LUNG CANCER RISK ESTIMATION WITH IMPERFECT DATA FROM MULTIPLE MODALITIES

By

Riqiang Gao

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Computer Science

March 31, 2022

Nashville, Tennessee

Approved:

Bennett Landman, Ph.D.

Yuankai Huo, Ph.D.

Thomas Lasko, M.D-Ph.D.

Ipek Oguz, Ph.D.

Zhoubing Xu, Ph.D.

Kim Sandler, M.D.

To my parents

ACKNOWLEDGMENTS

My advisor Dr. Bennett Landman has led me to the magic world of science and guided me on how to be a qualified researcher. He is my role model in my academic career. He is the most energetic and enthusiastic person that I have even closely worked with. His ways of managing a scientific lab greatly inspired me and I will follow most of them if I have the chance to establish my own lab in the future. Dr. Landman's expertise and vision are invaluable to me for my current and future work.

Dr. Yuankai Huo has acted the role of my unofficial second advisor. He was my guide of work and life in the United States especially when I first came to this country. He spent lots of time discussing projects with me and has provided invaluable advice. Dr. Thomas Lasko is an expert in biomedical informatics and computer science. He has provided plenty of valuable suggestions/discussions on multiple research projects, especially in the aspect of statistics. Dr. Pierre Massion and Dr. Kim Sandler are clinical experts in lung cancer. Most of my research questions come from the discussions with them and they have provided valuable suggestions and comments to shape my research direction and vision. Except being my committee member, Dr. Zhoubing Xu is my mentor during the internship in Siemens Healthineers. His talent, patience, and enthusiasm led me to have a productive and enjoyable experience. As one of the committee members, Dr. Ipek Oguz has provided sage wisdom to help me finish my dissertation work. Additionally, lots of other collaborators have offered valuable support and suggestions for training me.

I would like to thank the big family MASI lab. I have a fantastic experience in MASI that I will never forget throughout all my life. We have such a multicultural and harmonious atmosphere in the lab, where we discuss scientific questions and experiences to be a Ph.D. student. After getting off work, we are good friends. We gather to play card games, celebrate birthdays/holidays, etc. to achieve a work-life balance. I also thank all my friends (not limited to MASI) who support me going through the hard times and bring joy. You guys have made my off-work time more enjoyable.

I would like to thank Liang Gao who always sees the best in me. She loves and supports me, and calms me down when I was under pressure. We have experienced colorful life and believe there are more to be explored. Lastly, I want to thank my family. My parents Xuanliang Gao and Yuanliang

Fu unconditionally love and support me through all my life. I was born in an undeveloped region of China and the education level of my parents is not high due to historical reasons. However, the wisdom of my parents is much more beyond their circle and they have guided me on how to be a better man. I have not found accurate words to express my greatest gratitude to them.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Introduction

1.1 Overview

Lung cancer is one of the most prevalent cancers and has the highest mortality rate among all cancers as reported in the United States in recent years [1, 2]. The survival rates for lung cancer patients are highly dependent on the cancer stage at the time of diagnosis [3]. Recently, low-dose computed tomography (CT) of the chest has been regarded as one of the best available technologies for lung cancer screening [4]. The results of the National Lung Screening Trial (NLST) [4] led to the recommendation of the U.S. Preventive Services Task Force for annual lung cancer screening with low dose CT for people between 55 and 80 years old who have over 30 smoking pack-years and quit smoking within 15 years [5].

Lung cancer screening usually results in the detection of pulmonary nodules. Indeterminate pulmonary nodules (IPNs) management has become another critical procedure [6]. The medical community struggles with around 1.5 million IPNs every year in the United States [6], which may lead to an unsatisfactory situation that cancer-free patients suffer from invasive procedures while early cancer patients miss opportunities for intervention [7].

Low-dose CTs generate high-resolution imaging data that can be used to find small and low-contrast nodules [8]. Additionally, tabular data including clinical data elements (CDEs) (e.g., age, smoking history), human-curated imaging semantic features (ISFs) (e.g., nodule size, spiculation, lobe location), blood markers can provide complementary information with structural CT imaging. Those are two important data modalities that are used to predict lung cancer risk.

The focus of this dissertation is on the lung cancer risk estimation with imperfect data from multiple modalities (i.e., tabular data as non-image modality and CT images as the image modality). The key areas in which we seek to improve the existing literature are refined pipelines of quality assessment of raw CT images, learning from sequential CT images which may miss timepoints or be irregularly sampled, integrating image and non-image data in an end-to-end learning pipeline considering missing modality, imputing partial missing data in multiple modalities by modeling joint

distribution of multi-modal available data, calibration of predicting model, and careful evaluation in screening and incidental clinical cohorts with modern machine learning.

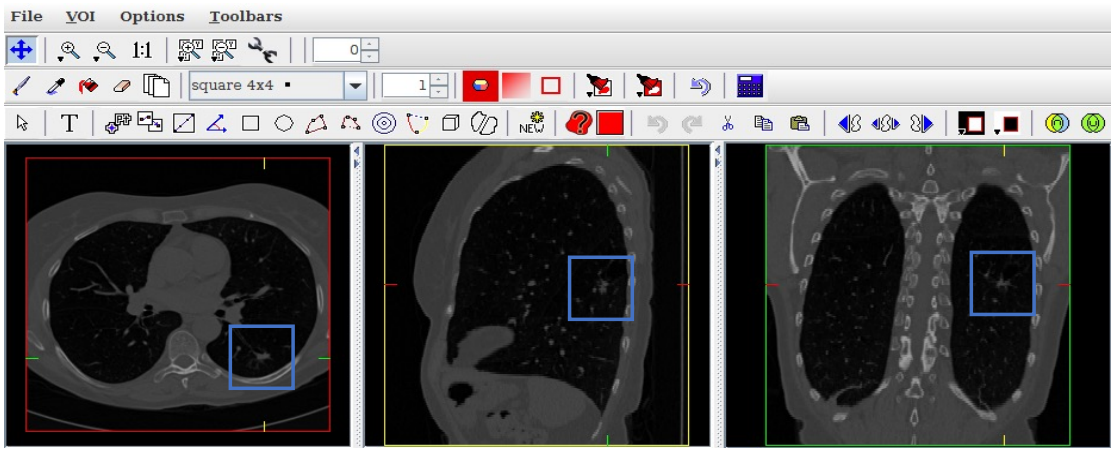
1.2 Image and Non-image Data used for Lung Cancer Risk Estimation

The human lungs (i.e., the left and right lungs) are in the thorax, which is the chest region of the body between the neck and the abdomen. The lungs are conical, with a narrowly rounded apex at the top, and a broad concave bottom on the convex surface of the diaphragm [9], as shown in Figure 1.1. The left lung has an indentation at its border that is called Cardiac Notch and the left lung shares space with the heart [10]. Each lung is divided into lobes (right lung consists of three lobes and left lung consists of two) by fissures, which are double folds of the pleura that help the expansion of the lung [11].

