Australia's National Science Agency

Optimization and Pilot Development of Pneumoconiosis Detection Software

The final project report for Coal Services Health and Safety Trust

Yulia Arzhaeva, Dadong Wang, Md Shariful Alam, Qiyu Liao, Olivier Salvado, Arcot Sowmya, Eun-Kee Park and Deborah Yates

Coal Services Health and Safety Trust Project No. 20656

CSIRO Report No. EP2022-4999

15 December 2022

Commercial-in-confidence

Health & Safety Trust

ST VINCENT'S HOSPITAL SYDNEY

UNSW SYDNEY

Imaging and Computer Vision Research Group, Data61, CSIRO

Contents

Figures

Tables

Acknowledgments

The authors would like to thank Resources Safety & Health Queensland (RSHQ), Dr Hyunrim Choi from Good Morning Hospital in South Korea, Coal Services Health (CSH) for providing de-identified sample X-ray images from their image database, and radiologist Dr Katrina Newbigin from Wesley Medical Imaging for providing examples of nodular dust related lung disease including cases of coal worker's pneumoconiosis and silicosis from the database at the Wesley Hospital Brisbane. We are very grateful to radiologists Drs Jesse Ende, Joanna Kao and Elizabeth Silverstone from St Vincent's Hospital Sydney for annotating lung fields in chest X-rays, and Dr Siavash Es'haghi from Lung Screen Australia Pty Ltd for providing annotation of pneumoconiosis lesions, to serve as the ground truth for our automated methods. Thanks are also due to Lung Screen Australia Pty Ltd for conducting the pilot study of automated pneumoconiosis detection and classification using our AI-Xrayder software. The authors would also like to extend their gratitude and appreciation to Prof. Robert Cohen, Director of Occupational Lung Disease from Northwestern University in the US, for providing advice at the initial stage of the project.

Executive summary

Pneumoconioses are preventable but incurable lung diseases caused by long-term inhalation of respirable dust such as coal, asbestos, and silica, and that from the inhalation of coal dust is more commonly known as black lung or Coal Workers' Pneumoconiosis (CWP). In Queensland, Australia, there has been a resurgence of this disease in the last three years. About 70 cases of mine dust lung diseases have been diagnosed in Financial Year 2022 and a total of 328 cases have been reported since 1984 [1]. NSW Dust Disease Register Annual Report 2021-2022 shows that there are 476 notifiable dust disease cases and 313 deaths [2]. Based on the data from National Health Commission of China, there were 15,898 new cases of pneumoconiosis reported in 2019 [3]. In the US, the prevalence of CWP among coal miners with 25 or more years of experience exceeds 10% in 2017 [4] compared to 2.1% in 1990 [5]. The Sim review [6] shows that poor dust control is to blame for the re-emergence of pneumoconiosis in Queensland, and patchy medical screening has failed in the early detection of this potentially fatal disease. For pneumoconiosis screening, chest radiographs are acceptable, widely available and relatively inexpensive. The current practice in Australia is that coal miners are required to undergo pre-employment chest X-rays, followed by routine X-ray screenings after the employment, and each X-ray requires two B-readers to review. However, the insensitivity of chest radiographs for detection of early or moderate pneumoconiosis limits their efficacy in screening. This also leads to low sensitivity and specificity of chest X-rays when read by a radiologist who is qualified as a B-reader, especially for the detection of pneumoconiosis at an early stage of the disease. Inter- and intra-reader variability in chest radiography has been acknowledged ever since chest radiography was first used to identify and classify pneumoconiosis. To date, there has been a lack of systematic, automated, and objective systems for detecting the presence and assessing the progression of pneumoconiosis for individual coal miners other than by expert radiologists.

With the advances in data storage and high performance computing technologies, deep learning has driven many artificial intelligence (AI) applications and services that are overwhelmingly successful, especially in image segmentation and classification. In the last five years, there have been lots of successful applications in medical imaging, such as CheXNet [7] for the detection of pneumonia from chest X-rays and CheXNeXt [8] for predicting diseases on X-ray images and producing heat maps to highlight lesions on X-ray images most indicative of each predicted disease.

In collaboration with St Vincent's Hospital at Sydney, and the University of New South Wales, this project aims at developing a deep learning-based automated pneumoconiosis detection and grading system based on the International Labour Organization (ILO) Classification guidelines. Extensive experiments have been conducted to test and validate the system in lab and also at a lung screening organisation via conducting a pilot study with previously unseen chest X-ray images.

Based on the deep learning models proposed in this report, we have developed a web-based software tool for automated pneumoconiosis detection and classification, named AI-Xrayder. The software has been used in our pilot study at Lung Screen Australa Pty Ltd.

Curation of Chest X-ray Image Dataset for Pneumoconiosis

We have collaborated with various organisations to collect chest X-ray images and associated ILO classifcations for the development of the automated pneumoconiosis detection and grading system. So far we have collected 1,842 chest X-ray images, including 1,182 ILO positive X-ray images and 660 normal X-ray images. These images are collected from Resources Safety and Health Queensland (RSHQ), NSW Coal Services Health (CSH), Wesley Medical Imaging (WMI) in Queensland, St Vincent's Hospital (SVH) in Sydney, Good Morning Hospital (GMH) in South Korea, International Labour Organization (ILO), and the latest NIOSH (The National Institute for Occupational Safety and Health, USA) B Reader study syllabus.

For all chest X-ray images collected above, there is an ILO category associated with each X-ray. For all X-ray images from RSHQ and GMH, lung zone-based ILO categories of each image are annotated by B readers. For chest X-ray images from RSHQ, pneumoconiosis lesions are also outlined and labelled by a B reader.

Apart from these images, we have also used some publicly available datasets to develop deep learning models in this project, such as CheXpert [11] containing 224,316 chest X-ray images curated by Stanford Maachine Learning Group, MIMIC-CXR [12, 13] including 377,110 images released by the Beth Israel Deaconess Medical Centre in Boston, ChestX-ray14 [14] including 112,120 X-ray images downloaded from National Institute of Health (NIH), Japanese Society of Radiological Technology (JSRT) dataset [15], Montgomery County X-ray Dataset and Shenzhen Hospital X-ray Dataset [16]. Some images from these datasets have been used for pre-training and testing our machine learning mdels.

Deep Learning-Based Pneumoconiosis Detection

Based on the investigation of different state-of-the-art deep learning methods, we have developed an innovative deep learning model for pneumoconiosis detection, named Masked Attention Convolutional Neural Networks (MA-CNN). By applying a mask attention constraint to the machine learning model, the proposed model is forced to learn image features from lung fields. Our 5-fold cross-validation results show that the proposed method can improve the pneumoconiosis detection accuracy, and we have achieved a sensitivity of 96.34%, a specificity of 98.52%, and an accuracy of 98.65% on the experimental dataset.

Lung segmentation is a crucial step in curating a training dataset for machine learning-based pneumoconiosis detection and classification. With the segmented lung field images, machine learning models can be trained to focus on the features inside the lung fields in a chest X-ray. However, we find that the most of state-of-the-art segmentation methods do not address the challenge of segmenting the lungs from chest X-ray images acquired at later stages of pneumoconiosis or when pneumoconiosis is accompanied by other diseases. Therefore, we have proposed and developed two new machine learning models, named BCL-UNet and MRUNet++, to solve the challenge of segmenting obscured or deformed lung fields. Our experimental results demonstrate the proposed method outperforms the state-of-the-art methods and can produce valid segmentation results from chest X-ray images even with obscured and deformed lung structures caused by severe diseases.

Deep Learning-Based Pneumoconiosis Classification

To classify radiographs of pneumoconioses, we have also developed and compared different deep learning models to identify the best performing model for the chest X-ray classification of pneumoconioses. These include a novel multi-scale network structure, Multi-Scale CNN (MS-CNN), and a lung zone-based pneumoconiosis classifier. Experiments demonstarte that MS-CNN is capable of learning both discriminative detailed textures in high-resolution images and global features in multi-channel low resolution images, and outperforms CheXNet when comparing both models on a benchmark dataset. The experiments show that MS-CNN has achieved an average accuracy of 85.04% in the X-ray classification of pneumoconioses, an average sensitivity of 95.73% and an average specificity of 93.97% in the detection of pneumoconiosis.

Pilot Study

To validate our deep learning models using unseen X-ray images, a pilot study has been conducted to test AI-Xrayder at Lung Screen Australia Pty Ltd. A total of 209 chest X-ray images are used for the study, a sensitivity of 82.57%, a specificity of 90%, and an accuracy of 86.12% are achieved in the detection of pneumoconiosis. For the grading of chest X-ray images into ILO categories 0, 1, 2 and 3, a classification accuracy of 74.64% is obtained.

Limitations and Future Work

Due to the low incidence of pneumoconiosis in Australia we were able to validate our tool only with a limited number of chest X-rays with pneumoconiosis, majority of them being grade 1. To make our tool more robust and suitable for clinical use, we will:

- Continue to work with our collaborators on additional acquisition of chest X-rays with features of pneumoconiosis;
- Further improve our deep learning models for detecting and grading pneumoconiosis into different categories of severity when more chest radiographs become available; and
- Contnue the pilot study with Lung Screen Australia Pty Ltd to validate the pneumoconiosis detection and classification software tool, AI-Xrayder, and collect feedback to further improve functionality and performance of the tool.

1 Human Research Ethics Approval

The study conducted in this project has been approved by the Human Research Ethics Committee of St Vincent's Hospital with an expiry date of 15 August 2022 [9]. Also, we received an approval for this study from CSIRO Health and Medical Human Research Ethics Committee on 12 October 2016. The CSIRO approval has been extended to 30 June 2023 [10]. This will ensure that this research project will be covered. All research panels have deemed this a low/negligible risk project.

2 Image Data Collection

We have collaborated with various organisations to collect chest X-ray images and associated ILO classifications to be used in the development of the automated pneumoconiosis detection and grading system. In this chapter, we summarise the chest X-ray datasets collected since the commencement of this project and during the Coal Services Health and Safety Trust Project 20647.

We have collected new pneumoconiosis X-ray images from the following sources:

- Good Morning Hospital (GMH) in South Korea – We have purchased 320 ILO positive digital X-ray images from GMH, which cover different ILO categories of pneumoconiosis.
- Resources Safety and Health Queensland (RSHQ) – We have signed a data sharing agreement with RSHQ to supply CSIRO ILO positive X-ray images collected from Queensland coal miners. RSHQ has provided 694 ILO positive chest X-rays and their associated B-readers' reports, and will provide CSIRO with more chest X-rays and associated reports as they become available to RSHQ.
- The latest NIOSH (The National Institute for Occupational Safety and Health) B Reader study syllabus for classification of radiographies of pneumoconiosis. This new syllabus includes 135 chest X-ray images covering different ILO categories. We have downloaded these images from the website of Centres for Disease Control and Prevention [17].

The X-ray images we collected during the project 20647 include those from:

- International Labour Organization (ILO) - We have purchased a digital set of 22 ILO Standard Radiographs. This set is used in the ILO Classification System for Pneumoconiosis on Chest Radiographs. For this project, we have selected 17 chest radiographs out of the 22. The selected images depict complete lung fields – either normal, or with small parenchymal abnormalities consistent with pneumoconiosis.
- Wesley Medical Imaging (WMI) - We have collected 64 chest X-rays belonging to normal individuals, and 25 chest X-rays with small parenchymal opacities consistent with pneumoconiosis, which belong to 25 de-identified male individuals.
- National Institute for Occupational Safety and Health (NIOSH) – We have downloaded the online B Reader Syllabus for preparing doctors to take the ILO Classification exam. 41 teaching images from this resource were selected for our study, using the same criteria as for selecting ILO Standard Radiographs.
- Coal Services Health (CSH) - We have acquired 511 chest X-ray images from CSH, including 505 chest X-rays exhibiting no signs of pneumoconiosis, 5 chest X-rays classified as ILO 0/1 that might have features consistent with pneumoconiosis, and one X-ray classified as ILO 2/2.
- St Vincent's Hospital (SVH) at Sydney – We have collected 100 chest X-ray images from SVH. Among these images, 35 images belong to normal individuals, and the other images are from

the patients with various lung diseases. We only used the normal X-rays for training our machine learning models.

The following table shows the numbers of new chest X-rays we collected since the commencement of this project in July 2020, and the chest X-ray images we collected during the project 20647.

Table 2-1 Summary of chest X-ray image data collection

Project	Current Project (No. 20656)			Project No. 20647				
Source	GMH	RSHQ	NIOSH	ILO	NIOSH	WMI	CSH	SVH
Number	320	694	135	17	41	89	511	35
Total	1,842 = 660 (Normal X-ray images) + 1,182 (ILO positive X-ray images)							

We have also used some publicly available datasets in this project to develop our machine learning-based algorithms for segmentation and classification. These include:

- Stanford Machine Learning Group – CheXpert is a large dataset of chest X-rays that contains 224,316 chest radiographs of 65,240 patients [11].
- MIMIC Chest X-ray (MIMIC-CXR) database – The database contains 377,110 images corresponding to 227,835 radiographic studies performed at the Beth Israel Deaconess Medical Centre in Boston, MA [12, 13].
- National Institute of Health - We have downloaded the ChestX-ray14 dataset from National Institute of Health [14]. This dataset includes 112,120 X-ray images of more than 30,805 unique patients collected from a hospital Picture Archiving and Communication System (PACS) with automatically text-mined image labels from their associated radiological reports. Among these X-ray images, 51,708 images contain one or more pathologies, and the remaining 60,412 images do not have any pathological findings.
- Japanese Society of Radiological Technology (JSRT) – We have downloaded the JSRT database that contains 247 digitized chest X-rays with annotated lung masks [15].
- Montgomery County X-ray Dataset - This dataset contains 138 posterior-anterior X-rays acquired from the Department of Health and Human Services of Montgomery County, MD, USA [16]. Among the 138 X-rays, 80 X-rays are normal, and the rest are abnormal with manifestations of tuberculosis.
- Shenzhen Hospital X-ray Dataset – This dataset has been collected by Shenzhen No. 3 Hospital in Shenzhen, Guangdong providence, China [16]. The dataset contains 326 normal and 336 abnormal X-rays showing manifestations of tuberculosis.

3 Deep Learning Based Lung Segmentation

Lung segmentation is a crucial step in curating a training dataset for machine learning based pneumoconiosis detection and classification. With the segmented lung field images, machine learning models can be trained to learn and focus on the features inside the lung fields in a chest X-ray.

Previously, we developed an automated lung segmentation method based on a deep learning network called UNet. In our last report we showed that the UNet-based method improved lung segmentation results compared to the previous state-of-the-art Pixel Classification approach. However, lungs with widespread or gross pathologies still present a significant challenge to automated methods for lung segmentation. Such pathologies are not uncommon for later stages of pneumoconiosis or when pneumoconiosis is accompanied by other diseases. In this chapter we present and discuss two proposed deep learning networks that improve segmentation of lung fields severely affected by pathologies.

3.1 Proposed BCL-UNet network

Deep Neural Networks (DNN)-based methods, particularly UNet, are considered state-of-the-art for many medical imaging tasks. UNet is a convolutional neural network (CNN) initially developed by Ronneberger et al. [23]. The network consists of an encoding, or a contractive, path along with a symmetric decoding, or an expansive, path that gives a U-shape to the network, as shown in the original paper [23] and in our previous report. In the encoding path, feature maps with reduced resolution are extracted. The feature maps are then upsampled using deconvolutional layers in the decoding path. There are connections between the layers of equal feature map size (known as skip connections) from the encoding path to the decoding path, that provide important high-resolution features to the deconvolution layers.

However, despite remarkable progress on segmenting the normal lung, performance of UNet is unsatisfactory on challenging chest X-ray (CXR) images [24]. In this study, we propose a DNN-based architecture that replaces the skip connections of UNet with a bidirectional convolutional-LSTM (BC-LSTM) module that allows exchange of more information between the encoder and decoder paths and also captures spatiotemporal information. For further improvement, we add a multiple kernel pooling (MKP) block at the lowest level of UNet to encode more spatial information by different sized pooling operations.

3.1.1 Overview of the BCL-UNet network

Motivated by the success of Bidirectional Convolutional-LSTM (BC-LSTM) and Multi-Kernel Pooling (MKP) [18, 19, 20, 21], we combine these two blocks with UNet to develop a new framework for lung segmentation, named BCL-UNet. In the encoding path of UNet the dimensionality of feature maps is reduced, which causes loss of some spatial information that could be important for lung segmentation. The recent success of the BC-LSTM module (see Figure 3-1 (a)) on various image

related tasks motivated us to combine this module with UNet to preserve spatial information and exchange more information between the encoding and decoding paths.

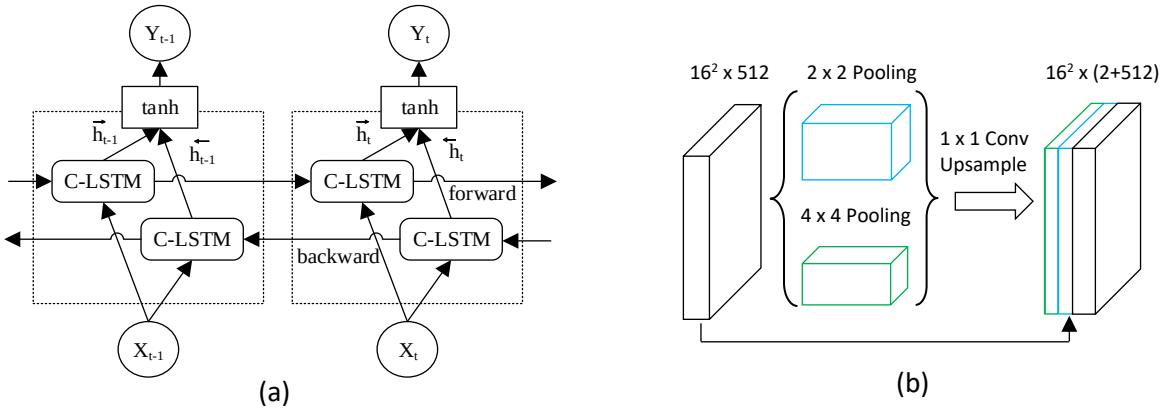


Figure 3-1 (a) Illustration of BC-LSTM module. It uses two convolutional LSTM (C-LSTM) modules with forward and backward paths. (b) Multi-Kernel Pooling (MKP) module. Information is encoded using two differently sized kernels.

