TOHOKU
UNIVERSITY

博 士 学 位 論 文

論文題目 Effect of Training Data Characteristics on Deep Learning Performance for COVID-19 Diagnosis using Chest X-ray Images
（深層学習における訓練データの性質が COVID-19 画像診断性能に与える影響に関する研究）

提 出 者　東北大学大学院医工学研究科
医工学専攻
学籍番号　B8WD9002
氏　　名　Zhang ZHANG

修了年度	2023 年度	課程	博士課程後期 3 年の課程

英文 Abstract

Title: Effect of Training Data Characteristics on Deep Learning Performance for COVID-19 Diagnosis using Chest X-ray Images

Author: Zhang ZHANG

Supervisor: Noriyasu HOMMA

Deep learning (DL)-based methods are increasingly applied within medical fields. Since DL needs large amount of training data that is often difficult to be collected from a single medical facility, data from multi-facilities with different settings can be used. A specific class imbalance, called intra-source imbalance, within the data collected from each medical facility might affect the performance, but received negligible attention and thus the impact remains unclear. This dissertation aims to clarify the impact by selecting the COVID-19 diagnosis using chest X-ray (CXR) images as a case study. To this end, we utilized two different datasets, both of which contain COVID-19 and non-COVID-19 images, to train a commonly used DL model, VGG-16. One dataset is intra-source imbalanced: each medical facility only provided COVID-19 data or non-COVID-19 data. The other dataset is intra-source balanced: it collected all data from a same medical facility. We removed lung regions from the CXRs and used the lungs-removed images to train and test VGG-16. Then, we made a cross-dataset test, that trains the model using one dataset and tests it on another dataset, to evaluate the performance of the models. As the results, for the imbalanced dataset, the model achieved the area under the receiver operating characteristic curve (AUC) of 0.99 on lungs-removed data which is unexpected for a lung-based illness. For the balanced dataset, the AUC value was 0.53 on lungs-removed data. In the cross-dataset test, the imbalanced dataset-trained model achieved 0.5 AUC on balanced dataset, which means nearly no classification capacity. In contrast, the balanced dataset-trained model achieved 0.84 AUC on the imbalanced dataset. The visualization results showed that areas outside lung regions strongly impacted the decisions of the imbalanced dataset-trained model, while the balanced dataset-trained model relied on area inside lung regions. This study reveals clear evidence that the intra-source balance of training data is vital for DL and suggests how to prepare a training dataset for DL-based methods.

.

和文アブストラクト

論文題目：　深層学習における訓練データの性質が COVID-19 画像診断性能に与える影響に関する研究

提出者氏名：張　彰　チョウ　ショウ

指導教員：　本間　経康

医療分野でも深層学習の応用が進み、多くの優れた成果が報告されている。深層学習は大量の学習データを必要とするが、希少疾患や新規疾患の場合、単一施設で大量の医療データを用意することは困難を伴うため、多施設からの寄せ集め的収集を余儀なくされることも多い。しかし、寄せ集めに起因するデータの不均衡は医療統計学的問題を生じやすいことが知られている。実際 COVID-19 感染爆発のように、未知の疾患でかつ迅速性が要求される場合、各医療施設から収集するデータの性質は十分検討できない場合もある。とくに、同一施設から収集される疾患群と対照群の不均衡（施設不均衡と呼ぶ）は、これまで注目されておらず、深層学習の性能にどの程度の悪影響を及ぼすか不明であった。本論文では、胸部 X 線画像における COVID-19 の深層学習を用いた診断を例に、この施設不均衡の影響を明らかにすることを目的とする。

代表的な深層学習モデルである VGG-16 と、COVID-19 疾患群と対照群を含む、2 つの異なる胸部 X 線データセットを使用した。1 つ目は、疾患群と対照群をそれぞれ異なる施設から寄せ集め的に収集した施設不均衡なデータセットである。2 つ目は、疾患群、対照群とも同一施設で収集した施設均衡なデータセットである。ただし、両データセットとも疾患群、対照群の症例数は同一（均衡）である。また、胸部 X 線画像の肺領域を特定し、原画像から肺を除去（黒埋め）した肺除去画像も作成した。COVID-19 の診断（鑑別）性能は受信者動作特性曲線下面積(Area under the curve, AUC)を用いて評価した。AUC は 0 から 1 の連続値で、1 に近いほど高性能を意味する。不均衡なデータセットを用いた場合の VGG-16 の鑑別性能は、原画像ならびに肺除去画像で訓練した結果が、いずれも AUC が 0.99 以上のほぼ完璧な鑑別性能を達成した。しかし、COVID-19 の多くは肺疾患を発症するため、肺除去画像での高性能鑑別は予想外であり、医学的に妥当であるとは言い難い。一方、均衡データセットの場合、原画像と肺除去画像を用いた訓練による AUC 値は、それぞれ 0.74 と 0.53 となり、予想通り肺除去による影響で鑑別性能が大きく低下したことが示唆された。さらに、VGG-16 の鑑別根拠を注目領域として可視化した結果、不均衡データセットで訓練した場合は、肺野以外の部分が鑑別に強く影響することが示されたのに対し、均衡データセットで訓練した場合は、主に肺野に注目していた。これらの結果は、深層学習ベースの医用画像診断において、データの施設均衡が重要であることをはじめて実証したものであり、高性能を実現するために訓練データセットが備えるべき新たな統計的性質を示唆するものである。

Declaration of Authorship

I, Zhang Zhang, declare that this thesis titled, "Effect of Training Data Characteristics on Deep Learning Performance for COVID-19 Diagnosis using Chest X-ray Images" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: _____

Date: _____

Acknowledgements

First and foremost, I would like to thank my supervisor, Prof. Noriyasu HOMMA at Tohoku University. Ph.D. study was not easy, but under his patient guidance, support, and encouragement, I enjoyed and appreciated this journey. He always appreciates me on my every little progress. He always provides essential supports for us to succeed in research. He is enthusiastic and he is always there for his students and willing to help. He encourages us to share our ideas and make great research together. I appreciate him creating a wonderful lab environment for us.

I would like to express my thank to Prof. Xiaoyong ZHANG, who gave me continuous support in my research during my Master study and Ph.D. study. I appreciate his time revising my research papers especially for the late-night work near deadlines. I appreciate all the supportive words and suggestions from him during my Ph.D. study. I would also like to thank Prof. Kei ICHIJI for his advice and suggestions on my research.

I would like to thank Prof. Yoshifumi SAIJO and Prof. Norihiro SUGITA for their constructive comments. Their comments gave me many ideas about my research and helped me to refine my research.

I would like to thank Global HAGI scholarship and the Pioneering Research Support Project for providing me financial support towards my study.

I would like to thank my lab mates for helping me to find my errors in papers and code. I appreciate my friends in Tohoku University for accompanying me when I was upset. I appreciate my old friends for listening to my complaints and giving me advice.

I would like to thank my family, especially my cousin, Mr. Jie SONG at Ritsumeikan

University. I appreciate him accompanying and supporting me when I felt anxious during my illness. I would like to thank my parents for their wise counsel and sympathetic ear. They are always there for me. Finally, I appreciate my wife, Ms. Zhiqiong FU, for all of her support and understanding.

TOHOKU UNIVERSITY

Abstract

Graduate School of Biomedical Engineering

Doctor of Philosophy

Effect of Training Data Characteristics on Deep Learning Performance for COVID-19 Diagnosis using Chest X-ray Images

by Zhang Zhang

Deep learning-based methods are increasingly developed, especially for use in medical fields. Since deep learning needs large amount of training data, medical data are often acquired in different settings among medical facilities as much as possible. A specific class imbalance, called *intra-source imbalance*, within the data collected from each medical facility might affect the performance of deep learning-based methods, but received negligible attention and thus the impact of the intra-source imbalance remains unclear. In this dissertation, we aim to clarify the impact by selecting the COVID-19 diagnosis using chest X-ray (CXR) images as a case study. To this end, we utilized two different CXR datasets, both of which contains COVID-19 data and non-COVID-19 data, to train and test VGG-16, a commonly used deep learning model. One dataset is an intra-source imbalanced dataset, because each medical facility only provided COVID-19 data or non-COVID-19 data. The other dataset is an intra-source balanced dataset, because all the data were collected from the same medical facility. We segmented lung regions from the CXRs, and then used the original data and lungs-removed data to train and test VGG-16, separately. Then, we made a cross-dataset test, that trains a model using one dataset and tested it on another dataset, to evaluate the performance of the VGG-16 models by using the area under receiver operating characteristic curve (AUC) value. Finally, we used Local Interpretable Model-agnostic Explanations (LIME) to visualize the explanations for the decisions made by the models in the cross-dataset test. As the results, for the imbalanced dataset, VGG-16 models trained by original data or lungs-removed data all achieved an AUC value over 0.99. Since COVID-19 is a lung-based illness, the result is unexpected and reveals an unreliability in terms of medical findings. For the balanced dataset, the AUC value was 0.74 on original data, and it was decreased to 0.53 when using lungs-removed data. In the cross-dataset test, when trained the model by the imbalanced dataset and tested on the balanced dataset, the AUC was 0.5 approximately, which means nearly no classification capacity. In contrast, the model

trained by the balanced dataset achieved 0.84 AUC value when tested on the imbalanced dataset. The visualization results showed that areas outside lung regions strongly impacted the decisions of the model trained by the imbalanced dataset, while the model trained by balanced dataset relied on area in the lung regions to make decisions. This study reveals clear evidence that the intra-source balance is vital for training data to minimize the risk of poor performance of deep learning-based methods and suggests how to prepare a training dataset for deep learning.

Contents

Abstract **v**

List of Figures **ix**

List of Tables **xvi**

List of Abbreviations **xvii**

1 Introduction **1**

2 Fundamental Studies and Related Research **6**

2.1 Fundamental studies of deep learning 7

2.1.1 Convolutional neural networks (CNNs) 7

Forward propagation 8

Back propagation 11

2.1.2 Deep learning models for image classification 15

AlexNet 15

VGG (Visual Geometry Group) 16

2.1.3 The lack of explainability 16
2.1.4 Data characteristics' impact on deep learning 19
2.2 COVID-19 Diagnosis Using Chest X-ray 22
2.2.1 Deep Learning-Based Methods for COVID-19 Diagnosis Using Chest X-ray 24

3 Comparison experiments and cross-dataset test 29
3.1 Datasets 30
3.2 Experiments 32
3.2.1 Comparison experiments 32
Results of Comparison experiments 33
3.2.2 Cross-dataset test 35
Results of cross-dataset test 35
3.3 summary 36

4 Visualization investigation in the cross-dataset test 43
4.1 LIME method 44
4.2 Visualization results 45
4.3 Summary 46

5 Discussion 52
5.1 Investigation for proper balance level 52
5.2 Investigation for the impact of unobserved features 53

6 Conclusion 63
6.1 summary 63
6.2 Future Directions 65

Bibliography 69

List of Figures

2.1 Architecture of a CNN named LeNet-5 network. 8
2.2 An example of convolution operation. At each location, an element of a convolution kernel is used as the weight for the element it overlaps in the previous feature map. The results are summed up to obtain the element at the location in the convolution result. 9
2.3 An example of max pooling. It calculates the largest value in each patch of each feature map. 11
2.4 An example of fully connected layers. Fully connected layers operate a linear combination on a flattened input where each input element is connected to all neurons. 12
2.5 Architecture of AlexNet. 15
2.6 Architecture of VGG-16. 16

2.7 Examples of results in MCC detection and classification using mammograms: (A) output of a deep learning-based method, (B) output of a clustering-based method. The deep learning-based method could obtain a high performance, but it was unable to explain the decision-making. On the other hand, the clustering-based method could provide not only the predicted class, but also several human-designed features to explain the decision. 17

2.8 An example of class balanced training data and class imbalanced training data. Red means positive class and green means negative class. The class imbalanced training data can lead classifiers to exhibit bias towards the majority class and ignore the minority class. 20

2.9 An example of data homogeneity. Red means positive class and green means negative class. The classifier trained by data from one medical facility could fail to classify data from the other medical facility. 22

2.10 Examples of CXR images: (A) a CXR image from a health case, (B) a CXR image from a COVID-19 case. Abnormalities can be found in the image from the COVID-19 case. 27

2.11 An overview of the previous study. CXR images with lung regions removed are utilized to investigate the reliability of deep learning models for COVID-19 classification. Deep learning models can achieve high accuracy when images with lung regions are removed, and the focused locations are outside the lung regions when deep learning models make a COVID-19 prediction. The result indicates the deep learning models are unreliable in terms of medical findings, but the cause of the unreliable performance is still unknown. 28

3.1 Examples of positive and negative CXR images in the two datasets: (A) a positive CXR image in the Qata-COV19 dataset, (B) a negative CXR image in the Qata-COV19 dataset, (C) a positive CXR image in the BIMCV dataset, (D) a negative CXR image in the BIMCV dataset. . 38

3.2 Overview of the comparative experiment. ROI hide-and-seek protocol operated (A) original images from the Qata-COV19 dataset or the BIMCV dataset to emphasize and hide the lung regions, respectively. (B) lungs-isolated images and (C) lungs-framed images were generated by emphasizing the lung regions, while (D) lungs-removed images and (E) lungs-boxed-out images were generated by hiding the lung regions. The original datasets and the modified datasets were utilized to train and test a VGG-16 model separately. 39

3.3 ROC curves for the VGG-16 models trained and tested on the original or the modified datasets from Qata-COV19: (A) original CXR images, (B) lungs-isolated images, (C) lungs-framed images, (D) lungs-removed images, and (E) lungs-boxed-out images. The deep learning models achieved high performance, even with hidden lung regions. . . 40

3.4 ROC curves for the models trained and tested on the original or the modified datasets from BIMCV: (A) original CXR images, (B) lungs-isolated images, (C) lungs-framed images, (D) lungs-removed images, and (E) lungs-boxed-out images. The performance degraded a lot when lung regions were hidden. 41

3.5 ROC curves for the cross-dataset test: (A) testing the Qata-COV19-trained model (the same model weights as in FIGURE 3.3 (A)) on the BIMCV dataset, (B) testing the BIMCV-trained model (the same model weights as in FIGURE 3.4 (A)) on the Qata-COV19 dataset. The model trained on original images from BIMCV dataset was able to classify original images from the Qata-COV19 dataset, while the model trained on original images from the Qata-COV19 dataset failed to classify original images from the BIMCV dataset. 42

4.1 An example of LIME method: (a) an input CXR image, (b) segmented superpixels based on the color similarity and proximity, (c) the weights of each superpixel, and (d) top-5 superpixels of the input CXR image. 47

4.2 The results when selecting top-5 superpixels as the saliency map. (A) Tested on BIMCV test subset, (B) tested on Qata-COV19 test subset. The model trained by the BIMCV dataset performed better than the model trained by the Qata-COV19 dataset. 48

4.3 The LIME explanations for classifying a positive case by (A) VGG-16 model trained by Qata-COV19 dataset (prediction: positive), and (B) VGG-16 model trained by BIMCV dataset (prediction: positive). Blue areas contribute to positive prediction and green areas contribute to negative prediction. Both of the models made a true decision, but the model trained by BIMCV focused more inside the lung regions while the model trained by Qata-COV19 focused on the label and background. 49

4.4 The LIME explanations for classifying another positive case by (A) VGG-16 model trained by Qata-COV19 dataset (prediction: positive), and (B) VGG-16 model trained by BIMCV dataset (prediction: negative). Blue areas contribute to positive prediction and green areas contribute to negative prediction. The model trained by Qata-COV19 correctly classified the image but focused on the label and background. The model trained by BIMCV focused more on the lung regions even it made a wrong decision. 50

4.5 The LIME explanations for classifying a negative case by (A) VGG-16 model trained by Qata-COV19 dataset (prediction: positive), and (B) VGG-16 model trained by BIMCV dataset (prediction: negative). Blue areas contribute to positive prediction and green areas contribute to negative prediction. The model trained by Qata-COV19 focused on the label and background information in the image and classified this negative image into positive class. 51

5.1 ROC curves for the VGG-16 models trained on data with different intra-source balance level: (A) original BIMCV dataset (100%), (B) data with 90% intra-source balance level, (C) data with 70% intra-source balance level, (D) data with 50% intra-source balance level, (E) data with 40% intra-source balance level, and (F) data with 30% intra-source balance level. When using a dataset with intra-source balance below 50%, the intra-source imbalance could influence the performance. 56

5.2 LIME explanations for the VGG-16 model trained by Qata-COV19 dataset. Blue areas contribute to positive prediction and green areas contribute to negative prediction. The markers 'D' and 'DCH' are always always the LIME explanations for positive predictions. (A) and (B) are positive CXRs from BIMCV dataset. (C) and (D) are negative CXRs from BIMCV dataset. 57

5.3 We selected CXRs without markers from Qata-COV19 dataset and used the CXRs to train a VGG-16 model. The model was tested on CXRs from BIMCV dataset. 58

5.4 The ROC curve for testing the VGG-16 model trained by selected data. The model trained by selected data achieved 0.57 AUC value on BIMCV dataset, which showed the model still failed to classify the CXRs in BIMCV dataset. 58

5.5 The SC values when testing different models on BIMCV dataset. The model trained by selected data focused more on the lung regions. . . . 59

5.6 The LIME explanations for classifying a negative case from BIMCV dataset by (A) VGG-16 model trained by selected data (prediction: negative), and (B) VGG-16 model trained by original Qata-COV19 dataset (prediction: positive). Blue areas contribute to positive prediction and green areas contribute to negative prediction. The model trained by selected data focused more on lung regions and made true prediction. 60

5.7 The LIME explanations for classifying a positive case from BIMCV dataset by (A) VGG-16 model trained by selected data (prediction: negative), and (B) VGG-16 model trained by original Qata-COV19 dataset (prediction: positive). Blue areas contribute to positive prediction and green areas contribute to negative prediction. The model trained by selected data did not focus on the marker and made a false prediction. 61

5.8 The LIME explanations for false positive predictions made by VGG-16 model trained by selected data (prediction: positive). Blue areas contribute to positive prediction. The negative CXRs from BIMCV dataset were classified in positive class and the model focus on the lung regions. 62

List of Tables

3.1 Our study used two image datasets (Qata-COV19, BIMCV); Qata-COV19 has images provided from various facilities and only for a single category, while BIMCV collected images from the same facility. 31

List of Abbreviations

AUC	**A**rea **U**nder the **C**urve
CAD	**C**omputer **A**ided **D**iagnosis
COVID-19	**CO**rona**VI**rus **D**isease 20**19**
CNN	**C**onvolutional **N**eural **N**etwork
CXR	**C**hest **X**-**R**ay
Grad-CAM	**Grad**ient-weighted **C**lass **A**ctivation **M**apping
GTC	**G**round **T**ruth **C**overage
IoU	**I**ntersection **o**f the **U**nion
LIME	**L**ocal **I**nterpretable **M**odel-agnostic **E**xplanations
MCC	**M**icro-**C**alcification **C**lusters
PCR	**P**olymerase **C**hain **R**eaction
ROC	**R**eceive **O**perating **C**haracteristics
ROI	**R**egion **O**f **I**nterest
SC	**S**aliency **C**overage
SLIC	**S**imple **L**inear **I**terative **C**lustering
VGG	**V**isual **G**eometry **G**roup

Chapter 1

Introduction

Over the last decades, medical imaging, e.g. computed tomography (CT), magnetic resonance image (MRI), positron emission tomography (PET), mammography, ultrasound, X-ray and so on, has shown the importance for the early detection, accurate diagnosis, and effective treatment of diseases. However, the medical image interpretation needs to be performed by human experts and sometimes it is difficult even for experts. Therefore, computer-aided diagnosis (CAD) systems, which applied computer vision and artificial intelligence (AI) including machine learning techniques on medical image interpretation, were proposed to help doctors in detection and differential diagnosis of many different types of abnormalities in medical images.

Prior to the proposal of CAD systems, computer systems have been used in picture archiving and communication systems (PACS) for the management of the medical images, but it seems unlikely to bring a significant clinical benefit to radiologists (Doi, 2007). To realize a major benefit in radiologists' daily work, it led to the

concept of automated computer diagnosis (Doi, 2007). Before 1980s, there were some attempts for automated computer diagnosis but they were not successful. After that, another approach, named computer-aided diagnosis, has spread widely and quickly. It assumed the computer output should be used as a "second opinion" but not the final decision. Now, machine learning has a potential to perform on par with medical experts and plays a key role in the daily work of doctors. Doctors rely on CAD systems for the detection, diagnosis, and treatment of diseases, and the research of CAD systems has been increased rapidly.

Different types of machine learning approaches are adapted in CAD systems, such as unsupervised learning and supervised learning. Unsupervised learning refers to the use of machine learning algorithms to identify patterns in datasets without labels. Clustering, which groups similar data into a cluster, is a typical unsupervised learning algorithm. For example, micro-calcification clusters (MCCs), important signs at an early stage of breast cancer, can be classified by using clustering (Zhang et al., 2020). Regarding the radiologists' workflow, the MCCs can be classified into grouped, regional, diffuse, segmental, and linear categories refer to the arrangement in the breast, and the spatial distribution categories are relative to the risk of malignancy. To mimic the workflow, a Gaussian mixture model-based method was proposed to extract the main features of MCCs, such as the area, the eccentricity, and the direction. Based on the extracted features, the MCCs can be clustered into five spatial distribution categories. The results showed the accuracy was 68%. On the other hand, supervised learning is defined by its use of labeled datasets to train algorithms for classification or prediction. For example, support vector machines have shown accurate results for various disease diagnosis tasks.