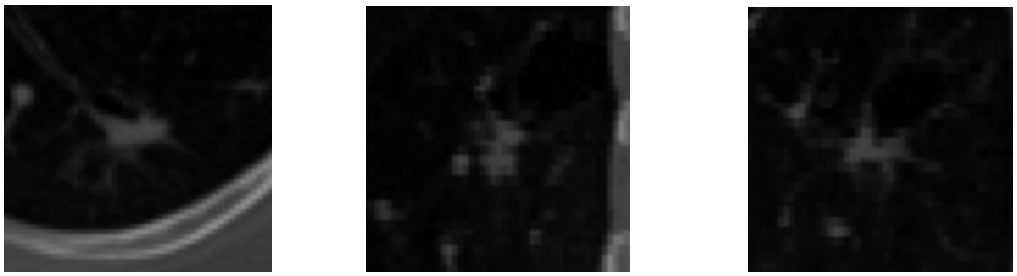
Lung cancer results from a malignant lung tumor characterized by uncontrolled cell growth in lung tissues [12]. There are two main types of lung cancer: small-cell lung carcinoma (SCLC) and non-small-cell lung carcinoma (NSCLC), where the dominate types of NSCLC are: adenocarcinoma, squamous-cell carcinoma, and large-cell carcinoma [13, 14]. Most NSCLC patients are smokers while adenocarcinomas can be existed in never-smoking ones. Compared with NSCLC, SCLCs are more responsive to radiation therapy and chemotherapy, but are harder to treat since they are easy to disseminate [15]. To assess the spread of lung cancer, cancer staging is widely used. The staging evaluation of NSCLC uses the TNM (tumor, node, metastasis) classification [16], which is based on the size of the primary tumor, lymph node involvement, and distant metastasis. The SCLC is classified as "limited stage" and "extensive stage" based on the tests for lung cancer (e.g., biopsies, imaging) [15, 17], where limited stage usually can be encompassed by radiation therapy while extensive stage has spread out of supraclavicular areas.

1.2.1 CT imaging and Related Clinical Data for Lung Cancer Detection

Computed tomography (CT) is a computerized x-ray imaging procedure, where a narrow beam of x-rays is quickly rotated around the body to produce signals processed by computers to generate cross-sectional slices [18]. When scanning, the x-ray tube rotates around the patient when the patient lies on the bed and the bed slowly passes through the body. The x-rays are picked up by the detectors and transmitted to a computer when leaving the body [18]. These cross-sectional



(a) Three-dimensional view with MIPAV, with a pulmonary nodule annotation



(b) Zoom in the pulmonary nodule

Figure 1.1: A 3D CT example visualized by MIPAV. A malignant nodule is highlighted in the (a) and zoomed in (b). Left: Axial view, middle: Sagittal view, right: Coronal view.

slices contain more detailed information than conventional x-rays, which can be stacked together to form a three-dimensional volume of the patient. The three-dimensional images are usually easier to identify abnormalities. Viewing the anatomy in three different planes (axial, sagittal, and coronal as shown in Figure 1.1 (a)) brings advantage to evaluate the disease of patients [19].

A pulmonary nodule is a small mass of tissue in the lung, which usually appears as round and white shadows in a CT scan (Figure 1.1 (b)). The Hounsfield scale of lung range -700 to -600, while the Hounsfield scale of nodules is around 20 [20]. Thus, the nodule is identifiable from the CT scan, as Figure 1.1. Most pulmonary nodules are not cancer, which can be caused by old infections or scar tissues [21]. A malignant pulmonary nodule on a chest radiograph is a commonly used indicator for lung cancer. The size and extent of the main tumor is an important indicator of lung cancer stage [16].

We consider a general diagnosis pipeline in clinical treatment [21], as in Figure 1.2. Patients have demographic information (e.g., age, gender) and some clinical data elements are collected from

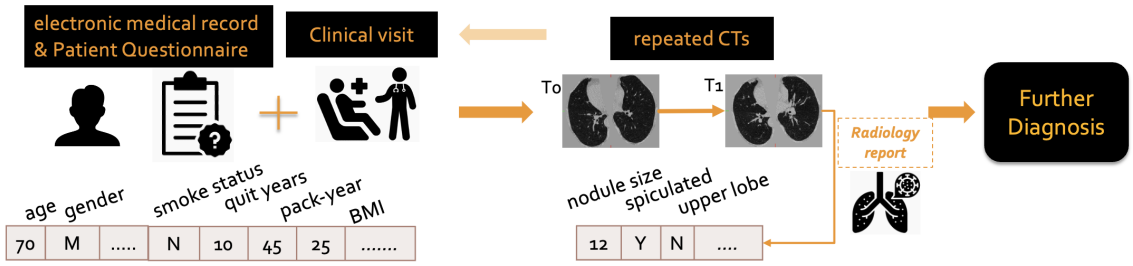


Figure 1.2: The illustration of a clinical treatment. Patients have data from multiple diagnosis stages. The details are described in text.

electronic medical records (EMR) and patient questionnaires during the clinical visit. Those data are used to determine if a CT scan is necessary. For each performed CT scan, a radiology report would typically be created to extract nodule characterization information. Then, such a process might recur according to patients' clinical status. Sometimes, the doctor may also suggest patients take a blood test in the follow-up diagnosis. According to the test results and symptoms, the doctor may suggest the patient take an actual diagnosis by looking at lung cells in the lab (e.g., biopsy).

In practice, factors such as network issues, accelerated acquisitions, motion artifacts, and imaging protocol design can impede the interpretation of collections of two-dimensional images, especially under the context of collaboration of multiple teams [22]. For example, the NLST dataset [4] integrates CT images from 33 sites across over 20 manufacturer and model combinations, which may trigger data issues when combined. Even within a single institution, such as the Vanderbilt University Institute for Imaging Science Center for Computational Imaging (VUIIS CCI) [23] database may encounter issues with incomplete transfer when the workload is heavy. Potential data quality issues can be exacerbated when human-based workflows use limited views of the data that may obscure digital artifacts that can drive erroneous decisions. In Chapter 3, we introduce quality assessment in more detail and our tools to handle quality assessment.

1.2.2 Chest CT Datasets

In this subsection, we introduce representative datasets in lung cancer diagnosis community.

Lung Image Database Consortium and Image Database Resource Initiative (LIDC-IDRI) [24]. The LIDC-IDRI is a public dataset including diagnostic and lung cancer screening chest CT images with marked-up annotated lesions. This dataset contains 1010 patients with 1308 studies

created by seven academic centers and eight medical imaging companies. Each patient contains images from chest CT and an associated annotation file recording image annotation performed by four radiologists. This dataset has been widely used for nodule detection, nodule segmentation [25], nodule classification. The LUNA16 dataset is a part of LIDC-IDRI [26].

