Multilabel Classification and Localization of Rare Pulmonary Diseases using Deep Learning

Author

Fariha Khaliq

Regn Number

00000317496

Supervisor

Dr. Syed Omer Gilani

DEPARTMENT BIOMEDICAL ENGINEERING AND SCIENCES

SCHOOL OF MECHANICAL & MANUFACTURING ENGINEERING

NATIONAL UNIVERSITY OF SCIENCES AND TECHNOLOGY

ISLAMABAD

OCTOBER 2022

Multilabel Classification and Localization of Rare Pulmonary Diseases using Deep Learning

Author

Fariha Khaliq

Regn Number

00000317496

A thesis submitted in partial fulfillment of the requirements for the degree of

MS BIOMEDICAL SCIENCES

Thesis Supervisor:

Dr. Syed Omer Gilani

Thesis Supervisor's Signature:

_____

DEPARTMENT BIOMEDICAL ENGINEERING AND SCIENCES

SCHOOL OF MECHANICAL & MANUFACTURING ENGINEERING

NATIONAL UNIVERSITY OF SCIENCES AND TECHNOLOGY, ISLAMABAD

OCTOBER 2022

Declaration

I certify that this research work titled "*Multilabel Classification and Localization of Rare Pulmonary Diseases using Deep Learning"* is my own work. The work has not been presented elsewhere for assessment. The material that has been used from other sources has been properly acknowledged/referred.

Signature of Student

Fariha Khaliq

00000317496

THESIS ACCEPTANCE CERTIFICATE

Certified that final copy of MS/MPhil thesis written by Ms. **Fariha Khaliq** (Registration No. **00000317496**), of **School of Mechanical and Manufacturing Engineering** has been vetted by undersigned, found complete in all respects as per NUST Statues/Regulations, is within the similarity indices limit and is accepted as partial fulfillment for the award of MS/MPhil degree. It is further certified that necessary amendments as pointed out by GEC members of the scholar have also been incorporated in the said thesis.

Signature: _____

Name of Supervisor: _____

Date: _____

Signature (HoD): _____

Date: _____

Signature (Dean/Principal): _____

Date: _____

Certificate for Plagiarism

It is certified that MS Thesis Titled **Multilabel Classification and Localization of Rare Pulmonary Diseases using Deep Learning** by **Fariha Khaliq** has been examined by us. We undertake the follows:

a. Thesis has significant new work/knowledge as compared already published or are under consideration to be published elsewhere. No sentence, equation, diagram, table, paragraph, or section has been copied verbatim from previous work unless it is placed under quotation marks and duly referenced.

b. The work presented is original and own work of the author (i.e., there is no plagiarism). No ideas, processes, results, or words of others have been presented as Author own work.

c. There is no fabrication of data or results which have been compiled/analyzed.

d. There is no falsification by manipulating research materials, equipment, or processes, or changing or omitting data or results such that the research is not accurately represented in the research record.

e. The thesis has been checked using TURNITIN (copy of originality report attached) and found within limits as per HEC plagiarism Policy and instructions issued from time to time.

Name & Signature of Supervisor

_____

Signature : _____

Copyright Statement

- Copyright in text of this thesis rests with the student author. Copies (by any process) either in full, or of extracts, may be made only in accordance with instructions given by the author and lodged in the Library of NUST School of Mechanical & Manufacturing Engineering (SMME). Details may be obtained by the Librarian. This page must form part of any such copies made. Further copies (by any process) may not be made without the permission (in writing) of the author.
- The ownership of any intellectual property rights which may be described in this thesis is vested in NUST School of Mechanical & Manufacturing Engineering, subject to any prior agreement to the contrary, and may not be made available for use by third parties without the written permission of the SMME, which will prescribe the terms and conditions of any such agreement.
- Further information on the conditions under which disclosures and exploitation may take place is available from the Library of NUST School of Mechanical & Manufacturing Engineering, Islamabad.

Acknowledgements

I am thankful to my Creator Allah Subhana-Watala to have guided me throughout this work at every step and for every new thought which You setup in my mind to improve it. Indeed, I could have done nothing without Your priceless help and guidance. Whosoever helped me throughout the course of my thesis, whether my parents or any other individual was Your will, so indeed none be worthy of praise but You.

I am profusely thankful to my beloved parents who raised me when I was not capable of walking and continued to support me throughout in every department of my life.

I would also like to express special thanks to my supervisor Dr. Syed Omer Gilani for his help throughout my thesis and also for Medical Image Analysis course which he has taught me. I can safely say that I haven't learned any other engineering subject in such depth than the one which he has taught. I would also like to pay special thanks for his tremendous support and cooperation. Each time I got stuck in something; he came up with the solution. Without his help I wouldn't have been able to complete my thesis. I appreciate his patience and guidance throughout the whole thesis.

I would also like to thank Dr. Asim Waris and Dr. Amer Sohail Kashif for being on my thesis guidance and evaluation committee.

Finally, I would like to express my gratitude to all the individuals who have rendered valuable assistance to my study.

Dedicated to my exceptional parents and adored sibling whose tremendous support and cooperation led me to this wonderful accomplishment.

Abstract

Chest radiography is the most common radiological examination used for the diagnosis of thoracic diseases. Currently, automated classification of radiological images is abundantly used in clinical diagnosis. However, each pathology has its own response characteristic receptive field regions, which is a key problem during the classification of chest diseases. In addition to extreme class imbalance, cases labelled as uncertain in the dataset further complicate this task. To solve this problem, we propose a semi-supervised learning approach known as U-SelfTrained. In this scheme, uncertain labels are left unlabeled in the dataset; first, the model is trained on labelled instances and then on unlabeled instances relabeling them with labels having a higher probability. Comprehensive experimentation was carried out on the CheXpert dataset, which consists of 223,816 frontal and lateral view CXR images of 64,740 patients with 14 diseases. The testing accuracy is 0.877 on the CheXpert dataset, which is currently the highest score achieved to date. This validates the effectiveness of this method for chest radiography classification. The practical significance of this study is the implementation of AI algorithms to assist radiologists in improving their diagnostic accuracy.

Key Words: *Chest Xray, Multi-label classification, semi-supervised learning, thorax disease, deep learning*

Table of Contents

List of Figures

List of Tables

CHAPTER ONE

CHAPTER 1: INTRODUCTION

Chronic lung diseases are the fourth leading cause of non-communicable diseases (NCD) and pose a particular challenge in low- and middle-income countries (Majkowska et al., 2020). Chronic lung diseases include several deadly illnesses with high prevalence rate, such as millions of people being affected by pneumonia worldwide annually and approximately fifty thousand people expire from pneumonia annually in the United States only. Diagnosing thoracic diseases at an early stage can help clinicians provide effective treatment (Li et al., 2018).

Chest radiography (CXR) is widely used for the detection of pulmonary diseases affecting the lungs. Technology helps physicians in the clinical diagnosis of several diseases such as pneumonia, cardiomegaly, edema, lesions, and lung opacity (Li et al., 2018). Moreover, the chances of survival of patients drastically increase due to the screening of diseases, such as lung cancer, using CXR. Since CXR is available in all clinics and is a cheap method for diagnosis, it is a highly valuable diagnostic method in comparison to other methods. Nevertheless, analyzing the results of CXR, the variety of sections complicates CXR analysis. Lack of consistency and specialized clinicians for the analysis of CXR and fatigue could lead to errors. As the interpretation of CXR reports differs from one specialist to another, it causes a major error in diagnosis due to inconsistency in interpretation (Zhao et al., 2021).