Among the supervised learning algorithms, deep learning has attracted more and more attention in recent years. One important advantage of deep learning is

that it does not need to manually design features from the images. In early days, feature engineering was an important step for developing a new CAD algorithm. Researchers should understand workflow of doctors for the task and design relevant features to mimic the workflow. The CAD algorithms often can extract the features automatically and combine the features into a computer score. However, to propose such a CAD system for complicated tasks, high-level expert knowledge is necessary, and the hand-crafted features may only work for limited cases and may be not robust. Therefore, image-based CAD algorithms without without the need of manually designed features were proposed, and deep learning-based CAD algorithms were the most representative image-based algorithms. Another advantage of deep learning is its groundbreaking performance. Liu et al. (2019) compared the performance of deep learning and health-care professionals in detecting diseases from medical images, and they reported deep learning algorithms have equivalent sensitivity and specificity to the professionals. Because of these advantages, deep learning algorithms are expected to solve new tasks, i.e., coronavirus disease 2019 (COVID-19) diagnosis.

On the other hand, the lack of accepted theoretical explanation remains the fundamental problem of deep learning, i.e., the black-box nature. The cause is that deep learning models lack transparency and explainability; it is difficult to know and understand how the model made a prediction, and the inner workings remain opaque to the outside observer. Without a sufficient understanding of a mechanism behind the machine-made prediction, it becomes very complicated to detect hidden risks in deep learning-based methods, i.e., training bias caused by mislabeled training data, especially for medical applications.

Because deep learning-based methods always apply huge data to train a standard architecture network as a black box, the performance of them strongly relies on

the training data. Therefore, many researches discussed the influence of data characteristics on the performance of deep learning-based methods. One important data characteristic focused by researchers is class imbalance, which means a difference in quantity among categories. When trained by a class imbalanced dataset, deep learning might over-classify the majority group due to its increased prior probability. Moreover, data from only one medical facility might have some limitations, such as over-representation of vulnerable populations. Using the data from one medical facility might lead to a hidden risk that the results could not depend on the desired properties of the data but instead depend on potentially unobserved aspects. Collecting high-variability images from different medical facilities are expected to solve this problem (McDermott et al., 2021). Therefore, researchers always collected as much data as possible from different medical facilities. Due to the class imbalance within each medical facility and the different settings among medical facilities, the issue of intra-source imbalance often occurs in medical datasets. This imbalance could also impact the performance of deep learning-based methods but receives negligible attention.

In this dissertation, for investigating the impact of intra-source imbalance on the performance of deep learning-based methods, we select the COVID-19 diagnosis in chest X-ray (CXR) images as a case study. The contents of this dissertation are organized as follows:

Chapter 1 is the introduction part, which introduce the background and purpose of this study. The outline of this dissertation is also given in this chapter. Chapter 2 is about the fundamental studies and related researches. We introduced the architecture of deep learning models and several important data characteristics when training deep learning models. In addition, previous studies on COVID-19

diagnosis in CXR images using deep learning are reviewed in this chapter. In Chapter 3, we talk about our comparison experiment and the cross-dataset test. In this chapter, our experimental results reveal the risk of unreliability when using intra-source imbalanced datasets in deep learning methods. Chapter 4 talks about the visualization investigation of the cross-dataset test in Chapter 3. The explanations for the decisions made by deep learning models are visualized by LIME (Local Interpretable Model-agnostic Explanations) method (Ribeiro, Singh, and Guestrin, 2016). The results intuitively show that the model trained by the intra-source imbalanced dataset classifies images based on the features representing data sources but not the features representing COVID-19. Chapter 5 is the discussion part. It has discussed the proper inter-source balance level for training data. The results showed when using a dataset with intra-source balance below 50%, the intra-source imbalance could influence the performance. Moreover, we demonstrated potentially unobserved aspects inside lung regions could also aspect deep learning performance. Chapter 6 draws conclusions of this dissertation and includes several possible extensions and directions we could further explore.

Chapter 2

Fundamental Studies and Related Research

COVID-19 has been widely spread worldwide and continues to have a devastating effect on the health and life of the global population. The polymerase chain reaction (PCR) test is the gold standard for detection nowadays, but it is time-consuming and laborious, and it is also suffering from the high cost. As one of the essential complements to PCR testing, chest X-ray (CXR) imaging has also demonstrated its effectiveness in current diagnosis. The CXR imaging is often part of the standard procedure for patients with respiratory complaints, and it is reported that some patients showed abnormalities in the CXR images before they eventually test positive for COVID-19 with the PCR test. Moreover, all the rapid triaging, availability, accessibility, and portability of CXR imaging indicated that it could be a preliminary tool

for COVID-19 screening. Nonetheless, one of the biggest bottlenecks of CXR screening is the need for experts to diagnose from the CXR images because the radiological signatures can be subtle.

The deep learning-based methods can actually enhance the diagnosis performance by radiologists (Homma et al., 2020b) and aid image diagnosis in the lung areas that is not easy even for the experts (Homma et al., 2020a). The success made by deep learning-based methods encouraged researchers to develop deep learning-based CAD systems which are expected to aid radiologists in detecting COVID-19 in CXR testing more rapidly and accurately. The purpose of this chapter is to introduce the fundamental studies of deep learning, including the structures, the commonly-used models, and the limitations of deep learning, as well as related research on deep learning-based methods for COVID-19 diagnosis in CXR images.

2.1 Fundamental studies of deep learning

Deep learning has attracted a high interest in the image classification and its architectures appear to solve problems that require complex highly-varying functions. In order to deal with complex problems, deep learning techniques learn characteristic hierarchies with features from higher levels of hierarchy formed by a composition of lower level features. Deep learning assimilates complex behaviors with expansive information sets to select effective characteristics automatically by convolutional neural network (CNN) structures (Affonso et al., 2017).

2.1.1 Convolutional neural networks (CNNs)

Neural network is comprised of neurons or units with some activation α and parameters $\theta = \{\omega, \beta\}$, where ω is a set of weights and β is a set of biases. The activation

can be computed as follows:

$$\alpha = \sigma(\boldsymbol{\omega}^T \boldsymbol{x} + \boldsymbol{\beta}),$$

where $\boldsymbol{x}$ is the input and σ is a nonlinear function.

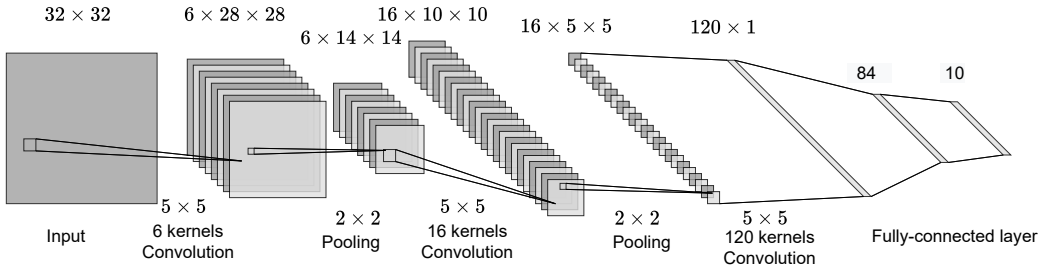


FIGURE 2.1: Architecture of a CNN named LeNet-5 network.

As a popular type of neural network, CNNs can learn complex features from available images and use these features to classify the images. A typical CNN structure would constitute a number of convolutional and pooling layer sets with fully-connected layers attached at the end for the classification (Garg and Mago, 2021). As shown in FIGURE 2.1, a convolutional layer contains several convolution kernels, which extract different features and generate different feature maps for the same input images. Every neurons of a feature map are connected to the neurons in the previous feature maps. To reduce the dimentionality of the features, pooling layers provide an approach to down sample feature maps by summarizing the presence of features in patches of the feature map. At the end of a CNN, the fully connected layers applies parameters to all neurons from the previous layer to generate an appropriate score for the prediction.

Forward propagation

Forward propagation refers to the calculation and storage of intermediate variables (including outputs) for a CNN in order from the input layer to the output layer. We

now introduce the computation of several important layers in forward propagation.

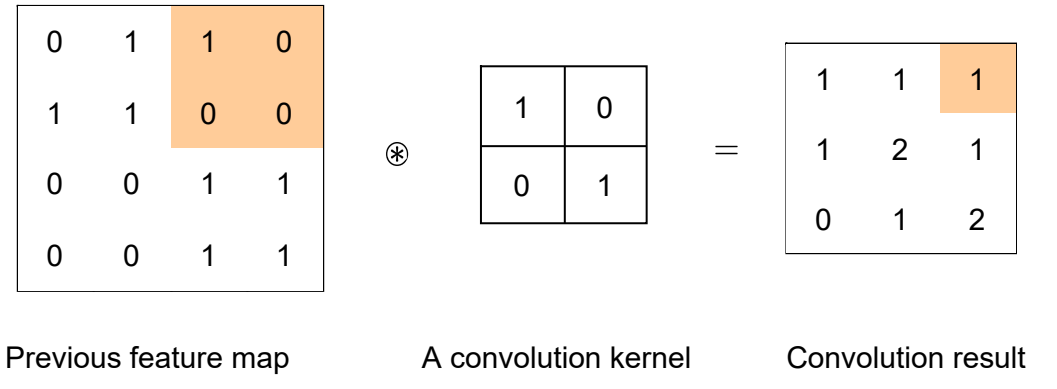


FIGURE 2.2: An example of convolution operation. At each location, an element of a convolution kernel is used as the weight for the element it overlaps in the previous feature map. The results are summed up to obtain the element at the location in the convolution result.

The convolutional layer is the central part of a CNN. To perform a convolution operation, the convolution kernels are flipped by 180 degrees and then slid across the previous feature maps in equal and finite strides to generate new feature maps. As shown in FIGURE 2.2, at each location, an element of a convolution kernel is used as the weight for the element it overlaps in the previous feature map. The results are summed up to obtain the element at the location in the convolution result. Different convolution kernels in a convolutional layer can help to form as many feature maps as desired. Specifically, new feature maps $\boldsymbol{X}^l$ can be obtained by first convolving the previous feature maps $\boldsymbol{X}^{l-1}$ with a set of kernels $\boldsymbol{\omega} = \{\boldsymbol{\omega}_1, \boldsymbol{\omega}_2, ..., \boldsymbol{\omega}_K\}$ and added biases $\boldsymbol{\beta} = \{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, ..., \boldsymbol{\beta}_K\}$, and then applying an element-wise nonlinear activation function σ, typically *sigmoid*, *tanh*, and *ReLU* activation, on the convolved results (Gu et al., 2018). The new feature maps can be computed as follows:

$$\boldsymbol{X}_k^l = \sigma(\boldsymbol{\omega}_k^{l-1} \circledast \boldsymbol{X}^{l-1} + \boldsymbol{\beta}_k^{l-1}).$$

where $\circledast$ is a convolution operation.

Pooling layers are always added after convolutional layers to summarised the features in a region of the feature maps. In forward propagation, $N \times N$ blocks in previous feature maps are reduced to a single value, where $N \times N$ is the pooling size. As shown in FIGURE 2.3, a commonly used pooling layer in CNNs, called max pooling layer, calculates the largest value in each patch of previous feature maps. The output of max pooling layers would be feature maps which contain the most prominent features in the previous feature maps. Another commonly used pooling layer is the average pooling layer, which calculates the average value in each patch of previous feature maps. The pooling results can be computed as follows:

Max pooling:

$$\boldsymbol{X}_k^l(i,j) = \max_{\substack{N(i-1)<a<Ni+1 \\ N(j-1)<b<Nj+1}} \boldsymbol{X}_k^{l-1}(a,b)$$

Average pooling:

$$\boldsymbol{X}_k^l(i,j) = \frac{1}{N^2} \sum_{a=Ni+1}^{N(i-1)} \sum_{b=Nj+1}^{N(j-1)} \boldsymbol{X}_k^{l-1}(a,b)$$

Fully connected layers are always the last layers of CNN architectures. As shown in FIGURE 2.4, fully connected layers operate a linear combination on a flattened input where each input element is connected to all neurons. Each element in the output of the last fully connected layer indicates the probability for the input image belonging to each class in the classification task. During the forward propagation, the output of a fully connected layer is given by:

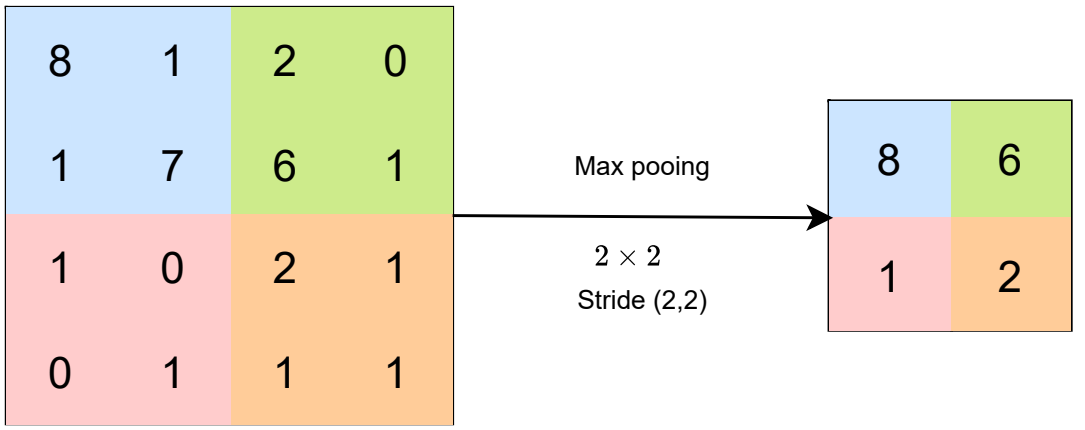


FIGURE 2.3: An example of max pooling. It calculates the largest value in each patch of each feature map.

$$\hat{\boldsymbol{Y}} = \sigma(\boldsymbol{\omega}\boldsymbol{X} + \boldsymbol{\beta}),$$

where $\boldsymbol{X}$ is the flattened input, $\hat{\boldsymbol{Y}}$ is the output, σ is the activation function, $\boldsymbol{\omega}$ is the weight matrix, and $\boldsymbol{\beta}$ is the bias matrix of the fully connected layer. To evaluate how well the CNN models the input data, loss functions are always used to measure the compatibility between output predictions through forward propagation and given ground truth labels. In classification tasks, cross-entropy loss is commonly used to evaluate the CNN models. Cross-entropy can be calculate as follows:

$$L = -\sum_{c=1}^{M} y_c \log(\hat{y}_c).$$

where M means the number of classes, y_c means the ground truth label, and $\hat{y}_c$ means the output prediction of a CNN model.

Back propagation

To achieve a better fit of for the dataset, back propagation performs a backward pass to update the CNN's parameters, including weights and biases, based on the loss obtained in the previous epoch.

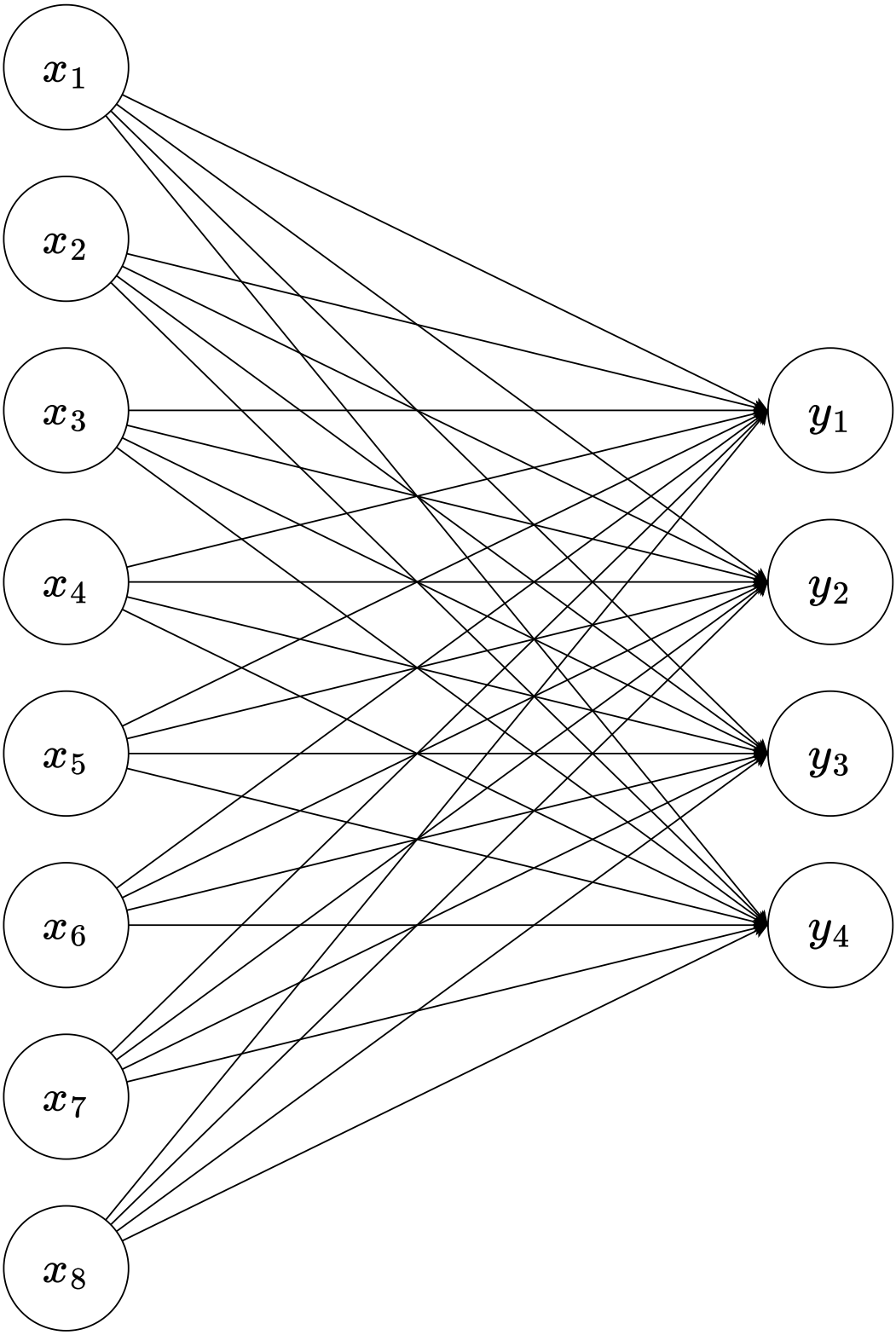


FIGURE 2.4: An example of fully connected layers. Fully connected layers operate a linear combination on a flattened input where each input element is connected to all neurons.

The back propagation in neural network computes the gradient of the loss function L for a single weight or bias by the chain rule. Assume there are functions $Y = f(X)$ and $Z = g(Y)$, we can compute the derivative of Z with respect of X by using the chain rule:

$$\frac{\partial Z}{\partial X} = \frac{\partial Z}{\partial Y}\frac{\partial Y}{\partial X}.$$

Therefore, in the back propagation, we can use the local gradient $\frac{\partial X^l}{\partial X^{l-1}}$ and the loss gradient from the previous layer $\frac{\partial L}{\partial X^l}$ to calculate the loss gradient $\frac{\partial L}{\partial X^{l-1}}$.

When the activation function is *ReLU*, the gradients in a fully-connected layer can be calculated as follows:

$$\frac{\partial L}{\partial x_i} = \sum_{c=1}^{M} \frac{\partial L}{\partial \hat{y}_c}\frac{\partial \hat{y}_c}{\partial \boldsymbol{X}} = \sum_{c=1}^{M} \frac{\partial L}{\partial \hat{y}_c}\omega_{i,c};$$

when $\hat{y}_c > 0$:

$$\frac{\partial L}{\partial \omega_{i,c}} = \frac{\partial L}{\partial \hat{y}_c}\frac{\partial \hat{y}_c}{\partial \omega_{i,c}} = \frac{\partial L}{\partial \hat{y}_c}x_i;$$

$$\frac{\partial L}{\partial \beta_c} = \frac{\partial L}{\partial \hat{y}_c}\frac{\partial Y}{\partial \beta_c} = \frac{\partial L}{\partial \hat{y}_c};$$

when $\hat{y}_c = 0$:

$$\frac{\partial L}{\partial \omega_{i,c}} = \frac{\partial L}{\partial \hat{y}_c}\frac{\partial \hat{y}_c}{\partial \omega_{i,c}} = 0;$$

$$\frac{\partial L}{\partial \beta_c} = \frac{\partial L}{\partial \hat{y}_c}\frac{\partial Y}{\partial \beta_{i,c}} = 0.$$

where $\omega_{i,c}$ and β_c is the weight and bias.

