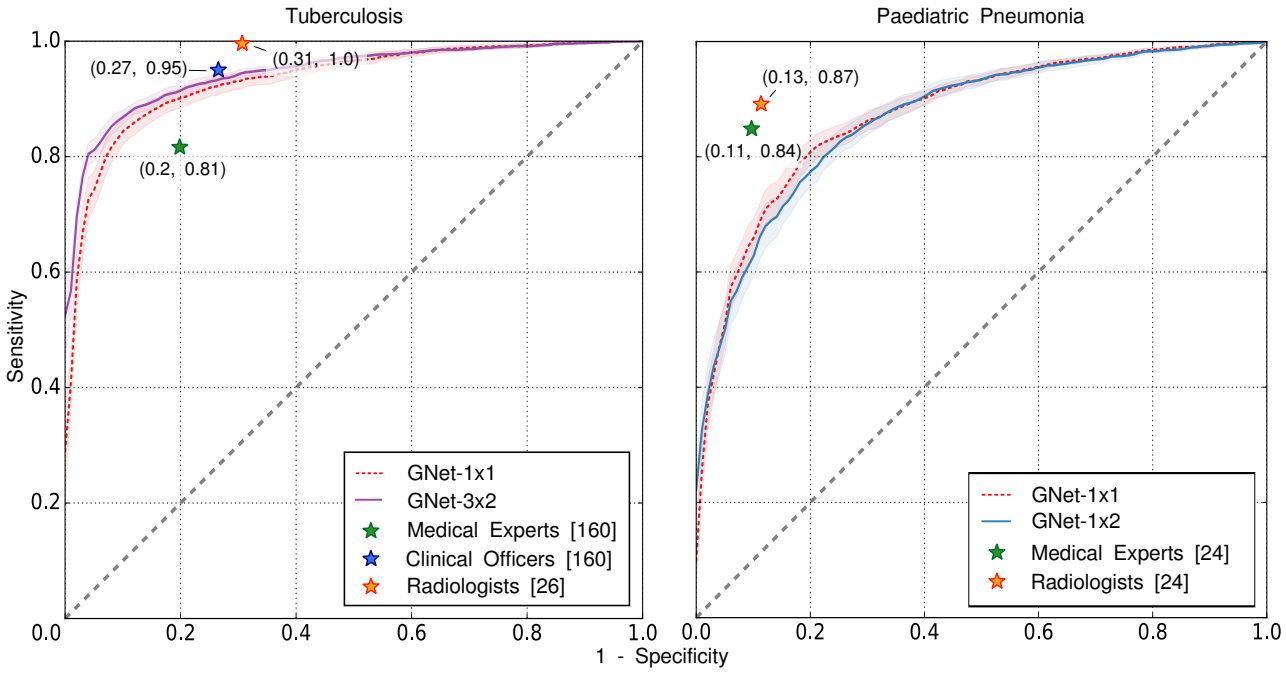


Figure 6.20: ROC curves for the MIL and holistic classification models compared against the performance of human readers. (Source: Personal collection).

direct comparison, with the absence of common datasets upon which performance can be gauged being a key limitation, they do provide an intuition of the operating levels currently upheld within the medical field which need to be met by automated diagnosis techniques.

For the reading of CXRs with PTB, it can be seen that the performance of the trained models versus human readers is very compelling. Here both the MIL and holistic setting are seen to greatly exceed the sensitivity and specificity of an individual chest physician as described in [160], while the 3x2 MIL model setting possesses only 1% less sensitivity (94%) than a panel of three clinical officers reading with a specificity of 73%. This second finding is an important result as the work of Vant Hoogt et al. [160] sought to establish the efficacy of training these and other clinical officers to perform reading in areas in which radiological experts were absent. Thus the close performance between these officers and the presented models highlights how the latter automated approach could provide an alternative means of CXR reading in areas with limited medical expertise. While a lack of common evaluation data once again limits the veracity of this previous statement, comparison of the 3x2 setting to a panel of two radiologists evaluated on the Montgomery PTB subset [29] sees the model only achieve 5% less sensitivity at the same specificity level.

Similarly compelling results are seen for the reading of CXRs with manifestations of paediatric pneumonia, with the GoogLeNet-1x1 model achieving the same sensitivity

Table 6.8: Comparison of CXR reading performance for human readers.

Pathology	**Data source**	**Sample size** (n)	**Reader** (Quantity)	**Sen**	**Spc**
TB	Vant Hoogt et al. 2011 [160]	1143	Chest physician (×1)	0.81	0.8
			Clinical officers (×3)	0.95	0.73
	Jaeger et al. 2014 [26]	138	Radiologists (×2)	1.0	0.87
Paediatric pneumonia	WHO*. 2005 [24]	208	Medical officer (×1)	0.84	0.89
			Radiologist (×1)	0.87	0.87

* World Health Organisation Chest Radiology Working Group

Sen = Sensitivity. Spc = Specificity

at only 14% less specificity compared to a medical officer extensively trained in the WHO standardised CXR interpretation methodology. For a radiologist with this same training, the holistic model possess 17% less specificity at the same sensitivity level of 87%. While these results may appear less impressive compared to those achieved for PTB, they are still appreciable for a disease which is often associated with high levels of inter-observer variability [161]. These findings do highlight the need for more work within this field, however, with no studies currently recorded which directly compare the performance of human readers and CAD technologies under controlled settings for this important pathology.

6.10 Summary

The holistic and MIL classification methods are evaluated across four CXR datasets which are representative of pathological findings for pneumonia and PTB.

Overall, holistic classification produces superior performances for the GoogLeNet and NIN models, obtaining average accuracies of 76.76% and 75.22% respectively. However, these results are not consistent across each dataset, with the MIL method outperforming holistic classification for the detection of PTB. An inspection of saliency maps produced during model evaluation also sees the MIL method enabling the correct detection of subtle findings which are missed under the holistic setting for both the PCV-13 and ChestX-ray14 datasets. The consideration of these combined findings seems to indicate that the MIL method is more appropriate for detecting smaller localised pathology signs, while the use of holistic classification is more appropriate for pathologies of a larger spatial extent.