Convolutional LSTM module (C-LSTM) preserves spatiotemporal information using convolution operations [21]. An important part of C-LSTM is a memory cell that stores the unit state. The information in the memory cell can be accessed, updated and cleared by the input, output and forget gates, also known as controlling gates. BC-LSTM uses two C-LSTMs to process the input data into forward and backward directions, which has shown improved prediction performance [18, 19].

A common way to encode contextual information in deep neural networks is by including pooling layers. Pooling aggregates statistics (min, max, average) of previously extracted features over a receptive field of a certain size. The standard pooling operation uses a single kernel of size 2 x 2. It has been shown that varying the size of receptive field used in pooling operations can improve the performance of image related problems such as detection, segmentation and classification [20]. Therefore, we propose to use two differently sized kernels, [2 x 2] and [4 x 4], to encode contextual information of different sizes. Figure 3-1(b) illustrates Multi-Kernel Pooling. Application of the kernels with different sizes creates feature maps of different sizes. Through upsampling we increase the dimension of the feature maps to match the feature maps of the original input. Finally, we concatenate upsampled feature maps with the input feature maps. A schematic visualization of our proposed framework that consists of three parts: UNet, BC-LSTM and MKP, is shown in Figure 3-2.

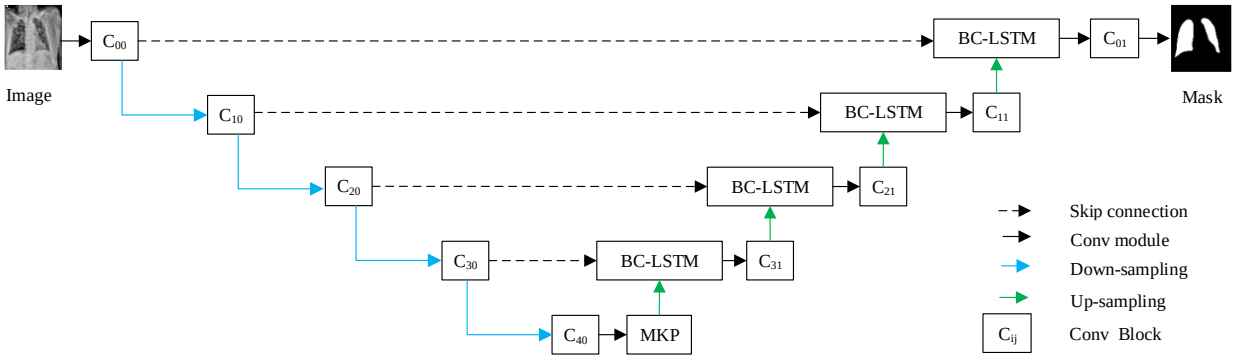


Figure 3-2 Architecture of the proposed BCL-UNet module, that combines UNet with BC-LSTM and MKP modules

3.1.2 Datasets

For the evaluation of BCL-UNet network we use three public datasets, namely Shenzhen, Montgomery and Japanese Society of Radiological Technology (denoted SMJ collectively) described in Chapter 2. We combine these three datasets and randomly divide them into two subsets for training (80%) and testing (20%). The training set is further divided into two subsets for training (90%) and validation (10%). We use a five-fold cross validation scheme for all experiments. Lung regions of the SMJ dataset contain mild abnormal lesions and most of the images do not contain dense opacities or other severe abnormalities. To evaluate the effectiveness of the proposed technique, an independent, challenging test dataset with 100 images is used, including 50 pneumoconiosis images from Good Morning Hospital, South Korea, denoted as GMH dataset, which covers different ILO categories of pneumoconiosis, and another 50 images with Covid-19 disease from publicly available datasets (denoted Covid-19) [22]. We obtained the ground truth lung masks for the test images annotated by two radiologists from St Vincent's Hospital, Sydney.

3.1.3 Evaluation Metrics

For each image, we computed the Dice coefficient (DC) and Jaccard Index (JI) metrics to evaluate the overlap between the estimated lung mask and the ground truth. These are popular metrics for segmentation evaluation, computed as follows:

$$DC(G,P) = \frac{2|PG|}{|P|+|G|} \quad (1)$$

$$JI(G,P) = \frac{|PG|}{|P|+|G|-|PG|} \quad (2)$$

where |G| is the number of pixels in the ground truth mask, |P| is the number of pixels in the estimated lung mask, and |PG| is the number of pixels in the overlap between the ground truth mask G and the estimated mask P. If the overlap between the ground truth and estimated masks is perfect, both DC and IJ would equal to one.

3.1.4 Experimental Setup

We employed five-fold cross validation and computed the mean and standard deviation of the metric values as percentages (%). To train the UNet model, we used Adam optimization algorithm with a learning rate of 4x10-4.

For the hidden and output layers, we used 'relu' and 'sigmoid' activation functions, respectively. All X-ray images were resized to 512 x 512 before being used for training and testing. We used standard data augmentation techniques such as rotation, flipping and zooming on training images only. The model was trained for 50 epochs with the batch size of 8.

3.1.5 Experimental Results

In the first experiment we replaced the skip connections of the UNet model with a BC-LSTM block. The produced results are demonstrated in Table 3-1. The results show that the BC-LSTM block is effective for the lung segmentation, and can improve the segmentation performance. To improve the accuracy further, we added the MKP block at the lowest level of the UNet model. The results shown in Table 3-1 suggest that our proposed method outperforms the standard UNet model for all test datasets in terms of both DC and JI.

Table 3-1 Lung segmentation performance on the four public and one private datasets using the proposed (UNet+ BC-LSTM+ MKP) and the standard UNet model. The best results are shown in bold

Database	Methods	DC (mean+-std. %)	JI (mean+-std. %)
SMJ	UNet	0.9556±0.45	0.9150±0.81
	UNet+ BC-LSTM	0.9583±0.93	0.9203±1.69
	UNet+ BC-LSTM+ MKP	**0.9604±0.46**	**0.9239±0.85**
GMH	UNet	0.9116±1.48	0.8387±2.34
	UNet+ BC-LSTM	0.9246±0.67	0.8601±1.15
	UNet+ BC-LSTM+ MKP	**0.9335±0.81**	**0.8754±1.42**
Covid-19	UNet	0.9290±0.52	0.8676±0.88
	UNet+ BC-LSTM	0.9339±0.94	0.8762±1.63
	UNet+ BC-LSTM+ MKP	**0.9431±0.20**	**0.8923±0.35**

To visualize the lung segmentation results, Figure 3-3 shows five CXR images from GMH, and Covid-19 datasets, their corresponding annotated ground truth (GT), segmented lung fields using our proposed (Prop) and the standard UNet models. From this figure, qualitative performance difference can be observed between the proposed and the standard UNet-based models. Compared to the standard UNet model, our proposed method can segment lungs from CXR images more accurately, this is particularly true for the CXR images with complex structures caused by severe diseases, for example, the X-ray images in Columns 1 and 3.

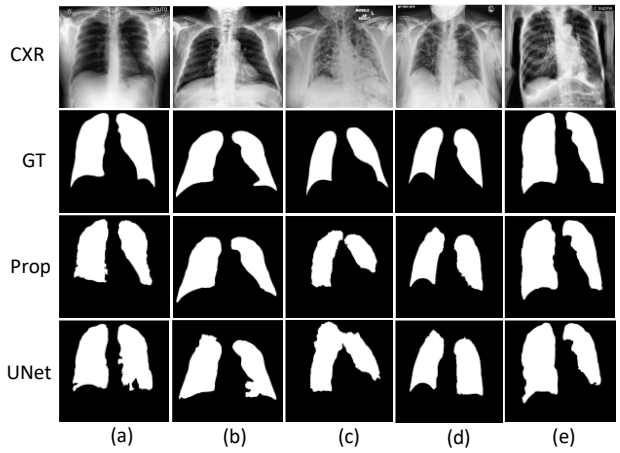


Figure 3-3 Qualitative comparison of the proposed method against the standard UNet model on five example CXR images (columns) from GMH and Covid-19 datasets

3.2 Overview of MRUNet++ network

MRUNet++ have been developed to segment the lung fields in chest X-ray images with complex structures due to pneumoconiosis and other lung diseases. We will briefly introduce this network and then discuss its qualitative and quantitative results.

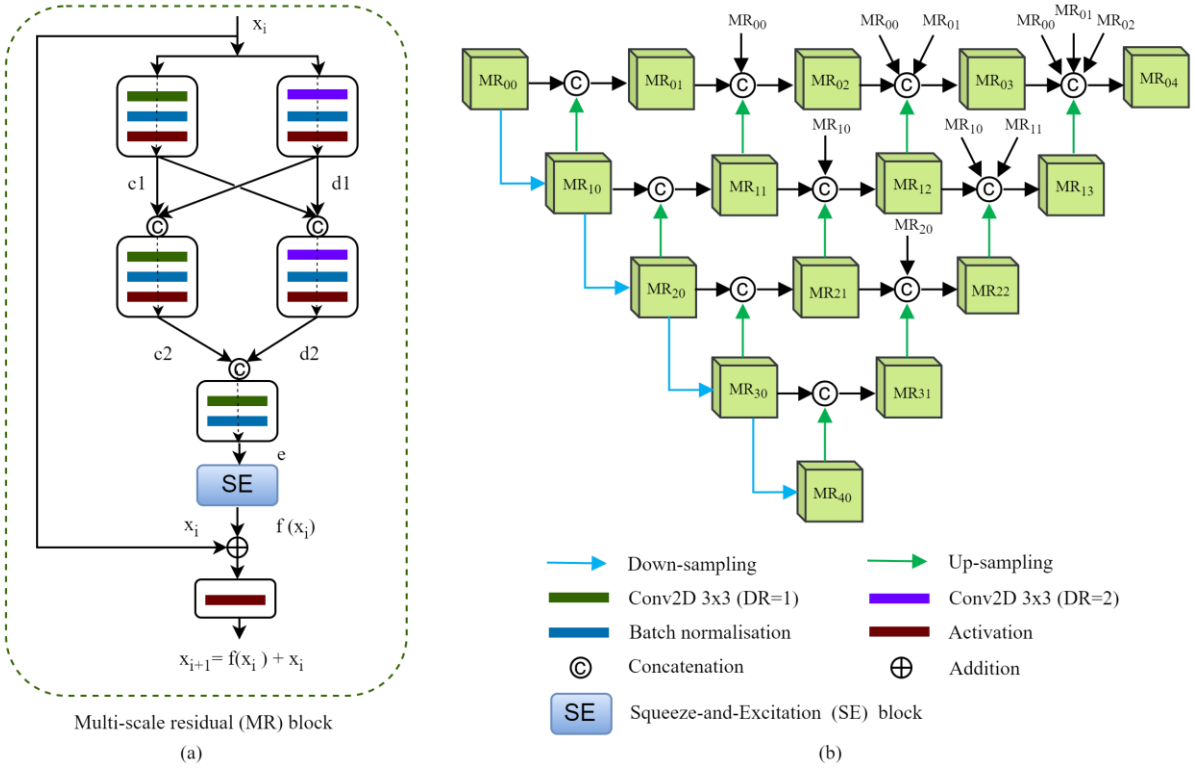


Figure 3-4 (a) Illustration of the i-th multi-scale residual (MR) block; (b) Architecture of the proposed MRUNet++ network for medical image segmentation. We replace the convolutional layers in the UNet++ architecture with the proposed MR blocks.

3.2.1 Overview of MRUNet++ network

MRUNet++ is based on the UNet++ model. Zhou et al. [24] developed the UNet++ model to solve the two main problems of the standard UNet, which are: (1) the exact depth level for the optimal architecture, and (2) a restrictive fusion scheme that forces aggregation of the same level feature maps**.** UNet++ is constructed from UNet by replacing skip connections with densely connected skip connections. These connections allow flexible deep feature transmission along skip connections and feature fusion at decoder nodes. Another challenge of many deep learning networks is the vanishing or exploding gradient problem that hampers the training process [25]. He *et al.* [25] addressed this problem by introducing residual learning in a residual neural network (ResNet). A residual unit is a component of ResNet where the activation from a previous layer is added to the activation of a deeper layer in the network. There are many combinations of the convolution layer, batch normalization and activation function in a residual unit, details of which can be found in [25].

In our proposed MRUNet++ network (Figure 3-4(b)) we replace the convolutional block of the UNet++ model with a multi-scale residual (MR) block which is depicted in Figure 3-4(a). Motivated by the success of multi-scale residual block to recover high quality images [26], we use a similar backbone for MRUNet++. The MR block consists of two-bypass networks which use different dilation rates (DR = 1 and DR = 2) and a convolution kernel of the same size [3 x 3]. To extract features at multiscale, the features between these bypass networks can be shared with each other. The operation of the bypass networks can be defined by the following transformations:

$$c1 = \sigma\,(ß\,(w^{1}_{3x3,DR=1} * x_i + b^1)) \quad (3)$$

$$d1 = \sigma\,(ß\,(w^{1}_{3x3,DR=2} * x_i + b^1)) \quad (4)$$

$$c2 = \sigma\,(ß\,(w^{2}_{3x3,DR=1} * [c1, d1] + b^2)) \quad (5)$$

$$d2 = \sigma\,(ß\,(w^{2}_{3x3,DR=2} * [d1, c1] + b^2)) \quad (6)$$

$$e = ß\,(w^{3}_{3x3,D=1} * [c2, d2] + b^3) \quad (7)$$

where σ(.) and ß(.) denotes the the activation function and batch normalisation function, respectively. Similarly, x_i , *w* and *b* represent input of the *i*-th MR unit, the weights and biases, respectively. The superscripts of *w* represent the number of the layers at which they are located, the subscripts of w represent the convolution filter size [3 x 3] and dilation rate, and [,] represents the concatenation operation.

A Squeeze-and-Excitation (SE) unit is inserted to focus more on relevant features as follows:

$$f(x^i) = Ɛ(e) \quad (8)$$

where ƒ(.) represents the residual learning function that performs a nonlinear transformation with a series of operations, and Ɛ(.) denotes SE function. Finally, to increase the gradient flow, residual connection is adopted for each block.

Finally, each MR can be defined in the following way:

$$x^{i+1} = \sigma(f(x^i)+x^i) \quad (9)$$

where x^i and x^{i+1} represent the input and output of the *i*-th MR block. The operation $f(x^i) + x^i$ is performed using elementwise addition.

3.2.2 Datasets

In this study, for the MRUNet++ model we use the two publicly available datasets described in Chapter 2, namely, Montgomery and Japanese Society of Radiological Technology, denoted collectively as MJ, for developing the machne learning model for lung segmentation. We use these two datasets because they come with ground truth lung masks for training and evaluation of the automated segmentation methods. The MJ datset is randomly divided into two subsets for training (80%) and testing (20%). The training set is further divided into two subsets for training (85%) and validation (15%).

Most of the abnormal chest X-rays of MJ dataset do not contain very dense opacities or other severe abnormalities in the lungs. To assess the effectiveness of the proposed methods for the lungs with a variable degree of pathology, an independent test dataset was used for evaluating the segmentation performance of the model trained on the MJ dataset. We selected 200 pneumoconiosis images with complex structure from the GMH dataset described in Chapter 2. This dataset covered different ILO categories of pneumoconiosis. Another 50 test images with Covid-19 disease were picked from publicly available datasets (denoted Covid-19) [22]. We obtained the ground truth lung masks for the test images annotated by two radiologists from St Vincent's Hospital, Sydney.

3.2.3 Experimental Setup

Our proposed network was trained and tested in a five-fold cross-validation manner. It means, that MJ dataset was divided into 5 non-overlapping subsets (folds), with 20% of data in each, and for each fold the rest of the dataset was used for training a network, and the fold was used for testing. For each fold, a trained network was also tested on 200 images from the GMH dataset.

3.2.4 Experimental Results for MRUNet++ network

Table 3-2 Comparative evaluation between the proposed MRUNet++ network and other state-of-the-art networks for lung segmentation measured by Dice Coefficient (DC) and Jaccard Index (JI) on the four datasets. The mean and standard deviation for each metric measured over five folds.

Method	MJ		GMH		COVID-19	
	DC	JI	DC	JI	DC	JI
U-Net	0.9546±1.93	0.9214±2.88	0.8475±2.38	0.7479±3.39	0.9140±0.98	0.8510±1.35
UNet++	0.9543±2.76	0.9248±3.62	0.8610±1.59	0.7650±2.21	0.9291±0.56	0.8718±0.86
AttentionUNet	0.9570±2.08	0.9257±2.82	0.8524±1.83	0.7512±2.65	0.9284±0.27	0.8699±0.48
ResUNet++	0.9655±1.04	0.9367±1.60	0.8598±2.45	0.7679±3.58	0.9317±0.40	0.8765±0.55
MultiResUNet	0.9617±1.93	0.9332±2.77	0.8224±2.45	0.7149±3.35	0.9235±0.89	0.8625±1.43
DCUNet	0.9570±2.69	0.9288±3.40	0.8303±0.74	0.7221±1.24	0.9256±0.99	0.8666±1.56
MRUNet++ (Proposed)	**0.9642±1.33**	**0.9345±2.15**	**0.8778±1.92**	**0.7893±2.98**	**0.9342±0.64**	**0.8797±1.05**

Quantitative comparison between the proposed MRUNet++ model and other state-of-the-art networks on test datasets is shown in Figure 3-5. The results for each network are averaged over the five folds. The results suggest that for MJ dataset the performance is comparable for all architectures. However, for COVID-19 and pneumoconiosis CXR images from the GMH dataset, our proposed method outperforms all other networks.