In a max-pooling layer, the gradients can be calculated as follows:

$$\frac{\partial L}{\partial \boldsymbol{X}_k^{l-1}(a,b)} = \begin{cases} \dfrac{\partial L}{\partial \boldsymbol{X}_k^{l}(i,j)} & (a,b) = \underset{\substack{N(i-1)<a'<Ni+1 \\ N(j-1)<b'<Nj+1}}{\operatorname{argmax}} \boldsymbol{X}_k^{l-1}(a',b') \\ 0 & \text{Others.} \end{cases}$$

In an average-pooling layer, the gradients can be calculated as follows:

$$\frac{\partial L}{\partial \boldsymbol{X}_k^{l-1}(a,b)} = \frac{1}{N^2} \frac{\partial L}{\partial \boldsymbol{X}_k^{l}(\lceil \frac{a}{N} \rceil, \lceil \frac{b}{N} \rceil)};$$

where $\lceil \rceil$ is a ceiling function to round up to the nearest integer.

In a convolutional layer, the gradients can be calculated as follows:

$$\frac{\partial L}{\partial \boldsymbol{X}^{l-1}} = (\frac{\partial L}{\partial \boldsymbol{X}^{l}} \circ \sigma^{'}(\boldsymbol{\omega}_k^{l-1} \circledast \boldsymbol{X}^{l-1} + \boldsymbol{\beta}^{l-1})) \circledast \mathrm{rot}_{180^\circ}(\boldsymbol{\omega}^{\mathrm{l}-1});$$

$$\frac{\partial L}{\partial \boldsymbol{\omega}^{l}} = \boldsymbol{X}^{l-1} \circledast (\frac{\partial L}{\partial \boldsymbol{X}^{l}} \circ \sigma^{'}(\boldsymbol{\omega}_k^{l-1} \circledast \boldsymbol{X}^{l-1} + \boldsymbol{\beta}^{l-1}));$$

$$\frac{\partial L}{\partial \boldsymbol{\beta}} = \frac{\partial L}{\partial \boldsymbol{X}^{l}} \circ \sigma^{'}(\boldsymbol{\omega}_k^{l-1} \circledast \boldsymbol{X}^{l-1} + \boldsymbol{\beta}^{l-1});$$

where $\circ$ means element-wise product.

In training steps, the parameters in a CNN can be updated according to the gradients, and the update speed can be controlled by the learning rate hyper parameter. From the learning step, we can see that the performance of a well-trained CNN model is critically dependent on the training data.

2.1.2 Deep learning models for image classification

The recent advancement in computing technology allowed to train much deeper and complex networks than in the past. CNNs have consistently been achieving state-of-the-art performance in classification tasks for both nature images and medical images. In this parts, I will introduce several important CNN models which are always applied to tasks in medical images.

AlexNet

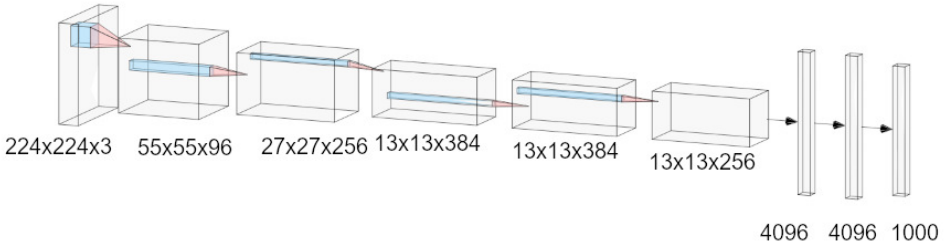


FIGURE 2.5: Architecture of AlexNet.

Krizhevsky, Sutskever, and Hinton (2017) designed a large deep CNN, called AlexNet, to classify ImageNet data (Deng et al., 2009). As shown in FIGURE 2.5, there are 5 convolutional layers and 3 fully-connected layers in the AlexNet. It also introduced the use of ReLU as activation function, dropouts to avoid the overfitting, and max-pooling layers instead of simple pooling. The AlexNet achieved top-1 and top-5 test set error rates of 37.5% and 17.0% on ImageNet LSVRC-2010, respectively. Due to its high performance on nature image tasks, the ALexNet has also been finetuned for medical image tasks, such as breast cancer histopathology image classification (Titoriya and Sachdeva, 2019), drowning diagnosis using post-mortem lung CT images (Homma et al., 2020a), and so on.

VGG (Visual Geometry Group)

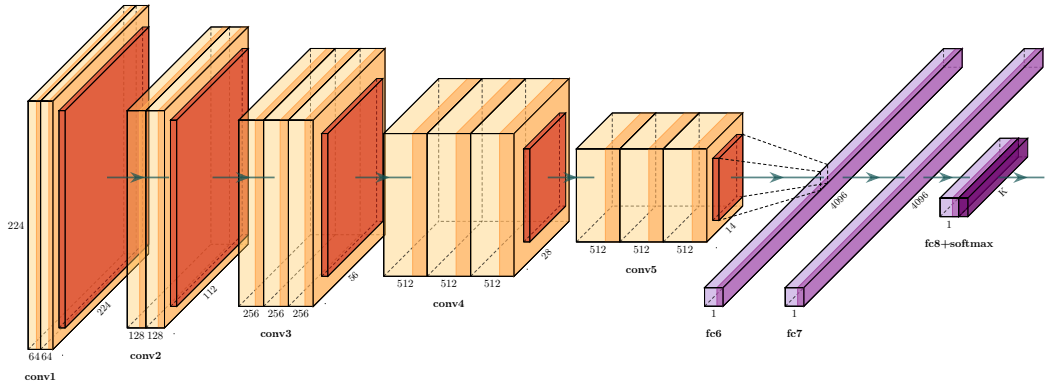
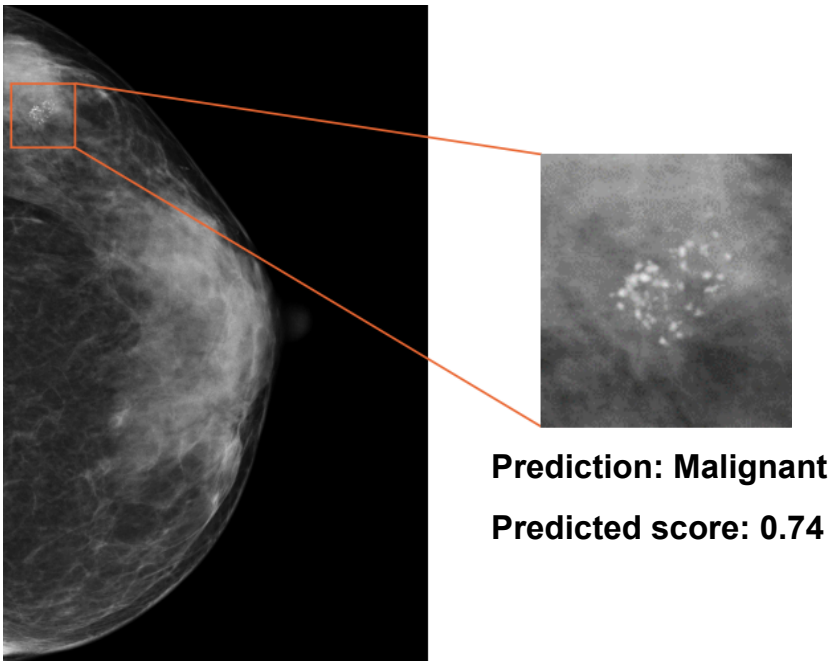


FIGURE 2.6: Architecture of VGG-16.

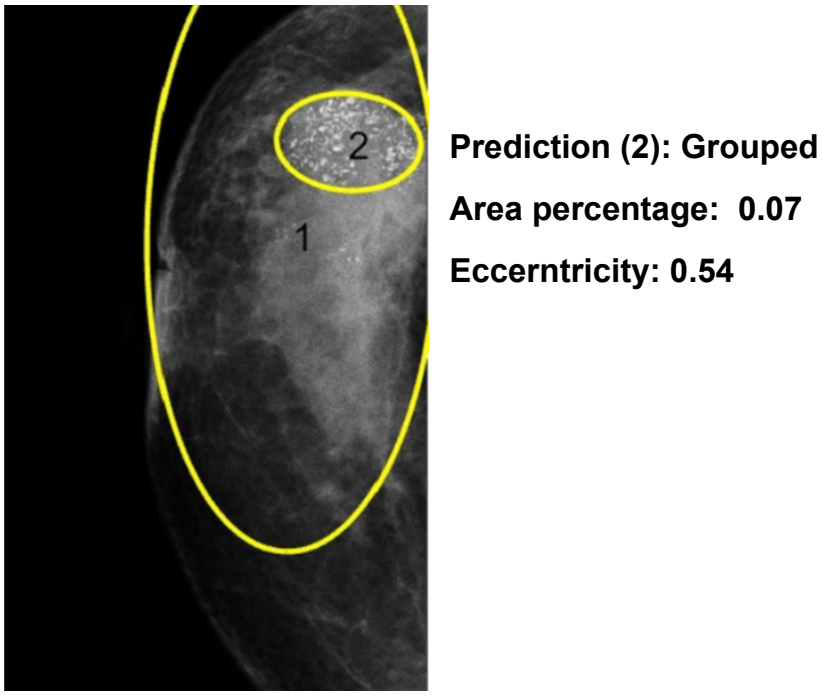
Simonyan and Zisserman proposed a deeper CNN named VGGNet (Simonyan and Zisserman, 2014). There are totally 6 different CNN configurations for VGGNet: A, A-LRN, B, C, D (VGG-16), and E (VGG-19) with 11, 11, 13, 16, 16, and 19 layers, respectively. As an example, FIGURE 2.6 shows the architecture of VGG-16. In VGGNet, the sizes of the kernels are 3×3 for all convolutional layers.

2.1.3 The lack of explainability

Although deep learning-based methods could achieve high performance in many medical tasks, they have been criticized for their lack of explainability of their prediction results compared to other methods. That is, it is difficult to know and understand how the model made a prediction, and the inner workings remain opaque to the outside observer (Castelvecchi, 2016). Without a sufficient understanding of a mechanism behind the machine-made prediction, it becomes very complicated to detect hidden risks in deep learning-based methods, i.e., training bias caused by mislabeled training data, especially for medical applications.



(A)



(B)

FIGURE 2.7: Examples of results in MCC detection and classification using mammograms: (A) output of a deep learning-based method, (B) output of a clustering-based method. The deep learning-based method could obtain a high performance, but it was unable to explain the decision-making. On the other hand, the clustering-based method could provide not only the predicted class, but also several human-designed features to explain the decision.

For example, when facing a task of breast cancer detection in mammograms, deep learning-based CAD systems (Ribli et al., 2018) could detect and classify masses and micro-calcification clusters (MCCs) from mammograms, but it was difficult to tell how the systems made the decisions. As shown in FIGURE 2.7 (A), deep learning-based methods often output a predicted class and a risk score without detailed information such as features of detected lesions. In contrast, when other types of CAD systems, which depended on human-designed features, were applied to the same task, the CAD systems could provide more information along with the prediction results. In one of our previous studies (Zhang et al., 2023c), we verified bilateral mammographic density differences, which means the absolute difference between left and right mammographic densities, as a risk factor to assess breast cancer risk. Such human-designed risk factors could give a reason for the predicted risk of breast cancer. In another study, we proposed a clustering-based CAD system to detect and classify MCCs in mammograms (Zhang et al., 2020). We mimicked the workflow of radiologists to classify MCCs into several classes based on the distributions. As shown in FIGURE 2.7 (B), clustering-based method can provide more features which could useful for explaining the classification of the detected MCCs. Since the features, such as the area percentage and the eccentricity, were human-designed, the decision-making could be easy to understand for radiologists.

On the other hand, it is hard to assess training biases without additional data because deep learning lacks explainability. Therefore, the training data could be very important for the reliability of deep learning-based methods.

2.1.4 Data characteristics' impact on deep learning

Since data characteristics are important for constructing a dataset for training a deep learning model, the question about how the data characteristics impact the performance of deep learning-based methods attracts researchers' attention.

A common problem in the development of deep learning-based methods is the class imbalance, an important data characteristic within the training data. In a binary classification task, class imbalance occurs when one class, the minority class, contains significantly fewer samples than the other class, the majority class (Johnson and Khoshgoftaar, 2019). Class imbalance is naturally inherent in many real-world tasks, especially in medical tasks, and in many problems (Yuan, Xie, and Abouelenien, 2018; Gao et al., 2020; Ibrahim, Torki, and El-Makky, 2018; Korkmaz, 2020), the class of interest is the minority class. As shown in FIGURE 2.8, the class imbalanced training data can lead classifiers to exhibit bias towards the majority class and ignore the minority class. Accuracy is one of the most frequently used metrics when evaluating classification results. When a dataset with 1% positive cases is used for evaluation, a native classifier can achieve 99% accuracy score by simply classifying all samples into the negative class. Such a model would provide no real value and could be dangerous when used in medical fields. Anand et al. (1993) analyzed the effect of class imbalance on the back propagation in shallow neural networks. They showed that when using training data with a class imbalance, the gradient is dominated by the majority class. So that the error of majority class could be decreased quickly but the error of minority class might be increased.

Many previous studies were proposed to address class imbalance, such as data-level methods, algorithm-level methods, and hybrid methods (Johnson and Khoshgoftaar, 2019). Data-level methods are always proposed to re-sample class imbalanced data. Masko and Hensman (2015) used random over-sampling (ROS), which

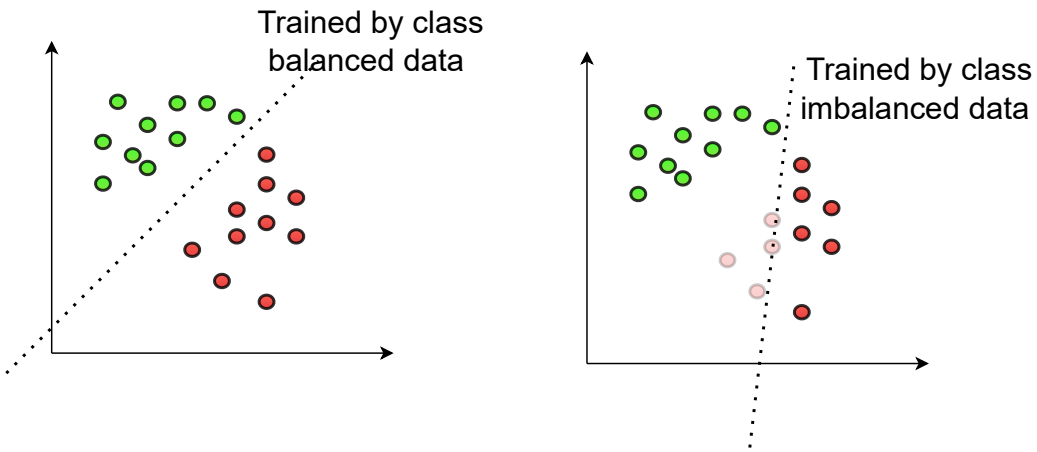


FIGURE 2.8: An example of class balanced training data and class imbalanced training data. Red means positive class and green means negative class. The class imbalanced training data can lead classifiers to exhibit bias towards the majority class and ignore the minority class.

means duplicating random images in minority classes until all classes had the same amount of images as the largest class, for addressing class imbalance. Their results showed that eliminating class imbalance with ROS can improve the classification performance. To decrease the impact of class imbalance, Lee, Park, and Kim (2016) pre-trained a CNN with class-normalized data, which is constructed by reducing the images in the majority class with random under-sampling (RUS), and fine-tuned the CNN with the original data. Their results showed that RUS can improve the performance on minority class. Pouyanfar et al. (2018) proposed a dynamic sampling method for classification tasks of class imbalanced data using CNNs. This method utilized class-wise performance to adjust the sampling rates for each class automatically.

Algorithm-level methods means modifying deep learning algorithms for the purpose of addressing class imbalance, such as new loss functions, cost-sensitive learning, and threshold moving. Wang et al. (2016) designed a mean false error

(MFE) loss function and a mean squared false error (MSFE) loss function, which are more sensitive to the errors from the minority class compared with commonly used mean squared error (MSE) loss function. Lin et al. (2017) designed a focal loss to address the extreme imbalance between foreground and background classes in training deep learning-based models for object detection tasks. Khan et al. (2017) proposed a cost-sensitive learning which can optimize the class-dependent costs and learn robust feature representations for both majority class and minority class automatically. Although threshold moving does not impact weights tuning and does not improve a model's ability to classify between classes, Buda, Maki, and Mazurowski (2018) showed it is an appropriate method to reduce classification bias that can be quickly implemented to already trained models.

Hybrid method means the method combines data-level and algorithm-level method. Huang et al. (2016) proposed quintuplet sampling and triple-head loss in their Large Margin Local Embedding (LMLE) method for training CNN models and the proposed LMLE method worked well with class imbalanced training data. Dong, Gong, and Zhu (2018) proposed class rectification loss (CRL) and hard sample mining in their batch-wise incremental minority class rectification model to address class imbalance in a large-scale image classification task. Their experimental results showed CRL and hard sample mining can also be implemented to other CNN models and improve the performance.

Because class imbalanced data are widely found in medical field, data-level, algorithm-level, and hybrid methods were applied to deep learning-based methods for medical tasks. Reza and Ma (2018) applied over-sampling to pathological breast cancer image classification with a CNN to counter the impact of class imbalance training data. Focal loss is also widely applied to deep learning-based methods for medical classification tasks, such as skin cancer classification (Le et al., 2020), lung

nodule classification (Tran et al., 2019), and femur fractures classification (Lotfy et al., 2019).

Another important data characteristic is data homogeneity. As shown in FIGURE 2.9, the classifier trained by data from one medical facility could fail to classify data from the other medical facility. Lo et al. (2021) compared a deep learning model trained by high-variability images with another model trained by images from a single medical center. Their results showed the model trained by high-variability images performed better in the cross-dataset test, which reveals that dataset homogeneity can have a significant impact on the generalization of CNN models. To address this problem in medical tasks, researchers always analyzed the cross-cohort generalizability (Bron et al., 2021) of their proposed deep learning-based methods.

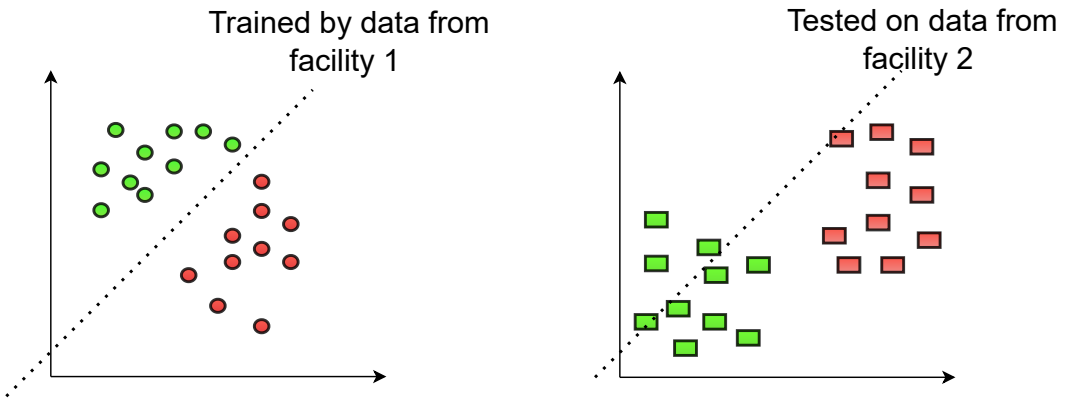


FIGURE 2.9: An example of data homogeneity. Red means positive class and green means negative class. The classifier trained by data from one medical facility could fail to classify data from the other medical facility.

2.2 COVID-19 Diagnosis Using Chest X-ray

The severe acute respiratory syndrome Coronavirus 2 (SARS-CoV-2) caused the Coronavirus disease 2019 (COVID-19) that has been widely spread worldwide and still

continues to have a devastating effect on the health and life of the global population (Wang et al., 2020a). The polymerase chain reaction (PCR) test is the gold standard for detecting SARS-CoV-2 nowadays (Wang et al., 2020b). Nevertheless, PCR testing is time-consuming and laborious, and it is also suffering from high cost (Love et al., 2021).

Radiological evaluation of patients with clinical–epidemiological suspect of COVID-19 is mandatory, especially in the emergency department while waiting for RT-PCR results, in order to have a rapid evaluation of thoracic involvement. As part of the standard procedure for patients with respiratory complaints, CXR imaging is used as a first-line triage tool because of the long waiting time for RT-PCR testing (Cozzi et al., 2020). In addition, using portable X-ray units can reduce the movement of patients and so minimizing the risk of cross-infection. As shown in figure 2.10, medical findings such as patchy or diffuse reticular-nodule opacity and consolidation can be found in CXR images from COVID-19 cases (Ng et al., 2020). It is reported that some patients showed abnormalities in the CXR images before they were eventually test positive for COVID-19 with RT-PCR test (Wong et al., 2020). Moreover, all the rapid triaging, availability, accessibility, and portability of CXR imaging indicated that it could be a complement to RT-PCR test. However, one of the biggest bottlenecks of CXR screening is the need for experts to diagnose from the CXR images because the radiological signatures can be subtle.