National Lung Screening Trial (NSLT) [4]. The NLST is a randomized controlled program for lung cancer detection that was designed to evaluate whether the screening with CT reduces mortality than screening with X-ray. About 54,000 participants are enrolled between 2002 and 2004. NLST is one of the largest lung CT datasets, which was collected from 33 centers. The eligibility criteria for NSLT include age between 55 and 74, $\geq$ 30 pack-year smoking history, $\leq$ 15 years quit smoking time. The exclusion criteria include: has lung cancer history, had cancer within 5 years, underwent chest CT within 18 months, recent medical problems (pneumonia or other acute respiratory infection, unexplained weight loss, coughing up blood), participation in other studies of cancer screening or prevention, removal of any portion of the lungs, inability to lie flat with arms raised over head, home oxygen supplementation, and metallic implants in the chest or back (descriptions from https://cdas.cancer.gov/learn/nlst/trial-summary/).

Vanderbilt Lung Screening Trial (VLSP) [27]. The VLSP is a comprehensive program including annual CT screening, managed by the department of Radiology, Vanderbilt University Medical Center. The VLSP eligibility criteria include age between 55 and 80, $\geq$ 30 pack-year smoking history, $\leq$ 15 years quit smoking time.

Consortium for Molecular and Cellular Characterization of Screen-Detected Lesions (MCL) [28]. In this project, images and biomarkers are collected from multiple hospital sites including Vanderbilt University Medical Center (VUMC), University of Pittsburgh Medical Center (UPMC), and Detection of Early Cancer Among Military Personnel (DECCAMP), University of Denver (UCD).

1.3 Computer Aided Diagnosis

Interpreting medical images for better computer-aided diagnosis (CAD) has been a long-standing challenge. In this section, we introduce the basic concepts of conventional CAD and model CAD with deep learning.

1.3.1 Conventional CAD: Radiomics

Generally, patient demographics include the information that can identify a patient (e.g., contact information, medical record number) and data allow for categorization of statistical analysis. The demographics are usually reported by the patients. In most research, including ours, the demographics are de-identified to protect patient privacy. In the research of lung cancer detection, lung cancer risk factors, sometimes termed as clinical data elements, are important resources and usually taken into consideration. The national lung screening trial used age, smoking status, pack-year, and quit smoking time as the inclusion/exclusion criteria [4]. PLCOm2012 model is refined selection criteria for lung cancer screening using 11 lung cancer risk factors [29]. The Mayo model included age, smoking status, and personal cancer history to predict the probability of malignancy of solitary pulmonary nodules [30]. The risk factors of the Brock model [31] include demographics of age, sex, family lung cancer history, emphysema.

Radiomics refers to methods that extract features from radiographic medical images with data characterization techniques [32]. The extracted radiomics features are designed to detect interesting characterization that cannot be directly measured by eye and can be used for personalized diagnosis or prognosis [32]. As discussed earlier, the characterization of pulmonary nodule usually reflects lung cancer condition. Radiomic features are widely used in evaluating lung cancer risk and guidelines for the management of indeterminate pulmonary nodules. For example, the nodule size, spiculation (whether the nodule is spiculated), and upper-lobe (whether the nodule is in the upper-lobe or not) are representative radiomic features and used in Mayo [30] and Brock [31] models. [31] show (1) a strong relationship between the nodule size and probability of lung cancer, (2) nodule spiculation might improve prediction indicated by the net reclassification improvement, (3) a larger number of nodules and cancers were observed in upper lobes. The correlation of nodule type (solid, nonsolid, and partial solid) and cancer probability is also validated by [31], and nodule type is included as a data element in the Brock model and Fleischner guidelines [33]. Blood tests are also common in clinical diagnosis. CYFRA 21-1 is a fragment of the protein cytokeratin-19, which was investigated as a potential lung cancer biomarker [34, 35].

The radiomic features and blood biomarkers are proved to be helpful for lung cancer risk estimation, while obtaining those data elements requires large human efforts, which increases the burden in buy imaging practices.

1.3.2 Modern CAD: Deep Learning

Prior to deep learning era, most of medical imaging models (e.g., radiomics) depend on morphological features to aid clinical decisions. For example, the Mayo model evaluate nodule malignancy using 3 clinical and 3 radiographic variables (age, smoking status, cancer history, nodule size, spiculation, upper lobe) with a multivariate logistic regression [30]. However, those features usually require large effort from human experts, making the image processing experience and hard to generalize to large scale dataset.

Recent development in modern machine learning (i.e., deep learning) has made great progress in medical image analysis [36]. Instead of using manually extracted radiomic features, deep learning models can obtain high-level feature from raw medical data automatically. In most of lung cancer risk estimation pipelines [37, 38], there are mainly two steps to diagnose a patient based on a CT scan: pulmonary nodule detection, classification based on nodule proposals. Pulmonary nodule detection refers to localize suspicious regions to be nodules and the classification stage usually extracts features from the nodule proposals (may also include image features from whole scan). Some methods such as [39, 40] only target at the classification of the malignancy of nodules, however, requires the pre-defined location of nodules. In this dissertation, our methods are motivated by the framework of [37], which diagnose the lung cancer without human localizing the nodules. The details of pulmonary nodule detection and classification stage are described in Chapter 2.3 and Chapter 2.4, respectively.

Sequential learning is a type of learning method that takes input of series data and captures the change across the sequence. Sequential models have been well developed with the marriage of deep learning. To utilize the serial CT scans, sequential learning models can be adapted in the context of lung cancer diagnosis. The technique background of sequential learning can be found in Chapter 2.5, and our proposed models can be found in Chapter 4 and 5.

Except for image modality data, there may be other data resource such as clinical data elements can be utilized. Multi-modal fusion is type of deep learning that can integrate multi-modal data for better prediction. We have several multi-modal based studies on lung cancer risk estimation in Chapter 6, 7, 9, 10.

	predicted Positive	predicted Negative
actual Positive	*True Positive (TP)*	*False Negative (FN)*
actual Negative	*False Positive (FP)*	*True Negative (TN)*

Figure 1.3: Confusion matrix of binary classification

1.4 Evaluation for Risk-Prediction Models

Discrimination, calibration are two essential concerns of a clinical decision making. Discrimination refers to the ability to separate samples of different classes. Calibration represents the agreement between observed and predicted class probability.

1.4.1 Discrimination

A prevalent metric for discrimination is called Area Under the Receiver Operating Characteristic curve (AUROC, sometimes just called AUC) [41]. A ROC graph is used for visualizing, organizing, and selecting model performance. In the following, we begin our analyses on classification of two classes (i.e., positive vs. negative, which matches lung cancer detection).

Generally, a binary classification model predicts a continuous value as the predicted risk of positive. By thresholding, the predicting can be converted to discrete output (i.e., positive or negative). Given a risk-prediction model and an instance, there are four situations as illustration in confusion matrix Figure 1.3.