To visualize the lung segmentation results, Figure 3-5 shows two challenging chest X-rays from the GMH and COVID-19 datasets, their corresponding annotated ground truth, segmented lung fields using the proposed and other models. In this figure qualitative performance differences can be observed between the proposed and other models. Compared to all other models, our proposed MRUNet++ can segment lungs from these chest X-rays more accurately, while other methods suffer from obscured and deformed lung structures caused by severe diseases

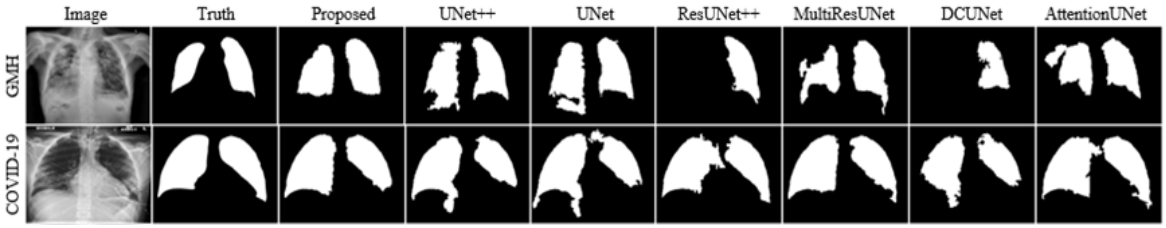


Figure 3-5 Qualitative comparison of the proposed methods against the against other state-of-the-art networks on two difficult chest X-ray images from GMH and COVID-19.

3.3 Summary

We have developed and implemented the two UNet-based deep learning networks, BCL-UNet and MRUNet++, and experimentally validated them on the lung segmentation task. The proposed two methods outperformed the state-of-the-art lung segmentation methods on a challenging pneumoconiosis dataset, and showed a comparable performance on a less challenging dataset that included normal images and abnormal chest X-rays without severe signs of pathology.

We have also compared the performance of MRUNet++ with the BCL-UNet network. For images with normal or mild disease conditions, the BCL-UNet network outperforms MRUNet++; however, for images with severe disease conditions MRUNet++ performs better than the BCL-UNet network. For MJ and COVID-19 datasets, the dice scores using the BCL-UNet network are 0.9716 and 0.9439, respectively; whereas, for the MRUNet++ network the scores are 0.9642 and 0.9342. However, for the GMH dataset, the dice scores for MRUNet++ and BCL-UNet are 0.8778 and 0.8445, respectively.

Using MRUNet++ or BCL-UNet for lung segmentation as the first step of automated detection of pneumoconiosis not only enables more accurate segmentation of pathological lungs but also assures that fewer images get rejected at this stage because of a failed lung segmentation.

4 Deep Learning-Based Pneumoconiosis Detection

We have continuously made efforts to improve the accuracy of pheumoconiosis detection by exploring state-of-the-art machine learning methods. Aiming at high detection accuracy, we have developed a Masked Attention CNN (Convolutional Neural Network), and introduced Multiple Instance Learning (MIL) to retain the original radiograph resolution and image details in machine learning for the pheumoconiosis detection. In this chapter, we report these methods, datasets used, experimental results and findings.

4.1 The Masked Attention CNN for Pneumoconiosis Detection

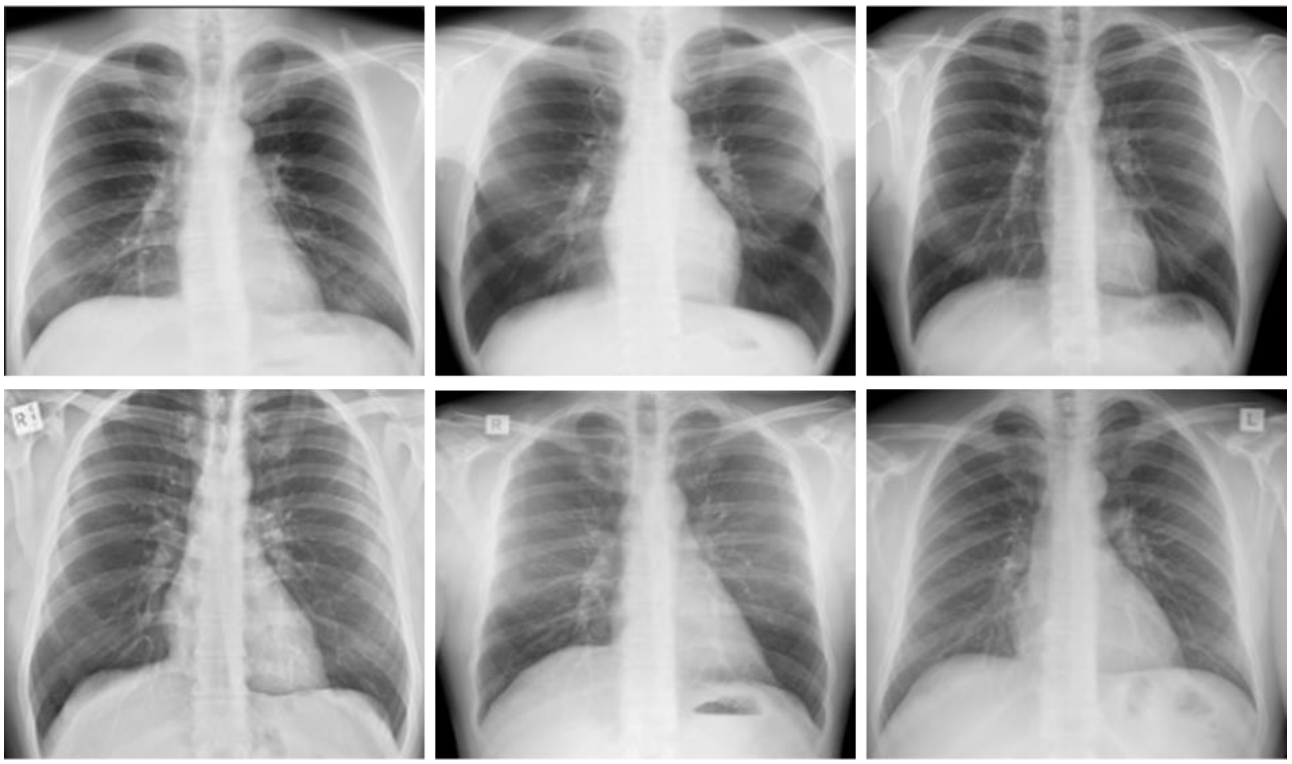
Traditionally, radiologists determine pneumoconiosis based on the similarity between a chest radiograph to be classified and the standard ILO chest radiographs in terms of image features of lesions. The image features of lesions in the pneumoconiosis chest radiograph mainly include small opaque proliferation, size and shape, small opacity aggregation into larger opacities, and pleural plaques in the lung field. Figure 4-1 shows the difference between normal and pneumoconiosis chest X-rays.

4.1.1 Method

We used the EfficientNet-b0 as the baseline to train a two-class classification model (normal vs. pneumoconiosis). The EfficientNet [27, 28] is a novel deep learning model with a scaling method that uniformly scales up all dimensions of depth, width, and resolution of Convolutional Neural Networks (CNNs) using a simple yet highly effective compound coefficient. Unlike conventional practice that arbitrarily scales these factors, the EfficientNet can uniformly scale the network width, depth, and resolution with a set of fixed scaling coefficients. EfficientNet has so far outperformed most general image classification models in both classification accuracy and efficiency. Figure 4-2 shows the EfficientNet architecture.

Understanding how a machine learning model makes a decision is important in the era of deep learning, and the GradCAM [29] is one of methods to visualize the outcome produced by the model. The method can be used to highlight the portion of image responsible for the model's decision. However, the machine learning model we trained does not always make decisions based on the features in lung fields. Figure 4-34-3 shows some example activiation maps which spread over the whole X-ray image, including the non-lung-field background masked as black regions in the top row images. Because the opacity only appears in the lung fields, the activation on the background doesn't make sense and means that the CNN is overfitting to the irrelevant image features such as the lung shape and heart size etc.

(a). Normal radiographs



(b). Pneumoconiosis radiographs

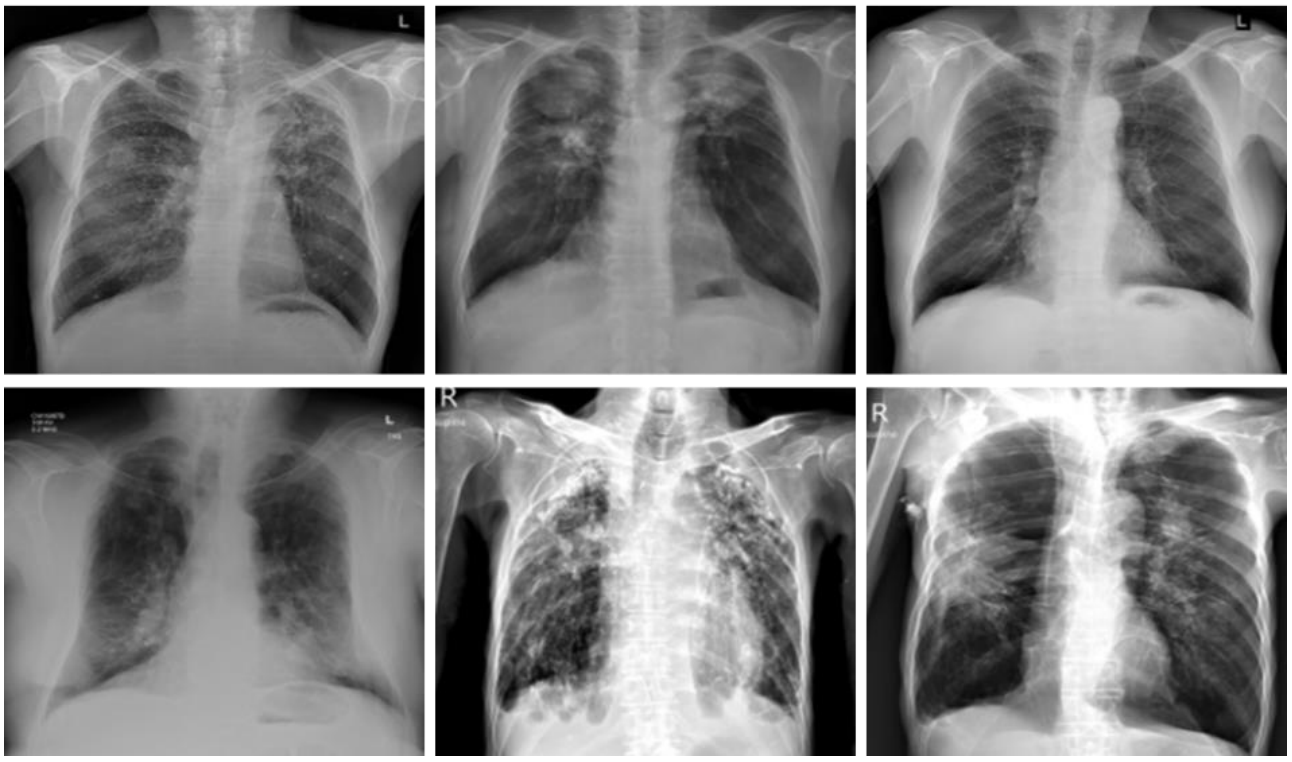


Figure 4-1 Samples of the pneumoconiosis dataset: (a) normal cases without pneumoconiosis; and (b) cases with pneumoconiosis

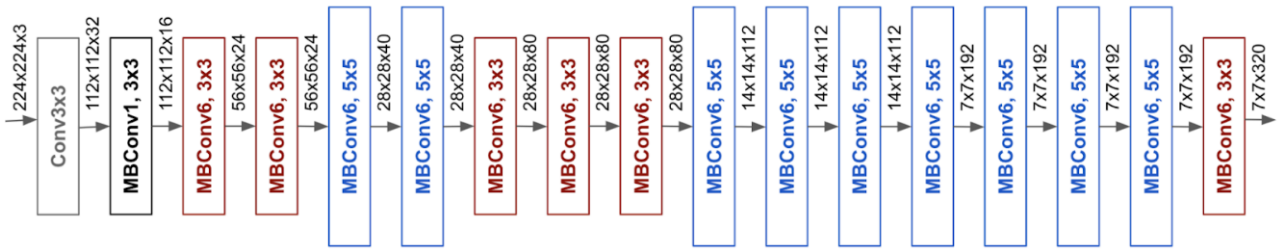


Figure 4-2 The EfficientNet architecture [35]

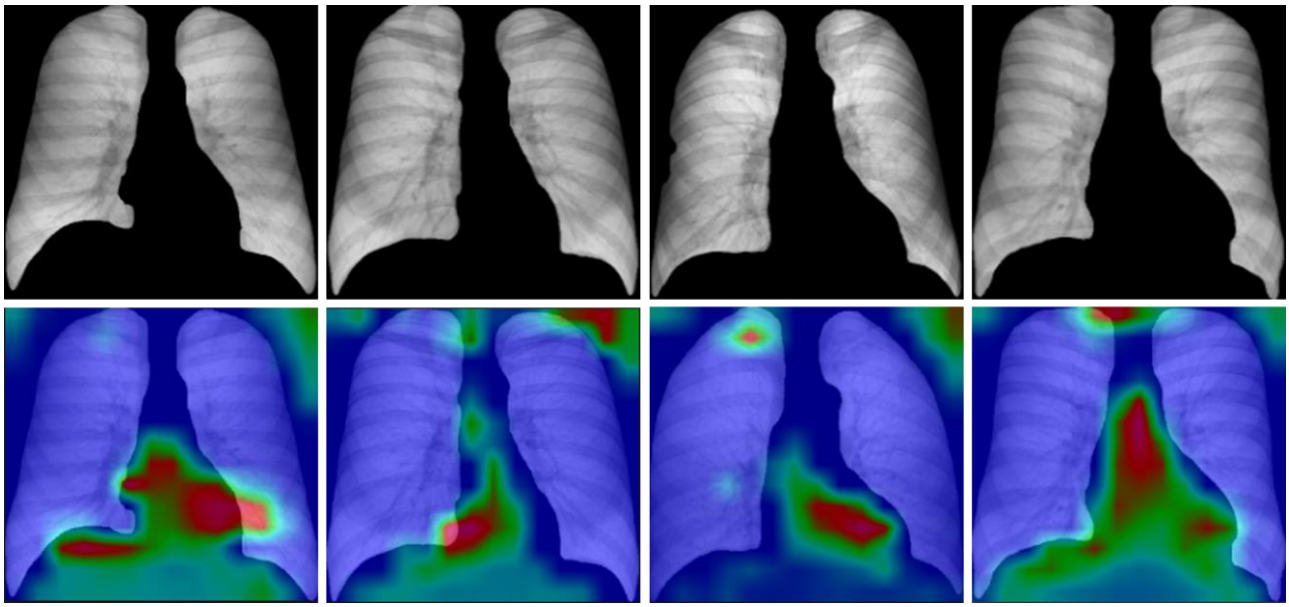


Figure 4-3 The activation map from CNN to the masked radiographs. Red region means higher activation from the CNN.

To address this problem, we proposed the Masked Attention CNN (MA-CNN). The workflow of the proposed method is illustrated in Figure 4-4. As the first step, we use the segmentation model described in Chapter 3 to generate the lung field masks of the input images. The input images are then masked, cropped, and reshaped to 448 x 448 using the lung field mask. The masked images are then fed into the CNN which produces k output activation maps (k is the number of classes, in our case, k = 2). The activation maps are masked with the lung-field masks to produce the masked attention maps. We use the categorization attention maps, and the masked attention maps to calculate the Attention Map Loss ($Loss_{am}$). The $Loss_{am}$ is the mean square loss between two same-sized attention maps (Att1 and Att2), which is defined as:

$$Loss_{am}(\text{Att1}, \text{Att2}) = \sum_{i,j,k}^{c,h,w} \left(\text{Att1}_{i,j,k}, \text{Att2}_{i,j,k}\right)^2 / c/w/h \quad (1)$$

where *c, w, h* are the channel number, width and height of the attention maps. In the other aspect, the square root of categorisation activation maps are normalized using L2 norm and classified using the $Loss_{cls}$, which is the typical cross-entropy classification loss between the output class scores and the one-hot class label. The final loss function can be depicted as follows:

$$\text{Loss} = Loss_{am} + Loss_{cls} \quad (2)$$

The final loss function combined the constraint from the one-hot class label and the factor that the activation of the CNN should be concentrated on the lung field. Our experiments illustrated the effectiveness of the Masked Attention CNN structure.