2.2.1 Deep Learning-Based Methods for COVID-19 Diagnosis Using Chest X-ray

The deep learning-based methods can actually obtain high performance in many medical tasks. The success made by deep learning-based methods encouraged researchers to develop deep learning-based methods that can aid radiologists in detecting COVID-19 in CXR images more rapidly and accurately.

Hemdan, Shouman, and Karar (2020) proposed an original COVIDX-Net framework to assist radiologists to automatically diagnose COVID-19 in CXR images. The COVIDX-Net includes seven different architectures of deep learning models: VGG-19, DenseNet-121, Inception-V3, ResNet-V2, Inception-ResNet-V2, Xception, and MobileNet-V2. They evaluated the framework on 25 CXR images from healthy patients and 25 CXR images from COVID-19 patients. The results showed that VGG-19 and DenseNet-121 achieved an accuracy of 90% for COVID-19 detection.

Brunese et al. (2020) applied VGG-16 model for COVID-19 detection using CXR images. They proposed an approach composed by three steps. The first one is to detect a chest X-ray as related to a healthy patient or to a patient with pulmonary diseases. The second step is aimed to discriminate between generic pulmonary disease and COVID-19. The last step is aimed to detect the interesting area in the chest X-ray to provide explainability. In the first and the second step, they used two differently fine-tuned VGG-16 models. They used datasets belonging to multiple medical facilities and there were totally 3520 CXR images from healthy patients, 250 CXR images from COVID-19 patients, and 2753 CXR images from patients with other pulmonary diseases. Their method achieved an average accuracy of 97% for COVID-19 detection on their test set.

Wang, Lin, and Wong (2020) proposed a new architecture of deep learning

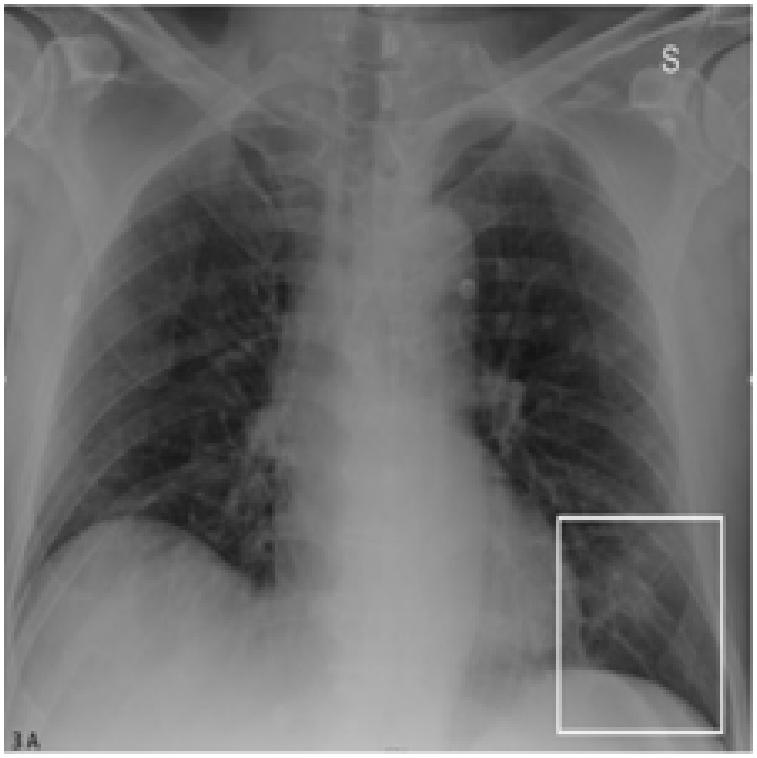
model named COVID-Net for COVID-19 detection. They used a large dataset, called COVIDx dataset, to train and evaluate the COVID-Net. The COVIDx dataset contains CXR images from five different data repositories. There were totally 266 COVID-19 patient cases, 8,066 normal cases, and 5,538 cases from patients with other pulmonary diseases. The COVID-Net achieved 93.3% test accuracy.

Although the deep learning models can achieve high performance on COVID-19 detection, the lack of accepted theoretical explanation remains the fundamental problem of deep learning, i.e., the black-box problem (Lei, Chen, and Zhao, 2018). The cause is that deep learning models lack transparency and explainability; it is difficult to know and understand how the model made a prediction, and the inner workings remain opaque to the outside observer (Quinn et al., 2022). Without a sufficient understanding of the machine-made prediction, it becomes very complicated to detect errors in models' performance, i.e., training bias caused by mislabeled training data, especially for medical applications. Therefore, the reliability of deep learning models remains a concern.

For assessing the reliability of deep learning models used for COVID-19 detection in CXR images, Sadre et al. (2021) proposed a region-of-interest (ROI) hide-and-seek protocol. As shown in Figure 2.11, to observe the reliability of these deep learning models, they removed lung regions from CXR images in a public CXR dataset and used them to train and test deep learning models. Then, a gradient-weighted class activation mapping (Grad-CAM) method (Selvaraju et al., 2020) was utilized to visualize which parts of the CXR images were focused on by the deep learning models. The experiment results showed that the deep learning models could achieve high performance even when using the images without lung regions, and the focused locations were outside the lung regions when deep learning models made a COVID-19 prediction. Results in this study (Sadre et al., 2021) indicated the deep

learning models are unreliable in terms of medical findings because the image features contributing to COVID-19 classification exist outside the lung regions, which is unexpected for a lung-based illness. However, the cause of the unreliable performance is still unknown.

(A)



(B)

FIGURE 2.10: Examples of CXR images: (A) a CXR image from a health case, (B) a CXR image from a COVID-19 case. Abnormalities can be found in the image from the COVID-19 case.

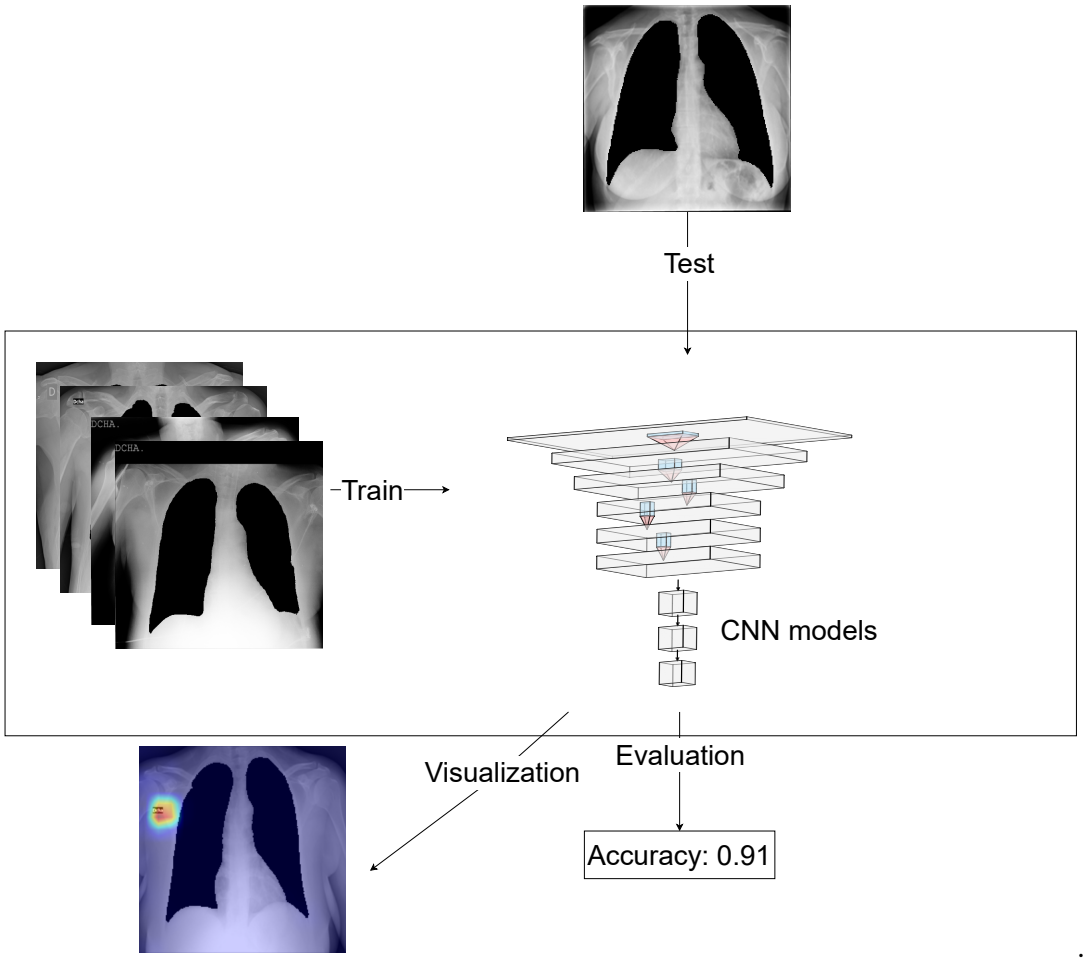


FIGURE 2.11: An overview of the previous study. CXR images with lung regions removed are utilized to investigate the reliability of deep learning models for COVID-19 classification. Deep learning models can achieve high accuracy when images with lung regions are removed, and the focused locations are outside the lung regions when deep learning models make a COVID-19 prediction. The result indicates the deep learning models are unreliable in terms of medical findings, but the cause of the unreliable performance is still unknown.

Chapter 3

Comparison experiments and cross-dataset test

The study (Sadre et al., 2021) mentioned that the unreliability of deep learning-based methods for COVID-19 diagnosis might be explained via data characteristics, because the previous studies collected as much data as possible from different medical facilities to develop deep learning-based methods for the urgent pandemic. The class imbalance, i.e., the difference in data quantity among categories, belongs among such data characteristics, and its impact on deep learning-based methods has attracted much attention from researchers. We noticed that when collecting the COVIDx dataset, every single repository, except COVID-19 Image Data Collection, only provided COVID-19 cases or non-COVID-19 cases. An intra-source imbalance, which means the class imbalance within the data collected from each medical facility, exists in the dataset and there are few investigations to analyze its impact on

deep learning models. Therefore, we organized two differently collected COVID-19 datasets and analyzed how the intra-source imbalance affects deep learning-based methods' performance(Zhang et al., 2023b). The both datasets consist of positive and negative categories, and they are well-balanced between the two categories. The data sources differ between the two datasets. One dataset (Qata-COV19) was collected from different medical facilities, and every single facility only provided positive or negative images. As one of the largest open-access COVID-19 dataset, the Qata-COV19 dataset has been used to train and test deep learning models in many previous studies. In another dataset (BIMCV), positive and negative CXR images were collected from a single medical facility. The ROI hide-and-seek protocol was implemented on the two datasets to investigate the effect of the intra-source imbalance on the deep learning models. Then, to evaluate the reliability of the deep learning models trained by each dataset, we made a cross-dataset test, which refers to training a deep learning model on one dataset and testing it on another dataset. Finally, we analyzed the relationship between the unreliability and the intra-source imbalance according to the experimental results.

3.1 Datasets

In this study, we used two CXR datasets collected from various public COVID-19 databases to investigate how the intra-source imbalance of training data impacts the deep learning models for the COVID-19 diagnosis. The intra-source imbalanced dataset is Qata-COV19 dataset, and the intra-source balanced dataset is BIMCV dataset.

As shown in TABLE 3.1, the Qata-COV19 dataset contains positive CXR images from five different public facilities and negative CXR images from seven other public

Dataset	Category	Data Source	Train	Test
Qata-COV19	Positive	BIMCV+ MHH SIRM COVID-chestxray dataset COVID-19 radiography dataset	3383	378
	Negative	RSNA Padchest dataset Guangzhou Women's Medical Center Indiana Network for Patient Care MC dataset Shenzhen Hospital ChestX-ray14 dataset	3383	378
BIMCV	Positive	BIMCV+	2222	239
	Negative	BIMCV-	2222	239

TABLE 3.1: Our study used two image datasets (Qata-COV19, BIMCV); Qata-COV19 has images provided from various facilities and only for a single category, while BIMCV collected images from the same facility.

facilities. It has been proposed by Yamac et al. (2021b) to help develop deep learning-based methods for COVID-19 diagnosis. The researchers gathered CXR images from different publicly available image sources and renamed all the CXR images. In comparison, the BIMCV dataset contains positive and negative CXR images from a single public facility, Valencian Region Medical ImageBank (Vayá et al., 2020; Vayá et al., 2021). Because they both have more negative images than positive images, we randomly under-sampled negative images to match the number of positive images in two datasets for addressing the class imbalance. In the Qata-COV19 dataset, one facility only provided CXR images in a single category. For example, BIMCV+ only provided positive images, and RSNA dataset only provided negative images for the Qata-COV19 dataset. Examples of positive and negative images in each dataset are shown in FIGURE 3.1. The important relationship between the Qata-COV19 dataset and the BIMCV dataset is that both shared the positive CXR images from BIMCV+

but did not share any negative CXR images.

In our study, the two datasets were used to clarify the influence of the intra-source imbalance on the reliability of deep learning models. All the images were resized to 512×512 pixels. We divided the data into training and test subsets, with a 90-10 split.

3.2 Experiments

3.2.1 Comparison experiments

As shown in FIGURE 3.2, we re-implemented the ROI hide-and-seek protocol on the Qata-COV19 dataset and the BIMCV dataset to clarify the relationship between intra-source imbalance and the reliability of deep learning models.

Firstly, we used a pre-trained U-Net model to segment lung regions from the original CXR images (FIGURE 3.2 (A)). The U-Net model was trained by Sadre et al. (2021) using the XLSor dataset for lung segmentation (Tang et al., 2019). According to the segmented lung regions, we generated the bounding boxes around the lung regions. For each CXR images, two lung regions and two bounding boxes were generated. Four types of modified images were generated by emphasizing and hiding the lung regions and the bounding boxes. Lungs-isolated images (FIGURE 3.2 (B)) and lungs-framed images (FIGURE 3.2 (C)) were generated by isolating the segmented lung regions and regions inside the bounding boxes from the original CXR images, respectively; lungs-removed images (FIGURE 3.2 (D)) and lungs-boxed-out images (FIGURE 3.2 (E)) were generated by removing the segmented lung regions and the regions inside the bounding boxes from the original CXR images, respectively. Because COVID-19 is a lung-based illness, the medical findings of it are expected to

exist inside lung regions. Therefore, CNN models trained by the original CXR images, lungs-isolated images, and lungs-framed images are expected to be capable of detecting COVID-19 in the test sets. On the other hand, CNN models trained by the modified images without lung regions, such as lungs-removed images and lungs-boxed-out images in this study, should not be able to classify COVID-19 images and non-COVID-19 images.

We used a VGG-16 pre-trained with ImageNet in this study. Before tuning, we replaced the first fully-connected layer with a global average pooling layer. The original CXR images or four types of the modified images were used to train different versions of the VGG-16 model separately, which aimed to classify the images into positive or negative class. We tuned all the weights and biases in the VGG-16 model during the training step.

In the first experiment, we trained VGG-16 models by using the original CXR images and modified images from the Qata-COV19 dataset separately to investigate the effect of lung regions on the performance of the VGG-16 models when using intra-source imbalanced dataset. In contrast, for investigating the effect when using an intra-source balanced dataset, we trained VGG-16 models by using original images and four types of modified images from the BIMCV dataset, separately.

Results of Comparison experiments

To evaluate the performance of the VGG-16 models, we utilized a receiver operating characteristics (ROC) curve (Fawcett, 2006). In statistics, the ROC curve is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. The ROC curve is created by plotting the true-positive rate against the false-positive rate at various threshold settings. Each point

on the ROC curve represents a sensitivity/specificity pair corresponding to a particular decision threshold. We calculate the area under the curve (AUC) to provide an aggregate measure of performance across all possible classification thresholds.

FIGURE 3.3 (A) showed the VGG-16 model achieved AUC of 0.99 when trained by the original CXR images in Qata-COV19 dataset. Other ROC curves in FIGURE 3.3 showed that lung regions had little effect on the performance: the VGG-16 model achieved relatively high performance even when lung areas were removed or boxed out, which showed the same results as in the previous study (Sadre et al., 2021). These results confirm the high risk of obtaining an unreliable deep learning model.

In contrast, the ROC curves in FIGURE 3.4 showed that when using the lungs-removed images or lungs-boxed-out images from BIMCV, the AUC values degraded a lot. In particular, the ROC curve suggested nearly no capacity for classification when lung regions were boxed out. The results showed that the classification of CXR images in the BIMCV dataset relies on the features representing COVID-19 characteristics in lung regions. Although the VGG-16 model achieved a relatively low performance on the BIMCV dataset, the performance could be more reliable in terms of the medical findings.

In the comparison experiments, we re-implemented ROI hide-and-seek protocol on the Qata-COV19 and BIMCV datasets and the results showed the VGG-16 model trained by the BIMCV dataset was more reliable. However, we need more analysis to demonstrate our hypothesis that intra-source imbalance lead to the unreliable performance of deep learning models.

3.2.2 Cross-dataset test

We utilized a cross-dataset test to demonstrate the intra-source imbalance lead to the unreliable performance of deep learning-based methods. Cross-dataset test is always used to evaluate the generalizability of classification models. In this study, cross-dataset test is used to evaluate the impact of the features representing data source characteristics on the deep learning models trained by intra-source imbalanced datasets.

We trained a VGG-16 model using the original images from the Qata-COV19 dataset (the same model weights as in FIGURE 3.3 (A)), and then tested it on the original images from the BIMCV dataset. Because the Qata-COV19 dataset contains COVID-19 positive images from the BIMCV dataset but without negative images from the BIMCV dataset, the cross-dataset test can evaluate how much the feature representing characteristics of BIMCV dataset impact the VGG-16 model. Moreover, we trained a VGG-16 model using the original images from the BIMCV dataset (the same model weights as in FIGURE 3.4 (A)), and then tested it on the original images from the Qata-COV19 dataset for comparison.

Results of cross-dataset test

As shown in FIGURE 3.5 (A), when testing the BIMCV-trained model on the original CXR images from the Qata-COV19 dataset, the AUC was nearly 0.5, and the performance was the same as a random classifier. The ROC curve shows that the model failed to classify the positive and negative images from BIMCV. The specificity in this test was 0, which showed that all the images from the BIMCV dataset were classified into the positive class even if they were negative. According to the specificity, all images from BIMCV were classified into the COVID-19 positive class. It showed that the model learned the features representing characteristics of BIMCV dataset

from the positive images in the training step, so that the negative images from the BIMCV dataset were also classified into positive class in the test step. It revealed that when using intra-source imbalanced datasets, the prediction bases are the features representing each data source characteristics, but not the features representing COVID-19 characteristics. The result demonstrates lacking balance in data sources leads to the unreliable performance. Especially, as shown in the cross-dataset test, the model trained by intra-source imbalanced datasets can be totally unable to make a diagnosis for other datasets.

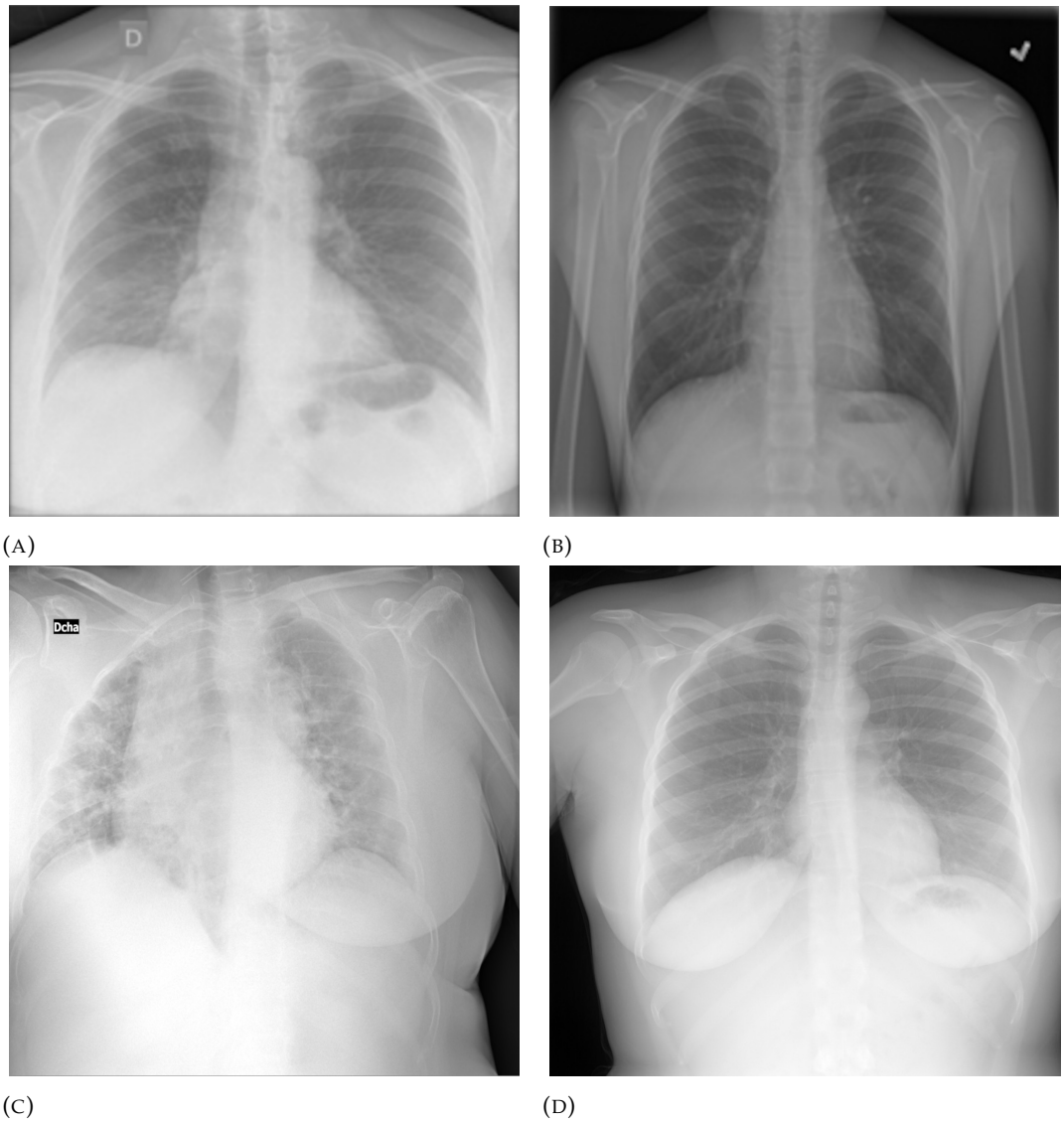
In comparison, as shown in FIGURE 3.5 (B), when testing the Qata-COV19-trained model on the original CXR images from the BIMCV dataset, the AUC was 0.84, and the model trained by BIMCV was able to classify positive and negative CXR images in the Qata-COV19 dataset. The model trained by the BIMCV dataset achieved a relatively high performance when testing on the Qata-COV19 dataset, which indicated it was more reliable.