The true positive rate (TPR), also termed as sensitivity, recall, is computed as follows:

$$TPR = \frac{TP}{TP+FN} = 1 - FNR \tag{1.1}$$

The true negative rate (TNR), also termed as specificity, selectivity, is computed as follows:

$$TNR = \frac{TN}{TN+FP} = 1 - FPR \tag{1.2}$$

The positive predictive value (PPV), also termed as precision, is computed as follows:

$$PPV = \frac{TP}{TP+FP} = 1 - FDR \tag{1.3}$$

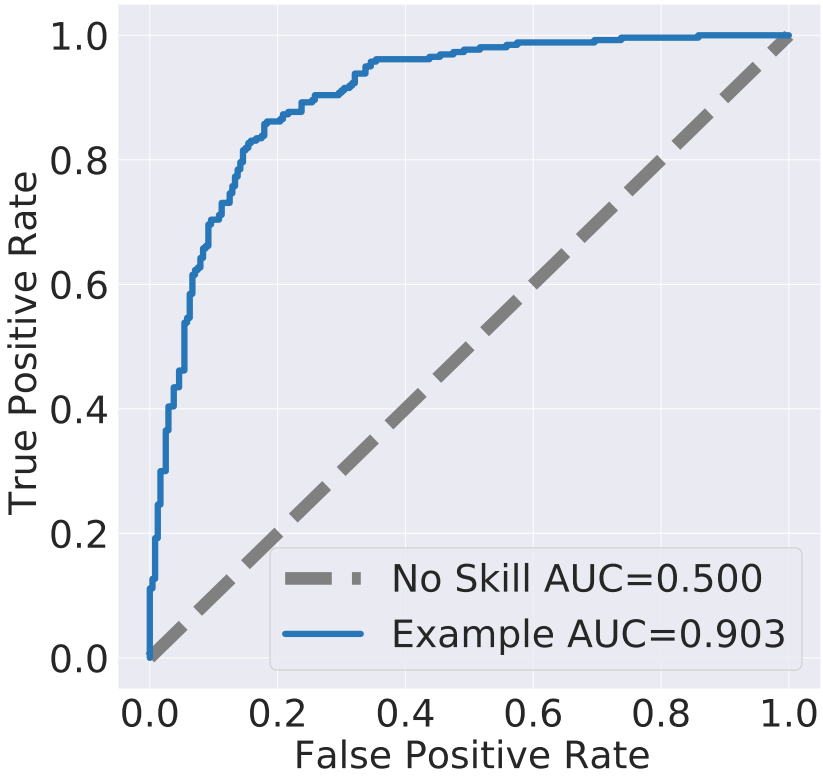


Figure 1.4: A ROC curve example. The random guessing performance (i.e., No Skill) is illustrated by the diagonal line. The orange line is an example model which achieves an AUC value of 0.903.

The ROC spaces are two-dimensional plots with the TPR as x-axis and the FPR as y-axis, one example shown in Figure 1.4.

Some classifiers, such as decision tree, only produce a discrete prediction value (e.g., 1 or 0). Such classifier only creates a single point in the ROC space.

1.4.2 Calibration

The expected output of risk-prediction models is associated with positive probability. Assessing the calibration of the predicted risk is an important aspect of the risk-prediction model. A model where the predicted occurrence probability matches the actual occurrence frequency has perfect calibration.

Expected Calibration Error (ECE) is one of the most popular metrics to measure the calibration

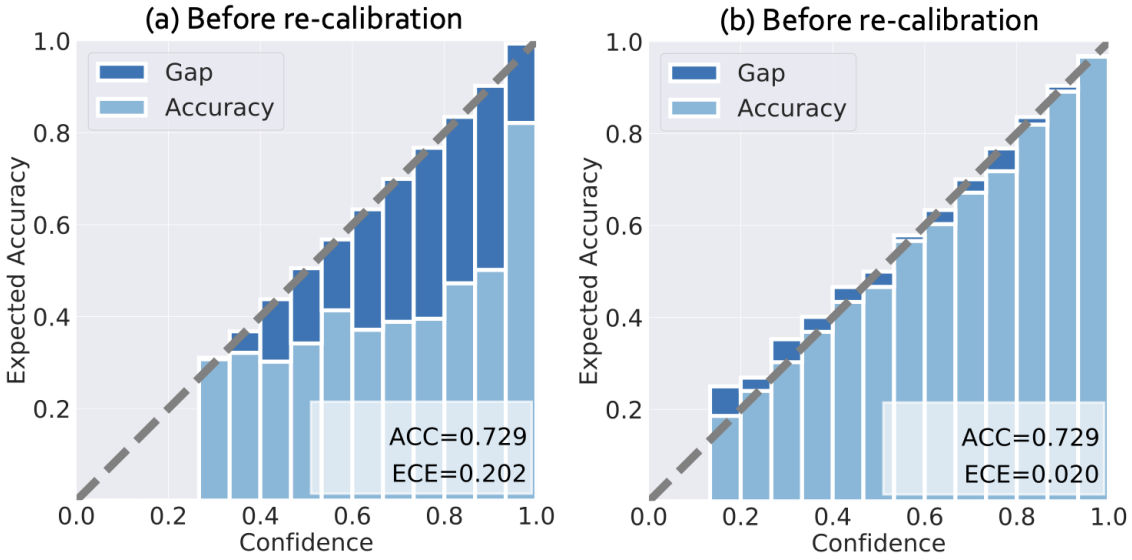


Figure 1.5: Reliability diagram examples. (a) is poorly calibrated example from a deep learning model (ECE=0.202). Using a re-calibration technique, the calibration performance can be greatly improved (ECE=0.02), while the accuracy remains the same.

performance. There are two main steps to compute ECE: (1) dividing the prediction value space into equal-space bins, (2) calculating the weighted average of the difference of accuracy and confidence. The ECE is defined as:

$$ECE = \sum_{m=1}^{M} \frac{|B_m|}{n} |acc(B_m) - conf(B_m)| \tag{1.4}$$

Where the n is the number of samples, B_m is the m-th bin.

Reliability diagrams [42] are the figures of the predicted probabilities (confidence) and the observed frequency (expected frequency), which shows the frequency of predicted probabilities happened in practice (that is calibration).

Figure 1.5 compares the performance of a deep learning model between before re-calibration and after re-calibration. Figure 1.5 (a) shows the discrimination performance (accuracy=0.729) and the calibration performance (ECE = 0.202) before the re-calibration. With a re-calibration technique, the calibration performance can be greatly improved (ECE: 0.202$-$ $>$0.020) while leave the discrimination performance unaffected.

1.5 Challenges

As an essential clinical problem, lung cancer risk estimation has its own challenges.

Challenge 1. Image quality assessment (IQA) is important for scientific inquiry, especially in medical imaging and machine learning. Potential data quality issues can be exacerbated when human-based workflows use limited views of the data that may obscure digital artifacts that can drive erroneous decisions. In practice, multiple factors such as network issues, accelerated acquisitions, motion artifacts, and imaging protocol design can impede the interpretation of collections of two-dimensional images as representing a complete three-dimensional view of anatomy. The medical image processing community has developed a wide variety of tools for the inspection and validation of imaging data. Yet, IQA of computed tomography (CT) remains an under-recognized issue and no user-friendly tool is commonly available to address these potential issues.

Challenge 2. With the rapid development of image acquisition and storage, multiple images per class are commonly available. The recurrent neural network (RNN) has been widely integrated with convolutional neural networks (CNN) to perform image classification on ordered (sequential) data. While the information encoded in the sequential data is not fully explored in lung cancer risk estimation. In addition, clinical imaging acquisition may be irregularly sampled, and such sampling patterns may be commingled with clinical usages. Existed methods have not well explored how to handle irregularly sampled longitudinal CTs in lung cancer risk estimation.