Computer-aided diagnosis (CAD) can be implemented to reduce the burden on radiologists while detecting uncertainty in CXR. Furthermore, several CAD systems have proven effective in diagnosing an extensive array of diseases (Wang et al., 2017). Several CAD techniques have been developed for the diagnosis and characterization of injuries in medical imaging, such as traditional projection radiography, computed tomography (CT), ultrasound, magnetic resonance imaging (MRI), and X-rays. Currently, organs such as breast, lungs, colon, brain, liver, kidneys, and other vascular and skeletal systems are being studied for CAD. CAD was implemented to provide a "second opinion" to aid radiologists in the image analysis (Kuo et al., 2021). Thus, it is necessary to develop a computer algorithm and determine how to use CAD technique output to assist radiologists in the diagnostic process. However, using a reliable methodology, such as receiver operating curve (ROC), to analyze the performance of a large-scale observer on radiologists is equally important as the

development of computer algorithms in the field of CAD. CAD techniques have been developed through team efforts by researchers with different backgrounds, such as physicists, radiologists, computer scientists, engineers, psychologists, and statisticians. CAD has a significant impact on medical imaging and the quantitative analysis of radiological images (Madjarov et al., 2012).

Convolutional neural networks (CNN) have shown prolific results in the medical field regarding disease classification. A CNN was built to imitate the alternating layers of cells present in the visual cortex of the human brain. The CNN consists of three layers: a convolutional layer, a pooling layer, and a fully connected layer. CNN implements a one-way model technique in which information is transmitted from the input layer to the output layer only; this is known as the feed-forward approach (Bressem et al., 2020). This feed-forward approach was implemented in both supervised and unsupervised deep learning models. CNN are widely used in deep learning approaches for biomedical image analysis. These models were designed to handle multiple array data, signal data, and 2D and 3D images. Some commonly used CNN include AlexNet, LeNet, R-CNN, Zfnet, GoogleNet, and ResNet (Rawat et al., 2017).

Currently, most research is conducted on single-label classification for CXR. However, we are interested in the multi-label classification of CXR while detecting uncertainties. In multilabel classification, one or more diseases may be present in an image (Hwang et al., 2019). Multilabel classification is an approach used to map data from single to multiple labels (Seyyed et al., 2020). These multilabels represent parts of the same label set comprising inconsistent labels. The aim of multilabel classification is to develop a classification model for previously unidentified samples (Madjarov et al., 2012). This complicates the problem because the algorithm must be able to detect multiple diseases in an image even if they overlap (Phillips et al., 2020). One of the main challenges in the multi-label classification of chest diseases from radiological images is that each disease has its own unique response characteristic receptive field region. In addition, the class imbalance of dataset labels further increases the complexity of the problem (Guan et al., 2021).

In this paper, we aimed to predict the probability of 14 different observations from multiview chest X-rays while classifying uncertain labels. We focused on

uncertainty labels in the dataset and determined the efficiency of the semi-supervised learning approach for the classification of uncertain labels during the training process.

CHAPTER TWO

CHAPTER 2: RELATED WORK

Limited techniques and research have targeted all fourteen pathology labels of chest diseases. Wang et al (2017) presented the first publicly available dataset of chest radiographs known as Chest Xray14 that provided a new dimension to the researchers. They used deep CNN approach and achieved promising results also stated that this dataset could be further extended by adding more disease labels into this dataset. A multi-scale channel attention module which integrates local channel to global channel statistics to solve problems when fusing different scale feature was proposed by Dai et al (2021).

Additionally, Chen et al (2020) proposed a new technique based on Graph Convolution Network (GCN) termed as CheXGCN. In this technique interdependence and cooccurrence of labels were integrated for multi-label classification of CXR image classification improving the recognition accuracy. Ho & Gwak (2019) proposed a unique framework to distinguish 14 pathologies by incorporating multiple features from both shallow and deep features and extracted the discriminant features from publicly available ChestX-ray14 dataset. Liu and his colleagues in (2020) introduced a semi-supervised technique for multilabel classification based on relation-driven which uses unlabeled data by predicting consistency of the given input under disturbance and generates high quality labels for unlabeled data using self-assembly model with an accuracy of 0.79.

Yao et al (2017) presented a two-stage end to end neural network model to exploit label dependencies. The model combines a densely connected image encoder with recurrent neural network decoder (RNN). Kumar et al (2017) proposed research in which loss function is more suitable for training CNN from scratch and presented an efficient CNN for global image classification. Rajpurkar et al (2017) proposed a well-known state of the art architecture known as CheXNet. The model fine tunes DenseNet-121 on global chest radiographs, which modifies last fully connected layer.

Guan et al (2018) proposed a supervised two branch CNN for the classification of thorax disease. The proposed model is trained by evaluating global and local cues learned in the local and global branches, thus the model achieved highest accuracy over the state-of-the-art models and techniques on CXR datasets. Contrarily to above

work Rubin et al (2018) proposed a new model DualNet architecture to classify Multiview i.e., frontal, and lateral CXRs images. It imitates usual clinical practice by considering multi-view images simultaneously. Allaouz & Ahmed (2019) proposed an approach which combines feature extraction and power of supervised multi-label classifier for the detection of CXR. They used pre-trained DenseNet121 model as feature extractor and various transformation methods like BR, LP, and CC with an AUC of 0.812.

Despite of excessive research and great success in development of medical image application techniques, still there is no good solution to the problem since the multilabel classification model before and after image transformation are inconsistent for multiple diseases. To resolve this problem, this paper undertakes Medical AI as the background and applies cutting-edge deep learning technology of semi-supervised learning for multi-label classification of medical images while tackling uncertainties in the dataset. This will drastically improve the efficacy and accuracy of clinical diagnosis.

CHAPTER THREE

CHAPTER 3: MATERIAL AND METHODS

3.1. Data set

In this research a novel deep learning thorax disease detection model is proposed based on CNN and DL techniques using a publicly available dataset introduced by Stanford Machine Learning (SML) Group. CheXpert is a large publicly available dataset for CXR interpretation, consisting of 223,816 CXR of 64,740 patients for the presence of 14 diseases labeled as positive (1), negative (0), and uncertain (-1). The data set contains multi-view chest radiographs to predict the probability of 14 different pathologies, figure 1 shows the example of multi-view chest radiographs from the dataset. The dataset contains 2D grayscale images in jpeg format along with Generic Gray Gamma 2.2 color profile. The dimension of the images varies according to the view. Dimensions of frontal view image ranges from 327 x 320 to 389 x 320 pixels whereas the lateral view images have a standard dimension of 320 x 320 pixels. The dataset contains 29000 PA view images, 162,000 AP view images, and 32000 LL view images (Garbin et al., 2021).

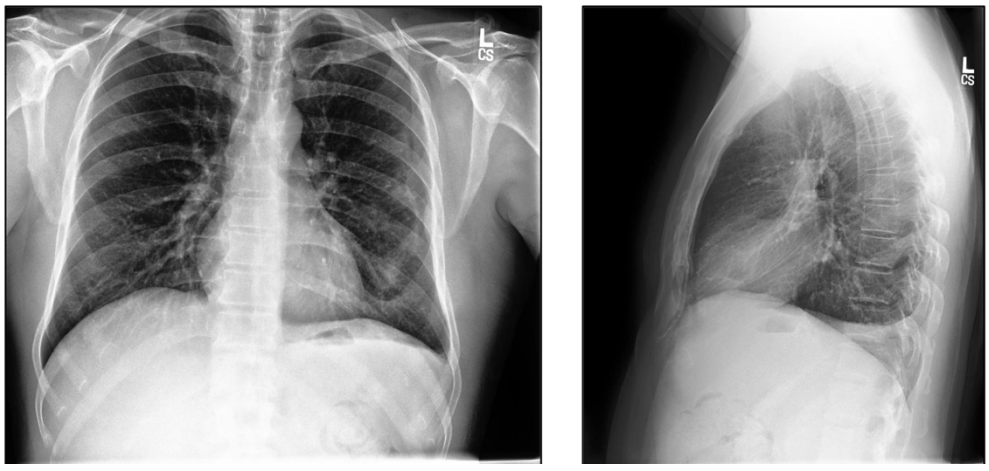

Figure 1 Example of CheXpert dataset: Frontal and Lateral view of chest radiographs

The training set contains 223,415 studies from 64,540. The studies were annotated by the consensus of five radiologists. Their annotations were multi-classed such that all present cases are treated as positive, absent as negative and unknown cases as uncertain in train set. The prevalence of the labels for different observation in

Table 1. The validation set contains 200 studies from 200 patients randomly sampled from the entire dataset with overlapping the patients with evaluation report. Each study was individually annotated by three-board certified radiologists, classifying each observation into one of present, uncertain likely, uncertain unlikely, and absent. The annotations were binarized such as all present and uncertain likely cases as positive and all absent and uncertain unlikely cases as negative. These binarized annotations are used to define as strong ground truth (Garbin et al., 2021).