3.3 summary

We report that the intra-source imbalance of training data leads to the unreliability of deep learning methods by re-implementing the ROI hide-and-seek protocol on two differently collected CXR datasets. Using a cross-dataset test, we show that the model trained by intra-source imbalanced datasets might classify images based on the features characterizing data sources; hence, it lacks the capability to diagnose other datasets. For the urgent COVID-19 pandemic, many previous studies collected as much data as possible from different medical facilities to train deep networks, but without enough validation. For example, many previous studies (Yamac et al., 2021a; Zaki, Amin, and Hamad, 2021) used the Qata-COV19 dataset to train and test deep learning models and obtained high performance on the test subset, but few of

them discussed about the reliability and generalizability. They might lack clinical applicability because of the intra-source imbalance of the training data.

Our study reveals the risk of unreliability when using intra-source imbalanced datasets in deep learning methods, not only for COVID-19 classification but also for other medical applications. Therefore, when developing deep learning methods, we should ensure the intra-source balance of the datasets before they are applied to train deep learning models.



(A) (B)

(C) (D)

FIGURE 3.1: Examples of positive and negative CXR images in the two datasets: (A) a positive CXR image in the Qata-COV19 dataset, (B) a negative CXR image in the Qata-COV19 dataset, (C) a positive CXR image in the BIMCV dataset, (D) a negative CXR image in the BIMCV dataset.

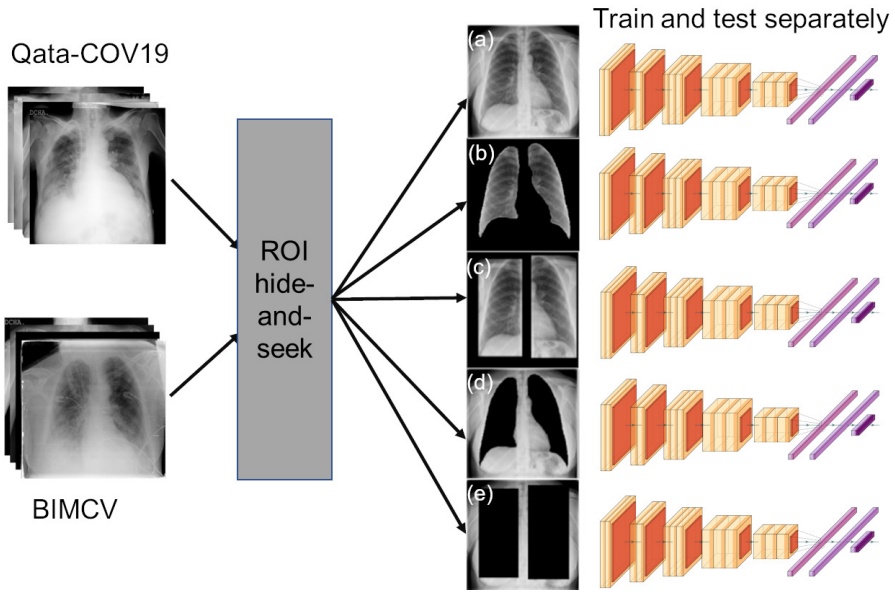


FIGURE 3.2: Overview of the comparative experiment. ROI hide-and-seek protocol operated (A) original images from the Qata-COV19 dataset or the BIMCV dataset to emphasize and hide the lung regions, respectively. (B) lungs-isolated images and (C) lungs-framed images were generated by emphasizing the lung regions, while (D) lungs-removed images and (E) lungs-boxed-out images were generated by hiding the lung regions. The original datasets and the modified datasets were utilized to train and test a VGG-16 model separately.

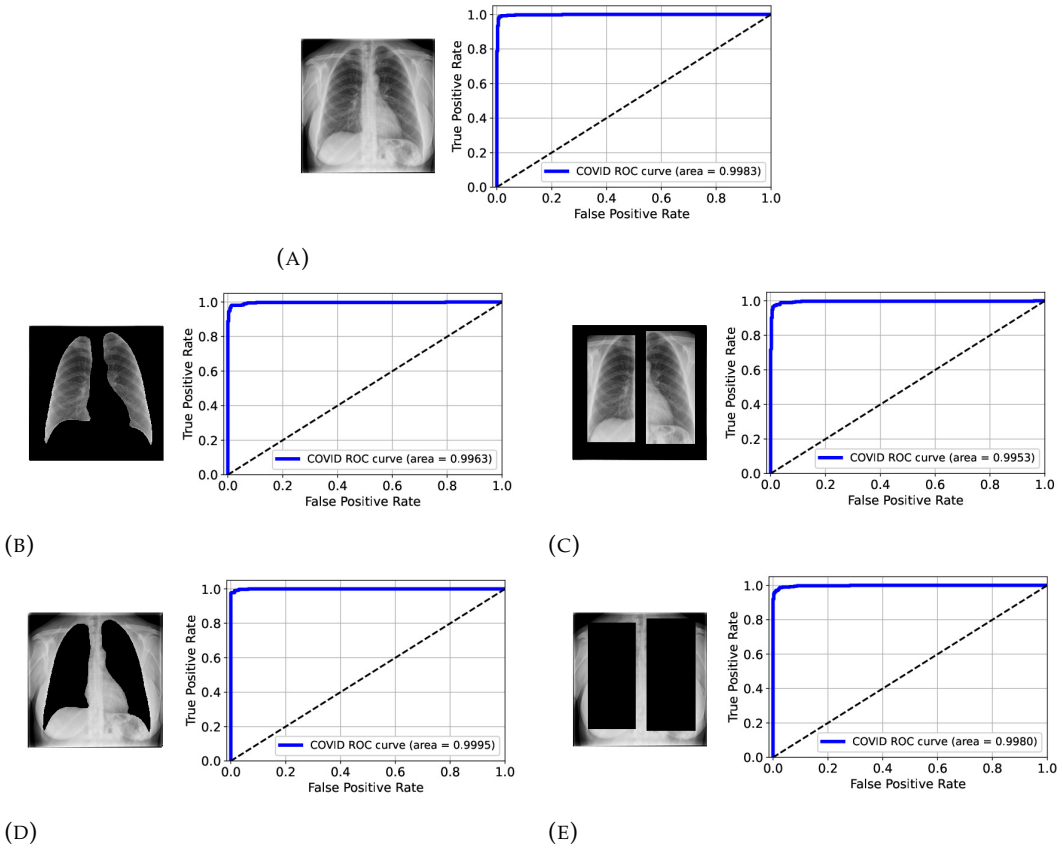


FIGURE 3.3: ROC curves for the VGG-16 models trained and tested on the original or the modified datasets from Qata-COV19: (A) original CXR images, (B) lungs-isolated images, (C) lungs-framed images, (D) lungs-removed images, and (E) lungs-boxed-out images. The deep learning models achieved high performance, even with hidden lung regions.

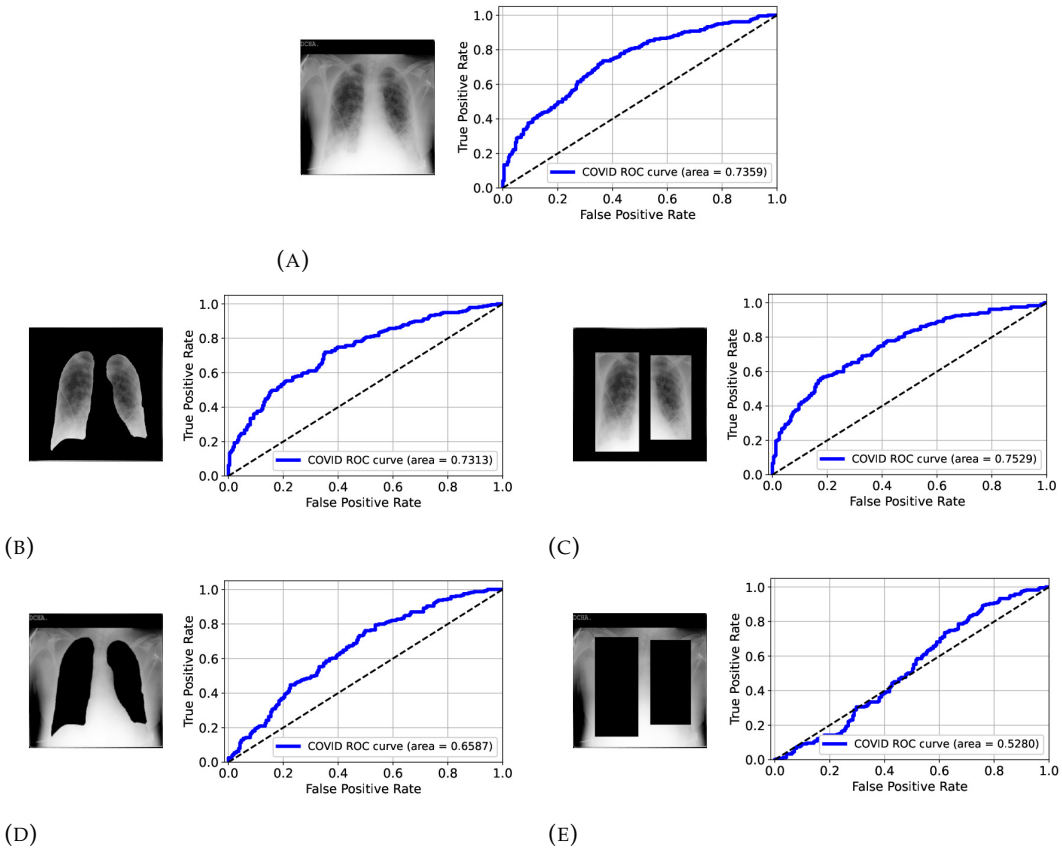
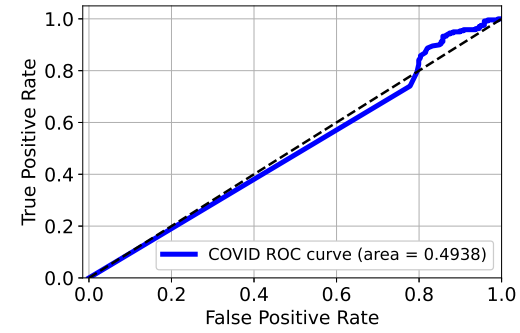
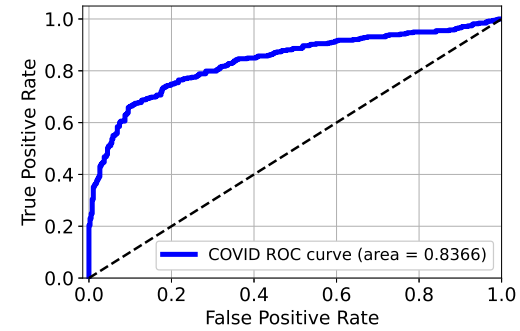


FIGURE 3.4: ROC curves for the models trained and tested on the original or the modified datasets from BIMCV: (A) original CXR images, (B) lungs-isolated images, (C) lungs-framed images, (D) lungs-removed images, and (E) lungs-boxed-out images. The performance degraded a lot when lung regions were hidden.



(A)



(B)

FIGURE 3.5: ROC curves for the cross-dataset test: (A) testing the Qata-COV19-trained model (the same model weights as in FIGURE 3.3 (A)) on the BIMCV dataset, (B) testing the BIMCV-trained model (the same model weights as in FIGURE 3.4 (A)) on the Qata-COV19 dataset. The model trained on original images from BIMCV dataset was able to classify original images from the Qata-COV19 dataset, while the model trained on original images from the Qata-COV19 dataset failed to classify original images from the BIMCV dataset.

Chapter 4

Visualization investigation in the cross-dataset test

In previous deep learning-based methods, visualization methods were always used to show the prediction basis of the deep learning models. For example, Brunese et al. (2020) utilized Gram-CAM to visualize which parts of the CXR images were focused on by the VGG-16 model. Wang, Lin, and Wong (2020) used GSInquire (Lin et al., 2019) tool to show the areas in lung regions are the main critical factors in the models' prediction.

In this chapter, to more intuitively show that deep learning models trained by intra-source imbalanced datasets might classify images based on the features characterizing data sources, we utilize Local Interpretable Model-agnostic Explanations (LIME) method, proposed by Ribeiro, Singh, and Guestrin (2016), to visualize the

prediction basis of the deep learning models(Zhang et al., 2023a).

4.1 LIME method

LIME method attempts to understand the model by perturbing the input of data samples and understanding how the predictions change. When used in image classification models, LIME randomly hides segmented superpixels, which means a group of pixels which have similar characteristics, in an input image and then it can determine which superpixel changes will have most impact on the prediction. It can reflect the contribution of each superpixel to the prediction result.

There are four steps to use LIME method to interpret a prediction by a CNN model. Firstly, generate superpixels from the input image. Superpixel algorithms can segment images into superpixels that adhere well to image boundaries. Then, we generate a dataset of perturbed samples by hiding some of the superpixels and derive the classification results in this dataset. Next, we learn a locally-weighted model from the classification results of the perturbed samples. Finally, we return the superpixels with the highest weights as the explanation. In this study, we use simple linear iterative clustering (SLIC), proposed by Achanta et al. (2012), to generate superpixels. As shown in FIGURE 4.1, an input CXR image can be segmented into superpixels and the segmentation is based on the color similarity and proximity. FIGURE 4.1 (c) shows the weights of each superpixel and FIGURE 4.1 (d) shows the top-5 superpixels of this example image. We can see the lung border is clear in the segmentation result and we can easily find a superpixel is inside lung regions or outside lung regions.

4.2 Visualization results

We use LIME method to visualize the explanations of the decisions made by the VGG-16 models in cross-dataset test. Intersection of Union (IoU), Ground Truth Coverage (GTC), and Saliency Coverage (SC) are used as evaluation metrics in this parts. We will also provide several examples to analyze the visualization results.

Because COVID-19 is a lung-based illness, we used the segmented lung regions as the ground truth G. For the explanations, we selected top-5 superpixels as the saliency map S. The Iou, GTC, and SC can be calculated as follows:

$$\begin{aligned} \text{IoU} &= \frac{|\text{G} \cap \text{S}|}{|\text{G} \cup \text{S}|} \\ \text{GTC} &= \frac{|\text{G} \cap \text{S}|}{|\text{G}|} \\ \text{SC} &= \frac{|\text{G} \cap \text{S}|}{|\text{S}|} \end{aligned}$$

where $|G \cap S|$ means the overlap area between the saliency map and lung regions, and $|G \cup S|$ means the area of union between the saliency map and lung regions.

FIGURE 4.2 shows the results of IoU, GTC, and SC. The model trained by the BIMCV dataset achieved higher IoU, GTC, and SC than the model trained by Qata-COV19 dataset even when tested on the Qata-COV19 test subset.

Then, we will show several examples to compare the explanation of the models trained by different datasets. FIGURE 4.3 shows the LIME explanations for classifying a positive case. Both of the models made a true decision, but the model trained

by BIMCV focused more inside the lung regions while the model trained by Qata-COV19 focused on the label and background. FIGURE 4.4 shows the LIME explanations of another positive case which was classified correctly by the Qata-COV19-trained model but mis-classified by the BIMCV-trained model. The model trained by BIMCV focused more on the lung regions even it made a wrong decision, while the model trained by Qata-COV19 focused on the label and background. FIGURE 4.5 shows the visualization results of negative case which was classified correctly by BIMCV-trained model but mis-classified by Qata-COV19-trained model. The model trained by Qata-COV19 focused on the label information in the image and classified this negative image into positive class.

From the examples we can find that the model trained by the Qata-COV19 dataset made decisions based on the markers more than on the lung regions. It shows the intra-source imbalanced dataset could lead the deep learning model focus on the features representing data source characteristics rather than the features representing COVID-19.

4.3 Summary

In this chapter, we utilized LIME method to visualize the explanations for the decisions made by the deep learning models used in cross-dataset test. The visualization results showed that the model trained by intra-source balanced dataset focused more on the lung regions, while the markers and background can strongly impact the decisions of the model trained by intra-source imbalanced dataset. The visualization results emphasized the conclusion that the intra-source imbalance of training data leads to the unreliability of deep learning-based methods.

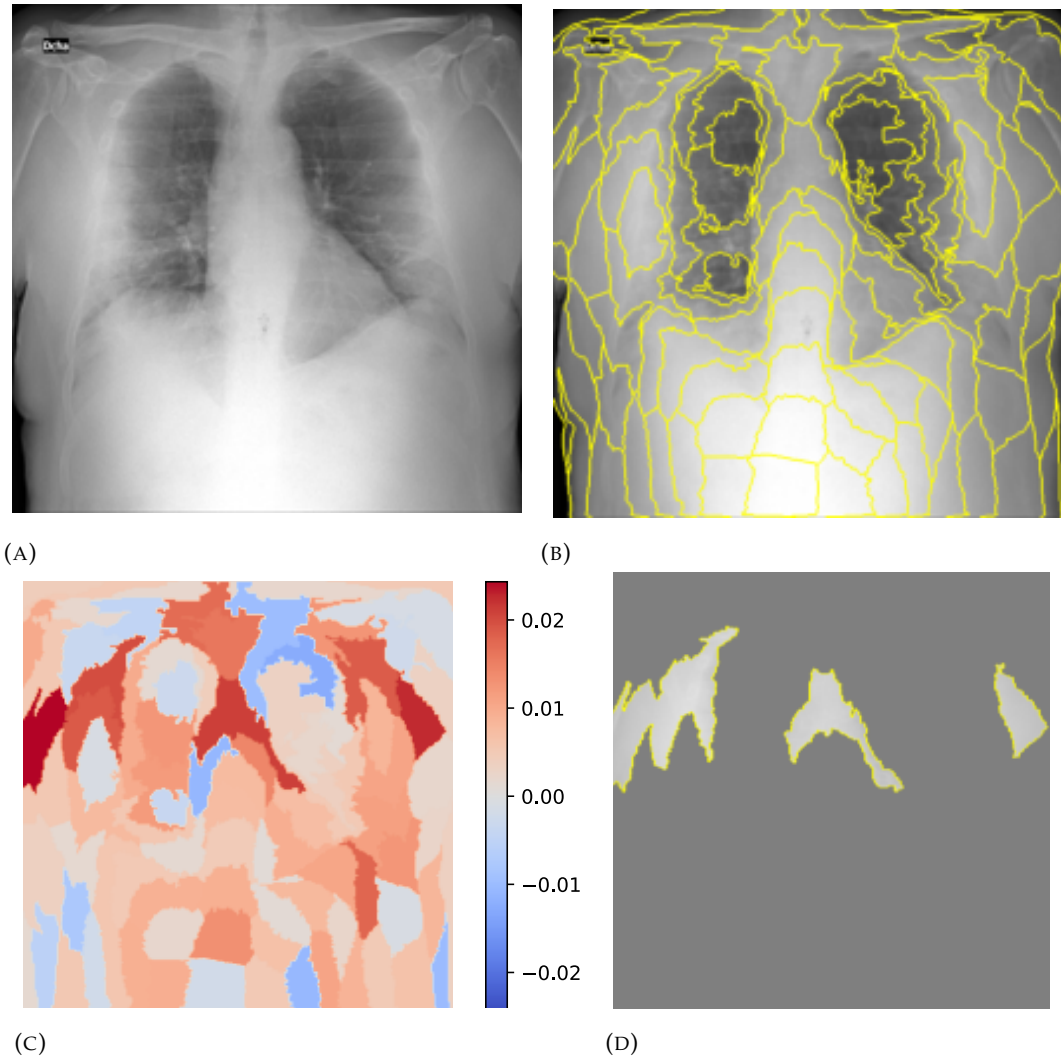


FIGURE 4.1: An example of LIME method: (a) an input CXR image, (b) segmented superpixels based on the color similarity and proximity, (c) the weights of each superpixel, and (d) top-5 superpixels of the input CXR image.