Experimentation further shows that the MIL method requires additional computation time compared holistic classification for both model training and inference. Here the use of the 3x2 patch setting increases the training time of the GoogLeNet model by a factor of 4.3 compared to the 1x1 setting for the ChestX-ray14 dataset. A similar increase is also observed for the Montgomery-Shenzhen dataset in which the training time is increased by a factor of 3.8 using the 3x2 patch setting. Inference sees a similar increase in computation time, with the 3x2 patch setting taking an average of 6.4 times longer than holistic classification to process an image. The effect of this increase is negligible to a human operator, however, with the 3x2 patch setting requiring only 17.73 ms for this inference.

Lastly, while a direct comparison cannot be made, the obtained classification results are compelling when considered against those of human readers within the literature. Here, under the MIL setting, the GoogLeNet-3x2 model performs with only 1% less sensitivity than a group of clinical officers trained for PTB detection. The GoogLeNet-1x1 model also operates at the same sensitivity and only 17% less specificity compared to an expert radiologist performing paediatric pneumonia screening.

Chapter 7

Conclusion

7.1 Overview

In the face of an ever increasing patient burden and limited medical expertise, countries of the developing world require effective strategies to cope against the high mortality rates associated with infectious lung diseases such as pulmonary TB and pneumonia. A favourable screening strategy in this regard is the reading of a patient's chest X-ray (CXR), which presents as a cheaper, faster and diagnostically more robust approach compared to alternative screening methods. Reading a CXR is difficult, however, with less experienced medical workers showing lower levels of sensitivity and higher levels of inter-reader disagreement compared to radiological experts.

Recent advances in computer-aided diagnosis (CAD) provide a digital solution to this need for expert reading. Here automated CXR reading has yielded competitive and affordable results when tested within clinical settings. To create forms of this technology which obtain high levels of performance and generalise well, current approaches to CAD utilise a branch of machine learning known as deep learning. Particular use is made of a model architecture known as a deep convolutional neural network (DCNN). Unfortunately DCNN models require large quantities of data in order to be trained, which poses a limitation within the medical imaging domain where the cost and effort required to prepare large image datasets are prohibitive.

To enable the creation of larger medical image datasets for CXR screening, the process of providing image-level labels is becoming more common. As a consequence of this labelling strategy, the training of a DCNN model is required to occur under a weakly-supervised setting - typically holistic image classification - where the learning signal to the trained model is diminished. For chest pathologies which are subtle and localised, the training signal is further attenuated and limits performance under the holistic image classification setting [124, 13].

This dissertation addresses the challenge of detecting and classifying subtle pathology within image-level labelled CXRs. In this regard, it proposes a multi-instance learning (MIL) method which enables local discriminative information to be used during weakly-supervised image classification. The findings surrounding this method and its comparison to holistic image classification are presented below.

7.2 Summary of Findings

The MIL method was formulated in Chapter 3, where two training stages were motivated to enable the detection of local discriminative information and to boost classification performance respectively. These stages are formulated in a manner that enables end-to-end learning to occur.

Using various experimental settings along with the GoogLeNet and NIN network architectures outlined in Chapter 4, the proposed MIL method was experimentally validated and compared to holistic image classification in Chapter 5. Here testing on two synthetic CXR image datasets containing small local objects was performed. The results saw the MIL method outperform holistic classification significantly, with the use of smaller patch regions for both networks leading to improved performance.

Chapter 6 extended the evaluation of the MIL and holistic classification methods onto four real-world medical datasets for CXR screening. Using five runs of 5-fold cross validation, the holistic method was seen on average to produce superior performances for pathologies with larger spacial extents, such as paediatric pneumonia and airspace opacity. On the other hand, the MIL method achieved the best performance for the smaller pathological signs associated with pulmonary TB. It was further observed that for two of the datasets associated with pneumonia that the MIL method was able to detect subtle pathological cases which were missed under the holistic setting.

During the evaluations in Chapter 6 the training and inference times for various model settings were also recorded. These values showed an increase in training time corresponding to an increasing number of patches used under the MIL classification setting. A similar increase was observed during inference, however, its effect was negligible owing to the rapid image classification enabled by DCNN models.

Finally, when compared against the performance of human readers recorded in the literature, the results obtained for both the MIL and holistic image classification were seen to be compelling.

7.3 Research Conclusions

Having reviewed the findings which have been obtained within this work, a discussion surrounding their meaning is put forward.

Initially, in consideration of the results produced for both the artificial and real-world datasets, it is concluded that the multi-instance learning method does increase the performance of a DCNN model screening for pathologies *which are subtle and localised*, relative to holistic classification. However, it is further seen that for pathologies of a larger spacial context, the use of only the local information obtained through the MIL method can also reduce overall classification performance.

It is also noted that the improved performance which can be enabled by the MIL method requires the trade-off of additional computation time or memory during model training. Once training is completed, however, the difference in computational time between the holistic and MIL methods during inference becomes negligible, with no additional memory requirements being imposed. This is seen to enable the MIL method to be deployed in the same manner as the holistic method while increasing performance. Thus the MIL method provides a feasible solution for improving automated CXR reading systems currently based on holistic classification techniques.

Lastly, from the comparable performance between the evaluated methods and human readers, it can be seen that these approaches form part of a technology which can solve an imposing problem faced by the developing world. Heeding the warning by Foster et al. [162], however, it is acknowledged that further studies across multiple additional datasets and clinical trials are required to establish the generalisability of the results presented within this work.