Challenge 3. Clinical data elements (CDEs) (e.g., age, smoking history), blood markers and chest computed tomography (CT) structural features have been regarded as effective means for assessing lung cancer risk. These independent variables can provide complementary information and we hypothesize that combining them will improve the prediction accuracy. In practice, not all patients have all these variables available. Existed methods have not well explored how to handle the missing data in lung cancer risk estimation.

Challenge 4. Discrimination and calibration are two essential aspects of clinical decision making. Currently, most methods focus on discrimination, i.e., use the AUC as the main metric to evaluate prediction performance. A predicting model with careful consideration on discrimination and calibration in clinical practice is not well explored.

1.6 Contributed Works

We explore the possibility of lung cancer risk estimation with imperfect data from multiple modalities. First, we developed a human-interactive semi-auto image quality assessment tool, which controls the image quality from large-scale dataset (Contribution 1, Challenge 1). We coordinate with clinical collaborators and manage update large-scale datasets. Second, we introduce new algorithms to handle imperfect longitudinal CT scans for lung cancer risk estimation by introducing new recurrent network structure and encoding time stamp information into model in a fashion way (Contribution 2, Challenge 2). Third, we develop a new end-to-end deep learning model integrating image feature and non-image biomarkers, with the ability of handling missing modality (Contribution 3, Challenge 3). We also extend the missing modality problem to partial missing and explore novel imputation techniques across-modality (Contribution 4, Challenge 3). Except for discrimination, calibration is essential for clinical decision making. We explore the confidence calibration in prediction models (Contribution 5, Challenge 4). Finally, we validate our proposed algorithm with specific clinical interests (e.g., re-classification) in screening and incidental cohorts (Contribution 6).

Contribution 1: human-interactive semi-auto image quality assessment tool

Image quality assessment (QA) is important for scientific inquiry, especially in medical imaging and machine learning. Potential data quality issues can be exacerbated when human-based workflows use limited views of the data that may obscure digital artifacts that can drive erroneous decisions. In practice, multiple factors such as network issues, accelerated acquisitions, motion artifacts, and imaging protocol design can impede interpretation of collections of two-dimensional images as representing a complete three-dimensional view of anatomy. The medical image processing community has developed a wide variety of tools for inspection and validation of imaging data. Yet, QA of computed tomography (CT) remains an under-recognized issue and no user-friendly tool is commonly available to address these potential issues. Here, we create and illustrate a pipeline specifically designed to identify and resolve issues encountered with large scale data mining of clinically acquired CT data. We describe our method and evaluation in Chapter 3.

Contribution 2: lung cancer detection with longitudinal CT scans

With the rapid development of image acquisition and storage, multiple images per class are commonly available for computer vision tasks (e.g., face recognition, object detection, medical

imaging, etc.). Recently, the recurrent neural network (RNN) has been widely integrated with convolutional neural networks (CNN) to perform image classification on ordered (sequential) data. By permutating multiple images as multiple dummy orders, we generalize the ordered "RNN+CNN" design (longitudinal) to a novel unordered fashion, called Multi-path x-D Recurrent Neural Network (MxDRNN) for image classification. To the best of our knowledge, few (if any) existing studies have deployed the RNN framework to unordered intra-class images to leverage classification performance. Specifically, multiple learning paths are introduced in the MxDRNN to extract discriminative features by permutating input dummy orders. Eight datasets from five different fields (MNIST, 3D-MNIST, CIFAR, VGGFace2, and lung screening computed tomography) are included to evaluate the performance of our method. More details will be introduced in Chapter 4.

The Long Short-Term Memory (LSTM) network is widely used in modeling sequential observations in fields ranging from natural language processing to medical imaging. The LSTM has shown promise for interpreting computed tomography (CT) in lung screening protocols. Yet, traditional image-based LSTM models ignore interval differences, while recently proposed interval-modeled LSTM variants are limited in their ability to interpret temporal proximity. Meanwhile, clinical imaging acquisition may be irregularly sampled, and such sampling patterns may be commingled with clinical usages. In Chapter 5, we propose the Distanced LSTM (DLSTM) by introducing time-distanced (i.e., time distance to the last scan) gates with a temporal emphasis model (TEM) targeting at lung cancer diagnosis (i.e., evaluating the malignancy of pulmonary nodules). Briefly, (1) the time distance of every scan to the last scan is modeled explicitly, (2) time-distanced input and forget gates in DLSTM are introduced across regular and irregular sampling sequences, and (3) the newer scan in serial data is emphasized by the TEM. Our model and studies will be introduced detailly in Chapter 5.

Contribution 3: multi-modal learning with missing modality data

Clinical data elements (CDEs) (e.g., age, smoking history), blood markers and chest computed tomography (CT) structural features have been regarded as effective means for assessing lung cancer risk. These independent variables can provide complementary information and we hypothesize that combining them will improve the prediction accuracy. In practice, not all patients have all these variables available. We propose a new network design, termed as multi-path multi-modal missing network (M3Net), to integrate the multi-modal data (i.e., CDEs, biomarker and CT image) consid-

ering missing modality with multiple paths neural network. Each path learns discriminative features of one modality, and different modalities are fused in a second stage for an integrated prediction. The network can be trained end-to-end with both medical image features and CDEs/biomarkers or make a prediction with single modality. More details about method and studies can be found at Chapter 6.

Contribution 4: multi-modal imputation with partial missing data

Ideally, patients can have sequential CT scans and clinical data element, while there are partial missing data can exist in both modalities. We posit that essential information missed in one modality can be maintained in another. We propose the Conditional PBiGAN (C-PBiGAN) to model the joint distribution across modalities by introducing 1) a conditional latent space in multi-modal missing imputation context; 2) a class regularization loss to capture discriminative information during imputation. Herein, we focus on lung cancer risk estimation, where risk factors and serial CT scans are two essential modalities for rendering clinical decisions. C-PBiGAN achieves superior predicting performance of downstream multi-modal learning tasks in three broad settings: 1) missing data in image modality, 2) missing data in non-image modality, and 3) both modalities have missing data. With C-PBiGAN, we validate that 1) CT images are conducive to impute missed factors for better risk estimation, and 2) lung nodules with malignancy phenotype can be imputed conditioned on risk factors. This contribution can be found in Chapter 7.

Contribution 5: calibration analyses of prediction models

Except for discrimination, confidence calibration is another aspect to evaluate classification/prediction models. we conduct a comparative study of representative calibration models. Starting from the application of multi-class classification with the widely used CIFAR10, we further extend the empirical experiments on lung cancer diagnosis with NLST and in-house datasets. In the experiments of CIFAR10, we include the calibration analyses of balanced training and imbalanced training (long-tail). In the lung cancer diagnosis, it is an imbalanced clinical problem in nature, we include external validation analyses across multiple sites. In conclusion, we compare the calibration performances across CIFAR10 and real lung cancer diagnosis and provide recommendations to choose a model. The details are in Chapter 8.