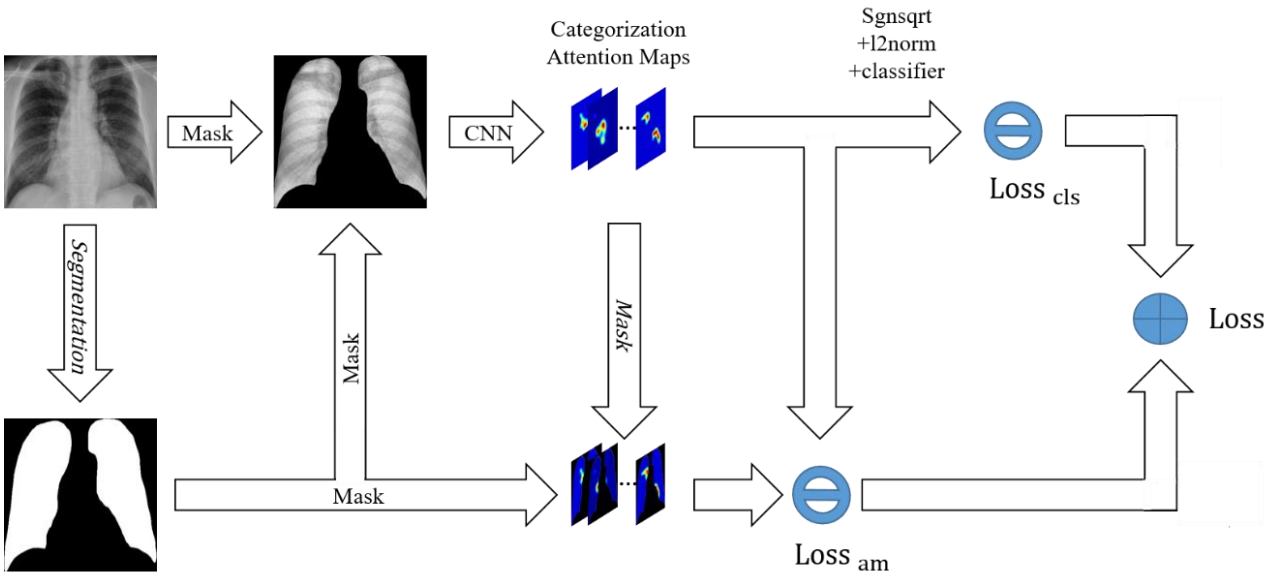


Figure 4-4 The workflow of the proposed masked CNN for black lung detection. The input image is masked using the segmentation method described in Chapter 3, and the output attention maps are masked to produce attention supervision with the attention map loss $Loss_{am}$, which is added to the class loss $Loss_{cls}$ to generate the final loss.

4.1.2 Experimental setup

We conducted experiments using the pneumoconiosis dataset described in Chapter 2, including 609 normal images and 683 pneumoconiosis images. We masked the images using the segmentation method described in Chapter 3, cropped the images using the outer bounding boxes of the lung masks, and reshaped them to the resolution of 512 × 512. To cross validate the experimental results, we randomly split the dataset into five folds and used four folds for training and one fold for testing for each experiment.

All experiments areimplemented in Python using the PyTorch platform for machine learning [30]. We use the ”RandomResizedCrop” function of PyTorch to augment all input images to the resolution of 448 × 448 for training. For testing, we resize the input image to 512 × 512 and center crop the image to 448 × 448. The training batch size is 8, and the weight of decay is 0.00001. For all layers of the proposed machine learning model, the initial learning rate is 0.01. We conducted testing at the end of each training epoch. If the testing accuracy does not increase for seven epochs, we reduce the learning rate by a factor of ten. If the learning rate is lower than 0.00001, we terminate the training process.

The evaluation metrics used in the study include the classification accuracy, sensitivity, specificity, and F1-Score as defined below:

$$Acccuracy = \frac{TP}{TP+FP} \tag{3}$$

$$Sensitivity = \frac{TP}{TP+FN} \tag{4}$$

$$Specificity = \frac{TN}{TN+FP} \tag{5}$$

$$F1\ score = 2 \times \frac{Accuracy \times Sensitity}{Accuracy + Sensitity} \quad (6)$$

where TN, TP, FP, and FN stand for True Negative, True Positive, False Positive, and False Negative, respectively.

4.1.3 Results

The following table shows the results produced by the proposed MA-CNN and the baseline CNN. The last two rows show the average improvement and the standard deviation of the improvement. The MA-CNN outperformed 0.13 ± 0.55%, 0.51 ± 1.14%, 0.45 ± 1.00%, and 0.30 ± 0.27% over the baseline CNN structure on Sensitivity, Specificity, Accuracy, and F1-Score for the five-fold cross validation.

Table 4-1 Comparison of experimental results produced by the proposed MA-CNN and the baseline CNN. The improvement row shows the changes the MA-CNN made over the baseline CNN. StanDev is the standard deviation of the improvements on the five folds.

		TP	TN	FP	FN	Sensitivity	Specificity	Accuracy	F1-Score
CNN	Fold-1	133	120	1	5	96.38%	99.17%	99.25%	97.79%
	Fold-2	124	113	6	4	**96.88%**	94.96%	95.38%	96.12%
	Fold-3	127	133	0	8	94.07%	100.00%	100.00%	96.95%
	Fold-4	133	111	4	3	97.79%	96.52%	97.08%	97.44%
	Fold-5	140	120	1	6	95.89%	**99.17%**	**99.29%**	97.56%
	SUM	657	597	12	26	96.19%	98.03%	98.21%	97.19%
MA-CNN	Fold-1	134	120	1	4	**97.10%**	99.17%	**99.26%**	**98.17%**
	Fold-2	123	116	3	5	96.09%	**97.48%**	**97.62%**	**96.85%**
	Fold-3	127	133	0	8	94.07%	100.00%	100.00%	96.95%
	Fold-4	133	112	3	3	97.79%	**97.39%**	**97.79%**	**97.79%**
	Fold-5	141	119	2	5	**96.58%**	98.35%	98.60%	**97.58%**
	SUM	658	600	9	25	**96.34%**	**98.52%**	**98.65%**	**97.48%**
Improvement		+1	+3	-3	-1	**0.13%**	**0.51%**	**0.45%**	**0.30%**
StanDev		±0.75	±1.36	±1.36	±0.75	±0.55%	±1.14%	±1.00%	±0.27%

We visualized the output attention maps of the proposed MA-CNN, and compared them with the output of the baseline CNN to verify the effectiveness of the mask attention constraint we proposed and applied to the machine learning model. Figure 4-5Figure 4-5 shows that the activation of the baseline CNN model concentrates more on the image background regions and is sparser than that of the proposed MA-CNN. The visual comparison between the activation maps produced by the baseline CNN model and the proposed metwork shows that the irrelevant activation has been suppressed using the proposed MA-CNN, and the activation of the proposed model is more focused on the lung fields.

Differently from the traditional methods that directly apply CNN to pneumoconiosis detection, the MA-CNN customises the CNN training phase by designing the constraint to force the CNN to learn from features from lung fields instead of the image background, lung shape, heart size, or any other irrelevant features. The experimental results show steady improvement in both the detection accuracy and network activation.

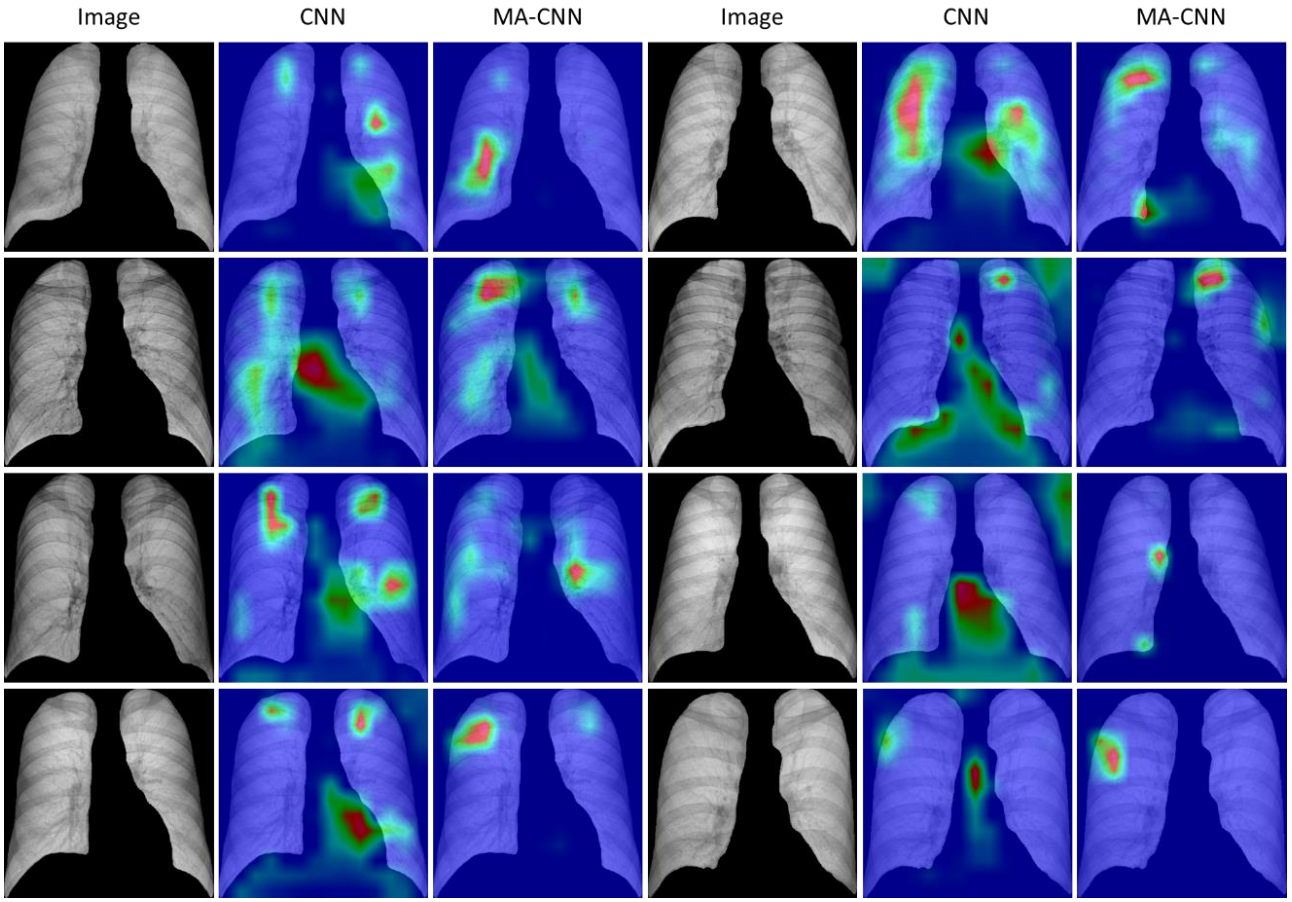


Figure 4-5 Visualization and comparison of the activation maps from the baseline CNN and the proposed MA-CNN. The Image column shows example original input images. The CNN column and MA-CNN column illustrate the activation maps from the baseline CNN and MA-CNN, respectively.

4.2 Multiple Instance Learning for Pneumoconiosis Detection

We also investigated Attention-based Multiple Instance Learning (A-MIL) [31] to evaluate its effectiveness for pneumoconiosis detection. With this method, each chest X-ray image needs to be split into small square image patches (instances) to make a bag of patches for the whole image. This bag of image patches acts as a batch in the training and testing phases. Firstly, the A-MIL architecture makes bags of input images by splitting each image into small patches. Then, the patches are passed to the feature extractor of the A-MIL model, which consists of a convolutional neural network block and several fully connected layers. Then, the instance-level features are passed to the classifier to get the instance-level attention weights. The attention weights are further used for attention aggregation to get the bag-level features. We apply a fully connected layer for the final classification of a chest X-ray into either the pneumoconiosis or normal class. The workflow of the A-MIL in our experiments is illustrated in Figure 4-6.

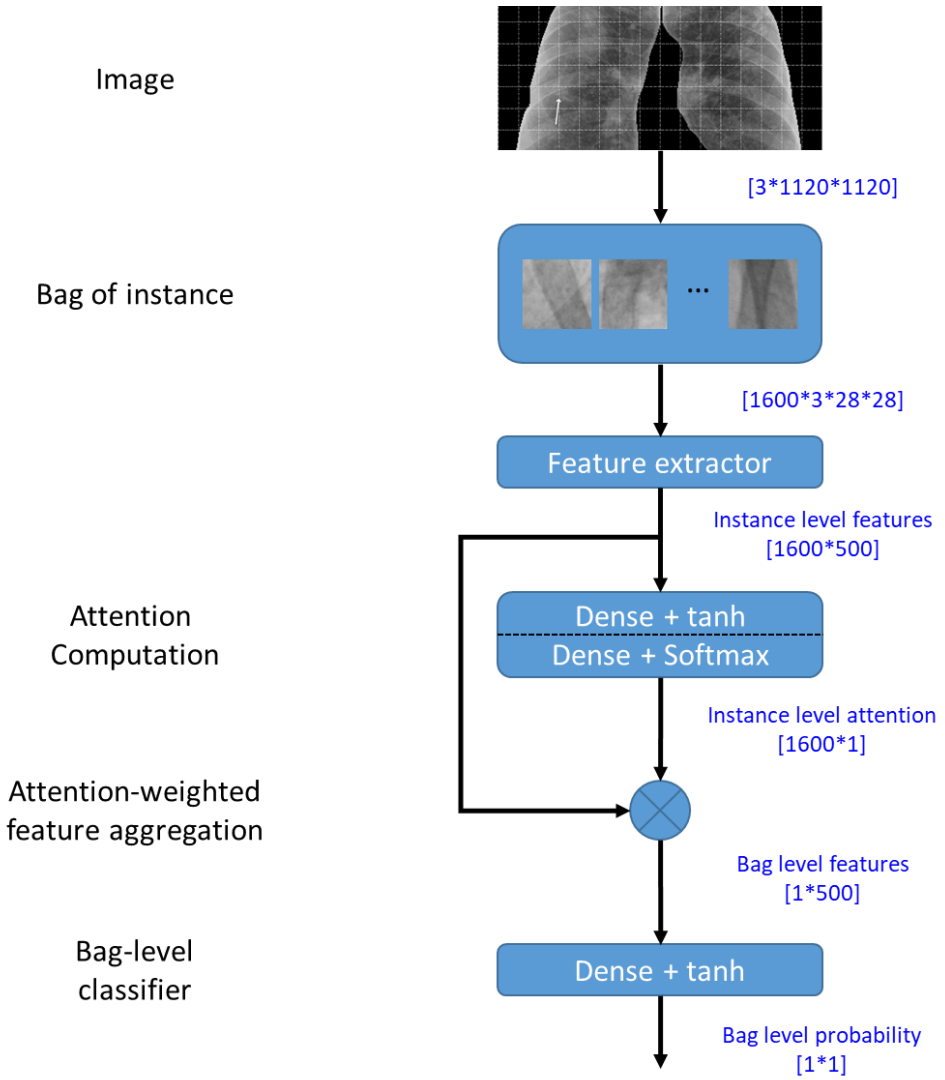


Figure 4-6 The framework of the A-MIL model.

4.2.1 Multiple Instance Learning (MIL)

Multiple Instance Learning (MIL) [32] provides the solution for a weakly supervised learning problem. In MIL, the task is to predict a classification label of a bag, which consists of multiple instances. If $\mathcal{D} = \{\mathrm{X}_1, \mathrm{X}_2, \dots, \mathrm{X}_\mathrm{n}\}$, where X_i's are the bags in dataset $\mathcal{D}$, and one bag contains m instances, i.e., $\mathrm{X}_i = \{i_1, i_2, \dots, i_\mathrm{n}\}$ where i_j is j-th instance with a binary label y_i in a bag X_i.

The most critical part of MIL is the instance-level pooling. Instance level pooling aggregates instance level features to obtain bag level features. The most popular instance pooling operations in MIL are the mean pooling and max pooling. Mean pooling operations average over all the instances to predict the bag label, whereas max pooling operations take the maximally activated instance label as the bag label. Both max pooling and mean pooling have their disadvantages. Max pooling only accounts for the maximum activation, which may be an outcome of an outlier. On the other hand, mean pooling weighs every instance equally, thus losing the information from the sparsely populated classes. In the method we selected, instance-level features $\mathrm{h}_1, \mathrm{h}_2, \mathrm{h}_3, .., \mathrm{h}_\mathrm{m}$ are pooled by taking their weighted average as shown in Equation 7 [31]. The coefficients of weighted average pooling are learned using a two-layer neural network with softmax activation.

The expression for attention computation is given in Equation 8.

$$z = \sum_{p=1}^{m} a_p h_p, \tag{7}$$

where,

$$a_p = \frac{\exp\{w^T \tanh(V h_p^T)\}}{\sum_{j=1}^{m} \exp\{w^T \tanh(V h_j^T)\}} \tag{8}$$

and $w \in \mathcal{R}^{l\times 1}$ and $V \in \mathcal{R}^{l\times m}$. In the above equation l is the number of instance-level features and a_p is the attention weight learned by the model.

4.2.2 Bag Preparation

We cropped each chest X-ray image with the resolution of 1,120 × 1,120 in the pneumoconiosis dataset into 28 × 28 patches with stride 28. This resulted in 1,600 patches which are packed into a single bag for each of X-ray images in the dataset.

4.2.3 A-MIL framework

The overall pipeline of the A-MIL framework is shown in Figure 4-6, which is inspired by [31]. Each patch in a bag is processed through a feature extractor to get instance-level features. The dense layer extracts 500 features from each instance. The attention computation block computes the attention score using these 500 features of each instance. These attention weights are further used for attention aggregation to get the bag-level features. A-MIL allows different weights for different instances in a bag. The attention aggregation computation makes the bag highly informative for the bag-level classifier. The detailed architecture of the instance-level feature extractor used in A-MIL is illustrated in the following table.

Table 4-2 Feature extractor used in A-MIL

Input dimension	Layer
[3 × 28 × 28]	Conv1
[20 × 24 × 24]	Maxpool
[20 × 12 × 12]	Conv2
[50 × 8 × 8]	Maxpool
[50 × 4 × 4]	Flatten
800	FC
500	Extracted features

4.2.4 Experiments

We conducted experiments using the same pneumoconiosis dataset as described in Section 4.1.2, which contained 609 normal images and 683 pneumoconiosis images. Each of the images was first segmented using the method described in Chapter 3, then cropped using the outer bounding boxes of the mask, and then reshaped to the resolution of 1,120 × 1,120. All of the images have one-hot class annotation only. We randomly split the dataset into 80% for training and 20% for testing. The A-MIL framework was trained with a batch size of 1, the learning rate of 0.001, and the binary cross-entropy as the loss function. We used data augmentation such as vertical and horizontal flip, rotation by 90◦, 180◦, and 270◦. All experiments were conducted using PyTorch platform in Python.