7.4 Suggestions for Future Work

Given the nature of the findings within this dissertation, future work should consider the combined use of both local and global scales of context for the classification of pathologies within weakly-labelled CXRs. This could be achieved through the formation of a fused DCNN architecture, similar to the work of [163], in which the branches of the network process image patches of varying scales in a multi-instance manner. Alternatively, the recurrent attention model proposed by Mnih et al. [164] can be adapted to adjust the size of the glimpses it takes when scanning an input CXR.

The suitability of the MIL method for detecting local discriminative information also merits evaluation on other chest pathologies which can contain findings of a small or

subtle nature. These pathologies may include lung nodules, pulmonary masses, pleural thickening or atelectasis. Furthermore, in the context of high resolution medical imaging in which an image cannot adequately be processed in a single pass by a DCNN, the method may offer a compelling form of classification if no region-of-interest labelling is present.

The MIL method was also only developed to facilitate binary classification between abnormal and normal chest X-rays. With the emergence of larger CXR datasets such as ChestX-ray14 in which multiple pathologies can be present within a single image, the extension of this method to facilitate multi-class classification is motivated. Due to the similarities in loss functions, classification layers and image labelling between the binary and multi-class classification settings, this extension is not foreseen to be difficult.

References

[1] WHO. *Global Tuberculosis Report 2016*. 2016.

[2] Klaus Schwab, Xavier Sala-Martin, Borge Brende, et al. *The Global Competitiveness Report*. Economic World Forum, 2014.

[3] Christa L. Fischer Walker, Igor Rudan, Li Liu, et al. Global burden of childhood pneumonia and diarrhoea. *The Lancet*, 381(9875):1405–1416, 2013.

[4] Holger R. Roth, Le Lu, Jiamin Liu, et al. Improving Computer-Aided Detection Using Convolutional Neural Networks and Random View Aggregation. *IEEE Transactions on Medical Imaging*, 35(5):1170–1181, 2016.

[5] Mingchen Gao, Ulas Bagci, Le Lu, et al. Holistic Classification of CT Attenuation Patterns for Interstitial Lung Diseases via Deep Convolutional Neural Networks. *Workshop on Deep Learning in Medical Image Analysis (MICCAI)*, (Cidi), 2015.

[6] Francesco Ciompi, Bartjan de Hoop, Sarah J. van Riel, et al. Automatic classification of pulmonary peri-fissural nodules in computed tomography using an ensemble of 2D views and a convolutional neural network out-of-the-box. *Medical Image Analysis*, 26(1):195–202, 2015.

[7] Jia Deng, Wei Dong, Richard Socher, et al. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2–9, 2009.

[8] Alex Krizhevsky, Ilya Sutskever, and Hinton Geoffrey E. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, pages 1–9, 2012.

[9] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems*, volume 27, pages 1–9, 2014.

[10] Hoo Chang Shin, Holger R. Roth, Mingchen Gao, et al. Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning. *IEEE Transactions on Medical Imaging*, 35(5):1285–1298, 2016.

[11] Nima Tajbakhsh, Jae Y. Shin, Suryakanth R. Gurudu, et al. Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning? *IEEE Transactions on Medical Imaging*, 35(5):1299–1312, 2016.

[12] Paras Lakhani and Baskaran Sundaram. Deep Learning at Chest Radiography: Automated Classification of Pulmonary Tuberculosis by Using Convolutional Neural Networks. *Radiology*, 284(2):574–582, 2017.

[13] Mark Cicero, Alexander Bilbily, Errol Colak, et al. Training and Validating a Deep Convolutional Neural Network for Computer-Aided Detection and Classification of Abnormalities on Frontal Chest Radiographs. *Investigative Radiology*, 52(5):281–287, 2017.

[14] Xiaosong Wang, Yifan Peng, Le Lu, et al. ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3462–3471. IEEE, 2017.

[15] Zhennan Yan, Yiqiang Zhan, Zhigang Peng, et al. Multi-Instance Deep Learning: Discover Discriminative Local Anatomies for Bodypart Recognition. *IEEE Transactions on Medical Imaging*, 35(5):1332–1343, 2016.

[16] Jonathan Gerrand, Quentin Williams, Dalton Lunga, et al. Paediatric Frontal Chest Radiograph Screening with Fine-Tuned Convolutional Neural Networks. In *Medical Image Understanding and Analysis (MIUA)*, volume 2, pages 850–861. Springer, 2017.

[17] Richard Mould. *A century of x-rays and radioactivity in medicine: with emphasis on photographic records of the early years*. CRC Press, 1993.

[18] Richard Daffner and Matthew Hartman. *Clinical Radiology, the Essentials*. Lippincott Williams & Wilkins, 4th edition, 2013.

[19] Cornelia Schaefer-Prokop, Ulrich Neitzel, Henk W. Venema, Martin Uffmann, and Mathias Prokop. Digital chest radiography: An update on modern technology, dose containment and control of image quality. *European Radiology*, 18(9):1818–1830, 2008.

[20] Matthew S. Kelly, Eric J. Crotty, Mantosh S. Rattan, et al. Chest radiographic findings and outcomes of pneumonia among children in Botswana. *Pediatric Infectious Disease Journal*, 35(3):257–262, 2016.

[21] Thure Holm. Technical Specifications for the World Health Imaing System for Radiology - The WHIS-RAD. Technical report, 1995.

[22] H. Melbye and K. Dale. Interobserver variability in the radiographic diagnosis of adult outpatient pneumonia. *Acta Radiologica*, 33(1):79–81, 1992.

[23] Gabrielle J. Williams, Petra Macaskill, Marianne Kerr, et al. Variability and accuracy in interpretation of consolidation on chest radiography for diagnosing pneumonia in children under 5 years of age. *Pediatric Pulmonology*, 48(12):1195–1200, 2013.

[24] Thomas Cherian, Eric Mulholland, JB Carlin, et al. Standardized interpretation of pediatric chest radiographs for the diagnosis of pneumonia in epidemiological studies. *Bulletin of the World Health Organisation*, 83(5):353–349, 2005.