Contribution 6: clinical validation with screening and incidental cohorts

Careful characterize data population depending on specific clinical cohorts (i.e., screening, or

incidental cohort), analysis how the framework collaboratively learning from multi-modal data can be adapted. We validate the multi-modal learning in the lung cancer screening and incidental lung nodule cohorts. Specifically, we have validated our model in relatively low risk (e.g., screening) and high risk (e.g., incidental nodules) populations separately to close address clinical questions. The validation details are illustrated in Chapter 9 and 10.

CHAPTER 2

Deep Learning for Lung Cancer Risk Estimation

2.1 CT Image Preprocessing

A CT scan is usually acquired with a series of Digital Imaging and Communications in Medicine (DICOM) images. Typically, scans are often represented as collections of two-dimensional images (e.g., DICOM) that should be re-composed / re-formatted to determine the three-dimensional structure or higher (e.g., Neuroimaging Informatics Technology Initiative (NIfTI)). Due to the heterogeneity of medical imaging, a proper image preprocessing pipeline can make the learning more robust.

Hounsfield unit. The general first step of CT image preprocessing is to convert the raw data into the Hounsfield unit (HU) [37]. The HU scale is a linear transformation of the original attenuation coefficient measurement into a value, in which the radiation density of distilled water under standard pressure and temperature (STP) is defined as 0 HU, and the radiation density of air at STP is defined as -1000 HU [43]. The HU value is defined as follow, given the average linear attenuation coefficient μ:

$$HU = \frac{\mu - \mu_w}{\mu_w - \mu_a} \times (HU_w - HU_a) \tag{2.1}$$

where $HU_w = 0$ and $HU_a = -1000$ represent the HU definition for water and air, respectively. The μ_w and μ_a represent the linear attenuation coefficient of water and air.

Lung Segmentation. A chest CT scan includes not only the lung but also other tissues, which may appear spherical like pulmonary nodules. To reduce these interfering factors, a convenient way is to segment the lung and rule out the other tissues [37]. Deep learning models have been used for lung segmentation [38, 44], however, which are data-hungry and annotation-need to train a robust segmentation. A simple way to segment the lung is using classical image processing (e.g., morphology operation) [37], and the detailed steps are shown as follows.

In Figure 2.1, the CT scan is in HU scale and resampled to $1 \times 1 \times 1$ mm resolution. According to the summary in [43], the HU of the lung ranges from -700 to -600. The HUs of tissues, bones

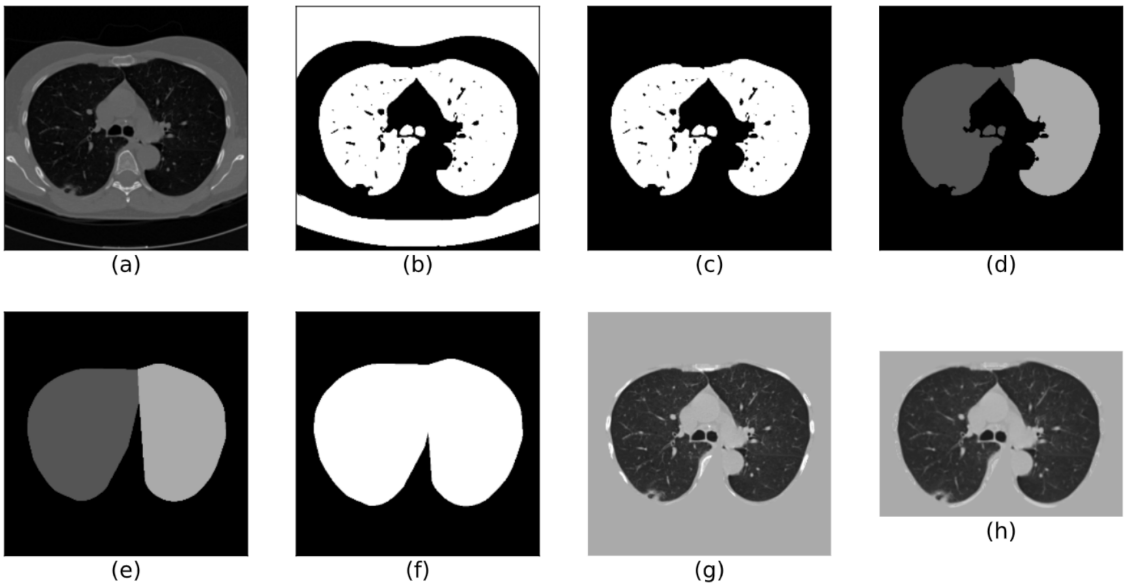


Figure 2.1: Preprocessing steps following Liao et al. (a) CT images in HU. (b) Binarized mask by thresholding. (c) Retaining the lung mask. (d) Mask after eroding and dilating. (e) Convex hulls of left and right lung masks. (f) Combining two masks by dilating. (g) Masked and normalized CT image. (h) Cropping the image and clip the bone. Figures are from Liao et al.

around the lung are large than -600. We follow the instruction of [37] using the threshold -600 HU to binarize the image as Figure 2.1(b). To keep the lung part of the image (located in the center), we calculate all 3-D connected components in the resulting binary 3-D matrix and remove those that touch matrix corner and whose volume is out of the range $[0.68L, 7.5L]$. Ideally, there is only one binary component left corresponding to the lung. By computing the minimum distance from each component to the image center and component area, we can remove the distracting ones and keep the lung mask component (as Figure 2.1(c)). Considering that there may be pulmonary nodules attached to the outer wall of the lung, the convex hulls are computed to include possible nodules. Two convex hulls for left and right lungs are obtained separately to avoid including unwanted tissues (like the heart). The lung mask is eroded until it is broken into two parts, corresponding to the left and right lungs. Then, the two parts are dilated back to original size, as shown in Figure 2.1(d). Figure 2.1(e) show the convex hulls of two lungs to include possible nodules on the wall, and Figure 2.1(f) is the obtained mask which is further dilated with 10 voxels involving surrounding tissues. For better storage and feeding to deep learning models, the CT scan has been transformed from the HU scale to UINT8. According to the HU values of the organ and tissues in the chest, we

eliminate the unwanted parts by clipping the HU data matrix within [-1200, 600]. Then, the voxel value is further linearly transformed to [0, 255]. A cleaned data matrix is achieved after multiplying the lung masks to the linearly transformed matrix and replacing intensity out of the mask as 170. Considering the ribs around the lungs have an HU value in the original CT, the intensity larger than 210 in the processed matrix is replaced with 170 to match the background. Finally, the data matrix is cropped according to the lung mask with 10 voxels extended in each dimension, as in 2.1(h).

2.2 Deep Learning Basics

Machine learning was defined as the study of computer algorithms that can make improvements automatically through experience [45]. Machine learning algorithms build models based on sample data (referred to as "training data") to make predictions or decisions, without the need for explicit programming. There are three broad categories in machine learning, divided by the nature of "feedback" in the learning system [46]:

Supervised learning: The algorithms are learned from data when their expected labels are available.