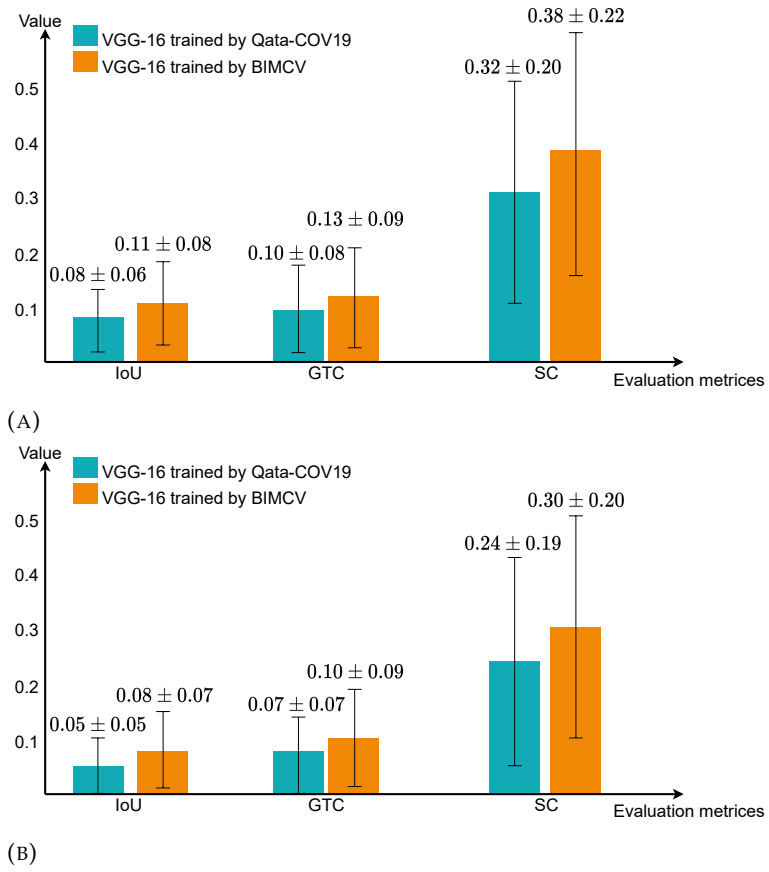
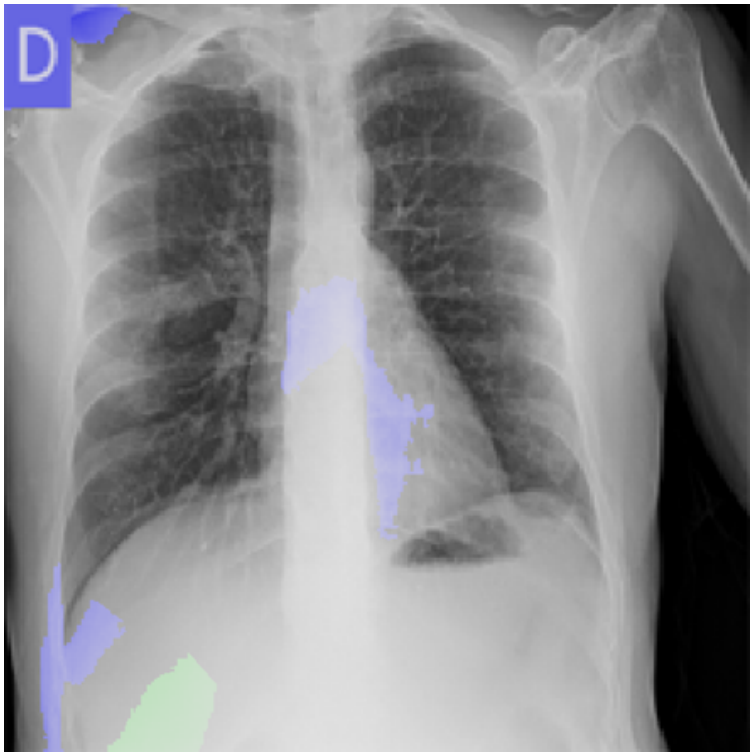
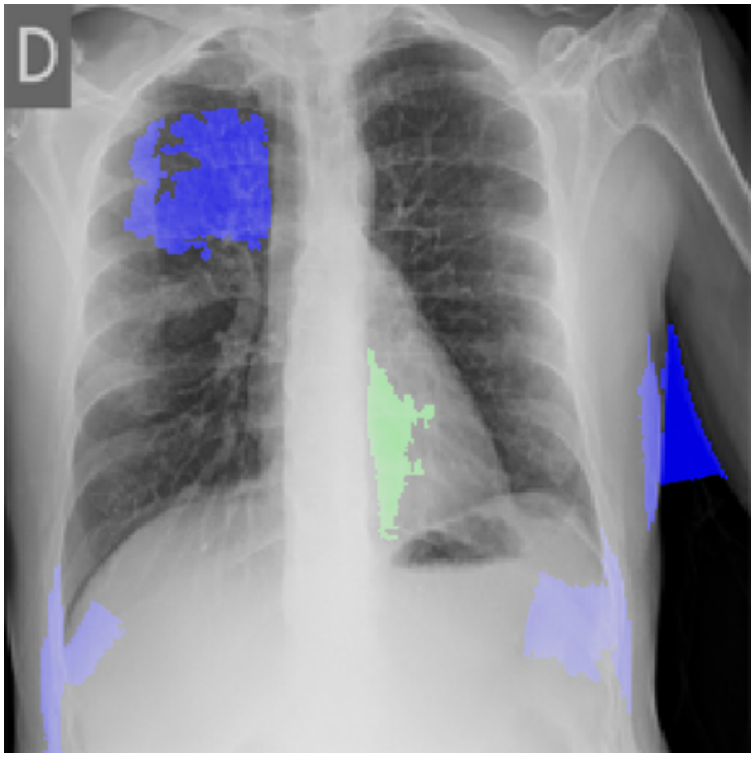


FIGURE 4.2: The results when selecting top-5 superpixels as the saliency map. (A) Tested on BIMCV test subset, (B) tested on Qata-COV19 test subset. The model trained by the BIMCV dataset performed better than the model trained by the Qata-COV19 dataset.

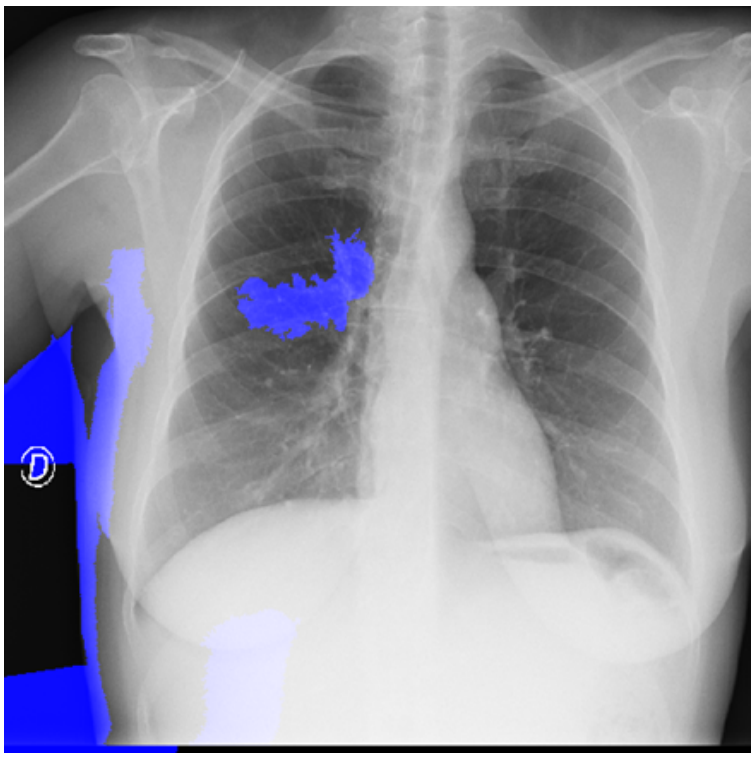


(A)

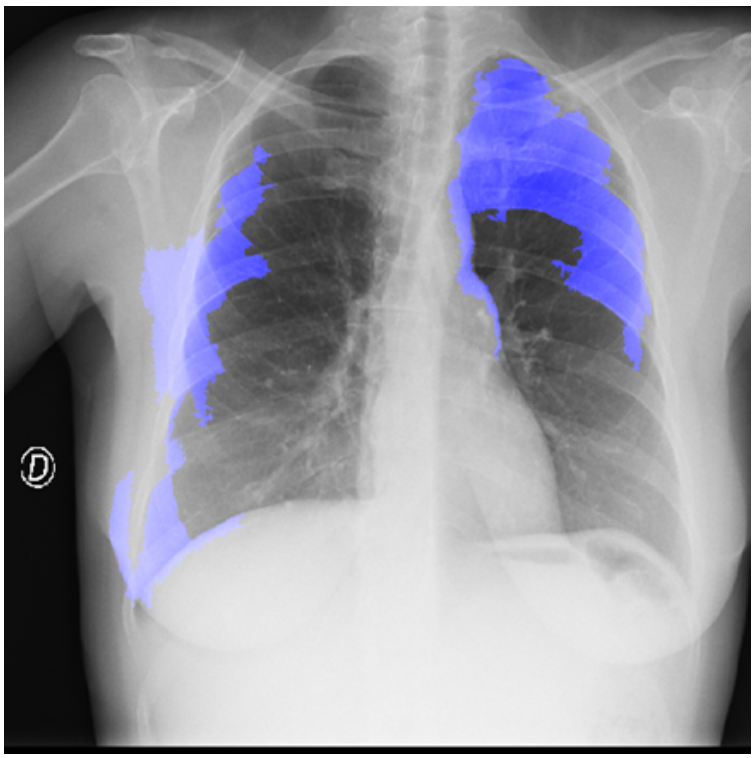


(B)

FIGURE 4.3: The LIME explanations for classifying a positive case by (A) VGG-16 model trained by Qata-COV19 dataset (prediction: positive), and (B) VGG-16 model trained by BIMCV dataset (prediction: positive). Blue areas contribute to positive prediction and green areas contribute to negative prediction. Both of the models made a true decision, but the model trained by BIMCV focused more inside the lung regions while the model trained by Qata-COV19 focused on the label and background.

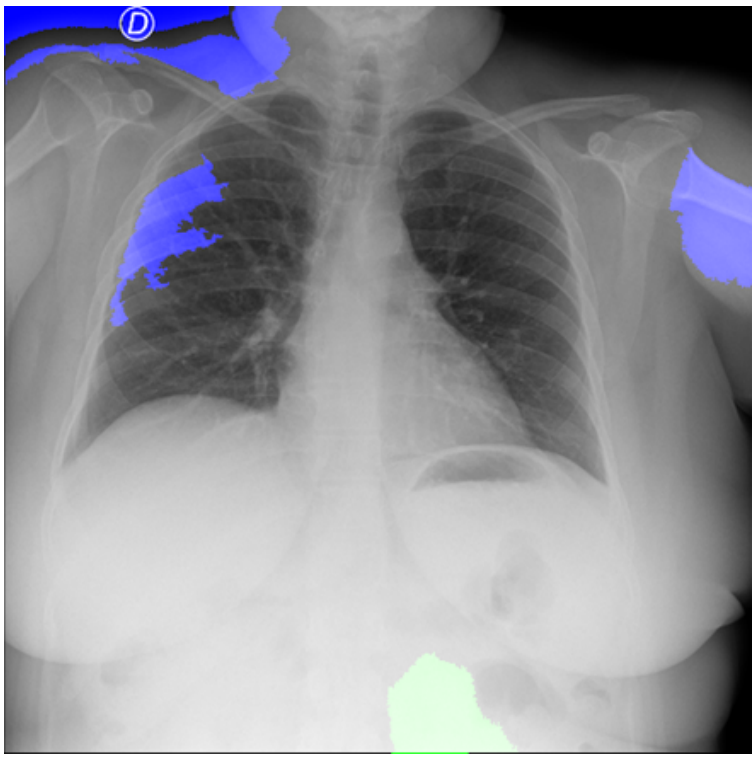


(A)

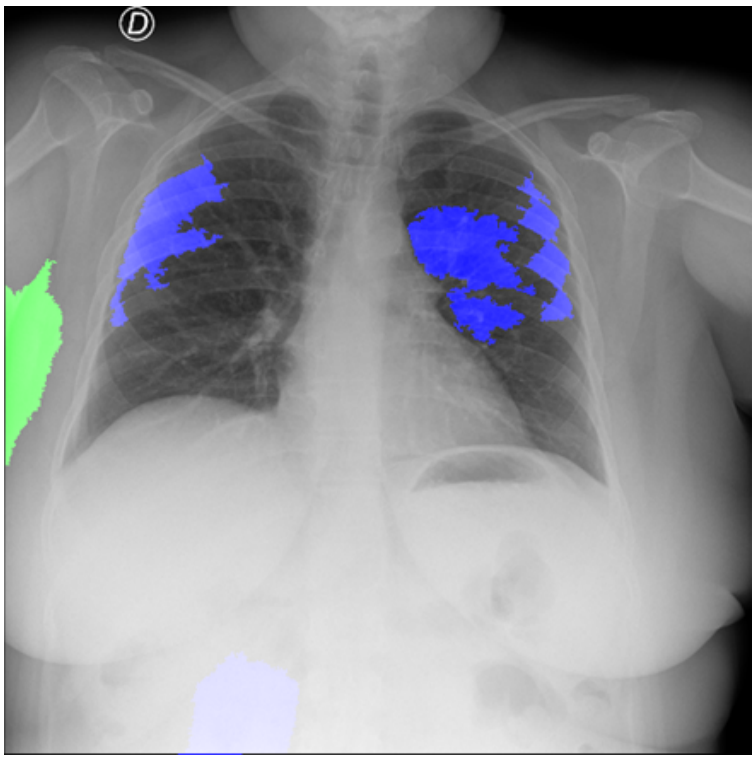


(B)

FIGURE 4.4: The LIME explanations for classifying another positive case by (A) VGG-16 model trained by Qata-COV19 dataset (prediction: positive), and (B) VGG-16 model trained by BIMCV dataset (prediction: negative). Blue areas contribute to positive prediction and green areas contribute to negative prediction. The model trained by Qata-COV19 correctly classified the image but focused on the label and background. The model trained by BIMCV focused more on the lung regions even it made a wrong decision.



(A)



(B)

FIGURE 4.5: The LIME explanations for classifying a negative case by (A) VGG-16 model trained by Qata-COV19 dataset (prediction: positive), and (B) VGG-16 model trained by BIMCV dataset (prediction: negative). Blue areas contribute to positive prediction and green areas contribute to negative prediction. The model trained by Qata-COV19 focused on the label and background information in the image and classified this negative image into positive class.

Chapter 5

Discussion

5.1 Investigation for proper balance level

In this study, we demonstrated the intra-source balance is vital for deep learning. However, intra-source imbalance often occurs in data collection, especially in medical fields. In this part, we will change the imbalance level and find out the impact of different intra-source balance level on deep learning models.

To find a proper balance level, we combined the two datasets and obtained several datasets with different balance level. Among the datasets, all the positive images are from BIMCV dataset. Negative images contains the images from BIMCV dataset and Qata-COV19 dataset. We set the balance level of each dataset to 30%, 50%, 70%, and 90%. Since the performance decreased a lot from 50% balance level to 30% balance level, we made an experiment to clarify the performance when using data with 40% balance level. For example, in the data with 30% balance level, 30% of

the negative images were randomly selected from BIMCV dataset, 70% of the negative images were randomly selected from Qata-COV19 dataset, and all the positive images were from BIMCV dataset. The data with different intra-source balance level were used to train VGG-16 model separately and we used the test subset in BIMCV to evaluate the performance.

FIGURE 5.1 showed the ROC curves in this experiment. As a result, the model trained on data with intra-source balance level over 50% could achieve over 0.71 AUC value, nearly the same performance on the test data with the model trained on the original BIMCV dataset. As shown in FIGURE 5.1 (E) and FIGURE 5.1 (F), when using dataset with 30% and 40% intra-source balance level, the AUC value was decreased to 0.66 and 0.58 respectively. In general, AUC over 0.7 indicates a good performance (Bekkar, Djemaa, and Alitouche, 2013). The result showed when trained on data with intra-source balance beyond 50%, the model could achieve a good performance. In contrast, when using data with balance level below 50%, the model might learn the features representing data sources and the intra-source imbalance could impact the deep learning performance.

5.2 Investigation for the impact of unobserved features

The markers inside CXR images are important features representing different data sources which can be observed in CXRs. The markers 'D' and 'DCH' represent data source in BIMCV, so they can be found in a part of positive CXRs and negative CXRs in BIMCV datasets. In Qata-COV19 dataset, the markers 'D' and 'DCH' can be found in positive CXRs, but not exist in negative CXRs. In the visualization results in Chapter 4, the markers are always the explanations for positive predictions, especially for false positive predictions as shown in FIGURE 5.2. The results indicate the markers

are important features representing data source and are learnt by the deep learning model.

There remains a question that if other unobserved features could represent different data sources and lead to the unreliability. As shown in FIGURE 5.3, to answer this question, we selected CXRs without markers from Qata-COV19 dataset and used the CXRs to train a VGG-16 model. We selected 1901 positive CXRs without markers from Qata-COV19 dataset and randomly selected 1901 negative CXRs from Qata-COV19 to keep the class balance in training data. We used the CXRs in BIMCV dataset to test the trained VGG-16 model.

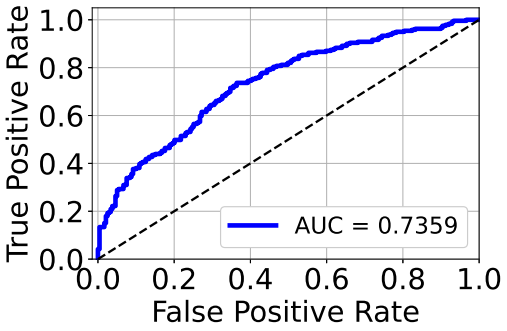
As shown in FIGURE 5.4, the model achieved 0.57 AUC value on test data from BIMCV dataset. Although there was little improvement on AUC value, the result showed the VGG-16 model trained by the CXRs without markers still failed to classify CXRs from BIMCV dataset. It indicated unobserved features inside CXRs could also lead to an unreliable performance of deep learning.

To visualize the explanations for the decisions, we also used LIME method in this experiment. FIGURE 5.5 shows the SC values when testing different models on BIMCV dataset. The model trained by selected data focused more on the lung regions. Then, we will show several examples. FIGURE 5.6 shows a true negative example. The model trained by selected data focused more on lung regions and made true prediction. FIGURE 5.7 shows a false negative example. The test data contains a marker 'D'. The model trained by original Qata-COV19 dataset focused on the marker and made a true prediction. The model trained by selected data did not focus on the marker and made a false prediction. When trained on the selected data, the model focused more on lung regions when made a positive prediction. As shown in FIGURE 5.8, the LIME visualization results showed areas inside lung regions were the explanations for not only true positive cases but also false positive

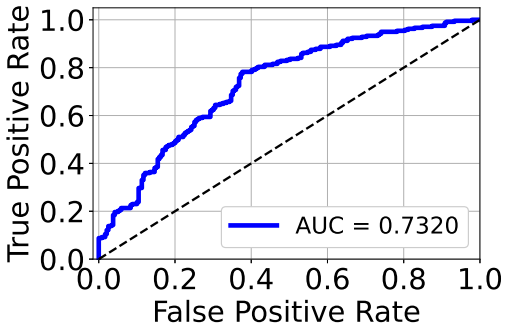
cases. The model can be unreliable even when it learned features inside lung regions and used them for classification.

This result demonstrated potentially unobserved aspects inside lung regions could also affect the deep learning performance. In addition, it reveals a hidden risk when validating deep learning reliability. In many previous studies for deep learning-based diagnosis of COVID-19, visualization results were used to explain the machine-made decisions and demonstrate the reliability of the deep learning model. For example, Wang, Lin, and Wong (2020) used GSInquire to visualize the explanations for COVID-19 predictions. They aimed to validate the diagnosis was not relying on improper information by qualitatively evaluating the visualization results with the lung regions. However, our result reveals that even a model focused on areas inside lung regions, it could be unreliable. Most visualization methods, such as Grad-cam and LIME, were proposed to visualize explanations for deep learning-based natural image classification. They aimed to find the locations of the features but not to find out what the features represent. Qualitatively evaluating the locations might be proper in natural image classification, but not enough in medical tasks according to our experimental results.