[25] Y Balabanova, R Coker, I Fedorin, et al. Variability in interpretation of chest radiographs among Russian clinicians and implications for screening programmes: observational study. *BMJ (Clinical research ed.)*, 331(7513):379–82, 2005.

[26] Stefan Jaeger, Sema Candemir, Sameer Antani, et al. Two public chest X-ray datasets for computer-aided screening of pulmonary diseases. *Quantitative imaging in medicine and surgery*, 4(6):475–477, 2014.

[27] UNICEF. *Committing to Child Survival - A Promise Renewed (2015 Progress Report)*. 2015.

[28] WHO. *The global plan to stop TB, 2011-2015*. 2010.

[29] Stefan Jaeger, Alexandros Karargyris, Sema Candemir, et al. Automatic tuberculosis screening using chest radiographs. *IEEE Transactions on Medical Imaging*, 33(2):233–245, 2014.

[30] Prabha Desikan. Sputum smear microscopy in tuberculosis: Is it still relevant? *Indian Journal of Medical Research*, 137(3):442–444, 2013.

[31] R. H.H.M. Philipsen, C I Sánchez, P Maduskar, et al. Automated chest-radiography as a triage for Xpert testing in resource-constrained settings: A prospective study of diagnostic accuracy and costs. *Scientific Reports*, 5, 2015.

[32] T J John, T Cherian, M C Steinhoff, E A Simoes, and M John. Etiology of acute respiratory infections in children in tropical southern India. *Reviews of Infectious Diseases*, 13(0162-0886 (Print)):463–469, 1991.

[33] David R. Murdoch, Susan C. Morpeth, Laura L. Hammitt, et al. The Diagnostic Utility of Induced Sputum Microscopy and Culture in Childhood Pneumonia. *Clinical Infectious Diseases*, 64(Supplement 3):S280–S288, 2017.

[34] Saskia F. Van Vugt, Theo J M Verheij, Pim A. De Jong, et al. Diagnosing pneumonia in patients with acute cough: Clinical judgment compared to chest radiography. *European Respiratory Journal*, 42(4):1076–1082, 2013.

[35] Nicholas Fancourt, Maria Deloria Knoll, Breanna Barger-Kamate, et al. Standardized Interpretation of Chest Radiographs in Cases of Pediatric Pneumonia From the PERCH Study. *Clinical Infectious Diseases*, 64(supplement_3):S253–S261, 2017.

[36] Nasreen Mahomed, Nicholas Fancourt, John de Campo, et al. Preliminary report from the World Health Organisation Chest Radiography in Epidemiological Studies project. *Pediatric Radiology*, 47(11):1399–1404, 2017.

[37] Sarita Shah, H Anh, T Thuy, et al. Population-based chest X-ray screening for pulmonary tuberculosis in people living with HIV/AIDS, An Giang, Vietnam. *International Journal of Tuberculosis and Lung Disease*, 12(4):404–410, 2008.

[38] Michael JA Reid and N. Sarita Shah. Approaches to tuberculosis screening and diagnosis in people with HIV in resource-limited settings. *The Lancet Infectious Diseases*, 9(3):173–184, 2009.

[39] Kyle Nelson, Christopher Morrow, Sarah Wingerter, Richard Bachur, and Mark Neuman. Impact of Chest Radiography on Antibiotic Treatment for Children with Suspected Pneumonia. *Pediatric Emergency Care*, 32(8):514–519, 2016.

[40] A Patel, M Mamtani, P Hibberd, et al. Value of chest radiography in predicting treatment response in children aged 3-59 months with severe pneumonia. *The international Journal of Tuberculosis and Lung Disease*, 12(11):1320–6, 2008.

[41] Bram van Ginneken. Fifty years of computer analysis in chest imaging: rule-based, machine learning, deep learning. *Radiological Physics and Technology*, 10(1):23–32, 2017.

[42] Kunio Doi. Computer-aided diagnosis in medical imaging: Historical review, current status and future potential. *Computerized Medical Imaging and Graphics*, 31(4-5):198–211, 2007.

[43] Gwilym Lodwick, Theodore Keats, and John Dorst. The Coding of Roentgen Images for Computer Analysis as Applied to Lung Cancer. *Radiology*, 81(2):185–200, 1963.

[44] Jun-Ichiro Toriwaki, Yasuhito Suenaga, Toshio Negoro, and Teruo Fukumura. Pattern recognition of chest X-ray images. *Computer Graphics and Image Processing*, 2(3-4):252–271, 1973.

[45] Bram van Ginneken, Bart ter Haar Romeny, and Max Viergever. Computer-aided diagnosis in chest radiography: A Survey. *IEEE Transactions on Medical Imaging*, 20(12):212–223, 2001.

[46] Pragnya Maduskar, Rick Philipsen, Jaime Melendez, et al. Automatic detection of pleural effusion in chest radiographs. *Medical Image Analysis*, 28:22–32, 2016.

[47] Tao Xu, Irene Cheng, Richard Long, and Mrinal Mandal. Novel coarse-to-fine dual scale technique for tuberculosis cavity detection in chest radiographs. *EURASIP Journal on Image and Video Processing*, 2013(1):3, 2013.

[48] Wei Yang, Zhentai Lu, Mei Yu, et al. Content-based retrieval of Focal Liver Lesions Using Bag-of-Visual-Words Representations of Single- and Multiphase Contrast-Enhanced CT images. *Journal of Digital Imaging*, 25(6):708–719, 2012.

[49] Juan Carrillo-de Gea and Gines Garca-Mateos. Detection of Normality/Pathology on Chest Radiographs using LBP. In *Bioinformatics*, pages 167–172, 2010.

[50] David G Lowe. Distinctive image features from scale invariant keypoints. *Int Journal of Computer Vision*, 60(2):91–110, 2004.