Unsupervised learning: No labels are given during learning; the machine learning algorithms are expected to learn data structure from inputs.

Reinforcement learning: The model learns from data by interacting with rewards feedback in a dynamic environment.

A limitation of conventional machine learning techniques is their ineffectiveness in processing raw natural data [47], where careful engineering and domain expertise is required when constructing a machine learning system. Representation learning refers to methods that learn from raw data and automatically discover features needed for next steps. Deep learning methods are multiple levels of representation learning. Each level is a simple but non-linear transformation of representation. Starting from the raw data, a higher, more abstractive feature is obtained with each level. With the composition of enough such transformations, complex functions can be learned during training [47].

The deep learning model training is performed by iteratively minimizing a loss function to adjust the model weights via gradient descent optimization and backpropagation. Key components are described as follows:

Loss function. The loss function measures the disagreement between the desired output (i.e., target) and real output (i.e., prediction). The model training aims to minimize the loss function.

Model Weights. Model weights are adaptable parameters that define the functions mapping the input data to output prediction. Typically, deep learning models can have millions of weights (or more) to construct the overall mapping function from data to output.

Gradient Descent. Gradient descent is an optimization procedure that minimizes the loss function by updating the model weights. The procedure starts off with the initial values of the model weights. Then, the derivative of the loss function is calculated. The weights are updated in the direction of the negative gradient. The process is repeated until the loss conforms to stop criteria.

Backpropagation. The working principle of the backpropagation algorithm is to calculate the gradient of the loss function relative to each weight through the chain rule, calculate the gradient once, and iterate backward from the last layer to avoid redundant calculation of the middle term in the chain rule.

In the following, we show how the deep learning methods are applied in automatic pulmonary nodule detection and image feature extraction.

2.3 Pulmonary Nodule Detection with Deep Learning

Object detection is terminology in computer vision and image processing that refers to algorithms detect instances of specific categories and semantic objects in digital images or videos [48]. With the development of deep learning and the availability of large-scaled annotated data, object detection has been well developed in computer vision. Almost all object detection methods can be divided to two categories: two-stage and one-stage. The two-stage methods generate a bunch of bounding boxes (i.e., proposals) in the first stage, and determine if proposals are the desired targets in the second stage, such as Faster RCNN [49], mask RCNN [50]. The one-stage methods include predicting the bounding boxes and class probabilities simultaneously (e.g., YOLO series [51, 52, 53]) or class probabilities predicted on default boxes without proposal generation (e.g., SSD [54]). Generally, the two-stage methods are more accurate and one-stage methods are faster.

Pulmonary nodule detection is a specific task of object detection. The emergence of large-scale datasets with nodule annotation (e.g., LUNA16 [26], LIDC [24]) accelerates the development of pulmonary nodule detection. [55] compared six classical nodule detection methods on ANODE09,

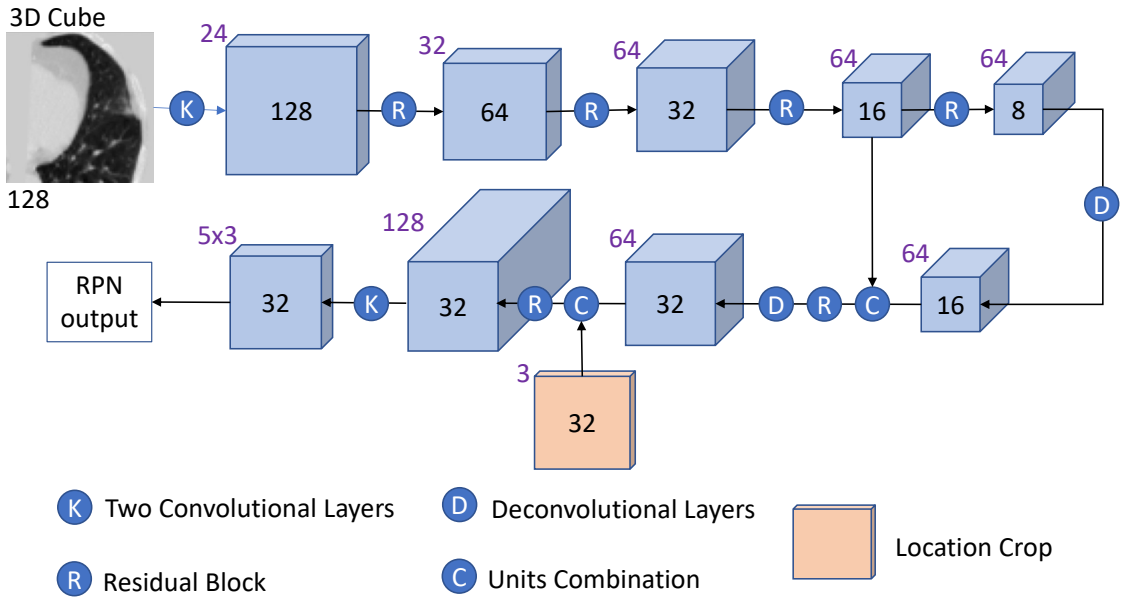


Figure 2.2: The illustration of nodule detection network (i.e., N-Net, follow the figure in Liao et al.). Each cube stands for a 4-D tensor. The gray text represents the number of channels and blank number represents the spatial size, where Length = Height = Width. Each Residual Block is consisted of three residual units. The Units Combination is concatenating feature maps in the channel dimension. The location crop carries the location of proposal, which is introduced in text in detail.

which contains 55 scans from a lung screening program. Some deep learning works [37, 56, 57] have been applied to nodule detection as the much larger datasets (e.g., LUNA16, LIDC) are available. [37] built a 3-D version of RPN [49] with a modified U-Net [58] structure to predict nodule bounding boxes. This detection method is one of the major components of the pipeline that won first place in the Kaggle Challenge (https://www.kaggle.com/c/data-science-bowl-2017). In the section, we introduce nodule detection of [37] in detail.

Generally, object detection networks take the whole image as input. However, the data size of a chest CT (when the resolution is $1 \times 1 \times 1mm^3$) is so large that extends beyond the GPU memory (i.e., TiTAN 12G). One solution to this challenge is to use a patch-based method. Small patches (size $128 \times 128 \times 128$) are cropped from the preprocessed matrix and feed into the network independently.

The illustration of the nodule detection network (termed as N-Net by following [37]) is shown in Figure 2.2. The N-Net consists of a U-Net [58] backbone and an RPN output layer. The U-Net structure enables the network can capture multi-scale features to handle the variable nodule size,

and the output format of RPN enables the network to generate proposals directly [37]. The value 170 will be padded if a patch extends beyond the edge of the whole image.