The initial experiments show that the classification accuracy is 82.5% which is lower than the accuracy of 97.6% obtained from the MA-CNN model. We will continuously investigate A-MIL model to improve its performance.

4.3 Summary

We have investigated the state-of-the-art machine learning methods such as the Attention Multiple Instance Learning, and proposed the MA-CNN for pneumoconiosis detection using segmented lung images.

With the proposed MA-CNN, our five-fold cross validation experiments show that the proposed machine learning model has improved pneumoconiosis detection performance. We have achieved a sensitivity of 96.34%, a specificity of 98.52%, an accuracy of 98.65%, and a F1-score of 97.48%. Visualisation of the activation maps shows the proposed MA-CNN model is focused on learning discriminative features in the lung fields, instead of the background or other regions. It should be pointed out that our experiments were conducted with the limited number of normal and pneumoconiosis images, so the robustness of the proposed MA-CNN model needs to be further evaluated with larger datasets. Future work in this area will include combination of the classification network with the localization of opacity, this will be used to classify a chest X-ray into a specific ILO category.

The Attention Multiple Instance Learning (A-MIL) model can effectively utilize chest X-ray images with their original resolutions, hence, detailed features of pneumoconiosis lesions can be learnt. Different from the original deep learning-based classification models, which have to take the downsampled X-ray images as input, leading to loss of some of image details, the A-MIL-based CNN model retains the original image resolution and classifies a chest X-ray image based on the lossless features in the image. We have investigated the A-MIL model and conducted some initial experiments for the detection of pneumoconiosis. The initial experimental results show that the detection accuracy is promising although it is lower than that of the proposed MA-CNN model. Future works to improve the A-MIL model may include:

1) Finetuning the training parameters for the A-MIL model, such as patch resolution, model depth, feature dimension, and data argumentation methods.
2) Improving the A-MIL model to accommodate flexible bag sizes – our initial experimental results show that this can further improve the detection accuracy. By removing patches without lung tissue in the bag of a chest X-ray image, the bag size, i.e. number of patches in the bag, will be reduced. Because the number of patches with lung tissue in different chest X-rays can be different, the bag sizes of images can be different. Another scenario is that the resolutions of chest X-ray images can be different because they may be acquired using different X-ray machines and with different settings. This will also result in different bag sizes. To develop the improved A-MIL model, patch level annotations in the training dataset are required.
3) Combining the bag level one-hot class annotation and the patch-wise annotation – this may enforce the A-MIL model to learn features of the pneumoconiosis lesions from pathology patches, and therefore improve the detection accuracy.

5 Deep Learning-Based Classification of Radiographs of Pneumoconioses Using Chest Radiograph Zones

Limited computing resources at pneumoconiosis screening sites make it impractical to develop and apply a deep learning-based model that works on the whole high-resolution CXR images. Soluitons include either reducing the capacity of the model (e.g., shallow learning) or down-sampling the images. However, pneumoconiosis diagnosis highly depends on subtle image features, therefore compromising the feature richness by model simplification or reduction of the image resolution is not desirable. In this project we developed a method that solves the problem by dividing the lung fields into six zones, classifying each zone separately and aggregating zone classification results into an image classification score. This chapter describes datasets, experimental setup for zone-based classification and our results.

5.1 Datasets

We used the chest X-rays with four different ILO catagories from the six datasets described in Chapter 2. given in the Table 6-1. We combined all these datasets to train and evaluate the model. The number of images with ILO category 0 is 632, and whereas only 61 images belong to category 3. The number of images for category 1 and 2 are 494 and 174, respectively. Only GMH and RSHQ datasets contains images with zone label . For other datasets, we assigned image label to the each of the zone label. For example, if a image belong to category 1 then all zones also belongs to catgory 1. Some images contains both small and large opacities.

Table 5-1 Datasets used in this study

Database	ILO category				Total
	0	1	2	3	
GMH	13	119	138	47	**317**
Syllabus_19	15	31	14	9	**69**
WMI	64	15	7	1	**87**
RSHQ	0	329	15	4	**348**
CSH	505	0	0	0	**505**
StVincent	35	0	0	0	**35**
Total	**632**	**494**	**174**	**61**	**1361**

5.2 Experimental setup

All experiments were conducted using Keras with TensorFlow as backend. All models were trained using the Adam optimizer with a learning rate of 0.0001, the number of epochs 120, batch size of 4, and categorical_crossentropy loss function. Five-fold cross-validation approach have been used to

evaluate the performance of our model. Average height and width information of all zones used in this experiment is presented in Table 5-2. We resized all the zones based on average height and width information to keep the same aspect ratio. We have used two different heights, 256 and 512, for all zones, the corresponding width have been selected using the aspect ratio metioned in Table 5-2.

Table 5-2 Average height and width information of all zones used in the experiment

Zone	Avg height	Avg width	Aspect ratio	Height	Width	Height	Width
RUZ	679	851	1:1.25	256	320	512	640
RMZ	678	913	1:1.34	256	340	512	680
RLZ	677	893	1:1.31	256	330	512	660
LUZ	686	818	1:1.19	256	300	512	600
LMZ	685	838	1:1.22	256	310	512	620
LLZ	684	608	1:0.88	256	230	512	460

Figure 5-1 demonstrates the architecture of our proposed zone-based classifier using a deep learning model. Firstly, we segmented the lung fields from the CXR images using UNetRes++ model described in Chapter 3. To segment the lungs we fed the downsized images to the model, and then upsized the resulting mask images to the original image size. After segmentation, each lung field was divided into three zones by dividing the vertical distance between the lung apex and the dome of the diaphragm into three equal parts and drawing a horizontal line at each division point. For an easy reference, the algorithm assigns each zone a label, which are Right Upper Zone (RUZ), Right Middle Zone (RMZ), Right Lower Zone (RLZ), Left Upper Zone (LUZ), Left Middle Zone (LMZ), and Left Lower Zone (LLZ).

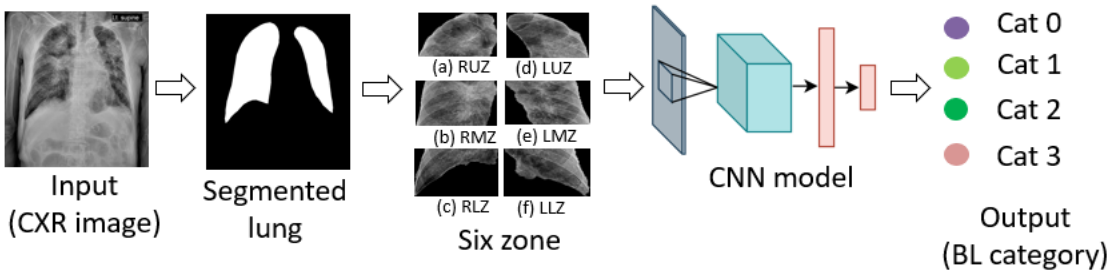


Figure 5-1 Overview of the CNN-based model for BL classification using zone labels

Six zone classifiers were trained to classify each zone of a X-ray image into an ILO category, such as a RUZ classifier. We used three CNN-based models, namely, ResNet152 [33], InceptionResNetV2 [34], and Xception [35] models to classify the profusion level of small opacities for each zone. All of these models were pre-trained on the ImageNet dataset [36]. We initialized the models with pre-trained weights and then finetuned the models using our training data. Each zone was classified as category 0, 1, 2, or 3. To obtain a classification label for the whole image, the predicted zone labels were combined in the following way: the highest category of the six zones determines the ILO category of the whole image. For example, if the six zones of an image are predicted as [RUZ: 1, RMZ: 0, RLZ: 0, LUZ: 2, LMZ: 1, LLZ: 0], the whole image is classified as ILO category 2.

Due to large variability in the appearance of pneumoconiosis images, it is difficult to train one robust network to achieve good results for all zones. Therefore, to further improve the pefromance we adopt ensemble learning where output results are combined based on majority voting between the three models.

5.3 Results

5.3.1 Zone-level classification

We have evaluated the performance of zone-based classifiers using two different zone sizes. Table 5-3 summarizes the classification results for six different zones with the zone height = 256. The zone width is selected using the aspect ratio from Table 5-2. We evaluated the performance of each model using classification accuracy, sensitivity, specificity, F1-score and Receiver Operating Characteristic curve (ROC) - AUC (Area under the Curve) score. The weighted averaged sensitivity, specificity, F1-score and AUC is calculated by taking the mean of a metric calculated per class while considering the number of actual occurrences of the class in the dataset. For multi-class classification, performances are comparable among different networks, however, Xception model achieved sligthtly better results compared to all other models in terms of all evaluation metrics. All networks achieved best results for LUZ in the left lung and the RUZ in right lung.

Table 5-3 Multi-class classification performance on zone level using three models with image height = 256

	ResNet152			InceptionResNetV2			Xception		
Zone	Accuracy	Precision	Recall	Accuracy	Precision	Recall	Accuracy	Precision	Recall
LLZ	0.7429	0.7317	0.7428	0.7339	0.7124	0.7339	0.7413	0.7315	0.7413
LMZ	0.7604	0.7469	0.7604	0.7732	0.7681	0.7732	0.7845	0.7826	0.7845
LUZ	0.8348	0.8256	0.8348	0.8317	0.8246	0.8317	0.8308	0.8262	0.8309
RLZ	0.7140	0.7170	0.7139	0.7214	0.7168	0.7214	0.7312	0.7314	0.7312
RMZ	0.7513	0.7490	0.7513	0.7543	0.7498	0.7543	0.7628	0.7495	0.7628
RUZ	0.8212	0.8104	0.8213	0.8222	0.8045	0.8222	0.8101	0.8010	0.8101
Avg.	**0.7708**	**0.7634**	**0.7708**	**0.7728**	**0.7627**	**0.7728**	**0.7768**	**0.7704**	**0.7768**

On the other hand, worst performances were achived by all networks for LLZ in the left lung and RLZ in the right lung. Peformance of LMZ is slightly better than RMZ for all models. There is a significant performance difference between upper zones and lower zones with the same model.
Table 5-4 summarizes multi-class classification performance with the zone height = 512. Like with the zone height = 256, performances of the three models are comparable for the six zones. However, Xception model achieved slightly better results compared to ResNet152 and InceptionResNetV2 models. At zone level, top zones for both lungs (LUZ and RUZ) have an outstanding performance in classification accuracy (> 0.8), while low accuracies were obserbed in the bottom two zones: LLZ and RLZ. The left middle zone achived slightly better results than the right middle zone. This could be due to that each of the right and left lungs has a hilum that lies roughly midway down the lung. When moving from the hilum to the periphery on the bottom, there is a gradual reduction of the anatomical lung markings. All these anatomical features weaken the accuracy in the two bottom zones. In contrast, the top left and right zones have more easily identifiable radiographic abnormalities in pneumoconiosis, resulting in higher accuracies. If we compare the performance

based on the zone size, we can see that the performances are comparable between the zone height of 256 and 512, though all networks achived slightly better results using the zone height of 512.

Table 5-4 Multi-class classification performance on zone level using three models with image height = 512

	ResNet152			InceptionResNetV2			Xception		
Zone	Accuracy	Precision	Recall	Accuracy	Precision	Recall	Accuracy	Precision	Recall
LLZ	0.7324	0.7084	0.7324	0.7480	0.7339	0.7480	0.7384	0.7329	0.7384
LMZ	0.7709	0.7587	0.7709	0.7656	0.7390	0.7656	0.7754	0.7628	0.7754
LUZ	0.8186	0.8091	0.8186	0.8309	0.8159	0.8309	0.8342	0.8305	0.8342
RLZ	0.7245	0.7141	0.7245	0.7244	0.7221	0.7244	0.7370	0.7351	0.7370
RMZ	0.7437	0.7467	0.7437	0.7520	0.7398	0.7520	0.7591	0.7529	0.7591
RUZ	0.8207	0.8036	0.8206	0.8244	0.8181	0.8244	0.8262	0.8166	0.8262
Avg.	**0.7685**	**0.7568**	**0.7685**	**0.7742**	**0.7615**	**0.7742**	**0.7784**	**0.7718**	**0.7784**

5.3.2 Image-level classification

The performance of pneumoconiosis classification at image level with the zone height of 256 is depicted in Table 5-5. Similarly to zone level classification results, Xception model outperforms all other models for binary classification in terms of all evaluation metrics. For multi-class classification, InceptionResNetV2 model outperforms ResNet152 and Xception in terms of accuracy, sensitivity and F1-score; however, Xception model achieves slightly better results for specificity and AUC. To further improve the performance we developed the ensemble model based on majority voting from these three models. The ensemble model offers a higher performance than the individual models for pneumoconiosis detection and classification in terms of all metrics.

Table 5-5 Pneumoconiosis detection and classification performance using image height = 256

	Binary-class classification					Multi-class classification				
Models	Accuracy	Precision	Recall	F1-score	AUC	Accuracy	Precision	Recall	F1-score	AUC
ResNet152	0.8626	0.8643	0.8626	0.8619	0.795	0.7029	0.7355	0.7029	0.7112	0.793
Inception ResNetV2	0.8581	0.8633	0.8582	0.8568	0.796	0.7133	0.7413	0.7133	0.7219	0.797
Xception	0.8759	0.8787	0.8760	0.8751	0.800	0.7030	0.7447	0.7030	0.7170	0.798
Ensemble model	**0.8921**	**0.8930**	**0.8922**	**0.8922**	**0.829**	**0.7561**	**0.7715**	**0.7561**	**0.7588**	**0.826**

Table 5-6 presents the performance of pneumoconiosis classification at image level with the zone height of 512. Simarly to the zone height of 256, for pneumoconiosis detection, Xception model outperforms all other models in terms of all metrics. However, for multi-class classification ResNet152 outperforms other two networks in terms of accuracy, sensitivity and F1-score. Xception model achieved best results for specificity and AUC. The ensemble model outperforms all three individual networks. However, the performance of ensemble model for different zone heights are comparable. The ensemble model performs slightly better with the zone height of 256 compared to the zone height of 512 for pneumoconiosis detection. For multi-class classification, the ensemble model performs slightly better with the zone height of 512 compared to the zone height of 256 in terms of accuracy and sensitivity. In terms of specificity, F1-score and AUC the same model performs better with the zone height of 256 compared to the zone height of 512.

.

Table 5-6 Pneumoconiosis detection and classification performance using image height 512

	Binary-class classification					Multi-class classification				
Models	Accuracy	Precision	Recall	F1-score	AUC	Accuracy	Precision	Recall	F1-score	AUC
ResNet152	0.8515	0.8524	0.851	0.8515	0.7939	0.7274	0.7376	0.7274	0.7272	0.7995
Inception ResNetV2	0.8589	0.8646	0.858	0.8576	0.7992	0.7184	0.7427	0.7185	0.7220	0.7990
Xception	0.8700	0.8701	0.870	0.8698	0.8032	0.7140	0.7377	0.7141	0.7198	0.7999
Ensemble model	**0.8811**	**0.8837**	**0.881**	**0.8812**	**0.8238**	**0.7599**	**0.7632**	**0.7599**	**0.7566**	**0.8215**

The average confusion matrix over the five folds for multi-class classification using the ensemble model and the zone height of 512 is presented in Table 5-7. The performance of pneumoconiosis classification could be affected by the presence of large opacities in some zones, so we have exclude some zones from our experiment, therefore, for ILO category 2 and category 3 we had smaller number of samples to train the model. Secondly, although we had the zone-wise ground truth for most training images, zone labels were assigned to their image level labels for a small number of training samples,. This could impact the ability of the models to learn from the training data. Finally, some zones contained artefacts that may negatively affect the classification performance. Examples of some zones with artefacts are presented in Figure 5-2.

Table 5-7 Confusion matrix for multiclass classification for ensemble model using zone height = 512

Predicted class	True class			
	Class 0	Class 1	Class 2	Class 3
Class 0	113.	10.8	0.8	0.
Class 1	19.6	64.8	12.4	2.
Class 2	0.8	7.6	23.2	3.4
Class 3	0.2	0.4	7.	4.6

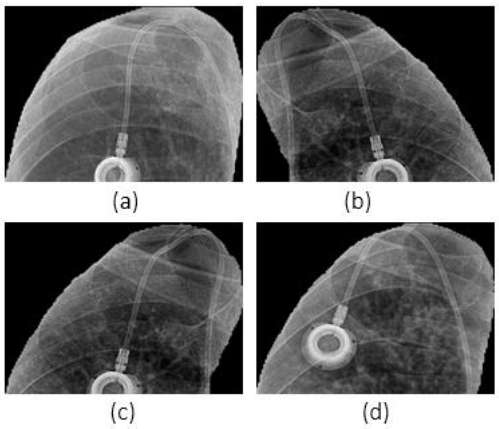


Figure 5-2 Example of some zones with artifacts

5.3.3 Ablation study

Additionally, we have conducted the same classification experiment directly on the whole images omitting the zone-level classification step and the fusion of the obtained zone labels into an image label. The whole chest X-ray image was resized to 580 x 512 and fed into the CNN model as shown in Figure 5-3. The three network architectures with the same hyperparameters as given in Section 5.2 were used. Table 5-8 displays the results for binary and multi-class classification obtained in this experiment. Compared with the results in Tables 5-5 and 5-6, the whole image-based classification performed better than the zone-based one, using the same models and datasets, for both binary and multi-class setups. This outcome further highlights the need of accurate labels for training data, such as accurate zone labels in our case.