[51] Qing Li, Weidong Cai, Xiaogang Wang, et al. Medical image classification with convolutional neural network. In *13th International Conference on Control Automation Robotics & Vision (ICARCV)*, pages 844–848. IEEE, 2014.

[52] Phillip H. Meyers, Charles M. Nice, Hal C. Becker, et al. Automated Computer Analysis of Radiographic Images. *Radiology*, 83(6):1029–1034, 1964.

[53] Uri Avni, Hayit Greenspan, and Jacob Goldberger. X-ray categorization and spatial localization of chest pathologies. In *Lecture Notes in Computer Science*, volume 6893 LNCS, pages 199–206. Springer, 2011.

[54] Stelmo Magalhães Barros Netto, Aristófanes Corrêa Silva, Rodolfo Acatauassú Nunes, and Marcelo Gattass. Automatic segmentation of lung nodules with

growing neural gas and support vector machine. *Computers in Biology and Medicine*, 42(11):1110–1121, 2012.

[55] Temesguen Messay, Russell C. Hardie, and Timothy R. Tuinstra. Segmentation of pulmonary nodules in computed tomography using a regression neural network approach and its application to the Lung Image Database Consortium and Image Database Resource Initiative dataset. *Medical Image Analysis*, 22(1):48–62, 2015.

[56] Tuan Anh Ngo and Gustavo Carneiro. Lung segmentation in chest radiographs using distance regularized level set and deep-structured learning and inference. In *International Conference on Image Processing (ICIP)*, volume 2015-Decem, pages 2140–2143. IEEE, 2015.

[57] Paola Campadelli, Elena Casiraghi, and Diana Artioli. A fully automated method for lung nodule detection from postero-anterior chest radiographs. *IEEE Transactions on Medical Imaging*, 25(12):1588–1603, 2006.

[58] Nico Karssemeijer, Johannes D M Otten, Andre L M Verbeek, et al. Computer-aided detection versus independent double reading of masses on mammograms. *Radiology*, 227(1):192–200, 2003.

[59] Berkman Sahiner, Heang Ping Chan, Lubomir M. Hadjiiski, et al. Effect of CAD on Radiologists' Detection of Lung Nodules on Thoracic CT Scans: Analysis of an Observer Performance Study by Nodule Size. In *Academic Radiology*, volume 16, pages 1518–1530, 2009.

[60] Monde Muyoyeta, Pragnya Maduskar, Maureen Moyo, et al. The sensitivity and specificity of using a computer aided diagnosis program for automatically scoring chest X-rays of presumptive TB patients compared with Xpert MTB/RIF in Lusaka Zambia. *PLoS ONE*, 9(4):16–18, 2014.

[61] P Maduskar, M Muyoyeta, H Ayles, et al. Detection of tuberculosis using digital chest radiography: automated reading vs. interpretation by clinical officers. *The International Journal of Tuberculosis and Lung Disease*, 17(12):1613–1620, 2013.

[62] Laurens Hogeweg, Christian Mol, Pim A. De Jong, et al. Fusion of local and global detection systems to detect tuberculosis in chest radiographs. In *Lecture Notes in Computer Science*, volume 6363, pages 650–657. Springer, 2010.

[63] Rosemary Arthur. Interpretation of the paediatric chest X-ray. *Paediatric Respiratory Reviews*, 13(1):438–447, 2003.

[64] Bram van Ginneken, Laurens Hogeweg, and Mathias Prokop. Computer-aided diagnosis in chest radiography: Beyond nodules. *European Journal of Radiology*, 72(2):226–230, 2009.

[65] Leandro Luís Galdino Oliveira, Simonne Almeida e Silva, Luiza Helena Vilela Ribeiro, et al. Computer-aided diagnosis in chest radiography for detection of childhood pneumonia. *International Journal of Medical Informatics*, 77(8):555–564, 2008.

[66] Rafael T. Sousa, Oge Marques, Fabrizzio Alphonsus A.M.N. Soares, et al. Comparative performance analysis of machine learning classifiers in detection of childhood pneumonia using chest radiographs. In *Procedia Computer Science*, volume 18, pages 2579–2582, 2013.

[67] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, et al. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, 2017.

[68] Jürgen Schmidhuber. Deep Learning in neural networks: An overview. *Neural Networks*, 61:85–117, 2015.

[69] Kevin P. Murphy. *Machine Learning: A Probabilistic Perspective*. MIT Press, 2nd edition, 2012.

[70] Ian J Goodfellow, Yoshua Bengio, and Aaron Courville. Deep Feedforward Neural Networks. In *Deep Learning*, pages 167–227. 2015.

[71] D. H. Hubel and T. N. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology*, 160(1):106–154, 1962.

[72] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. Convolutional Networks. In *Deep Learning*, pages 330–371. 2015.

[73] Kunihiko Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36(4):193–202, 1980.

[74] Yan LeCun, Bernhard Boser, J. S. Denker, et al. Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Computation*, 1(4):541–551, 1989.

[75] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, volume 86, pages 2278–2323. IEEE, 1998.

[76] Shih-Chung B Lo, Shyh-Liang a Lou, Jyh-Shyan Lin, et al. Artificial convolution neural network techniques and applications for lung nodule detection. *IEEE Transactions on Medical Imaging*, 14(4):711–718, 1995.

[77] Jyh Shyan Lin, Akira Hasegawa, Matthew T Freedman, and Seong K Mun. Differentiation between nodules and end-on vessels using a convolution neural network architecture. *Journal of Digital Imaging*, 8(3):132–141, 1995.

[78] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: prevent NN from overfitting. *Journal of Machine Learning Research*, 15:1929–1958, 2014.

[79] Alex Krizhevsky and Geoffrey Hinton. Learning Multiple Layers of Features from Tiny Images. Technical report, 2009.