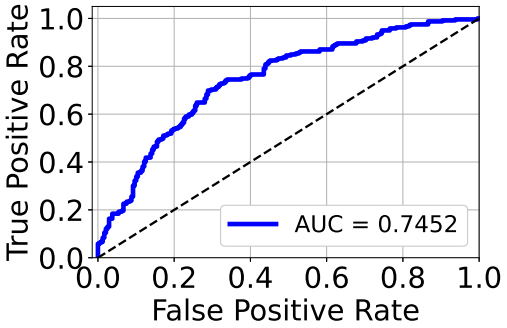
(A) 100%



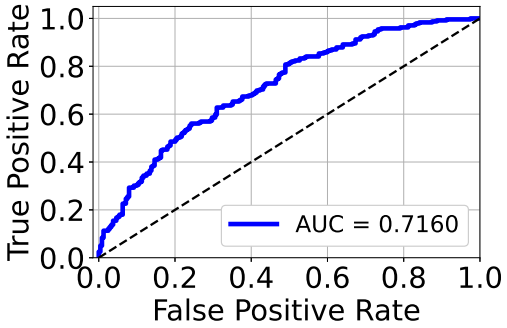
(B) 90%



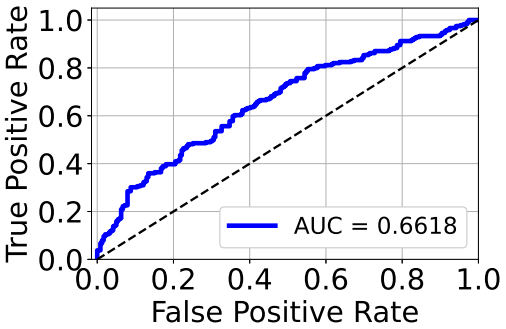
(C) 70%



(D) 50%



(E) 40%



(F) 30%

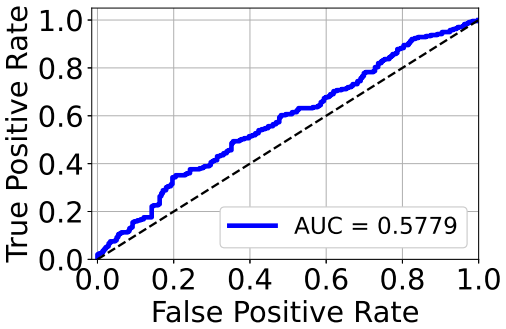
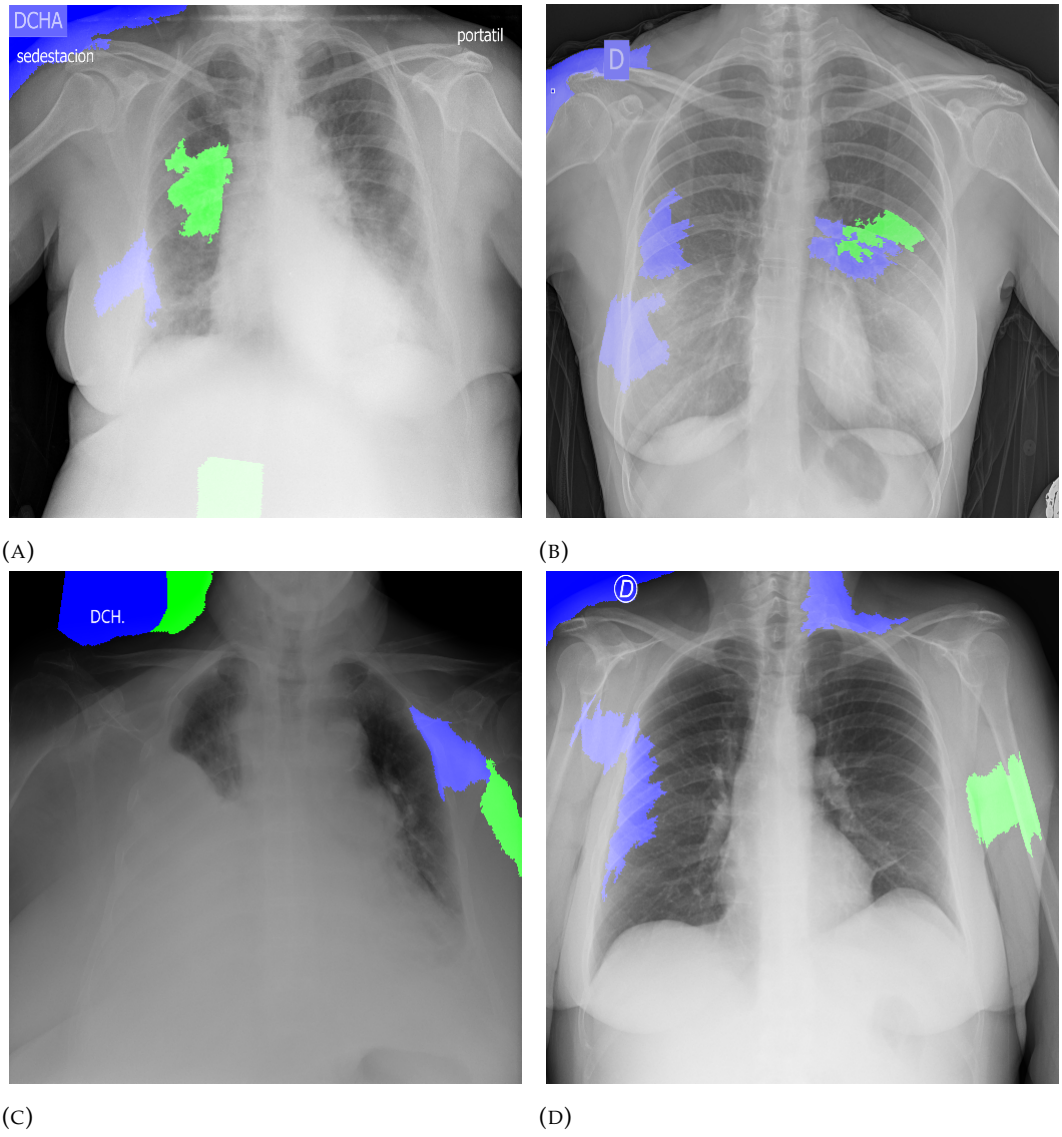


FIGURE 5.1: ROC curves for the VGG-16 models trained on data with different intra-source balance level: (A) original BIMCV dataset (100%), (B) data with 90% intra-source balance level, (C) data with 70% intra-source balance level, (D) data with 50% intra-source balance level, (E) data with 40% intra-source balance level, and (F) data with 30% intra-source balance level. When using a dataset with intra-source balance below 50%, the intra-source imbalance could influence the performance.



(A) (B)

(C) (D)

FIGURE 5.2: LIME explanations for the VGG-16 model trained by Qata-COV19 dataset. Blue areas contribute to positive prediction and green areas contribute to negative prediction. The markers 'D' and 'DCH' are always always the LIME explanations for positive predictions. (A) and (B) are positive CXRs from BIMCV dataset. (C) and (D) are negative CXRs from BIMCV dataset.

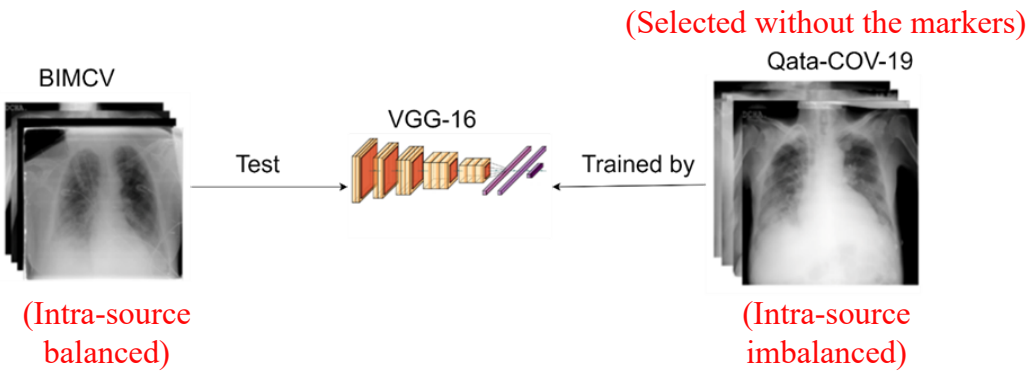


FIGURE 5.3: We selected CXRs without markers from Qata-COV19 dataset and used the CXRs to train a VGG-16 model. The model was tested on CXRs from BIMCV dataset.

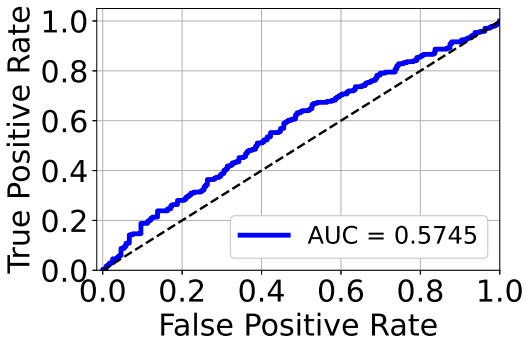


FIGURE 5.4: The ROC curve for testing the VGG-16 model trained by selected data. The model trained by selected data achieved 0.57 AUC value on BIMCV dataset, which showed the model still failed to classify the CXRs in BIMCV dataset.

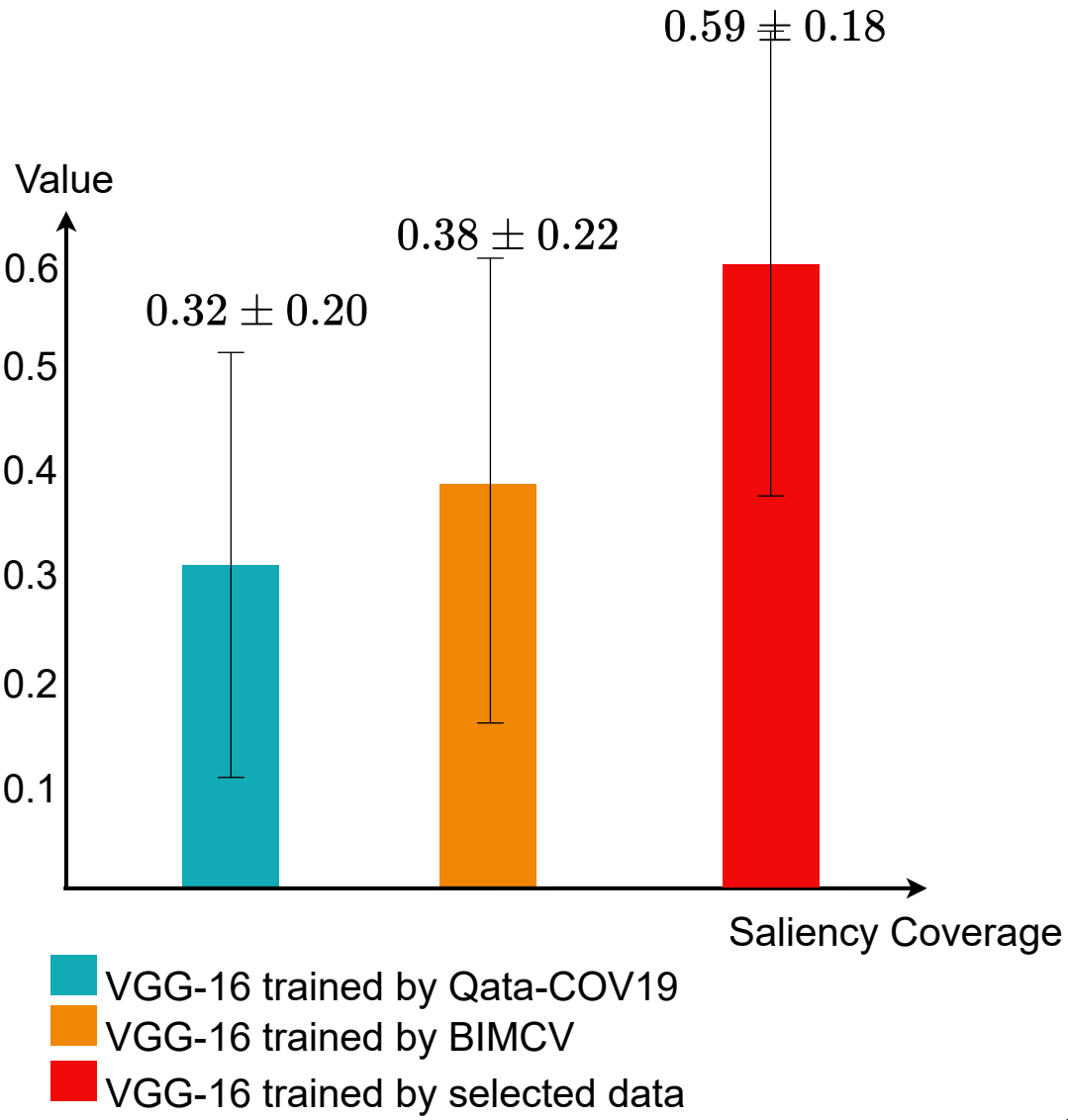
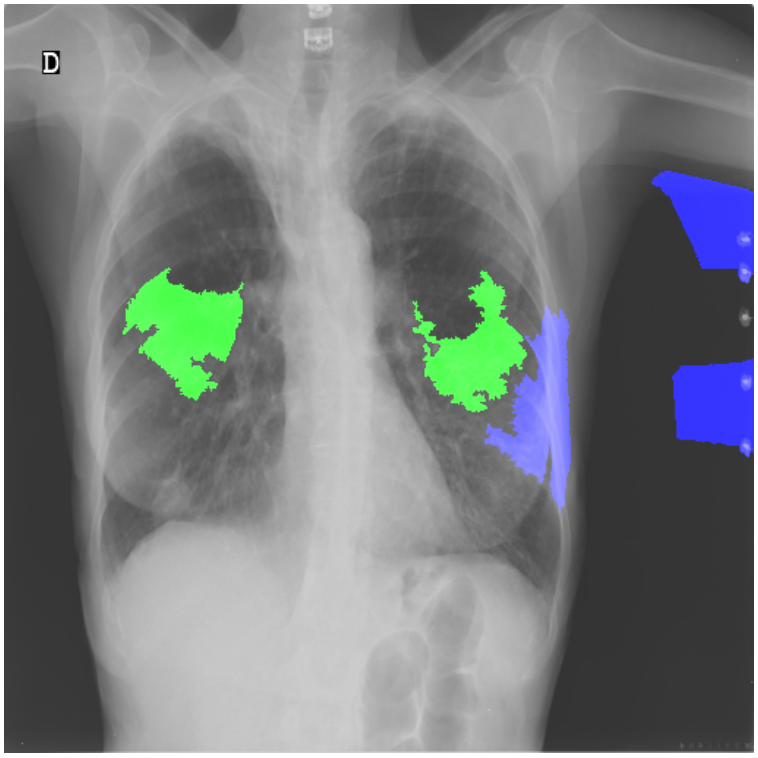
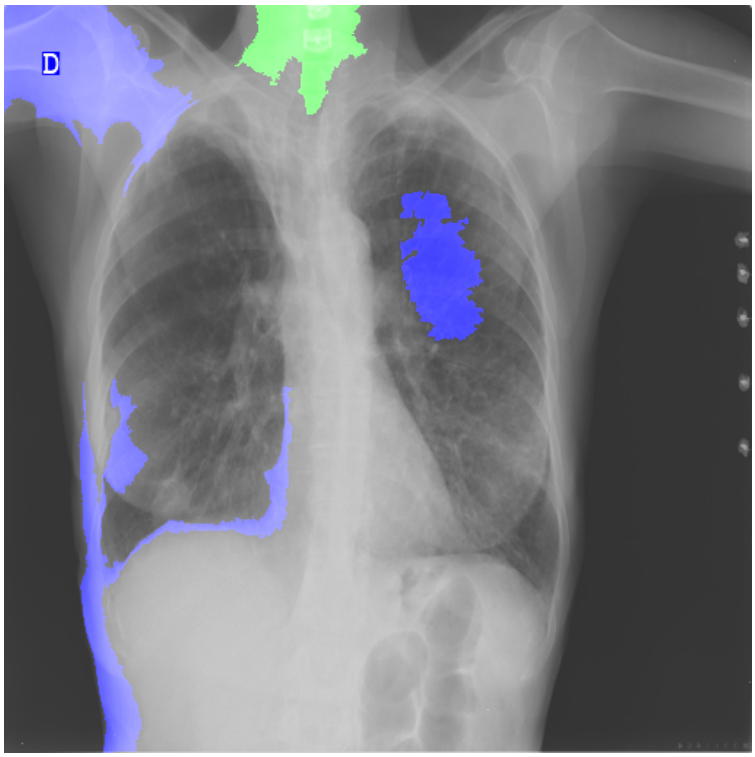


FIGURE 5.5: The SC values when testing different models on BIMCV dataset. The model trained by selected data focused more on the lung regions.

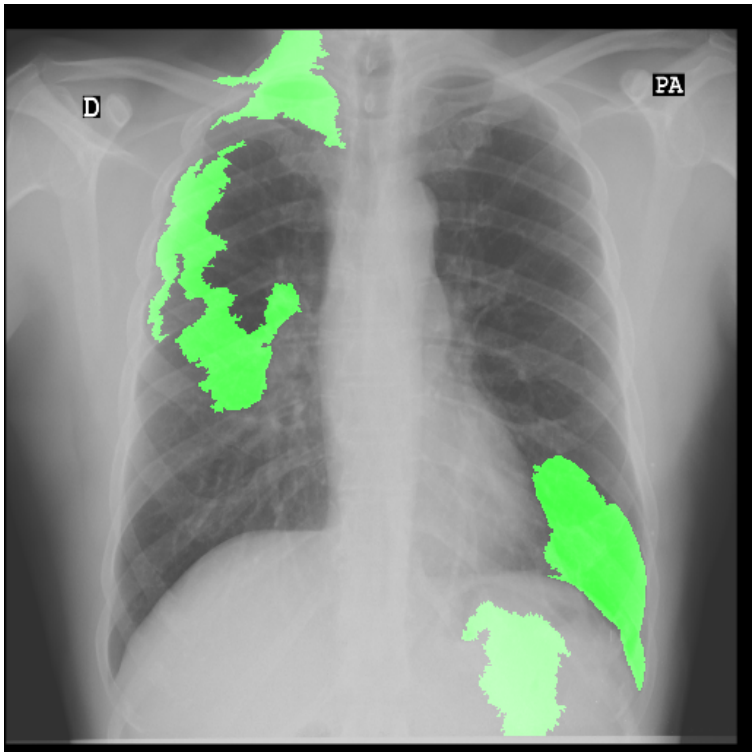


(A)

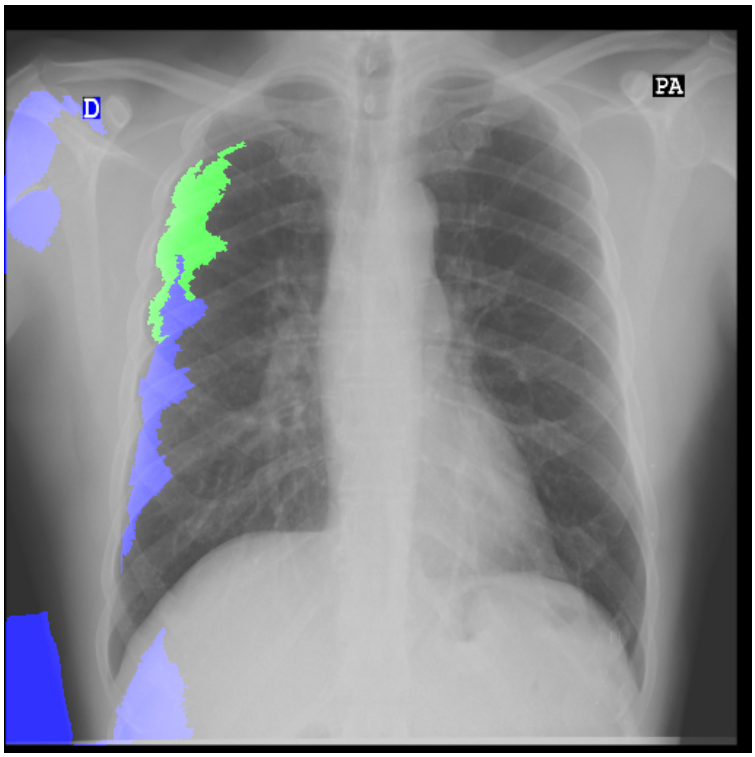


(B)

FIGURE 5.6: The LIME explanations for classifying a negative case from BIMCV dataset by (A) VGG-16 model trained by selected data (prediction: negative), and (B) VGG-16 model trained by original Qata-COV19 dataset (prediction: positive). Blue areas contribute to positive prediction and green areas contribute to negative prediction. The model trained by selected data focused more on lung regions and made true prediction.