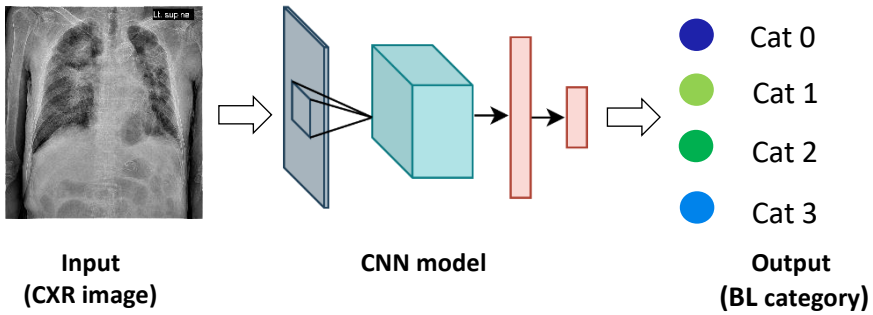


Figure 5-3 Overview of the CNN-based model for BL classification without zone labels

Table 5-8 Pneumoconiosis classification performance using three CNN-based models on the whole image

	Binary-class classification				Multi-class classification			
Models	Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score
Xception	0.9349	0.9360	0.9350	0.9349	0.8196	0.8220	0.8196	0.8169
ResNet152	0.9401	0.9408	0.9402	0.9400	0.8204	0.8228	0.8204	0.8183
IncepV3	0.9291	0.9296	0.9291	0.9290	0.8130	0.8186	0.8130	0.8121

5.4 Summary and future work

In this chapter, we applied three different network architectures to develop the zone-based deep learning models, namely, ResNet152, InceptionResNetV2 and Xception. The models were trained on the two different zone sizes. The performances of the three network architectures were comparable for zone- and image-level classification. For all models, the upper lung zones achieved the best results and the lower lung zones achieved the worst results for both lungs.

Performances were also comparable for the zone heights of 256 and 512. To further improve the classification performance, we utilized an ensembling technique based on maximum voting among the three networks architectures. For both binary and multi-class classification and the two different zone sizes, the ensemble model outperformed all three individual networks significantly in terms of

all evaluation metrics. The whole image-based classification, however, outperformed, the ensemble model, possibly, due to the absence of inaccurate zone labels for training the models for part of the dataset.

To improve the zone-based classification performance, future works may include:

1) Generating some artificial images with ILO category 2 and 3 to solve the problem of data imbalance;
2) Removing the artefacts of the images, or excluding images with the artefacts; and
3) Obtaining zone-wise annotations of all CXR images.

6 Deep Learning-Based Classification of Radiographs of Pneumoconioses Using Multiscale CXRs

In this chapter we propose a novel multi-scale network structure, Multi-Scale CNN (MS-CNN), for pneumoconioses categorisation into four classes (ILO categories 0, 1, 2 and 3). MS-CNN is based on the EffecientNet-b0 model described in Chapter 4, which can enhance the ability of deep learning models to capture discriminative detail features in high-resolution images. Additionally, we investigate the effect of different input image resolutions on the performance of EffecientNet-b0, the effect of transfer learning using different publicly available large image datasets, and the effect of various image pre-processing techniques on the classification perfromamce. The final experiments show that the proposed MS-CNN can significantly improve the classification performance of the baseline model, EffecientNet-b0, and outperforms state-of-the-art deep learning model, CheXNet.

6.1 Input image resolution

The deep-learning-based pneumoconiosis radiograph classification suffers from the dilemma between highly detailed features, like small opacity and pleural plaques in the lung field, and the very low resolution that a typical deep convolutional network can effectively use. For example, VGG [37] is only capable of ingesting 224 x 224 three-channel 8-bit input images, and ResNet usually uses 224 x 224 and can be transferred into 384 x 384 and 448 x 448 images [38]. For clinical diagnosis with chest X-ray images doctors often use the standard full resolution images for the diagnosis of pneumoconiosis, which are stored in Digital Imaging and Communications in Medicine (DICOM) libraries [39]. DICOM images usually contain more than 2,000 pixels in each dimension and three-channel 16-bit grayscale for each pixel, which is difficult to be fully utilized in the current deep learning networks.

Figure 6-1 shows the relationship between input resolution, the four-class classification accuracy of EffecientNet-b0, and the standard deviation for five-fold validation. The EfficientNet-b0 model was explained in Chapter 4. Its structure is designed for 224 and 384 square input images and degrades when the input resolution is very high (>800). The best four-class classification accuracy can be achieved when the input size is between 600 to 700 square. The lowest deviation is for 400 to 500 square. The accuracy decreased, and the standard deviation increased when the input resolution was higher than 700 square. For pneumoconiosis detection and classification, we expect the deep learning model can operate on higher resolution images so that the more detailed image features such as pneumoconiosis lesions could be extracted.

Figure 6-1 The relationship between the four-class classification accuracy of EffecientNet-b0 on the pneumoconioses image dataset and the input resolution. The green ranges are the standard deviation of the five-fold validation.

6.2 Transfer learning using large-scale chest X-ray image dataset

Recent works using deep neural networks to classify CXR images [14] usually use the model parameters pretrained on ImageNet [36]. For our application, we use a comparably smaller-scale image dataset with 1,345 chest X-ray images of different ILO categories of pneumoconiosis. The domain gap between ImageNet and our chest X-ray dataset is wide because ImageNet images are natural three-channel (colour) images while, iin contrast, our dataset only contains high-resolution grayscale chest X-ray images. The wide domain gap usually causes the model representation to degrade and make training of the model more difficult to converge. In this section, we utilised a bridging CXR dataset to mitigate the problems caused by the dataset domain gap.

6.2.1 Transfer learning using ChestX-ray14

To address the domain gap between the ImageNet and our pneumoconiosis dataset, we introduced another chest X-ray image dataset, ChestX-ray14 [14], as a bridge to perform transfer learning from the ImageNet to the pneumoconiosis dataset. ChestX-ray14 is a publicly available dataset comprising 112,120 frontal-view chest X-ray images of 30,805 unique patients (collected from 1992 to 2015) with the text-mined fourteen common disease labels. The ChestX-ray14 images are grayscale and reshaped to 1,024 square dimension. The images are visually similar to the images in our pneumoconiosis dataset, which means that it is an ideal bridging dataset for transfer learning.

The transfer learning procedure and the image samples from ImageNet, ChestX-ray14, and our pneumoconiosis dataset are illustrated in Figure 6-2. We use the EfficientNet-b0 model pretrained with ImageNet to fine-tune a fourteen-class classification model on the ChestX-ray14 dataset with different input image sizes (1,024 x 1,024 and 384 x 384). Then we use the 14-class classification model to train the models for pneumoconiosis classification.

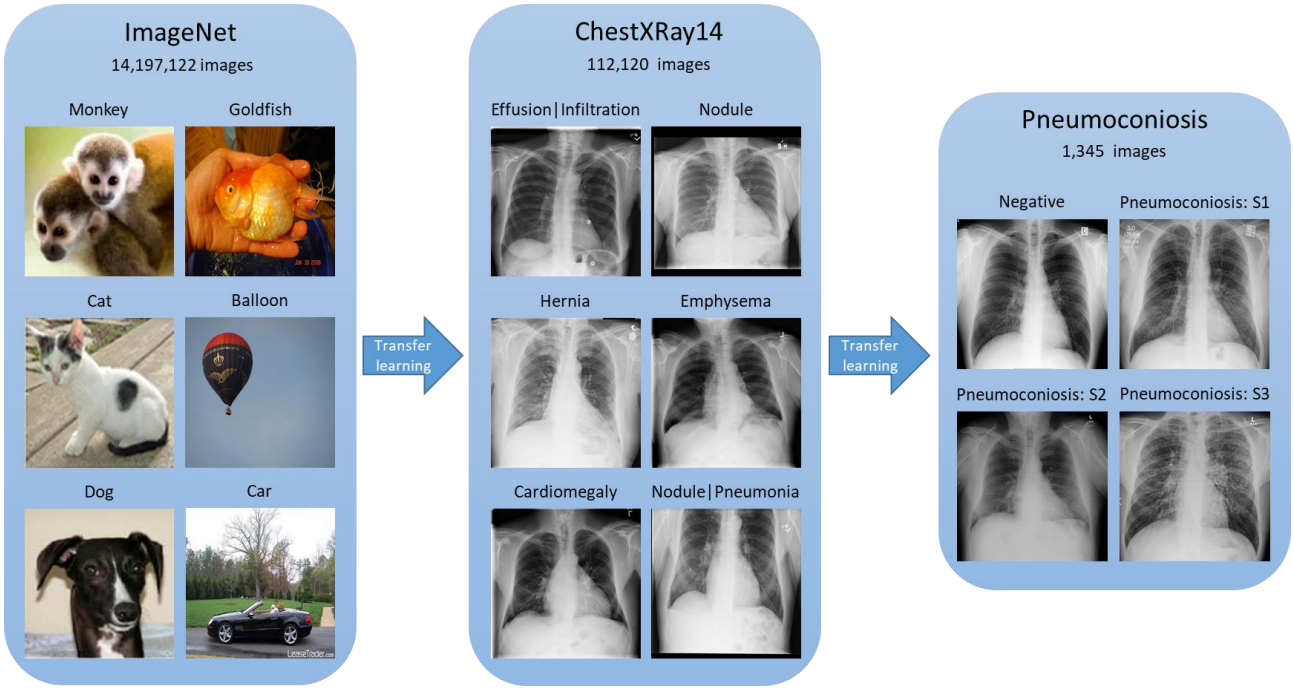


Figure 6-2 The transfer learning procedure from ImageNet bridging with ChestX-ray14 to our proposed Pneumoconiosis dataset.

6.2.2 Experiment

Experiments were conducted using the proposed transfer learning on ChestX-ray14 compared with the pre-trained model on the ImageNet. We used three different input image resolutions with five-fold cross-validation for a fair comparison. The training setup is the same, except for the initial model and the input resolution. All experiments were conducted using the Pytorch platform and implemented in Python. The training batch size is 16. We used Adam optimizer and Cosine scheduler (T_0 = T_mult = 2) for learning rate adjustment. For all layers, the initial learning rate is 0.001. We conducted evaluation at the end of each training epoch and recorded the best classification accuracy in 100 training epochs. We used 0.05 brightness and 0.05 contrast jittering, 5 degrees random rotation, random resize crop to 50% of the original input image size, and random horizontal flipping as augmentation followed by image normalization with mean = 0.5 and std = 0.25.

Table 6-1 Comparison of pneumoconiosis classification accuracies with different transfer learning datasets and different input image dimensions

	resolution	Fold					AVE	DEV
		1	2	3	4	5		
Pre-trained on ImageNet	224	84.44	80.37	79.93	82.77	85.02	82.51	5.34
	384	85.19	79.63	80.67	83.90	83.90	82.65	5.64
	448	85.19	81.11	82.90	84.64	83.52	83.47	2.55
Pre-trained on ChestX-ray14	224	85.19	82.96	80.67	82.02	86.52	**83.47(+0.94)**	5.61(+0.27)
	384	84.07	81.11	81.41	83.52	85.02	**83.03(+0.38)**	**2.90(-2.74)**
	448	84.81	84.44	84.01	85.39	83.90	**84.51(+1.04)**	**0.38(-2.17)**

The experimental results are shown in Table 6-1. The model pre-trained on ChestX-ray14 achieved better results than that pre-trained on ImageNet for all the input image resolutions. The average four-class classification accuracy improvement is 0.79%, which is repeatable and statistically significant. Another notable improvement is that the deviation (Dev) is lower using the model transferred from ChestX-ray14, which means that the model is more robust to data perturbations.

6.3 Data preparation

Before the images are processed by the CNN model, we need to do some pre-processing to reveal their discriminative features and eliminate confusing information. We investigated different techniques to pre-process images, including lung field mask application, cropping/resizing, and whether to keep the original ratio aspect of the images. Numerous experiments have been conducted to explore the optimal pre-processing techniques for the pneumoconiosis X-ray image classification.

6.3.1 Lung Mask

The original CXR images are grayscale, and pneumoconiosis lesions only appear in the lung fields, which means that the rest of the image contains less relevant information than the lung fields. Given such a priori knowledge, we performed the lung field segmentation and generated a lung mask using the method described in Chapter 3. We tried two methods of applying masks to the original CXR images so that the deep learning model knows where to pay its attention. The first method is to apply the lung field mask on an image and discard the background by setting the pixel values to “0”, which we name “Mask”. The second method is to apply the mask and discard the background in only one of the three input channels, which makes the image artificially colored, and we call it “Color” in the rest of the report. We compared the classification performance of the two masking methods with the original images in this section.

6.3.2 Image resizing

The original dimension of the images in the pneumoconiosis datasets ranges from 1,767 square to 5,376 square. Some image shapes are not square, so we need to normalize the image resolution. For data preparation, we tried four different pre-processing methods and their combinations. As shown in Figure 6-3, with the original image and its lung field mask, we pad it into a square and keep its aspect ratio without cropping or resizing (the second column of Figure 6-3). Or we can reshape the image into a square instead (the fourth column of Figure 6-3) without considering the original aspect ratio, named “Reshape”. The other way of pre-processing is to crop its lung region with the bounding box, named “Crop” (the third column of Figure 6-3). We have four combinations of image resizing methods. Combined with the three types of masking methods described in Section 6.3.1, we tried 12 methods for data pre-processing as shown in Figure 6-3.

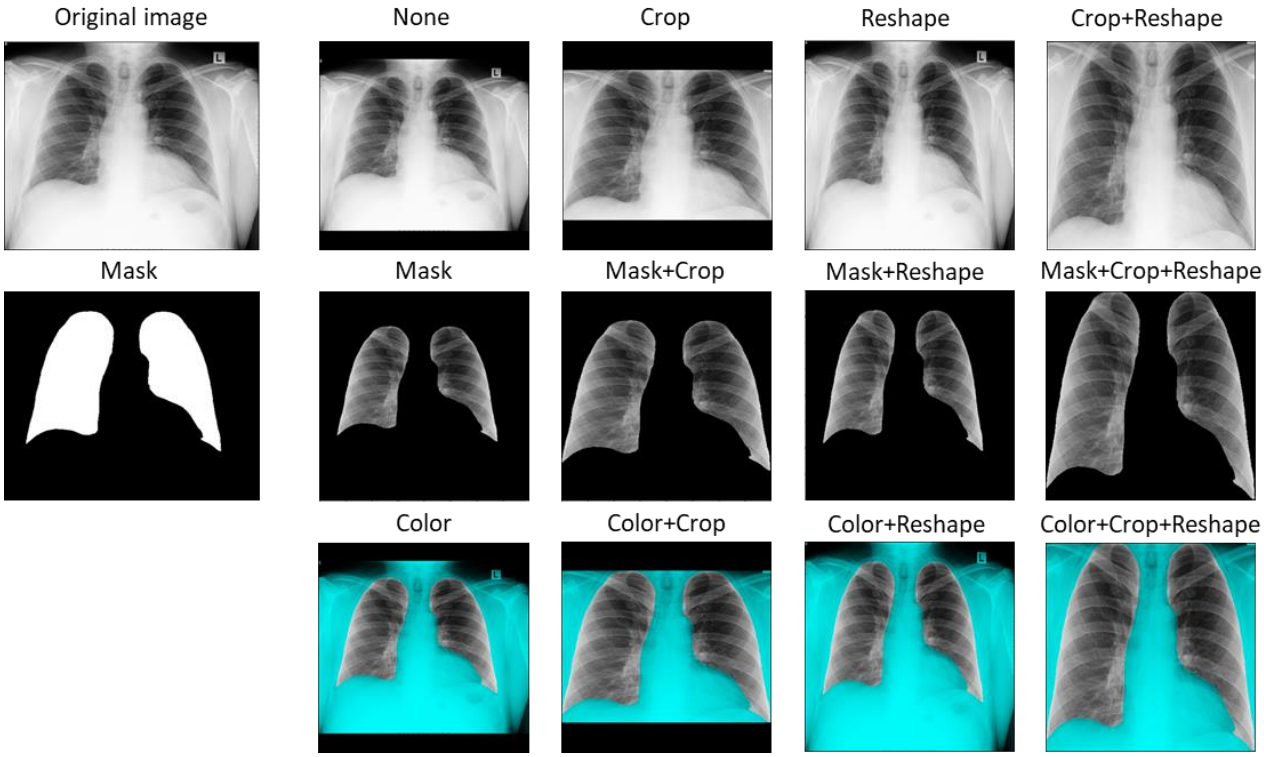


Figure 6-3 Samples of the combination of different image pre-processing methods.

6.3.3 Experiment

The experimental setup, hyperparameters and data augmentation are the same as described in Section 6.2.2. We used previously described masking and reshaping methods to train a model using 448 x 448 input resolution with five-fold cross-validation. We calculated the average four-class classification accuracy and its standard deviation to verify the effectiveness of each data pre-processing method. Higher accuracy means higher performance and better representation ability of the model trained with the correspondent dataset, while lower deviation usually means higher robustness over data perturbation.

The results are listed in Table 6-2. Colored images without reshaping or cropping achieved the best results in the five-fold cross-validation experiment, and the best overall classification accuracy. Colored cropped images produced the second-best overall performance. On the opposite side, masked images achieved the lowest classification accuracy, which means that the background can provide some useful information, such as image quality, contrast, brightness, etc., to the model and assist the model in making decisions.

We have evaluated the average performance of each pre-processing step with the original image. The results shown in Table 6-3 demonstrate that “Mask” is not improving the classification result. The reason could be that it discards useful background information. Reshaping is also not helpful, for it distorts the original image by changing the aspect ratio, making pneumoconiosis lesions more difficult to detect. Replacing one image channel with the lung field mask (“Color”) and cropping using segmentation bounding boxes improves the classification accuracy significantly because it keeps the hint from the background and enhances the lung field information. In the rest of this chapter we will use the images with color and cropping pre-processing unless noted otherwise.