[80] Andrej Karpathy. CS231n Convolutional Neural Networks for visual recognition. URL `https://cs231n.github.io/convolutional-networks/cs231n.github.io/neural-networks-1/{#}actfun.`, 2017. Accessed: 2018-03-20.

[81] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. Regularization for Deep Learning. In *Deep Learning*, pages 228–273. 2015.

[82] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. Machine Learning Basics. In *Deep Learning*, pages 98–160. 2015.

[83] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. Optimization for Training Deep Models. In *Deep Learning*, pages 274–329. 2015.

[84] John Duchi, Elad Hazan, and Yoram Singer. Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. *Journal of Machine Learning Research*, 12:2121–2159, 2011.

[85] Tijmen Tieleman and Geoffrey Hinton. Lecture 6.5-rmsprop: "Divide the gradient by a running average of its recent magnitude.", 2012.

[86] Diederik Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations*, pages 1–15, 2015.

[87] Jocelyn Sietsma and Robert J.F. Dow. Creating artificial neural networks that generalize. *Neural Networks*, 4(1):67–79, 1991.

[88] Min Lin, Qiang Chen, and Shuicheng Yan. Network In Network. In *International Conference on Learning Representations*, pages 10–20, 2014.

[89] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *International Conference on Learning Representations*, pages 1–10, 2015.

[90] Christian Szegedy, Pierre Sermanet, Scott Reed, et al. Going deeper with convolutions. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9. IEEE, 2015.

[91] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the Inception Architecture for Computer Vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826. IEEE, 2016.

[92] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 770–778. IEEE, 2016.

[93] Fei-Fei Li, Justin Johnson, and Serena Yeung. Lecture 09: CNN Architectures. URL `http://cs231n.stanford.edu/slides/2017/cs231n{_}2017{_}lecture9.pdf`, 2017. Accessed: 2017-12-17.

[94] Andre Esteva, Brett Kuprel, Roberto A. Novoa, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, 2017.

[95] Dan C. Ciresan, Alessandro Giusti, Luca M. Gambardella, and Jurgen Schmidhuber. Mitosis detection in breast cancer histology images with deep neural networks. In *International Conference on Medical Image Computing and Computer-assisted Intervention*, pages 411–418. Springer, 2013.

[96] Marios Anthimopoulos, Stergios Christodoulidis, Lukas Ebner, Andreas Christe, and Stavroula Mougiakakou. Lung Pattern Classification for Interstitial Lung Diseases Using a Deep Convolutional Neural Network. *IEEE Transactions on Medical Imaging*, 35(5):1207–1216, 2016.

[97] Wenlu Zhang, Rongjian Li, Houtao Deng, et al. Deep convolutional neural networks for multi-modality isointense infant brain image segmentation. In *NeuroImage*, volume 108, pages 214–224, 2015.

[98] Pouya Bashivan, Irina Rish, Mohammed Yeasin, and Noel Codella. Learning Representations from EEG with Deep Recurrent-Convolutional Neural Networks. arXiv Preprint: `http://arxiv.org/abs/1511.06448`, 2015.

[99] Rahimeh Rouhi, Mehdi Jafari, Shohreh Kasaei, and Peiman Keshavarzian. Benign and malignant breast tumors classification based on region growing and CNN segmentation. *Expert Systems with Applications*, 42(3):990–1002, 2015.

[100] Shadi Albarqouni, Christoph Baur, Felix Achilles, et al. AggNet: Deep Learning From Crowds for Mitosis Detection in Breast Cancer Histology Images. *IEEE Transactions on Medical Imaging*, 35(5):1313–1321, 2016.

[101] Dina Demner-Fushman, Marc D. Kohli, Marc B. Rosenman, et al. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, 2016.

[102] Hoo-Chang Shin, Lauren Kim, Ari Seff, Jianhua Yao, and Ronald M. Summers. Interleaved text/image Deep Mining on a large-scale radiology database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1090–1099, 2015.

[103] Hossein Azizpour, Ali Sharif Razavian, Josephine Sullivan, Atsuto Maki, and Stefan Carlsson. From generic to specific deep representations for visual recognition. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 36–45, 2015.

[104] B Zhou, A Lapedriza, J Xiao, A Torralba Advances in neural . . . , and Undefined 2014. Learning deep features for scene recognition using places database. In *Advances in Neural Information and Processing Systems*, pages 487–495, 2014.

[105] Jianxiong Xiao, James Hays, Krista A Ehinger, and Antonio Torralba. SUN Database : Large-Scale Scene Recognition from Abbey to Zoo. In *Computer Vision and Pattern Recognition (CVPR)*, pages 3485 – 3492. IEEE, 2010.

[106] Yaniv Bar, Idit Diamant, Lior Wolf, et al. Chest Pathology Detection Using Deep Learning With Non-Medical Training. *IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 294–297, 2015.

[107] Varun Gulshan, Lily Peng, Marc Coram, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Journal of the American Medical Association*, 316(22):2402–2410, 2016.

[108] Jaime Melendez, Bram Van Ginneken, Pragnya Maduskar, et al. A novel multiple-instance learning-based approach to computer-aided detection of tuberculosis on chest X-rays. *IEEE Transactions on Medical Imaging*, 34(1):179–192, 2015.

[109] Alvin Rajkomar, Sneha Lingam, Andrew G. Taylor, Michael Blum, and John Mongan. High-Throughput Classification of Radiographs Using Deep Convolutional Neural Networks. *Journal of Digital Imaging*, 30(1):95–101, 2016.

[110] Hoo-Chang Shin, Kirk Roberts, Le Lu, et al. Learning to Read Chest X-Rays: Recurrent Neural Cascade Model for Automated Image Annotation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2497–2506. IEEE, 2016.

[111] Andrej Karpathy and Fei Fei Li. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3128–3137. IEEE, 2015.

[112] Thomas G. Dietterich, Richard H. Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, 89(1-2):31–71, 1997.