Table 1Number of studies which contain these 14 observations in training dataset

Pathology	Positive (%)	Negative (%)	Uncertain (%)
No Finding	16627 (8.86)	171014 (91.14)	0 (0.0)
Enlarged Cardiom.	9020 (4.81)	168473 (89.78)	10148 (5.41)
Cardiomegaly	23002 (12.26)	158042 (84.23)	6597 (3.52)
Lung Lesion	6856 (3.65)	179714 (95.78)	1071 (0.57)
Lung Opacity	92669 (49.39)	90631 (48.3)	4341 (2.31)
Edema	48905 (26.06)	127165 (67.77)	11571 (6.17)
Consolidation	12730 (6.78)	150935 (80.44)	23976 (12.78)
Pneumonia	4576 (2.44)	167407 (89.22)	15658 (8.34)
Atelectasis	29333 (15.63)	128931 (68.71)	29377 (15.66)
Pneumothorax	17313 (9.23)	167665 (89.35)	2663 (1.42)
Pleural Effusion	75696 (40.34)	102526 (54.64)	9419 (5.02)
Pleural Other	2441 (1.3)	183429 (97.76)	1771 (0.94)
Fracture	7270 (3.87)	179887 (95.87)	484 (0.26)
Support Devices	105831 (56.4)	80912 (43.12)	898 (0.48)

Several CXR studies in the dataset contains multiple pathology labels. Several CXR studies in the dataset contained multiple pathology labels. The correlation between labels, is illustrated in figure 2b and the distribution of classes in single label is illustrated in figure 2a.

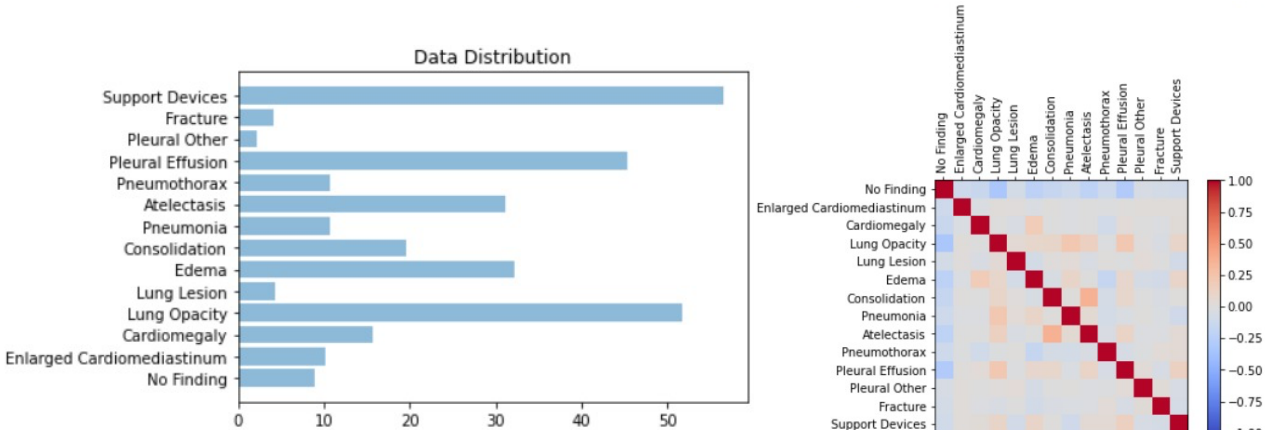

Figure 2(A) Distribution of classes in a single label in CheXpert dataset
(B) Correlation between labels in the dataset.

3.2. Proposed Model

In this research a self-trained deep learning model is proposed for the classification of 14 different lung pathologies-based CNN and DL techniques. The proposed model implements different experimental setups. During the experimentation, training data is augmented to increase the robustness of the models. The data augmentation is utilized during the training of the proposed model. The generalized pipeline of the multi-label classification of chest-radiograph model is illustrated in Fig 3.

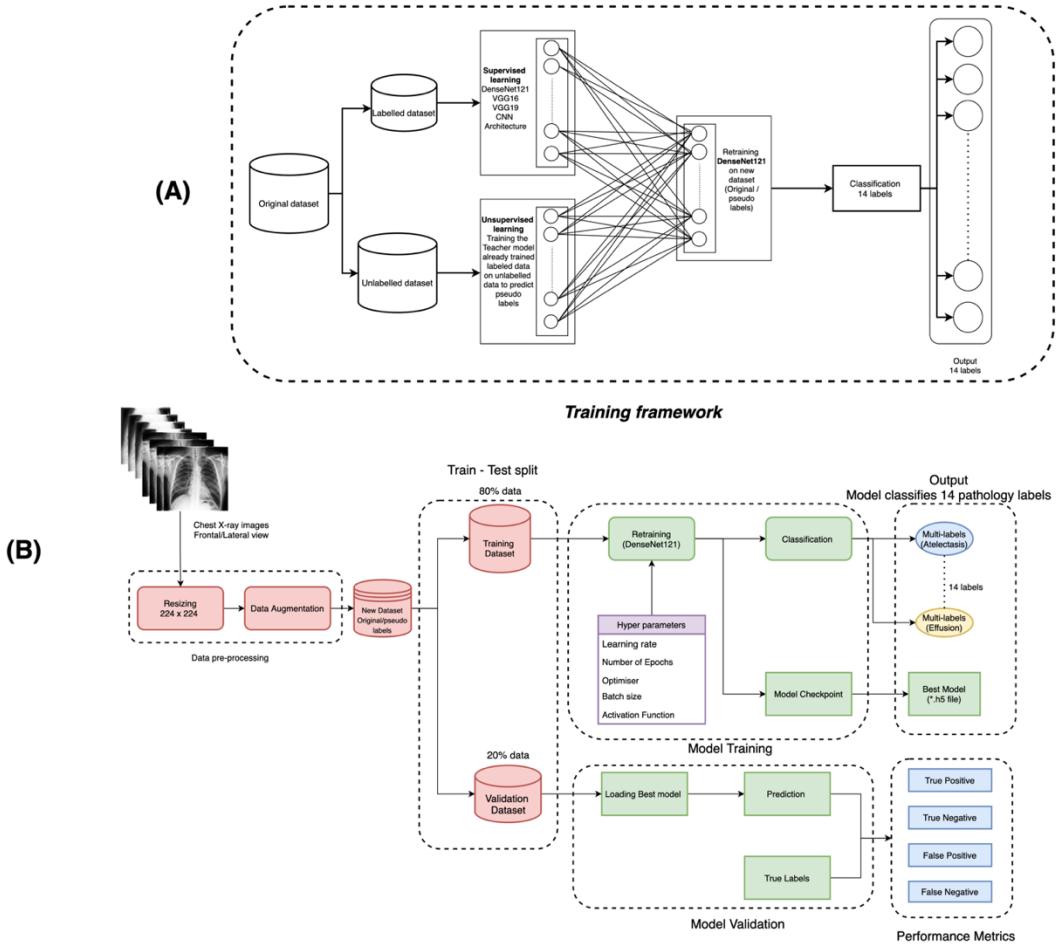

Figure 3 (A) Illustrates the training framework that illustrates that first the dataset is divided into labeled and unlabeled data. Labeled data is fed to supervised model to classify the instances in 14 pathology labels. Further unlabeled data is Fed to Teacher Model to predict pseudo labels. Final step the best model is retained on new dataset that consists of original and pseudo labels to classify the instances. (B) Generalized pipeline of multilabel classification framework. Illustrates complete training and validation process.