Table 6-2 Comparison between different data pre-processing techniques and their combination

	Reshape	Crop	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	AVE	DEV
None	×	×	85.19	78.15	81.78	**83.52**	**85.77**	82.88	9.42
	√	×	**85.56**	80.74	81.41	**83.52**	84.27	83.10	4.01
	×	√	84.44	81.48	**83.27**	84.27	84.64	**83.62**	**1.71**
	√	√	84.81	**82.22**	81.41	83.15	83.90	83.10	**1.80**
Mask	×	×	83.70	80.00	78.44	79.03	82.40	80.71	5.07
	√	×	82.96	81.11	79.18	80.90	82.77	81.38	**2.39**
	×	√	83.33	80.37	81.78	79.03	83.90	81.68	4.10
	√	√	84.07	80.00	80.3	81.27	82.4	81.61	2.77
Color	×	×	**85.93**	**82.22**	**82.16**	83.15	**85.39**	**83.77**	3.17
	√	×	84.44	81.48	81.04	81.64	**85.39**	82.80	3.90
	×	√	83.70	80.74	82.53	**87.27**	84.27	**83.70**	5.80
	√	√	**86.67**	**81.85**	**82.16**	83.15	84.27	**83.62**	3.80

Table 6-3 Comparison of performance improvement made by each of the pre-processing techniques

	No	Yes	Improving
Mask	83.18	81.37	×
Color	83.18	83.47	√
Reshape	82.72	82.6	×
Crop	82.44	82.88	√

6.4 Multi-Scale CNN (MS-CNN) for global-local feature fusion

To take full advantage of the original high-resolution DICOM CXR images using the deep convolutional network structure, we propose a novel multiple-scale image classification system that combines general information from the thumbnail images and detailed information from the high-resolution image. This novel CNN structure is customised for chest X-ray image-based lung disease categorisation.

6.4.1 Method

The framework of the proposed Multi-Scale CNN (MS-CNN) is illustrated in Figure 6-4. The original image and its mask, generated as described in Chapter 3, are cascaded to produce the simulated colored images, as described in section 6.3.1. The masked color images are used to train a first-stage local feature extractor. We have replaced the last fully connected layer with a convolutional layer using the same input and output dimensions and 1 x 1 kernel, so that the CNN can generate feature maps to represent the network's activation for each class. We apply a global average pooling layer and obtain the class scores. Label smoothing Softmax operator is used to calculate the loss between the ground truth label and the output class score, and train the local feature extractor.

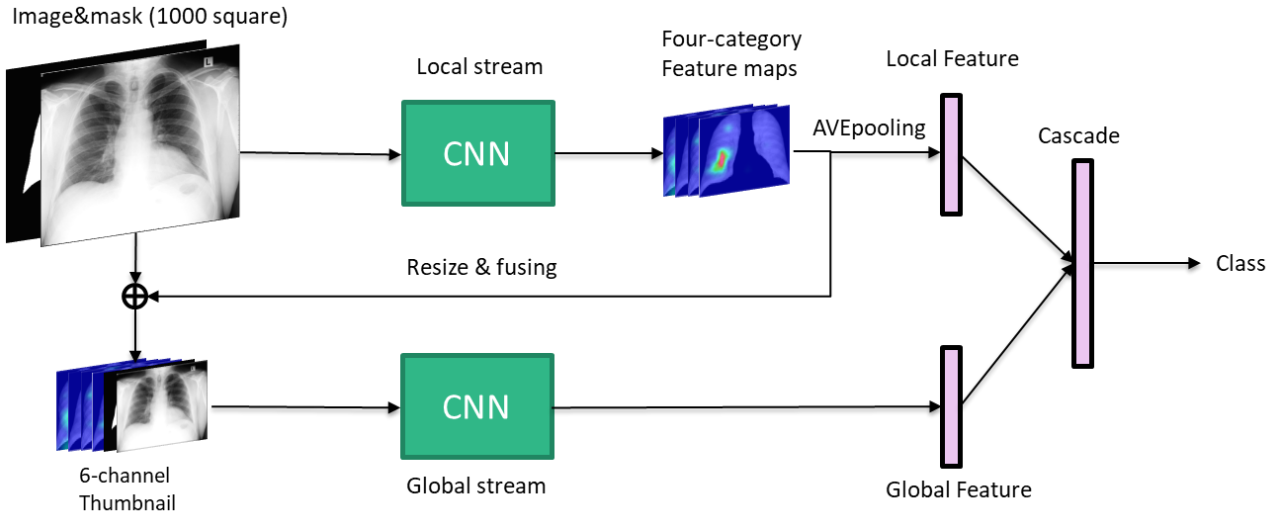


Figure 6-4 the framework of Multi-Scale CNN

The output feature maps from the local stream are resized and cascaded to the thumbnail of the original image and mask to form the 6-channel input image for the global stream CNN. The output feature of the global stream CNN is cascaded with the local feature to produce the fused feature for classification. Finally, we use a fully connected layer to obtain the output confidence score of each class.

6.4.2 Experimental setup

We used the model transferred from the ChestX-ray14 dataset and finetuned the model with our pneumoconiosis dataset. Firstly, we finetuned the model of the local stream alone. In the next step, we loaded and fixed the parameters in pretrained local steam and finetuned the global stream. The hyperparameters of the training are the same as described in Section 6.2.2.

6.4.3 Feature fusion setting

We have tested different methods of fusing the feature of the local stream and the global stream, e.g., cascading the output label, the last layer image feature, the second and the third last layer image feature from the two streams, to determine the best way of feature fusion. We did not use the ChestX-ray14 pretraining in this experiment, and the input and resized image resolutions for the two scales of the model are 1,120 and 384 square.

Table 6-4 Classification accuracy using different layers of cascaded features

	Fold					AVE	DEV
	1	2	3	4	5		
label	85.93	82.96	81.04	86.14	84.64	84.14	4.61
Last layer feature	86.30	82.96	81.04	86.14	84.64	84.23	4.97
Second last layer feature	84.81	82.22	81.04	85.39	84.27	83.55	3.39
Third last layer feature	84.81	82.59	81.04	85.39	83.90	83.55	3.08

Table 6-4 shows the average four-class classification accuracy of five-fold cross-validation using different fused features. According to the results in the table, the cascaded last layer feature and the label achieve similar accuracy, but the last layer feature performs slightly better. In the rest of this section we used the cascaded last layer feature as the multiple-scale feature for the classification.

6.4.4 Global stream input resolution

We investigated the relationship between the input resolution and the model's classification performance. Considering the computation efficiency, we used 1,120 as the input size of the local stream, and compared 224, 384, and 448 for the global stream. The four-class five-fold validation classification accuracies for each input resolution are listed in Table 6-5.

Table 6-5 Classification accuracy using different global stream resolutions

Resolution	Fold					AVE	DEV
	1	2	3	4	5		
1120-224	85.56	80.74	82.90	84.27	85.02	83.70	3.73
1120-384	85.56	84.07	82.53	84.27	88.76	85.04	5.49
1120-448	85.56	83.70	82.90	85.39	87.27	84.96	2.92

Among the three tested input resolutions for the global stream, 384 achieved the best accuracy. We used the 1120-384 as the input resolution for the two streams in the rest of our experiments.

6.4.5 Ablation experiments

The ablation experiment results are shown in Table 6-6. The input resolution for the first three rows is 384, and for MS-CNN, the input resolution is 1120-384. The ChestX-ray14 transfer learning improved the accuracy by 0.38%. Colored images improved it by further 1.41%, and MS-CNN further improved the accuracy by 0.6%. The proposed framework obtained a notable 2.39% improvement over the EfficientNet-based CNN.

Table 6-6 Ablation experiment results

Methods	Five-fold validation average accuracy
EfficientNet-b0	82.65%
EfficientNet-b0 + ChestX-ray14	83.03%
EfficientNet-b0 + ChestX-ray14 + Color	84.44%
EfficientNet-b0 + ChestX-ray14 + Color + MS-CNN	85.04%

6.4.6 Binary classification

We used our proposed methods to train a binary classification model for pneumoconiosis detection. Experiments were conducted using the same dataset and the five-fold cross validation method. The accuracy, sensitivity, precision, specificity and F1 score of the model are listed in Table 6-7.

Table 6-7 Binary classification performance

	fold-1	fold-2	fold-3	fold-4	fold-5	AVE
Accuracy	0.938	0.968	0.951	0.941	0.970	0.949
Precision	0.923	0.985	0.945	0.932	0.976	0.946667
Sensitivity	0.962	0.952	0.962	0.958	0.967	0.957333
Specificity	0.912	0.984	0.938	0.923	0.974	0.939667
Fscore	0.943	0.969	0.953	0.944	0.972	0.952

6.4.7 MS-CNN on ChestX-ray14

Besides the Pneumoconioses image dataset, we also tested our proposed MS-CNN on the benchmark CXR dataset, ChestX-ray14. We described the dataset in section 6.2. Experiments were conducted to compare the performance of MS-CNN with state-of-the-art deep elarning model CheXNet [14]. The experimental results for the detection of each disease are listed in Table 6-8. The proposed MS-CNN outperformed CheXNet in the detection of nine out of fourteen diseases. Especially for fibrosis (+5.68%) and nodule (+5.37%), which depend highly on the details in image texture, and are the key factors for Pneumoconioses detection. The MS-CNN achieved an average accuracy of 85.41% for the classification of 14 diseases and outperformed the CheXNet by 1.27%.

Table 6-8 Comparison between the proposed MS-CNN and the CheXNet on ChestX-ray14 dataset

Label	CheXNet	MS-CNN	Improvement
Atelectasis	80.94	**83.59**	**+2.65**
Cardiomegaly	**92.48**	91.27	-1.21
Consolidation	79.01	**81.24**	**+2.23**
Edema	88.78	**90.18**	+1.40
Effusion	86.38	**88.84**	**+2.46**
Emphysema	93.71	**94.85**	+1.14
Fibrosis	80.47	**86.15**	**+5.68**
Hernia	**91.64**	91.50	-0.14
Infiltration	**73.45**	71.50	-1.95
Mass	**86.76**	86.01	-0.75
Nodule	78.02	**83.39**	**+5.37**
Pleural_Thickening	**80.62**	**80.62**	+0.00
Pneumonia	**76.8**	76.35	-0.45
Pneumothorax	88.87	**90.29**	+1.42
AVE	84.14	**85.41**	+1.27

6.5 Conclusion and future work

In this chapter, we proposed a novel deep learning-based framework for Pneumoconioses categorisation. Firstly, we apply transfer learning using a large-scale CXR dataset, ChestX-ray14, to

address the challenge of cross-domain learning between the natural images in ImageNet and our Pneumoconioses CXR images. Secondly, we use multi-channel masking to provide the model with the mask information while restoring the discriminative background information. Finally, we proposed the multi-scale CNN for learning detailed texture and the global outline in parallel and improving the accuracy of pneumoconiosis classification. Our experiments show that the model is more potential and interpretable than the previous methods.

The future works may include:

1) Instead of the two-scale CNN, using more scales in MS-CNN to enhance the model representation;
2) Using annotated pneumoconiosis lesions to generate synthetic ILO positive images to improve the model's robustness; and
3) Using more recent large-scale CXR datasets to pretrain the MS-CNN model, e.g., VinDr-CXR [43], CheXpert [44], and MIMIC-CXR [45].

7 Pilot Study

In this chapter a pilot study is described that aims to validate our automated pneumoconiosis prediction methods in clinical environment. We have developed a software tool named AI-Xrayder that utilizes the deep learning-based pneumoconiosis detection and classification methods described in Chapters 5 and 6. AI-Xrayder has been deployed to Lung Screen Australia Pty Ltd, a lung screening organisation specialising in occupational lung disease screening.

7.1 Aims and design

A pilot, or validation, study is required to ensure that the software works as expected, robustly and with similar accuracy, in a real environment as during an initial laboratory testing. In the first phase of the pilot study, an uncurated set of radiographs collected by Lung Screen and read by at least two B-readers is also processed by AI-Xrayder. Further in this section the outcomes on this dataset are presented and compared with the laboratory results from the previous sections. Our primary aim is to select the best performing method for a longer term second phase of the pilot study.

The secondary, longer term aim is to run AI-Xrayder with the best performing method for an extended period of time (3 to 6 months) alongside the Lung Screen's usual screening routine, targeting systematic and outlier errors and collecting more radiographs with pneumoconiosis as well as normal radiographs typical for coal miners. This will be followed by finetuning and re-training the deep learning-based models with the newly collected data.

Additionally, we are seeking a feedback from Lung Screen on AI-Xrayder's usability and reliability.

7.2 Implementation

AI-Xrayder software is implemented as a web service and consists of a client (web browser) that uses HTTP (Hypertext Transfer Protocol) to make requests of a web server, through an internet or locally, if the client and server are located on the same machine. The web server initates a child process that runs a machine learning model in a separate computing environment. A diagram in Figure 7-1 displays the three parts of the software and the flow of information between them.

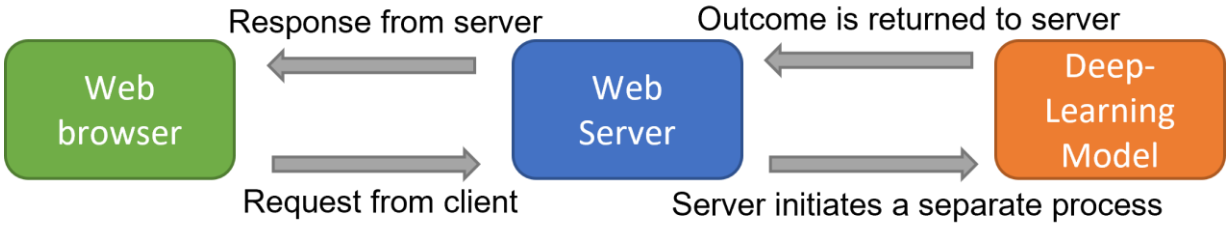


Figure 7-1 A diagram of AI-Xrayder architecture

The client part of AI-Xrayder is implemented using AngularJS – free and open-source JavaScript-based framework for developing single-page web applications [42]. The server is implemented using Node.js, an open-source JavaScript runtime environment, and Express – a popular Node.js web framework [43]. The machine learning models are implemented as described in corresponding chapters of this report and run in a Python environment with necessary libraries installed.

7.3 User experience

Lung Screen has been provided with the written instructions on how to install and launch AI-Xrayder. They reported that the installation and operation of AI-Xrayder was smooth and error-free.

For the pilot study only we mandate that AI-Xrayder is installed on the same machine that is used to open AI-Xrayder web application. This avoids security issues with sending sensitive data, such as medical images, over a network, and also saves computational time and resources.

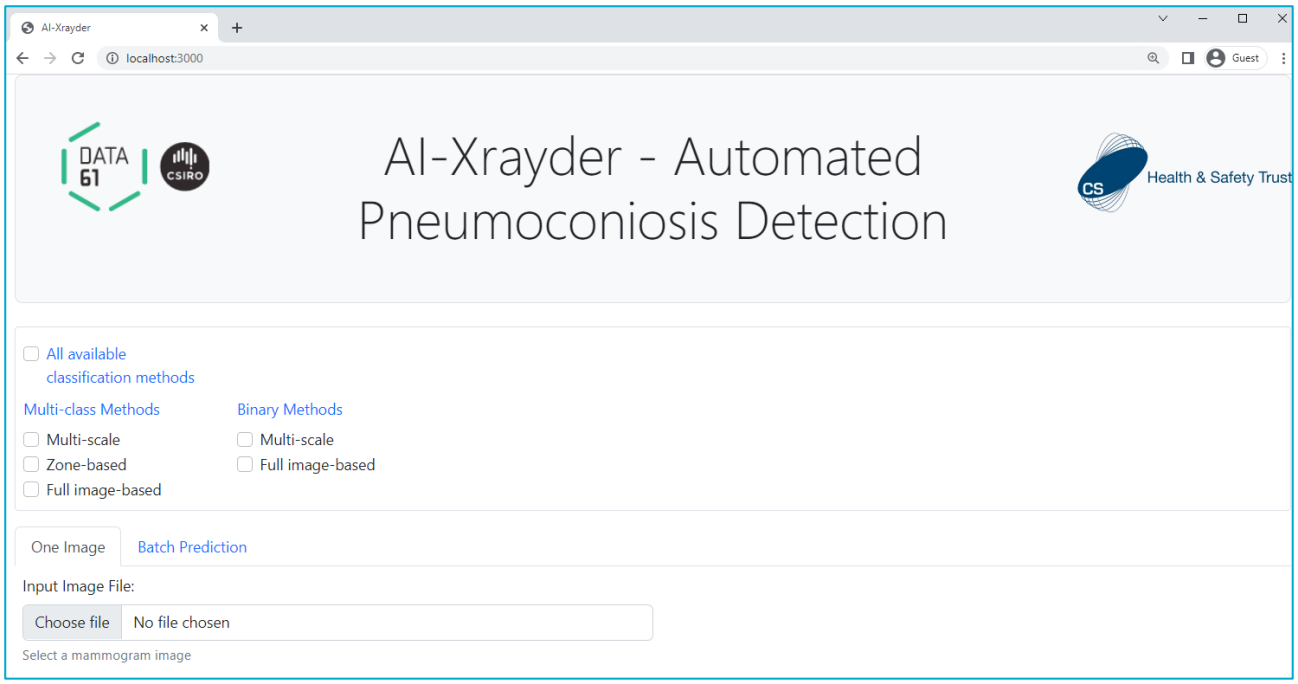


Figure 7-2 AI-XRayder user interface as seen in a Google Chrome web browser

The web page presented to a user is shown in Figure 7-2. A user can choose which classification methods they want to apply, or all of them, and select between One Image and Batch Prediction options (tabs). In One Image tab, when an input image file is selected, an uploaded image is displayed in the browser, and the chosen classification methods are applied to this image when a user clicks Predict button (see Figure 7-3).

The methods offered are divided into Multi-class and Binary. Multi-class and Binary Multi-scale methods are described in Chapter 6, while Zone-based and Full image-based multi-class and binary classification methods are described in Chapter 5.