[113] B Babenko. Multiple instance learning: algorithms and applications. Technical report, 2008.

[114] Stuart Andrews, Ioannis Tsochantaridis, and Thomas Hofmann. Support Vector Machines for Multiple-Instance Learning. In *Advances in Neural Information Processing Systems (NIPS)*, pages 561—-568, 2003.

[115] U. K. Lopes and J. F. Valiati. Pre-trained convolutional neural networks as feature extractors for tuberculosis detection. *Computers in Biology and Medicine*, 89:135–143, 2017.

[116] Ross Girshick, Jeff Donahue, Trevor Darrell, U C Berkeley, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Computer Vision and Pattern Recognition (CVPR)*, pages 580–587. IEEE, 2014.

[117] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Region-Based Convolutional Networks for Accurate Object Detection and Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(1):142–158, 2016.

[118] Ian Sheshadri, Aashish, Endres. Describing Objects by their Attributes. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1778–1785. IEEE, 2012.

[119] Li-jia Li, Hao Su, Eric P Xing, and Li Fei-fei. Object Bank : A High-Level Image Representation for Scene Classification & Semantic Feature Sparsification.

In *Advances in Neural Information and Processing Systems (NIPS)*, pages 1378–1386, 2010.

[120] Saurabh Singh, Abhinav Gupta, and Alexei A. Efros. Unsupervised discovery of mid-level discriminative patches. In *European Conference on Computer Vision (ECCV)*, pages 73–86. Springer, 2012.

[121] Pedro F Felzenszwalb, Ross B Girshick, David Mcallester, and Deva Ramanan. Object Detection with Discriminatively Trained Part Based Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1–20, 2009.

[122] Chunshui Cao, Xianming Liu, Yi Yang, et al. Look and think twice: Capturing top-down visual attention with feedback convolutional neural networks. In *International Conference on Computer Vision (ICCV)*, pages 2956–2964. IEEE, 2015.

[123] Tianjun Xiao, Yichong Xu, Kuiyuan Yang, et al. The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In *Computer Vision and Pattern Recognition (CVPR)*, pages 842–850. IEEE, 2015.

[124] Emanuele Pesce, Petros-Pavlos Ypsilantis, Samuel Withey, et al. Learning to detect chest radiographs containing lung nodules using visual attention networks. arXiv Preprint: `http://arxiv.org/abs/1712.00996`, 2017.

[125] Qingji Guan, Yaping Huang, Zhun Zhong, et al. Diagnose like a Radiologist: Attention Guided Convolutional Neural Network for Thorax Disease Classification. *arXiv preprint*, arXiv:1801, 2018.

[126] Yan Xu, Tao Mo, Qiwei Feng, et al. Deep learning of feature representation with multiple instance learning for medical image analysis. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, number 1, pages 1626–1630, 2014.

[127] Peng Tang, Xinggang Wang, Bin Feng, and Wenyu Liu. Learning multi-instance deep discriminative patterns for image classification. *IEEE Transactions on Image Processing*, 26(7):3385–3396, 2017.

[128] Jiajun Wu, Yinan Yu, Chang Huang, and Kai Yu. Deep multiple instance learning for image classification and auto-annotation. In *Computer Vision and Pattern Recognition (CVPR)*, pages 3460–3469. IEEE, 2015.

[129] Robert E Schapire. The boosting approach to machine learning: an overview. In *Nonlinear Estimation and Classification*, volume 171, pages 149–171. Springer, 2003.

[130] W. J. Youden. Index for rating diagnostic tests. *Cancer*, 3(1):32–35, 1950.

[131] Roi Livni, Shai Shalev-Shwartz, and Ohad Shamir. On the Computational Efficiency of Training Neural Networks. In *Advances in Neural Information and Processing Systems (NIPS)*, pages 1–15, 2014.

[132] Elisa Sayrol, Amania Salvador, Jordi Torres, Eva Mohedano, and Kevin McGuinness. Memory usage and computational considerations. URL `http://imatge-upc.github.io/telecombcn-2016-dlcv/slides/D2L1-memory.pdf`, 2016. Accessed: 2018-02-15.

[133] Alfredo Canziani, Eugenio Culurciello, and Adam Paszke. An Analysis of Deep Neural Network Models for Practical Applications. arXiv Preprint: `https://arxiv.org/pdf/1605.07678.pdf`, 2016.

[134] Pat Langley. Machine Learning as an Experimental Science. *Machine Learning*, 3:5–8, 1988.

[135] Pierre Sermanet, David Eigen, Xiang Zhang, et al. OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks. arXiv Preprint: `http://arxiv.org/abs/1312.6229`, 2013.

[136] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer Assisted Intervention*, pages 234–241, 2015.

[137] Ron Kohavi. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In *International Joint Conference on Artificial Intelligence*, volume 14, pages 1137–1143, 1995.

[138] Jaime Melendez, Bram van Ginneken, Pragnya Maduskar, et al. On Combining Multiple-Instance Learning and Active Learning for Computer-Aided Detection of Tuberculosis. *IEEE Transactions on Medical Imaging*, 35(4):1013–1024, 2016.

[139] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Artificial Intelligence and Statistics (AISTATS)*, volume 9, pages 249–256, 2010.

[140] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, et al. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. arXiv Preprint: `http://arxiv.org/abs/1711.05225`, 2017.

[141] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European Conference on Computer Vision (ECCV)*, pages 630–645. Springer, 2016.

[142] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q Weinberger. Densely Connected Convolutional Networks. In *Conference on Computer Vision and Pattern Recognition*, pages 4700–4708. IEEE, 2017.

[143] Li Yao, Eric Poblenz, Dmitry Dagunts, et al. Learning to diagnose from scratch by exploiting dependencies among labels. arXiv Preprint: `http://arxiv.org/abs/1710.10501`, 2017.