3.2.1. Data pre-processing

3.2.1.1. Size of the image

Comparing to the common ImageNet classification problems, considerably minor spatial extent of several diseases inside X-ray image with a typical dimension of 3000 x 2000 pixels inflict challenges in both the capacity of computing hardware and the development of deep learning algorithm. In CheXpert the images were extracted from DICOM files. Since the Dataset was present in two formats such that high resolution and down sampled format. We down sampled the dataset by resizing all the images as 224 x 224 pixels without losing any significant details compared to the original size which varied from 327 x 320 to 389 x 320 pixels in frontal view and 320 x 320 pixels in lateral view.

3.2.2. Data augmentation

CNN models usually overfits when a small number of samples are provided to the model. Thus, a large number of images are required for intensive training and enhancing the overall performance of the model. There was extreme class imbalance in the dataset to balance the classes we implemented data augmentation that includes random horizontal flipping as shown in Table 2. The augmentation technique utilized on the training data set improves the generalization and robustness of the proposed multi-label classification model.

Table 2 Data augmentation parameters

Augmentation	Parameters
Flip	[0.5]

3.2.3. Multi-label classification framework

Multi-label classification has gained a lot of interest in field of computer vision and has been implemented to solve the problems of image and video annotation. Multi-label is different from single-label classification, in multi-label classification the classifier assigns more than one label to an image or no label at all. Whereas different approaches are implemented to tackle multi-label classification (Rubin et al., 2018). The aim of this research is to fit the uncertain labels in the data to an algorithm by transforming the multi-label classification problem into one or more labels

(binary/multi-class) and then combine their results to form the multi-label predictions (Allaouz and Ahmed, 2019).

Irvin et al (2019) proposed a simple approach to handle the uncertain **u** labels by ignoring these labels, which is a baseline technique in comparison to other techniques which are explicitly incorporated the uncertainty labels. This approach is known as ***(U-Ignore)***. In this technique the sum of masked binary cross-entropy is optimized over the observations, masking the loss for the observations which are labeled as uncertain in the dataset. The loss for an instance X is given in following equation (1),

$$L\,(X, y) = -\sum_{o} \mathbb{I}\,\{y_o \neq u\}\,[y_o \log p(Y_o = 1|X) + (1 - y_o) \log p\,(Y_o = 0|X)]$$

In eq (1) X denotes the input image, whereas y denotes the vector of labels of length 14 in this this study, and the sum is taken of all the 14 pathologies. Using *U-Ignore* independently can produce biased model if the cases are note deleted randomly. As, in the dataset, in some pathologies uncertain labels are twice the positive labels such that for Consolidation, the value uncertain label is (12.78%) twice the positive labels (6.78%) along with pneumonia has uncertain labels (8.34%) four times the positive labels (2.34%) in the dataset. Therefore, if this approach is implemented independently a large amount of data is deleted which generate the biased and less credible results. After being motivated by the results Irvin et al (2019) in this research to tackle uncertain labels we have implemented ***"Self-training approach".***

3.2.3.1. Self-training approach

Self-training technique is a semi-supervised approach for tackling multi-label classification problems giving training data that have partial annotations of their labels, also known as (U-SelfTrained). In this technique firstly, we train model implementing U-Ignore technique which means that first model ignores all the **u** labels during training (Rajan et al., 2021). In next step model make predictions and re-label each uncertain label with the prediction generated by the model. We do not replace any instance of 0 and 1s which eliminates all the biases from the results. In this approach we set up the loss as the mean of the binary cross-entropy losses over the observations. This study follows the approach introduced by (Yarowsky 1995), who trained the model on labeled instances and then predicted unlabeled instances labelling

them when the labels with higher probability and repeated until convergence. In self-training approach the training process mainly involves data distillation as discussed below.

3.2.3.2. Data Distillation

Data distillation is a generalized method used in self-training process that labels the unlabeled dataset with training large set of models. Data distillation involves the following four steps (Radosavovic et al., 2018).

i. First step is to train model on manually labeled data (Just like traditional supervised learning)

ii. Second step is training the trained model on unlabeled data to make predictions.

iii. In third step pseudo labels are ensembled with manually annotated labels and new dataset is formed.

iv. In last step the model is retrained on new dataset that contains both manual and pseudo labels.

In this study we implemented the same training protocol proposed by the Rajan et al 2020. In this process the information is distill from the trained model $\mathcal{F}(\theta)$ which is also known as Teacher model to develop student model $\mathcal{G}$ with the parameters $\emptyset$ which can achieve improved predictions. This is obtained by using labeled data denoted by (X_ℓ, Y_ℓ) and with additional unlabeled data $(X_\mathcal{U})$ as shown in figure 4. The empirical evidence of this approach is in the recent work of Xie et al (2020). To reduce the effect of uncertainties in teacher model during training the sharpening of instances is performed in teacher model as follow in equation (2):

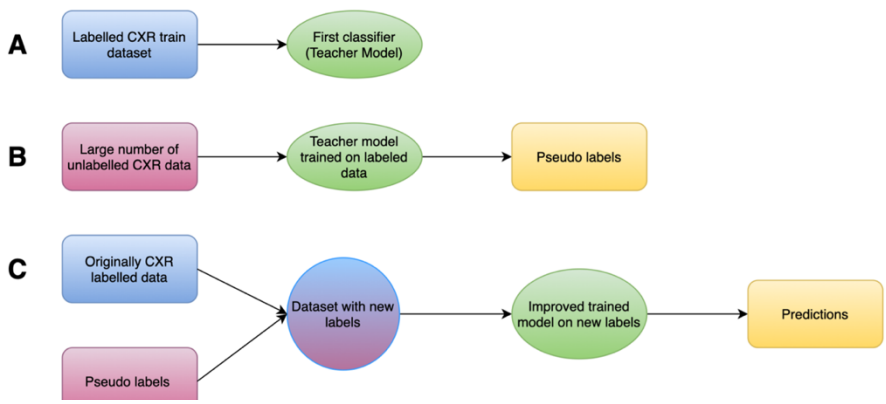

Figure 4 Generalized description of data distillation

$$\widehat{Y}_{u,\gamma} = (1 - \gamma)\widehat{Y}_u + \gamma \mathbb{I}\,[\widehat{Y}_u \geq 0.5]$$

$\mathbb{I}$ denotes the activation function, we have implemented sigmoid as the activation function $(1e8 \times (\widehat{Y}_u - 0.5))$ whereas γ is the hyper parameter. The sharpening drives the probabilities of the predictions for each of the labels closer to 1 when it is greater than 0.5 and closer to 0 when the probability is less than 0.5.

3.2.4. Proposed CNN Architecture

In this step new CXR dataset with original and pseudo labels is fed to the model to classify CXR into one or multiple possible diseases. In this research we have used multiple models but the final CNN which gave the best results was DenseNet121. We preferred DenseNet121 over other models since it can be used as a "feature extractor" (Takeuchi et al., 2019). The Dense Convolutional Neural Network (DenseNet) is new CNN yet has outperformed many CNNs like VGG16 and VGG19 by providing state-of-the-art results on highly complex problems. The fundamental idea of DenseNet is to make sure that maximum flow of information within layers in the network by connecting all layers directly with each other. Figure 5 illustrates how DenseNet is different from traditional architectures by introducing $\frac{l\times(L+1)}{2}$ connections in an L-layer network (Ho and Gwak, 2019). The architecture of DenseNet is composed of stack of dense blocks followed by transition layers (Allaouzi and Ahmed, 2019). A dense block consists of series of units and each unit integrates two convolutions, followed by Batch Normalization and ReLU activations.