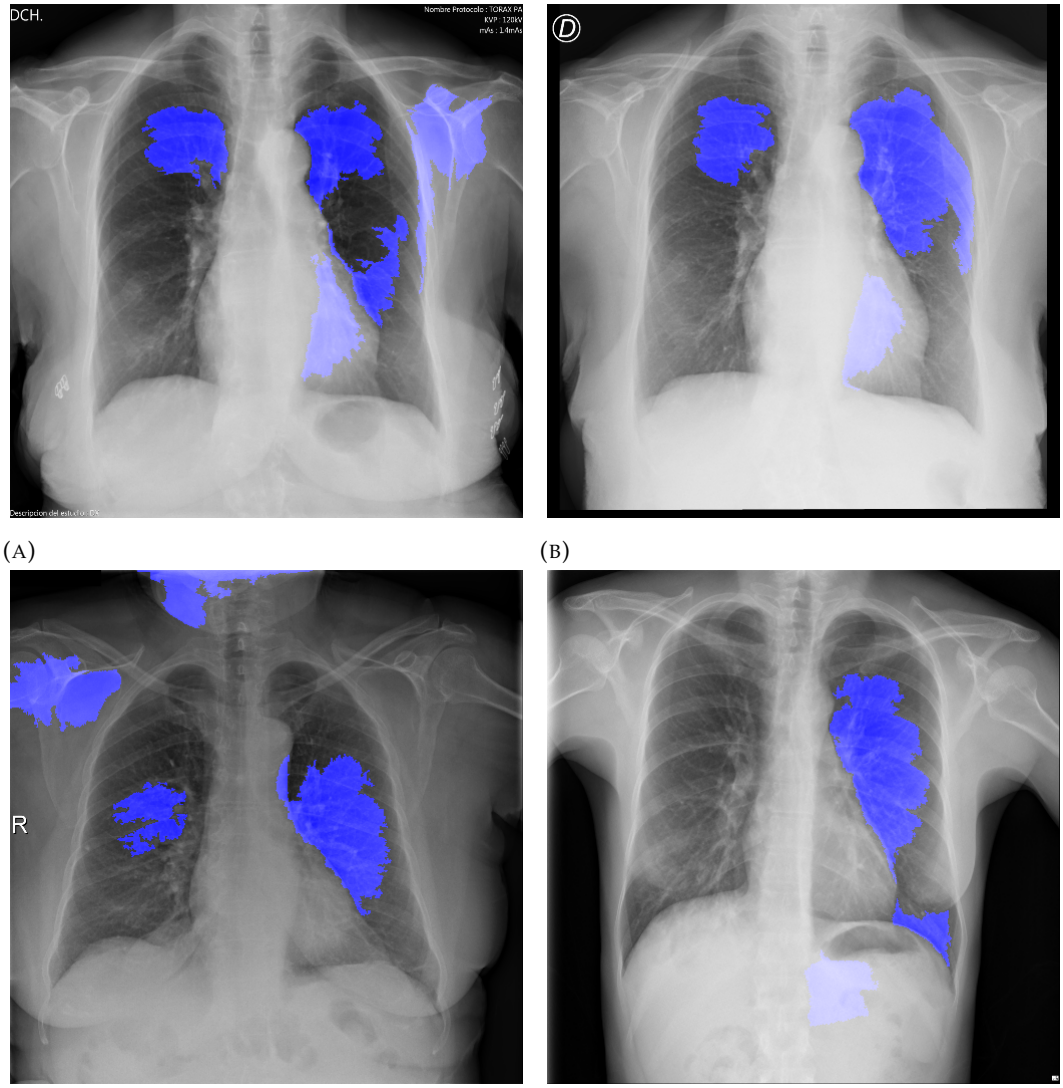


(A)



(B)

FIGURE 5.7: The LIME explanations for classifying a positive case from BIMCV dataset by (A) VGG-16 model trained by selected data (prediction: negative), and (B) VGG-16 model trained by original Qata-COV19 dataset (prediction: positive). Blue areas contribute to positive prediction and green areas contribute to negative prediction. The model trained by selected data did not focus on the marker and made a false prediction.



(A) (B)

(C) (D)

FIGURE 5.8: The LIME explanations for false positive predictions made by VGG-16 model trained by selected data (prediction: positive). Blue areas contribute to positive prediction. The negative CXRs from BIMCV dataset were classified in positive class and the model focus on the lung regions.

Chapter 6

Conclusion

6.1 summary

Deep learning-based methods are strongly relied on the training data. Therefore, the impact of training data characteristics on deep learning is very important for the application of deep learning-based methods, especially in medical fields. In this study, we used COVID-19 diagnosis in chest X-ray images as a case study to investigate the impact of intra-source imbalance, an important data characteristic, on the deep learning-based methods. Although McDermott et al. (2021) mentioned that potentially unobserved aspect could influence the deep learning performance, this study demonstrated the intra-source imbalance could lead to the same hidden risk. Each chapter of this dissertation are summarized as follows.

Chapter 1 has introduced the background of deep learning and the data characteristics. The purpose of this study, to investigate the impact of intra-source imbalance on the deep learning-based methods, has also been given in this chapter.

Chapter 2 has introduced the fundamental studies of deep learning and previous studies about COVID-19 diagnosis in chest X-ray images using deep learning. In the first part, we have introduced CNN architecture, training steps of CNNs, and several important data characteristics. In the second part, we have discussed the previous studies on COVID-19 diagnosis in chest X-ray images using deep learning. The ROI hide-and-seek protocol, used in Chapter 3, has been also introduced in this part.

Chapter 3 has introduced our experiments to investigate the impact of intra-source imbalance. We have introduced two different collected datasets, one is an intra-source balanced dataset and the other one is an intra-source imbalanced dataset. We have introduced the comparison experiments on the datasets by re-implementing ROI hide-and-seek protocol and the cross-dataset test. The results reveals the risk of unreliability when using intra-source imbalanced datasets in deep learning methods.

Chapter 4 has introduced our visualization investigation. We have utilized LIME method to generate explanations for the decisions made by deep learning models. The visualization results have intuitively shown that deep learning models trained by intra-source imbalanced datasets classified CXR images based on the features characterizing data sources rather than the features characterizing COVID-19.

Chapter 5 has discussed the proper inter-source balance level when constructing a dataset for deep learning. Moreover, it has discussed the impact of potentially unobserved aspects. The results have demonstrated potentially unobserved aspects

inside lung regions could also aspect deep learning performance.

6.2 Future Directions

In addition to the topics we covered in this dissertation, there are several possible extensions and directions we could further explore.

1. In previous study, many methods have been proposed to improve the performance of deep learning-based methods on COVID-19 diagnosis in CXR images, but most of them were trained and tested on intra-source imbalanced dataset. This research reveals the intra-source imbalance impact on the deep learning-based methods. However, the model trained by the intra-source balanced dataset achieved reliable but not high performance. Thus, it is necessary to evaluate the previous methods on intra-source balanced dataset or propose new methods to achieve a reliable and high performance on COVID-19 diagnosis in CXR images.

2. Although intra-source imbalance might lead to an unreliable performance, sometimes it is inevitable, especially when collecting medical data. Therefore, when collecting data from imbalanced facility, we need to pre-process the data to minimize the risk of unreliable performance. It is desired to further research on the impact of different settings on the data and the appropriate normalization methods for intra-source imbalanced data.

3. The hidden risk might led by the different deployment environment in each medical facility. Therefore, developing data standards could be another avenue to solve the problem. Increased use of data standards would make it

easier to collect proper data for deep learning-based methods. Moreover, better descriptions of contents, potential confounding and biases, and how the data were created should be provided to help to ensure the data is proper.

List of Publications

Academic articles

(1) Zhang Zhang, Xiaoyong Zhang, Kei Ichiji, Yumi Takane, Satoru Yanagaki, Yusuke Kawasumi, Tadashi Ishibashi, and Noriyasu Homma (2020). “Adaptive Gaussian mixture model-based statistical feature extraction for computer-aided diagnosis of micro-calcification clusters in mammograms". SICE Journal of Control, Measurement, and System Integration 13.4, pp. 183–190.
(2) Zhang Zhang, Xiaoyong Zhang , Kei Ichiji, Ivo Bukovsky, and Noriyasu Homma. "How intra-source imbalanced datasets impact the performance of deep learning for COVID-19 diagnosis using chest X-ray images". Scientific Reports (Submitted on September 5, 2022).
(3) Zhang Zhang, Xiaoyong Zhang, Jiaqi Chen, Yumi Takane, Satoru Yanagaki, Naoko Mori, Kei Ichiji, Katsuaki Kato, Mika Yanagaki, Akiko Ebata, Minoru Miyashita, Takanori Ishida and Noriyasu Homma. "Risk Analysis of Breast Cancer by Using Bilateral Mammographic Density Differences: A Case-control Study". The Tohoku Journal of Experimental Medicine (Submitted on July 4, 2023)

Domestic Conferences

(1) Zhang, Zhang, Xiaoyong Zhang, Kei Ichiji, Makoto Osanai, and Noriyasu Homma (2018). “Computer-Aided Diagnosis of Micro-Calcification Clusters in Mammograms Using an Adaptive Gaussian Mixture Model”. Proceedings of SSI2018, SS17-02.

(2) Zhang Zhang, Xiaoyong Zhang, Jiaoyang Wang, Yuwen Zeng, Kei Ichiji, Noriyasu Homma (2021). “Quantitative Evaluation of Explainable AI for COVID-19 Detection in Chest X-ray Images”. The 29th Symposium on Fuzzy, Artificial Intelligence, Neural Networks and Computational Intelligence, Online, 345-348.

International Conferences

(1) Zhang Zhang, Xiaoyong Zhang , Kei Ichiji, Ivo Bukovsky, and Noriyasu Homma (2022). "How intra-source imbalance affects deep learning performance for COVID-19 diagnosis using chest X-ray images". International Symposium on Human Welfare Engineering for Smart-Aging.

(2) Zhang Zhang, Xiaoyong Zhang, Kei Ichiji, Ivo Bukovsky, Shuoyan Chou and Noriyasu Homma (2023). "How Different Data Sources Impact Deep Learning Performance in COVID-19 Diagnosis using Chest X-ray Image". 2023 14th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI)

Awards

(1) Zhang, Zhang, Xiaoyong Zhang, Kei Ichiji, Makoto Osanai, and Noriyasu Homma (2018). SSI Excellent Paper Award, SICE System and Information Division.

Bibliography

Achanta, Radhakrishna et al. (2012). "SLIC superpixels compared to state-of-the-art superpixel methods". In: *IEEE transactions on pattern analysis and machine intelligence* 34.11, pp. 2274–2282.

Affonso, Carlos et al. (2017). "Deep learning for biological image classification". In: *Expert Systems with Applications* 85, pp. 114–122. ISSN: 0957-4174.

Anand, Rangachari et al. (1993). "An improved algorithm for neural network classification of imbalanced training sets". In: *IEEE Transactions on Neural Networks* 4.6, pp. 962–969.

Bekkar, Mohamed, Hassiba Kheliouane Djemaa, and Taklit Akrouf Alitouche (2013). "Evaluation measures for models assessment over imbalanced data sets". In: *J Inf Eng Appl* 3.10.

Bron, Esther E et al. (2021). "Cross-cohort generalizability of deep and conventional machine learning for MRI-based diagnosis and prediction of Alzheimer's disease". In: *NeuroImage: Clinical* 31, p. 102712.

Brunese, L. et al. (2020). "Explainable deep learning for pulmonary disease and coronavirus COVID-19 detection from X-rays". In: *Computer Methods and Programs in Biomedicine* 196, p. 105608.

Buda, Mateusz, Atsuto Maki, and Maciej A Mazurowski (2018). "A systematic study of the class imbalance problem in convolutional neural networks". In: *Neural networks* 106, pp. 249–259.

Castelvecchi, Davide (2016). "Can we open the black box of AI?" In: *Nature News* 538.7623, p. 20.

Cozzi, Diletta et al. (2020). "Chest X-ray in new Coronavirus Disease 2019 (COVID-19) infection: findings and correlation with clinical outcome". In: *La radiologia medica* 125.8, pp. 730–737.

Deng, Jia et al. (2009). "Imagenet: A large-scale hierarchical image database". In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee, pp. 248–255.

Doi, Kunio (2007). "Computer-aided diagnosis in medical imaging: historical review, current status and future potential". In: *Computerized medical imaging and graphics* 31.4-5, pp. 198–211.

Dong, Qi, Shaogang Gong, and Xiatian Zhu (2018). "Imbalanced deep learning by minority class incremental rectification". In: *IEEE transactions on pattern analysis and machine intelligence* 41.6, pp. 1367–1381.

Fawcett, T. (2006). "An introduction to ROC analysis". In: *Pattern recognition letters* 27, pp. 861–874.

Gao, Long et al. (2020). "Handling imbalanced medical image data: A deep-learning-based one-class classification approach". In: *Artificial intelligence in medicine* 108, p. 101935.

Garg, Arunim and Vijay Mago (2021). "Role of machine learning in medical research: A survey". In: *Computer Science Review* 40, p. 100370.

Gu, Jiuxiang et al. (2018). "Recent advances in convolutional neural networks". In: *Pattern recognition* 77, pp. 354–377.

Hemdan, E E D., M A. Shouman, and M E Karar (2020). "Covidx-net: A framework of deep learning classifiers to diagnose COVID-19 in X-ray images". In: *arXiv e-prints* arXiv: 2003, p. 11055.

Homma, N. et al. (2020a). "A deep learning aided drowning diagnosis for forensic investigations using post-mortem lung CT images". In: *42th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 1262–1265.

Homma, N. et al. (2020b). "Human ability enhancement for reading mammographic masses by a deep learning technique". In: *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 2962–2964.

Huang, Chen et al. (2016). "Learning deep representation for imbalanced classification". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5375–5384.

Ibrahim, Mai, Marwan Torki, and Nagwa El-Makky (2018). "Imbalanced toxic comments classification using data augmentation and deep learning". In: *2018 17th IEEE international conference on machine learning and applications (ICMLA)*. IEEE, pp. 875–878.

Johnson, J M. and T M Khoshgoftaar (2019). "Survey on deep learning with class imbalance". In: *Journal of Big Data* 6, pp. 1–54.

Khan, Salman H et al. (2017). "Cost-sensitive learning of deep feature representations from imbalanced data". In: *IEEE transactions on neural networks and learning systems* 29.8, pp. 3573–3587.

Korkmaz, Selcuk (2020). "Deep learning-based imbalanced data classification for drug discovery". In: *Journal of Chemical Information and Modeling* 60.9, pp. 4180–4190.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E Hinton (2017). "Imagenet classification with deep convolutional neural networks". In: *Communications of the ACM* 60.6, pp. 84–90.

Le, Duyen NT et al. (2020). "Transfer learning with class-weighted and focal loss function for automatic skin cancer classification". In: *arXiv preprint arXiv:2009.05977*.

Lee, Hansang, Minseok Park, and Junmo Kim (2016). "Plankton classification on imbalanced large scale database via convolutional neural networks with transfer learning". In: *2016 IEEE international conference on image processing (ICIP)*. IEEE, pp. 3713–3717.

Lei, D., X. Chen, and J. Zhao (2018). "Opening the black box of deep learning". In: *arXiv*, preprint arXiv:1805.08355.

Lin, Tsung-Yi et al. (2017). "Focal loss for dense object detection". In: *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988.

Lin, Zhong Qiu et al. (2019). "Do explanations reflect decisions? A machine-centric strategy to quantify the performance of explainability algorithms". In: *arXiv preprint arXiv:1910.07387*.

Liu, Xiaoxuan et al. (2019). "A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis". In: *The lancet digital health* 1.6, e271–e297.

Lo, Jui-En et al. (2021). "Data Homogeneity Effect in Deep Learning-Based Prediction of Type 1 Diabetic Retinopathy". In: *Journal of Diabetes Research* 2021.

Lotfy, Mayar et al. (2019). "Investigation of focal loss in deep learning models for femur fractures classification". In: *2019 International Conference on Electrical and Computing Technologies and Applications (ICECTA)*. IEEE, pp. 1–4.

Love, J. et al. (2021). "Comparison of antigen-and RT-PCR-based testing strategies for detection of SARS-CoV-2 in two high-exposure settings". In: *PloS one* 16, e0253407.

Masko, David and Paulina Hensman (2015). *The impact of imbalanced training data for convolutional neural networks*.

McDermott, Matthew BA et al. (2021). "Reproducibility in machine learning for health research: Still a ways to go". In: *Science Translational Medicine* 13.586, eabb1655.

Ng, Ming-Yen et al. (2020). "Imaging profile of the COVID-19 infection: radiologic findings and literature review". In: *Radiology: Cardiothoracic Imaging* 2.1.

Pouyanfar, Samira et al. (2018). "Dynamic sampling in convolutional neural networks for imbalanced data classification". In: *2018 IEEE conference on multimedia information processing and retrieval (MIPR)*. IEEE, pp. 112–117.

Quinn, T.P. et al. (2022). "The three ghosts of medical AI: Can the black-box present deliver?" In: *Artificial Intelligence in Medicine* 124, p. 102158.

Reza, Md Shamim and Jinwen Ma (2018). "Imbalanced histopathological breast cancer image classification with convolutional neural network". In: *2018 14th IEEE International Conference on Signal Processing (ICSP)*. IEEE, pp. 619–624.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin (2016). ""Why should i trust you?" Explaining the predictions of any classifier". In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144.

Ribli, Dezső et al. (2018). "Detecting and classifying lesions in mammograms with deep learning". In: *Scientific reports* 8.1, p. 4165.

Sadre, R. et al. (2021). "Validating deep learning inference during chest X-ray classification for COVID-19 screening". In: *Scientific reports* 11, pp. 1–10.

Selvaraju, R R. et al. (2020). "Grad-cam: Visual explanations from deep networks via gradient-based localization". In: *International Journal of Computer Vision* 128, pp. 336–359.

Simonyan, K. and A Zisserman (2014). "Very deep convolutional networks for large-scale image recognition". In: *arXiv preprint* arXiv:1409, p. 1556.

Tang, You-Bao et al. (2019). "Xlsor: A robust and accurate lung segmentor on chest x-rays using criss-cross attention and customized radiorealistic abnormalities generation". In: *International Conference on Medical Imaging with Deep Learning*. PMLR, pp. 457–467.

Titoriya, Ankit and Shelly Sachdeva (2019). "Breast cancer histopathology image classification using AlexNet". In: *2019 4th International conference on information systems and computer networks (ISCON)*. IEEE, pp. 708–712.

Tran, Giang Son et al. (2019). "Improving accuracy of lung nodule classification using deep learning with focal loss". In: *Journal of healthcare engineering* 2019.

Vayá, M I. et al. (2020). *BIMCV COVID-19+: a large annotated dataset of RX and CT images of COVID-19 patients*. *IEEE Dataport* https://dx.doi.org/10.21227/w3aw-rv39.

– (2021). *BIMCV COVID-19-: a large annotated dataset of RX and CT images of no COVID-19 patients*. *IEEE Dataport* https://dx.doi.org/10.21227/m4j2-ap59.

Wang, C. et al. (2020a). "A novel coronavirus outbreak of global health concern". In: *The lancet* 395, pp. 470–473.

Wang, L., Z Q. Lin, and A Wong (2020). "COVID-Net: A tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images". In: *Scientific Reports* 10, p. 19549.

Wang, Shoujin et al. (2016). "Training deep neural networks on imbalanced data sets". In: *2016 international joint conference on neural networks (IJCNN)*. IEEE, pp. 4368–4374.

Wang, W. et al. (2020b). "Detection of SARS-CoV-2 in different types of clinical specimens". In: *Jama* 323, pp. 1843–1844.

Wong, H Y F. et al. (2020). "Frequency and distribution of chest radiographic findings in patients positive for COVID-19". In: *Radiology* 296, E72–E78.

Yamac, M. et al. (2021a). "Convolutional sparse support estimator-based COVID-19 recognition from X-ray images". In: *IEEE Transactions on Neural Networks and Learning Systems* 32, pp. 1810–1820.

Yamac, M. et al. (2021b). *Qatar University and Tampere University COVID-19 (QataCOV19) Data set*. *Kaggle* https://www.kaggle.com/aysendegerli/qatacov19-dataset.

Yuan, Xiaohui, Lijun Xie, and Mohamed Abouelenien (2018). "A regularized ensemble framework of deep learning for cancer detection from multi-class, imbalanced training data". In: *Pattern Recognition* 77, pp. 160–172.

Zaki, M., K. Amin, and A. M. Hamad (2021). "COVID-19 Detection Based on Chest X-Ray Image Classification using Tailored CNN Model". In: *IJCI. International Journal of Computers and Information* 8, pp. 100–108.

Zhang, Zhang et al. (2020). "Adaptive Gaussian mixture model-based statistical feature extraction for computer-aided diagnosis of micro-calcification clusters in mammograms". In: *SICE Journal of Control, Measurement, and System Integration* 13.4, pp. 183–190.

Zhang, Zhang et al. (2023a). "How Different Data Sources Impact Deep Learning Per- formance in COVID-19 Diagnosis using Chest X-ray Image". In: *2023 14th IIAI Interna- tional Congress on Advanced Applied Informatics (IIAI-AAI)*. IEEE CPS, pp. 508–513.

Zhang, Zhang et al. (2023b). "How intra-source imbalanced datasets impact the performance of deep learning for COVID-19 diagnosis using chest X-ray image". In: *Scientific Reports*, Submitted on September 5, 2022.

Zhang, Zhang et al. (2023c). "Risk Analysis of Breast Cancer by Using Bilateral Mammographic Density Differences: A Case-control Study". In: *The Tohoku Journal of Experimental Medicine*, Submitted on July 4, 2023.