[144] Ashia C. Wilson, Rebecca Roelofs, Mitchell Stern, Nathan Srebro, and Benjamin Recht. The Marginal Value of Adaptive Gradient Methods in Machine Learning. In *Advances in Neural Information and Processing Systems (NIPS)*, pages 4151–4161, 2017.

[145] Nancy a Obuchowski. Receiver operating characteristic curves and their use in radiology. *Radiology*, 229(1):3–8, 2003.

[146] Janez Demšar. Statistical Comparisons of Classifiers over Multiple Data Sets. *Journal of Machine Learning Research*, 7:1–30, 2006.

[147] Remco R Bouckaert. Choosing between two learning algorithms based on calibrated tests. In *Proceedings of the 20th International Conference on Machine Learning*, pages 51–58, 2003.

[148] Steven Salzberg. On Comparing Classifiers : Pitfalls to Avoid and a Recommended Approach. *Data Mining and Knowledge Discovery*, 328:317–328, 1997.

[149] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision (ECCV)*, pages 818–833. Springer, 2014.

[150] Alex Kendall and Yarin Gal. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? In *Advances in Neural Information Processing Systems (NIPS)*, pages 5580–5590, 2017.

[151] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for Simplicity: The All Convolutional Net. In *International Conference on Learning Representations (ICLR)*, pages 1–14, 2015.

[152] Maxime Oquab, Leon Bottou, Ivan Laptev, and Josef Sivic. Is object localization for free? - Weakly-supervised learning with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 685–694. IEEE, 2015.

[153] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, et al. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *International Conference on Computer Vision (ICCV)*, pages 618–626. IEEE, 2016.

[154] Martín Abadi, Ashish Agarwal, Paul Barham, et al. TensorFlow: A System for Large-Scale Machine Learning. In *InOSDI*, volume 16, pages 265–283, 2016.

[155] R Collobert, C Farabet, and K Kavukcuoglu. Torch7: A Matlab-like Environment for Machine Learning. In *Advances in Neural Information and Processing Systems (NIPS) - BigLearn Workshop*, 2011.

[156] Tianqi Chen, Mu Li, Yutian Li, et al. MXNet: A Flexible and Efficient Machine Learning Library for Heterogeneous Distributed Systems. arXiv Preprint: `http://arxiv.org/abs/1512.01274`, 2015.

[157] Yangqing Jia, Evan Shelhamer, Jeff Donahue, et al. Caffe: Convolutional Architecture for Fast Feature Embedding. In *International Conference on Multimedia (ICM)*, pages 675–678. ACM, 2014.

[158] Koray Kavukcuoglu. Research Blog: "DeepMind moves to TensorFlow". URL `https://research.googleblog.com/2016/04/deepmind-moves-to-tensorflow.html`. Accessed: 2018-03-05.

[159] Shabir A Madhi, Michelle J Groome, Heather J Zar, et al. Effectiveness of pneumococcal conjugate vaccine against presumed bacterial pneumonia hospitalisation in HIV-uninfected South African children: a case-control study. *Thorax*, 70(12):1–7, 2015.

[160] Anna H. Van't Hoog, H K Meme, H Van Deutekom, et al. High sensitivity of chest radiograph reading by clinical officers in a tuberculosis prevalence survey. *International Journal of Tuberculosis and Lung Disease*, 15(10):1308–1314, 2011.

[161] George H Swingler. Observer variation in chest radiography of acute lower respiratory infections in children : a systematic review. *BMC Medical Imaging*, 1(1), 2001.

[162] Kenneth R Foster, Robert Koprowski, and Joseph D Skufca. Machine learning, medical diagnosis, and biomedical engineering research - commentary. *BioMedical Engineering OnLine*, 13(1):1–10, 2014.

[163] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional Two-Stream Network Fusion for Video Action Recognition. In *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on*, pages 1933–1941, 2016.

[164] Volodymyr Mnih, Nicolas Heess, Alex Graves, and Koray Kavukcuoglu. Recurrent Models of Visual Attention. In *Advances in Neural Information Processing Systems (NIPS)*, volume 27, pages 2204–2212, 2014.

Appendix A

MHREC - Clearance Certificate

R14/49 Mr Jonathan Gerrand

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)

CLEARANCE CERTIFICATE NO. M1611133

NAME: (Principal Investigator)	Mr Jonathan Gerrand
DEPARTMENT:	School of Electrical and Information Engineering
PROJECT TITLE:	Training Mehods for Convolutional Neural Networks Perfoming Automatic Chest Reading on Frontal Pneumonia Chest X-Rays
DATE CONSIDERED:	Adhoc
DECISION:	Approved unconditionally
CONDITIONS:	Sub-Study under Primary Study M120933 Prof Shabir Madhi
SUPERVISOR:	Adam Pantanowitz
APPROVED BY:	Professor A Dhai, Co-Chairperson, HREC (Medical)
DATE OF APPROVAL:	12/12/2016

This clearance certificate is valid for 5 years from date of approval. Extension may be applied for.

DECLARATION OF INVESTIGATORS

To be completed in duplicate and **ONE COPY** returned to the Research Office Secretary in Room 301, Third floor, Faculty of Health Sciences, Phillip Tobias Building, 29 Princess of Wales Terrace, Parktown, 2193, University of the Witwatersrand. I/we fully understand the conditions under which I am/we are authorized to carry out the above-mentioned research and I/we undertake to ensure compliance with these conditions. Should any departure be contemplated, from the research protocol as approved, I/we undertake to resubmit the application to the Committee. **I agree to submit a yearly progress report**. The date for annual re-certification will be one year after the date of convened meeting where the study was initially reviewed. In this case, the study was initially reviewed in November and will therefore be due in the month of November each year. Unreported changes to the application may invalidate the clearance given by the HREC (Medical).

Principal Investigator Signature _____ Date _____

PLEASE QUOTE THE PROTOCOL NUMBER IN ALL ENQUIRIES