Moreover, each component generates a set number of feature vectors. These parameters are known as growth rate, controls the amount of new information that is transmitted by layers. Transition layers are the layers between these dense blocks which implement down-sampling of the features passing the network. A detailed description of DenseNet121 architecture is illustrated in figure 5. Impressed by the results of DenseNet121 on ChestX-ray-14 dataset we have trained the DenseNet121 model on our CheXpert dataset, using the weights we obtained by calculating the weights for each class to balance the dataset. Since, the dataset was extremely unbalanced and using initial weight obtained from the pre-trained network, on ImageNet did not give desired results.

We used a batch size of 16, and number of epochs up to 100, binary cross-entropy as a loss function whereas the best model was selected based on the validation loss. We used Adam optimizer with variable learning rate with an initial learning rate of 10^{-8} which is decreased by 10 each time the validation loss is obtained after an epoch. In next step, we freeze the best weights from the lower convolutional layers and replace the last fully connected layer with fully connected layer of a 14-dimensional output with sigmoid as the activation function. Each iteration in training phase aims to optimize the validation losses through the following equation,

$$l\left(Y, Y^{(p)}\right) = \sum_{c=1}^{14} Y_c \ log\left(Y_c^{(P)}\right) + \ (1 - Y_c) \ log\left(1 \ - Y_c^{(P)}\right)$$

In equation (1) Y is the ground truth vector and $Y_{(p)}$ is the predicted label vector which value in binary; 1 and 0, this represents the presence and absence of the corresponding labels.

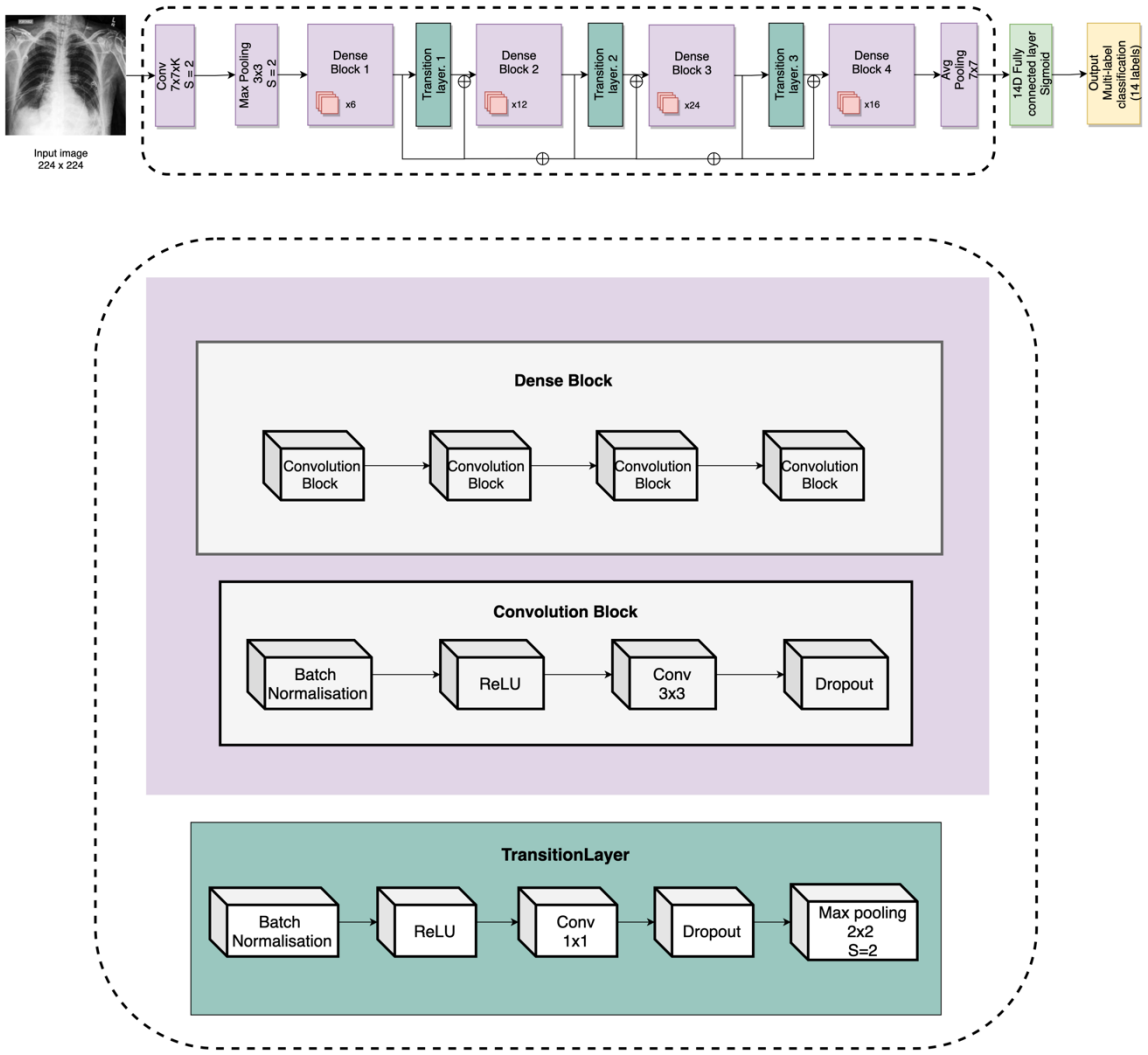

Figure 5 Shows the detailed architecture of DenseNet121 model.

3.2.5. Training

In this research we experimented with different convolutional neural network (CNN) architectures specifically VGG16, VGG19, and DenseNet121 with different parameters, and found that Dense 121 architecture provided the state-of-the-art results. Therefore, we used DenseNet121 for the final experimentation. The data is split as 80 percent for training and 20 percent for testing. The images were fed into the network with size 224 x 224 pixels. Adam optimizer was used with default β-parameters of β1 = 0.9, β2 = 0.999 and variable learning rate which decreases with the validation loss

during training. The batch size was 16 images which was fixed during training. We have trained model with custom weights for 100 epochs saving checkpoints every 6283 iterations. The model was trained on Intel Core i7 CPU with 32 GB RAM, 512 SSD memory and NVIDIA GeForce RTX 2080 Ti with 11GB GDDR6. Since, in this study we have implemented self-training approach for experimentation the main step of involved in training process is data distillation as illustrated in Figure 4. Table 5 shows the training description and figure 6a illustrates the training and validation accuracy whereas figure 6b demonstrates training and validation loss.