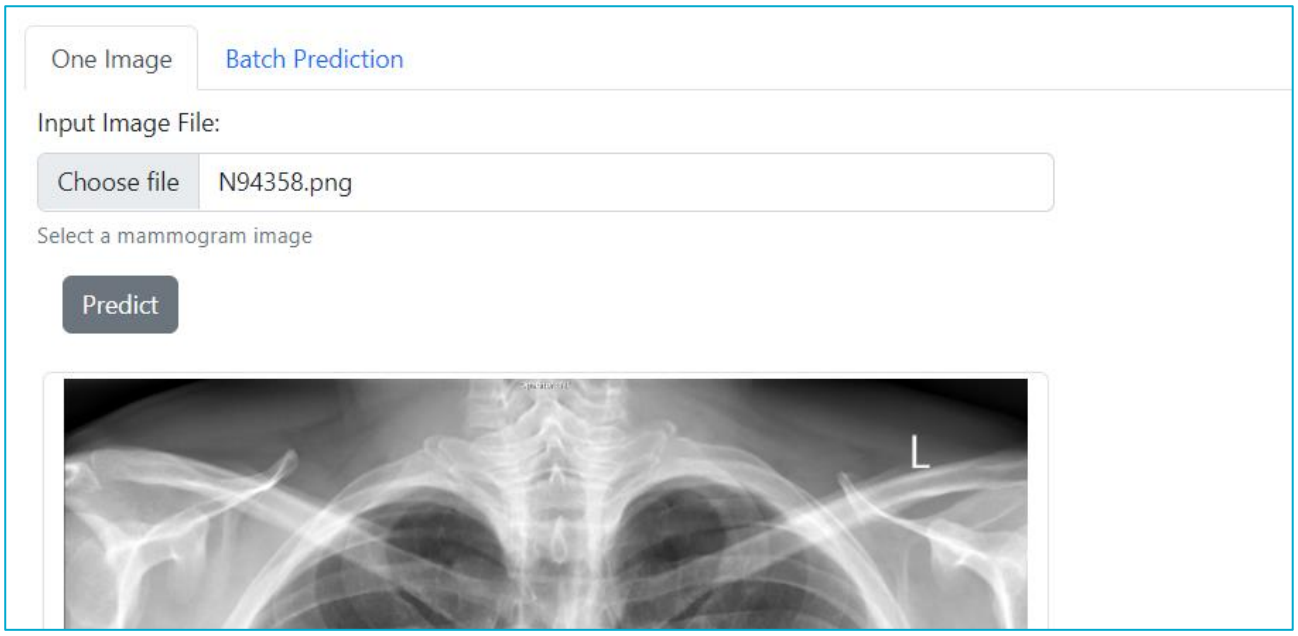


Figure 7-3 User interface showing one image is selected for classification

Batch prediction means classifying all the images in an input directory. When Batch Prediction option is selected a user has a choice of utilizing a directory configured in advance, or uploading images from a directory they choose. Figure 7-4 shows both options. Batch prediction from a preconfigured directory is advantageous when a web server and input image directory are on the same machine or local network. Not having to upload images saves computational times and resources. A configuration file is a text file in JSON format, with fields for input and output directories to be configured by the user.

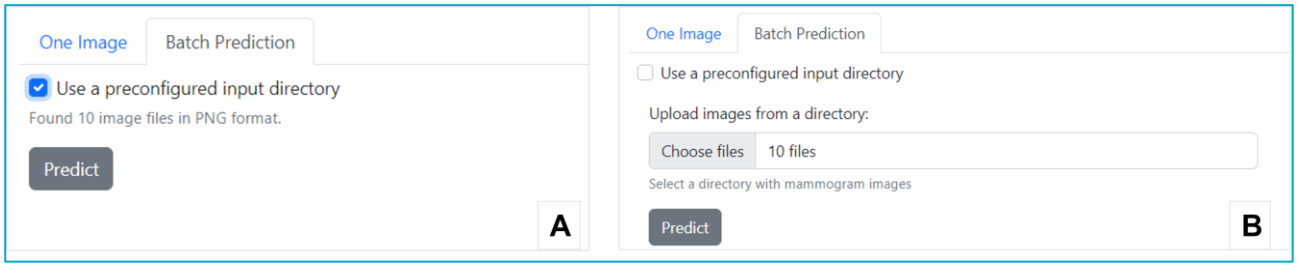


Figure 7-4 A. Input images from a preconfigured directory are used for classification. B. Images from a directory selected by a user are uploaded to the web server for classification.

The results of classification are stored in a text file in a tablular (CSV) format and can be conveniently opened with MS Excel or similar applications. For each multi-class method, the primary category and the next likely category, with the corresponding probabilities, are outputted. For the binary methods, the most likely class – normal or abnormal – and its probability are outputted and stored in the file.

7.4 Study Results and Analysis

Lung Screen tested 209 chest X-rays with AI-Xrayder. Their class distribution is shown below:

Table 7-1 X-ray images used in the pilot study

Chest X-rays number	Class 0	Class 1	Class 2	Class 3
Total = 209	100	100	6	3

The classification results with each multi-class and binary method are given in Table 7-2. The confusion matrices are shown in Figure 7-5 for multi-class classification and Figure 7-6 for binary classification. Clearly, all the tested methods have performed better in the laboratory settings, especially the zone-base and full-image based methods. For multi-scale methods, the difference of performance in clinic and laboratory is about 8-10%, which could be explained either by differences in images - we have used chest X-rays from four different sources to train the methods while the pilot study was conducted on images from one source, RSHQ, or by a bias in image annotations that haven't been discovered yet. We will move to the second phase of the pilot study using the multi-scale model as our preferred method, and will endeavour to improve the classification performance in clinic by (1) working closely with radiologists to understand a possible bias, and (2) building up a more representative training set that consists of screening chest X-rays only.

Table 7-2 The pilot study classification results for multi-class and binary methods

Evaluation metrics		Multi-scale CNN	Zone-based	Full-image based
Multi-class Classification	Accuracy	**74.64%**	43.06%	59.33%
Binary Classification	Accuracy	**86.12%**	--	62.68%
	Sensitivity	82.57%	--	**98.17%**
	Specificity	**90.00%**	--	24.00%

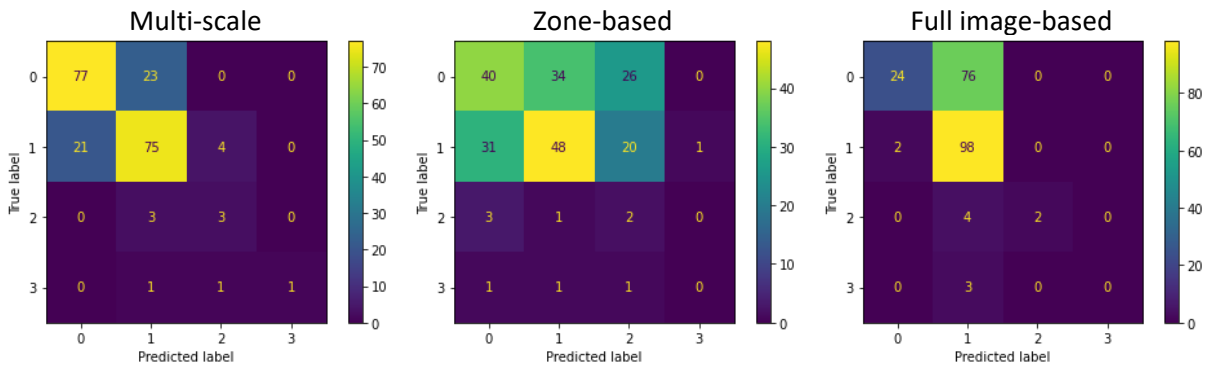


Figure 7-5 Confusion matrices for multi-class classification methods

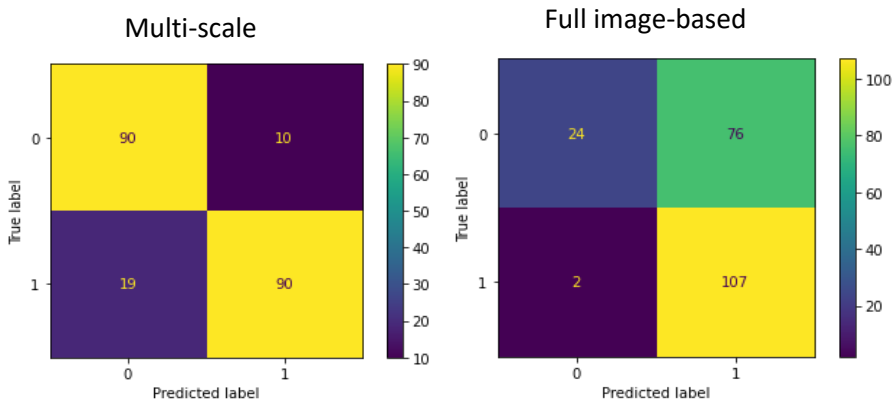


Figure 7-6 Confusion matrices for binary classification method

References

1. Queensland Government web page, https://www.business.qld.gov.au/industries/mining-energy-water/resources/safety-health/mining/accidents-incidents/mine-dust-lung-diseases, last accessed 2022/12/05.
2. https://www.nsw.gov.au/sites/default/files/2022-10/nsw-dust-disease-register-annual-report-2021-22.pdf, last accessed 2022/12/05.
3. National Health Commission. Statistical bulletin of China's health development. 2020. http://www.nhc.gov.cn/guihuaxxs/s10748/202006/ebfe31f24cc145b198dd730603ec4442.shtml, last accessed 2022/12/05.
4. Blackley DJ, Halldin CN, Laney AS. Continued increase in prevalence of coal workers' pneumoconiosis in the United States, 1970-2017. Am J Public Health. 2018;108(9):1220-1222. doi:10.2105/AJPH.2018.304517.
5. Laney AS, Attfield MD. Coal workers' pneumoconiosis and progressive massive fibrosis are increasingly more prevalent among workers in small underground coal mines in the United States. Occup Environ Med. 2010;67(6):428-431. doi:10.1136/oem.2009.050757.
6. Sim M., Glass D., Hoy R., Roberts M., Thompson B., Cohen R.: Review of Respiratory Component of the Coal Mine Workers' Health Scheme for the Queensland Department of Natural Resources and Mines, Final Report. Monash Centre for Occupational and Environmental Health, Monash University (2016).
7. Rajpurkar, P., Irvin J., Zhu K., Yang, B., Mehta H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., Lungren, M., and Ng, A.: CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. arXiv preprint arXiv:171105225 (2017).
8. Rajpurkar P, Irvin J, Ball RL, Zhu K, Yang B, Mehta H, Duan T, Ding D, Bagul A, Langlotz CP, Patel BN, Yeom KW, Shpanskaya K, Blankenberg FG, Seekins J, Amrhein TJ, Mong DA, Halabi SS, Zucker EJ, Ng AY, Lungren MP. "Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists", PLoS Med. 2018 Nov 20;15(11):e1002686. doi: 10.1371/journal.pmed.1002686. PMID: 30457988; PMCID: PMC6245676.
9. Yulia Arzhaeva, Dadong Wang and Deborah Yates, "The study protocol for the Coal Services Health and Safety Trust Project No. 20647 - Development of Automated Diagnostic Tools for Pneumoconiosis Detection from Chest X-Ray Radiographs", St Vincent's Hospital Research Office approved in August 2017.
10. Yulia Arzhaeva, Dadong Wang, CSIRO Human Research Ethics Low Risk Research Project Application Form for the project - "Development of Automated Diagnostic Tools for Pneumoconiosis Detection from Chest X-Ray Radiographs", CSIRO Human Research Ethics: Low Risk Review Panel approved in Oct. 2016.
11. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., Seekins, J., Mong, D. A., Halabi, S. S., Sandberg, J. K., Jones, R., Larson, D. B., Langlotz, C. P., Patel, B. N., Lungren, M. P., & Ng, A. Y. (2019). CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In 33rd AAAI Conference on Artificial Intelligence, AAAI 2019.
12. Johnson AE, Pollard TJ, Berkowitz SJ, Greenbaum NR, Lungren MP, Deng CY, Mark RG, Horng S. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Scientific Data, 2019;6.

13. Johnson AE, Pollard TJ, Greenbaum NR, Lungren MP, Deng C-Y, Peng Y, Lu Z, Mark RG, Berkowitz SJ, Horng S. MIMIC-CXR-JPG, a large publicly available database of labelled chest radiographs. arXiv preprint arXiv:1901.07042, 2019.
14. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., and Summers, M.: 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), pp. 3462-3471. Hawaii, USA (2017).
15. Japanese Society of Radiological Technology (JSRT) Database: http://db.jsrt.or.jp/eng.php, cited Jan. 22, 2018.
16. S. Jaeger, S. Candemir, S. Antani, Y.X.J. Wáng, P.X. Lu, and G. Thoma, "Two public chest X-ray datasets for computer-aided screening of pulmonary diseases," Quantitative imaging in medicine and surgery, vol. 4(6), pp. 475-477, 2014.
17. Study Syllabus for Classification of Radiographs of Pneumoconioses, Centers for Disease Control and Prevention, https://www.cdc.gov/niosh/learning/b-reader/start/1.html, last accessed 2022/12/12.
18. H. Song, W. Wang, S. Zhao, J. Shen, and K.-M. Lam, "Pyramid dilated deeper ConvLSTM for video salient object detection," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 715-731.
19. R. Azad, M. Asadi-Aghbolaghi, M. Fathy, and S. Escalera, "Bi-directional ConvLSTM U-Net with densley connected convolutions," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019, pp. 406-415.
20. Z. Gu et al., "Ce-net: Context encoder network for 2D medical image segmentation," IEEE transactions on medical imaging, vol. 38, no. 10, pp. 2281-2292, 2019.
21. S. Xingjian, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," in *Advances in neural information processing systems*, 2015, pp. 802-810.
22. J. P. Cohen, P. Morrison, L. Dao, K. Roth, T. Q. Duong, and M. Ghassemi, "Covid-19 image data collection: Prospective predictions are the future," arXiv preprint arXiv:2006.11988, 2020.
23. Ronneberger, O., P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. in International Conference on Medical image computing and computer-assisted intervention. 2015. Springer.
24. Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: Redesigning skip connections to exploit multiscale features in image segmentation," IEEE transactions on medical imaging, vol. 39, no. 6, pp. 1856-1867, 2019.
25. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770-778.
26. J. Li, F. Fang, K. Mei, and G. Zhang, "Multi-scale residual network for image super-resolution," in Proceedings of the European conference on computer vision (ECCV), 2018, pp. 517-532
27. Tan M, Le Q V., "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks", International Conference on Machine Learning, 2019: 6105-6114.
28. Tan M., and Le, Q V., "EfficientNet: Improving Accuracy and Efficiency through AutoML and Model Scaling", Google AI Blog: https://ai.googleblog.com/2019/05/efficientnet-improving-accuracy-and.html#:~:text=EfficientNet%2DB0%20is%20the%20baseline,than%20the%20best%20existing%20CNN, viewed on 22 Feb. 2021.

29. Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D., “GradCam: Visual explanations from deep networks via gradient-based localization”, in Proceedings of the IEEE international conference on computer vision, 2017, pp. 618-626.
30. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., et al, “PyTorch: An Imperative Style, High-Performance Deep Learning Library”, In Advances in Neural Information Processing Systems, 2019, vol. 32, pp. 8024–8035.
31. Ilse, M., Tomczak, J., & Welling, M., “Attention-based deep multiple instance learning. In International conference on machine learning”, in Proceedings of Machine Learning Research, 2018, pp. 2127-2136.
32. Patil, A., Tamboli, D., Meena, S., Anand, D., and Sethi, A., “Breast Cancer Histopathology Image Classification and Localization using Multiple Instance Learning”, 2019, in Proceedings of the IEEE International WIE Conference on Electrical and Computer Engineering (WIECON-ECE), pp. 1-4.
33. He, K., et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition, 2016.
34. Szegedy, C., et al. "Inception-v4, inception-resnet and the impact of residual connections on learning." Thirty-first AAAI conference on artificial intelligence, 2017.
35. Chollet, F.,"Xception: Deep learning with depthwise separable convolutions." Proceedings of the IEEE conference on computer vision and pattern recognition, 2017.
36. Deng J., et al. “ImageNet: A large-scale hierarchical image database”, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Florida, USA, June 20-25, 2009.
37. Simonyan K, Zisserman A., “Very deep convolutional networks for large-scale image recognition”, arXiv preprint arXiv:1409.1556, 2014.
38. Lin T Y, RoyChowdhury A, Maji S., “Bilinear CNN models for fine-grained visual recognition”, in Proceedings of the IEEE international conference on computer vision, 2015, pp. 1449-1457.
39. DICOM, https://www.dicomstandard.org/
40. Nguyen H Q, Lam K, Le L T, et al., “VinDr-CXR: An open dataset of chest x-rays with radiologist’s annotations”, Scientific Data, 2022, 9(1), pp. 1-7.
41. Irvin J, Rajpurkar P, Ko M, et al., “CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison, in Proceedings of the AAAI conference on artificial intelligence”, 2019, 33(01), pp. 590-597.
42. Jain N, Bhansali A, and Mehta D, “AngularJS: A modern MVC framework in JavaScript,” Journal of Global Research in Computer Science, vol. 5, no. 12, pp. 17-23, 2014.
43. Express - Node.js web application framework, [Online]. Available: https://expressjs.com/.

As Australia's national science agency and innovation catalyst, CSIRO is solving the greatest challenges through innovative science and technology.

CSIRO. Unlocking a better future for everyone.

Contact us

1300 363 400
+61 3 9545 2176
csiroenquiries@csiro.au
www.csiro.au

For further information

Data61
Dr Dadong Wang
+61 2 9325 3223
dadong.wang@csiro.au
www.data61.csiro.au