Table 3 Results on CheXpert dataset

Model	Training Time	Metrics			
		Training Accuracy	Training Loss	Test Accuracy	Test Loss
DenseNet121	10 hours and 30 minutes	98.57	0.09	0.877	0.14

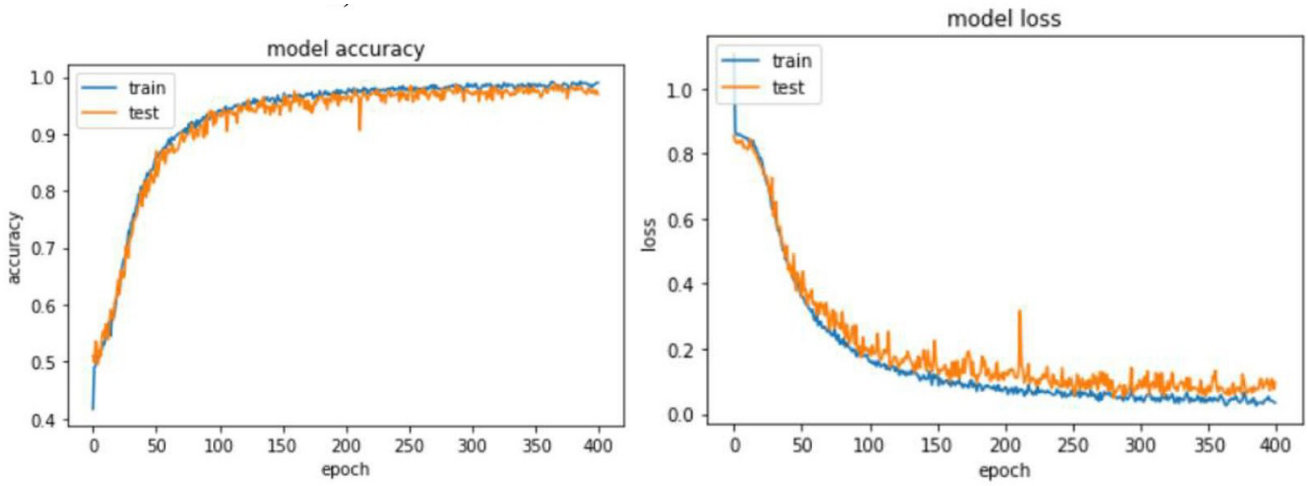

Figure 6 (A) Illustrates the training and testing accuracies of the proposed model. (B) shows the training and testing loss of the proposed model.

3.2.5.1. Balancing of Class Weights

Previously, All the research conducted on multi-label classification of chest-radiographs on chestX-ray8 and ChestX-ray14 used initial weights obtained from pre-trained network on ImageNet to train DenseNet121. The approach showed promising results on ChestX-ray14 and ChestX-ray8 however, the desired results were not

obtained when the same approach is implemented on CheXpert dataset. So, in this research inspired by Rajpurkar et al (2017) we calculated the class weights for each class through equation (3) to provide the custom weights to the model to achieve better accuracy. Table 3 shows the class weights for each class after balancing the balancing the class weights.

$$log\ loss = \frac{1}{N}\sum_{i=1}^{N}\left[-\left(w_0\left(y_i \times log\ (\hat{y}_i)\right) + w_1((1-y_i)\ \times\ log(1-\hat{y}_i))\right)\right]$$

Whereas in eq (3) w denotes weight imbalance between positive and negative imbalance samples for class 0 and 1, N is the number of values, y_i is the actual values of the target class, and $\hat{y}_i$ is the predicted probability of the target class.

3.2.5.2. Tuning of Hyperparameters

From the previous research conducted on CheXpert it is evident that hyperparameters plays a crucial role in optimizing the training phase. We performed multiple iterations with different hyperparameters proposed in literature to achieve higher accuracy on CheXpert dataset. From the experimentation it is revealed that hyperparameters implemented on chestX-ray14 did not improve accuracy on CheXpert. The hyperparameters involved in optimizing training phases are most importantly learning rate, Number of epochs, optimizer, Batch size, L2 regularization, and activation function. Table 5 shows the best hyperparameters used to train model on CheXpert dataset to achieve higher accuracy.

Table 4 Description of Hyperparameters

Hyperparameters	Description
learning rate	0.003
Epochs	400
Batch size	32
Optimizer	Adam
Activation Function	Sigmoid
Loss	binary_crossentropy
Drop out	0.2

3.2.6. Evaluation Metrics for Multi-label Classification Task

Multi-label classification is different from single-label classification where the prediction is either correct or wrong. Multi-label classification problems require special evaluation metric since all the labels are considered. In MLC prediction can be

fully correct (positive), partially correct (uncertain) or fully wrong (Negative). The evaluation metrics of MLC are categorized into two groups.

3.2.6.1. Example based metrics

In this method average difference between the predicted and ground-truth classes for each test instance is evaluated and later averaged over all examples in test set. Following are the commonly used example-based metrics:

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}}$$

$$\text{F-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision+Recall}}$$

3.2.6.2. Labeled based metrics

This category uses two types of averaging method. Prior, is called macro-average where the binary evaluation metric is computed for each individual class and later averaged over all classes. Whereas the second metric is called micro-average binary evaluation metrics is computer for all the samples and classes. Receiver operating curve (ROC) also known as AUC is widely used in MLC task since it helps in eliminating subjectivity in the threshold selection process, as continuous probability derived scores are transformed into binary presence or absence by summarizing overall performance of the model over all possible thresholds.

CHAPTER FOUR

CHAPTER 4: RESULTS

In this section we will discuss results obtained from the experimentation conducted on the CheXpert dataset. Moreover, we will also compare our results and classification approach with literature. This section will also highlight the limitation of previous classification techniques and how our proposed approach is a better technique for the classification of uncertain (**u**) labels.

We have identified that ignoring uncertainty labels during training or removing from dataset is not effective approach to handle the uncertainty labels in dataset and in particular it is ineffective in case of cardiomegaly. In case of cardiomegaly most of the uncertain labels are marginal such as "marginal cardiac enlargement" which if ignored would probably cause poorly on the cases which are difficult to differentiate. So, explicitly implementing a supervised learning approach such as (U-uncertain, U-zeros, U-ones, and U-ignore) would either distinguish non-diseased cases as diseased, diseased cases as non-diseased cases, or completely ignore uncertain cases. Implementing a semi-supervised approach "U-SelfTrained" to differentiate between borderline cases from non-borderline cases could allow the model to better delineate borderline cases by labelling them with the prediction with higher probability and repeated until convergence. Our proposed semi-supervised learning approach has achieved highest averaged AUC of 0.863 for all the 14 pathology labels in comparison to literature while classifying the uncertain labels in the dataset.

Table 5 Detailed classification report of 14 pathology labels present in the CheXpert dataset

Labels	Precision	Recall	F1-score
No Finding	0.72	0.84	0.77
Enlarged Cardiomediastinum	0.63	0.32	0.42
Cardiomegaly	0.58	0.70	0.64
Lung Opacity	0.74	0.71	0.73
Lung Lesion	0.71	0.95	0.81
Edema	0.69	0.73	0.71
Consolidation	0.43	0.76	0.55
Pneumonia	0.64	0.45	0.52
Atelectasis	0.66	0.65	0.66

Pneumothorax	0.64	0.51	0.57
Pleural Effusion	0.77	0.76	0.77
Pleural Other	0.59	0.96	0.73
Fracture	0.52	0.99	0.68
Support Devices	0.85	0.83	0.84
Micro average	0.69	0.72	0.70
Macro average	0.66	0.72	0.67
Weighted average	0.70	0.72	0.70
Sample average	0.66	0.69	0.65

In this research we implemented self-training technique also known as semi-supervised learning approach on CheXpert dataset. Table 7 provides the summary of classification performance results in terms of F1-score, Precision, Recall, and Accuracy. Figure 6 illustrates the AUC values and the ROC curves obtained by the classifier on 14 pathology labels. It is evident from the ROC curve and AUC values that our model has provided the start of the art results outperforming the previously published results in literature. The AUC values are almost similar and high for all 14 pathology labels.

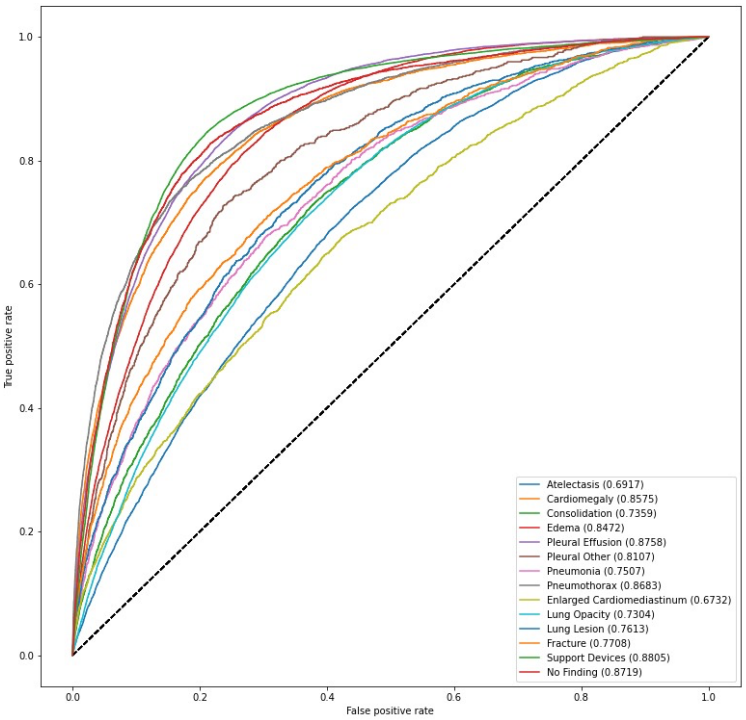

Figure 7 The AUC results and ROC curves obtained on CheXpert dataset.

While the dataset suffered from extreme class imbalance as illustrated in figure 2 and Table 1. To tackle the class imbalance, we used the class weight to balance the dataset. Previously, the research published on CheXpert dataset mostly covers to frontal CXRs or binary labels excluding the uncertain values and lateral views. Moreover, for the visualization of the areas of CXR predicted by the model to be the most indicative was done using Gradient-weighted Class Activation Mappings (Grad-CAMs). Figure 7 shows the examples of (Grad-CAMs).

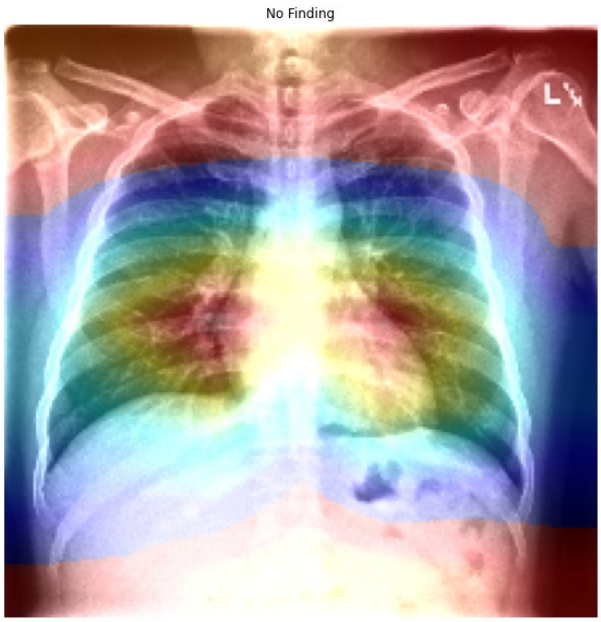

Figure 8 Example of Heat Map generated using gradient-weighted class activation mapping (Grad-CAMs)

4.1. Comparison with start of the art methods

In this section, we compare the results of our proposed approach with state-of-the-art results on the CheXpert dataset. For a fair comparison, we used the same training and test split as used by other researchers in the literature, with 80 percent for training and 20 percent for testing. However, in previous methods the researchers ignored u label cases or considered u label cases as “positive cases” whereas we considered the u labels as unlabeled cases which were labelled by the model with the labels with higher probability. Consequently, this approach improves the performance of our model. The proposed semi-supervised learning approach yielded the best per-label AUC in five pathologies: fracture, lung lesion, pleural other, pneumonia, and pneumothorax. The highest AUC of 0.88 among all pathology labels was attained for

support devices, whereas the highest AUC of 0.87 among all pathology labels in comparison to literature was achieved for pneumothorax. Most importantly, this comparison proves the validity of the assumption of our paper that using a semi-supervised approach where the model labels the unlabeled cases with predictions with higher probability is the most suitable and accurate approach for the multi-label problems, this will drastically improve the classification performance as seen in testing accuracy table 3. The proposed semi-supervised learning technique exceeds the results published by Allaouzi and Ahmed (2019) and Ho and Gwak. (2019) with an average of 1% as described in Table 6.

Table 6 Comparison of proposed approach with literature

Pathologies	Allaouzi & Ahmed (2019) Frontal view	Ho and Gwak (2019) Frontal view	Our Proposed Model Frontal/lateral view
Atelectasis	0.70	**0.71**	0.69
Cardiomegaly	**0.87**	0.79	0.85
Consolidation	0.74	**0.75**	0.74
Edema	**0.86**	**0.86**	0.84
Enlarged Cardio	**0.68**	0.55	0.67
Fracture	0.68	0.73	**0.77**
Lung Lesion	0.74	0.80	**0.76**
Lung Opacity	0.75	**0.78**	0.73
No Finding	**0.88**	0.85	0.87
Pleural Effusion	**0.90**	0.89	0.87
Pleural Other	0.74	0.68	**0.81**
Pneumonia	0.76	0.66	**0.75**
Pneumothorax	0.84	0.83	**0.87**
Support Devices	0.86	**0.91**	0.88
Average	0.78	0.77	**0.79**

Ho and Gwak (2019) implemented the feature extraction technique using the model DenseNet121 taking the original CheXpert dataset for the classification of 14 pathology labels. For the classification of uncertain labels, Ho and Gwak used an SVM model that efficiently worked for labels with high class instances, whereas it did not perform well for labels that suffered from extreme class imbalance. In literature the researchers have implemented the publicly available code of CheXNet of CheXpert dataset to determine the benchmark for the 14 pathology labels of CheXpert which shows that CheXNet is not efficient at classifying CheXpert labels as it classifies Chest X-ray14 efficiently.

Allaouzi & Ahmed (2019) proposed three different techniques binary relevance, label powerset and classifier chains using U-Ignore approach for the

classification of uncertain samples. All three techniques performed well on the CheXpert dataset whereas Binary Relevance provided the highest test accuracy of 0. 812.. Even though the technique provided the promising results, but a major drawback of this study is that the researchers deleted images labelled as uncertain samples along with lateral view images which reduced the size of the dataset drastically. So, it is evident that using BR technique along U-Ignore approach to classify uncertain samples is not an efficient approach.

CHAPTER FIVE

CHAPTER 5: DISCUSSION

Multilabel classification has drastically gained a lot of importance in medical imaging since the pandemic as several thorax disease cases report the presence of multiple diseases. In last 5 years different approaches and datasets are released with promising results however, a lot of research is done on Chest Xray-14 dataset and a little and limited research is conducted on CheXpert dataset. CheXpert is a large dataset that also comprises of not only positive and negative labels but also uncertain labels. One thing which common in all the previously published research is that despite of implementation of different classification techniques all the published studies have used an overparametrized DenseNet121 model with pretrained weights. However, we differentiate this study not only by implementing a semi-supervised approach but also training same model DenseNet121 with custom class weights to tackle the extreme class imbalance in the dataset. To be fair in comparison, we used the same train and test split as other methods with 80% for training and 20% for testing. As a result, we have noticed the using custom weights instead of pretrained weights did affect the performance results of the model showing and increase in accuracy. Our proposed method of semi-supervised learning outperformed current state of the models which implemented either U-Ones, U-zeros, and U-Uncertain approach to classify uncertain samples. There are a few limitations of CheXpert dataset to performing this experiment. First, There is an extreme class imbalance within the labels and more training data for similar labels need to be available. Second, no patient history is available to have the model to access the history of the patient. This is the first detailed study of the CheXpert dataset. Previously, all published studies were conducted on chest X-ray14 and simultaneously run the same on the model on CheXpert without optimizing the parameter or highlighting the limitations of the labels present in CheXpert. However, this study not only focused on the classification of all 14 pathologies present in CheXpert, but also highlighted the limitations of the dataset.

CHAPTER SIX

CHAPTER 6: CONCLUSION

In this paper, we propose a semi-supervised learning framework, Self-Training Model (U-SelfTrained) to address multi-label disease classification in chest radiography. The task was carried out using DenseNet-121 with custom weights. Through the implementation of custom class weights class imbalance problem is also addressed. The evaluation of the model was conducted using the performance metrics like F1-score, Recall, Precision, and Average AUC. Self-training approach not only accurately predicts positive and negative cases but also it classifies uncertain cases accurately. Moreover, U-SelfTrained approach classifies the uncertain cases with prediction with higher probability which classifies them either positive or negative cases. The quantitative results demonstrates that our method achieves the state-of-the-art results 0.863 AUROC respectively. To further substantiate this research in future train the DenseNet-121 with our custom weights on more balanced data to avoid the problem of imbalanced label distribution. This research has crucial practical significance of implementation of AI in aiding radiologists to improve the work efficiency and diagnostic accuracy which will reduce the chance of misdiagnosis drastically and improve the quality of diagnosis in thorax disease detection. In future we plan to research the efficacy of Ensemble technique on common diseases present in both the datasets Chest Xray14 and CheXpert by merging them. We further plan on developing a model based on natural language processing and Deep learning in which model will have the access to patient history provided by the radiologist along with the images to classify the samples in datasets available in future.

REFERENCES

[1] Garbin, C., Rajpurkar, P., Irvin, J., Lungren, M. P., & Marques, O. (2021). Structured dataset documentation: a datasheet for CheXpert. *arXiv preprint arXiv:2105.03020*.

[2] Sundaram, S., & Hulkund, N. (2021). Gan-based data augmentation for chest X-ray classification. *arXiv preprint arXiv:2107.02970*.

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

[4] Majkowska, A., Mittal, S., Steiner, D. F., Reicher, J. J., McKinney, S. M., Duggan, G. E., ... & Shetty, S. (2020). Chest radiograph interpretation with deep learning models: assessment with radiologist-adjudicated reference standards and population-adjusted evaluation. *Radiology*, *294*(2), 421-431.

[5] Takeuchi, D., Thai, R., & Tran, K. (2019). Exploring model architectures and view-specific models for chest radiograph diagnoses.

[6] Seyyed-Kalantari, L., Liu, G., McDermott, M., Chen, I. Y., & Ghassemi, M. (2020). CheXclusion: Fairness gaps in deep chest X-ray classifiers. In *BIOCOMPUTING 2021: proceedings of the Pacific symposium* (pp. 232-243).

[7] Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., ... & Ng, A. Y. (2019, July). Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 33, No. 01, pp. 590-597).

[8] Phillips, N. A., Rajpurkar, P., Sabini, M., Krishnan, R., Zhou, S., Pareek, A., ... & Lungren, M. P. (2020, November). CheXphoto: 10,000+ photos and transformations of chest X-rays for benchmarking deep learning robustness. In *Machine Learning for Health* (pp. 318-327). PMLR.

[9] Guan, Q., Huang, Y., Luo, Y., Liu, P., Xu, M., & Yang, Y. (2021). Discriminative Feature Learning for Thorax Disease Classification in Chest X-ray Images. IEEE Transactions on Image Processing, 30, 2476-2487.

[10] Bressem, K. K., Adams, L. C., Erxleben, C., Hamm, B., Niehues, S. M., & Vahldiek, J. L. (2020). Comparing different deep learning architectures for classification of chest radiographs. *Scientific reports*, *10*(1), 1-16.

[11] Kuo, P. C., Tsai, C. C., López, D. M., Karargyris, A., Pollard, T. J., Johnson, A. E., & Celi, L. A. (2021). Recalibration of deep learning models for abnormality detection in smartphone-captured chest radiograph. *NPJ digital medicine*, *4*(1), 1-10.

[12] Bajwa, N., Bajwa, K., Rana, A., Shakeel, M. F., Haqqi, K., & Khan, S. A. (2020). A generalized deep learning model for multi-disease Chest X-Ray diagnostics. *arXiv preprint arXiv:2010.12065*.

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

[14] Zhao, J., Li, M., Shi, W., Miao, Y., Jiang, Z., & Ji, B. (2021). A deep learning method for classification of chest X-ray images. In Journal of Physics: Conference Series (Vol. 1848, No. 1, p. 012030). IOP Publishing.

[15] Madjarov, G., Kocev, D., Gjorgjevikj, D., & Džeroski, S. (2012). An extensive experimental comparison of methods for multi-label learning. Pattern recognition, 45(9), 3084-3104.

[16] Kumar, P., Grewal, M., & Srivastava, M. M. (2018, June). Boosted cascaded convnets for multilabel classification of thoracic diseases in chest radiographs. In International conference image analysis and recognition (pp. 546-552). Springer, Cham.

[17] Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., ... & Ng, A. Y. (2017). Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. arXiv preprint arXiv:1711.05225.

[18] Khanh Ho, T. K., & Gwak, J. (2019). Multiple feature integration for classification of thoracic disease in chest radiography. Applied Sciences, 9(19), 4130.

[19] Li, Z., Wang, C., Han, M., Xue, Y., Wei, W., Li, L. J., & Fei-Fei, L. (2018). Thoracic disease identification and localization with limited supervision. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 8290-8299).

[20] Rawat, W., & Wang, Z. (2017). Deep convolutional neural networks for image classification: A comprehensive review. Neural computation, 29(9), 2352-2449.

[21] Hwang, E. J., Park, S., Jin, K. N., Im Kim, J., Choi, S. Y., Lee, J. H., ... & Park, C. M. (2019). Development and validation of a deep learning–based automated detection algorithm for major thoracic diseases on chest radiographs. JAMA network open, 2(3), e191095-e191095.

[22] Rubin, J., Sanghavi, D., Zhao, C., Lee, K., Qadir, A., & Xu-Wilson, M. (2018). Large scale automated reading of frontal and lateral chest x-rays using dual convolutional neural networks. arXiv preprint arXiv:1804.07839.

[23] Rajan, D., Thiagarajan, J.J., Karargyris, A. and Kashyap, S., 2021, February. Self-training with improved regularization for sample-efficient chest x-ray classification. In Medical Imaging 2021: Computer-Aided Diagnosis (Vol. 11597, pp. 418-425). SPIE.

[24] Xie, Q., Luong, M. T., Hovy, E., & Le, Q. V. (2020). Self-training with noisy student improves imagenet classification. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 10687-10698).

[25] Radosavovic, I., Dollár, P., Girshick, R., Gkioxari, G. and He, K., 2018. Data distillation: Towards omni-supervised learning. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4119-4128).

[26] Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. and Batra, D., 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE international conference on computer vision (pp. 618-626).

[27] Dai, Y., Gieseke, F., Oehmcke, S., Wu, Y. and Barnard, K., 2021. Attentional feature fusion. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (pp. 3560-3569).

[28] Chen, B., Li, J., Lu, G., Yu, H. and Zhang, D., 2020. Label co-occurrence learning with graph convolutional networks for multi-label chest x-ray image classification. IEEE journal of biomedical and health informatics, 24(8), pp.2292-2302

[29] Liu, Q., Yu, L., Luo, L., Dou, Q. and Heng, P.A., 2020. Semi-supervised medical image classification with relation-driven self-ensembling model. IEEE transactions on medical imaging, 39(11), pp.3429-3440

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

[31] Guan, Q., Huang, Y., Zhong, Z., Zheng, Z., Zheng, L. and Yang, Y., 2018. Diagnose like a radiologist: Attention guided convolutional neural network for thorax disease classification. arXiv preprint arXiv:1801.09927.

[32] Yarowsky, D., 1995, June. Unsupervised word sense disambiguation rivaling supervised methods. In 33rd annual meeting of the association for computational linguistics (pp. 189-196).