The location of nodule proposal may affect the confidence of a nodule and judgment of nodule malignancy, as the nodule location is a risk factor in some clinical models [30, 31]. The N-Net encodes the location information in the following way: For each image patch, the location crop (orange cube in Figure 2.2) is calculated as the same size as the output feature map (i.e., $32 \times 32 \times 32$). The three channels in the location crop correspond to the normalized $x-y-z$ coordinates. Given the whole image size (n_x, n_y, n_z), cropping starting index (s_x, s_y, s_z), the intensity (x, y, z) at coordinate index (i_x, i_y, i_z) of the $128 \times 128 \times 128$ patch are calculated with:

$$x = 2 \times \frac{4 * i_x + s_x}{n_x - 1} \tag{2.2}$$

$$y = 2 \times \frac{4 * i_y + s_x y}{n_y - 1} \tag{2.3}$$

$$x = 2 \times \frac{4 * i_x + s_x}{n_x - 1} \tag{2.4}$$

The values lie in the range of $[-1, 1]$ indicates the relative position of the pixel in the original whole image. The i_x is multiplied by 4 because the location crop is downscaled by a ratio of 4 (i.e., 128 to 32). Let (G_x, G_y, G_z, G_r) represent the target information of a nodule, where G_x, G_y, G_z denote $x-y-z$ axes index and G_r is size length. Similarly, (A_x, A_y, A_z, A_r) are the corresponding information of an anchor. The label (nodule or not nodule) of an anchor is defined by Intersection over Union (IoU), that is, anchor has a IoU with target box larger than 0.5 is positive sample or IoU with target box smaller than 0.02 is negative. Other anchors (i.e., whose IoU in $[0.02, 0.5]$) are ignored in the training. The binary cross-entropy (BCE) is used for the classification loss:

$$L_c = p \cdot \log(\hat{p}) + (1 - p) \cdot \log(1 - \hat{p}) \tag{2.5}$$

where p is the nodule label (0 for not a nodule and 1 for nodule), $\hat{p}$ is the predicted probability for an anchor. The regression difference of target nodule and anchor are defined as:

$$d_{ind} = \frac{G_{ind} - A_{ind}}{A_r} \tag{2.6}$$

$$d_r = \log \frac{G_r}{A_r} \tag{2.7}$$

where $ind \in x, y, z$. The regression loss is calculated as:

$$L_r = \sum_{i \in \{x,y,z,r\}} S(d_i, \hat{d}_i) \tag{2.8}$$

where $\hat{d}_i$ is the predicted value, and S is the smoothed L1 loss. As the negative samples do not need a regression loss, the total loss function is

$$L = L_c + \mathbf{1}\{p == 1\} \cdot L_r \tag{2.9}$$

where $1\{state\}$ is a bool function, which returns 1 of state is true, otherwise, returns 0.

There are some other training strategies introduced in [37]. As the number of small sized nodules is much larger than the number of large nodules, Positive Sample Balanced is used to avoid the network biased to small nodules. Consider (1) there are much more negative samples than positive ones and (2) a few negative samples that have similar appearance as nodules are hard to classify, Hard Negative Mining are used to improve training.

2.4 Image Feature Extraction and Cancer Classification

Generally, automatic lung cancer diagnosis using whole CT scan includes two steps: (1) detect all suspicious nodules from a scan and (2) evaluate the malignancy of the whole scan. The nodule detection has been introduced in Chapter 2.3. In this section, we introduce malignancy classification for a whole scan based on the detection results.

Some deep learning works have been proposed to evaluate malignancy given a nodule rather than a whole scan. [59] propose a multi-scale CNN to capture nodule heterogeneity by extracting discriminative features with a new deep learning net structure. [60] introduce a multi-task learning framework with proposed margin ranking loss to model the nodule heterogeneity and address the discrimination capability on ambiguous cases. [39] proposed a Lung Cancer Prediction Convolution Neural Network (LCP-CNN) that shows that clinically validated deep learning techniques are superior to Mayo and Brock models when nodule location from a CT scan is provided. Xu et al. [61] built an RNN upon pre-trained CNN model to predict lung cancer treatment response with

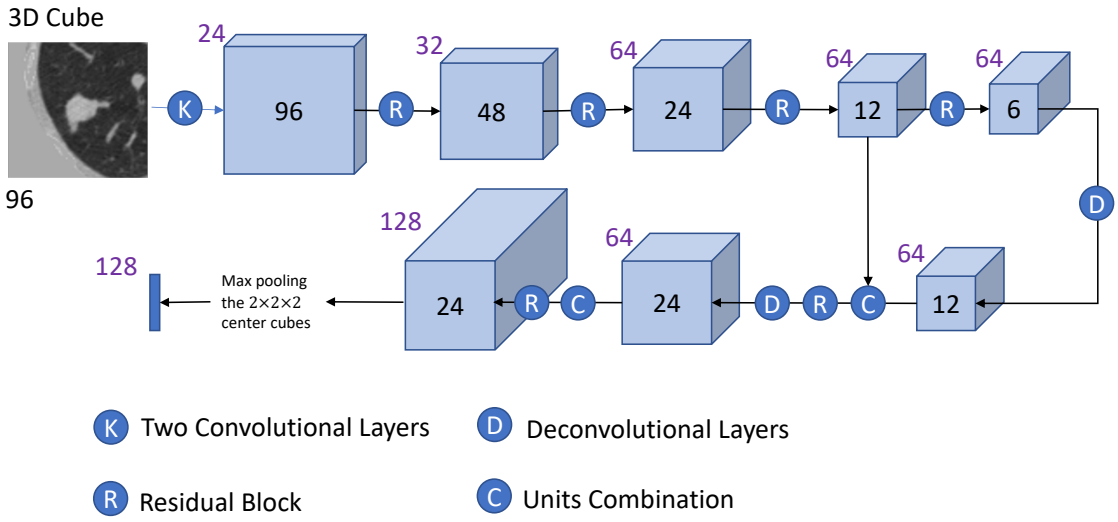


Figure 2.3: Cancer / non-cancer classification framework given a pre-processed image. The N-Net (Figure 2.3) are trained iteratively for detection and feature extraction tasks. A multi-instance learning technique is used to combine the feature or prediction from selected five proposals.

longitudinal medical imaging.

To get nodules from a scan with certainty usually requires human efforts. Recently, works [37, 38] are developed to diagnose with whole scans without human intervention. We use the framework of [37] as an example like the following.

To cover the possible nodules in the detection step, five nodule proposals are selected based on the confidence score. Based on the validation of [37], five proposals with top confidence scores are enough to include most nodules. In the training, as a way of data augmentation, nodules proposals are picked randomly from the detected nodule candidate pool. To avoid overfitting, the N-Net in the nodule detection phase is re-used for feature extraction. The cancer/non-cancer classification framework is shown in Figure 2.3. Instead of feeding with a $128 \times 128 \times 128$ data matrix in detection, a $96 \times 96 \times 96$ data patch is cropped for each selected nodule, and the data patch center is the center of the nodule. The output tensor size of the N-Net in this feature extraction phase is $24 \times 24 \times 24 \times 128$. A 128-d feature vector is obtained for each nodule proposal by max-pooling the $2 \times 2 \times 2$ center cubes of the output tensor. A multi-instance learning algorithm can be applied to integrate the prediction of the top five proposals.

Multi-instance learning (MIL) is a type of supervised learning, where the label is based on the bags of multi-instances, rather than instances independently. In the binary classification setting, a

bag of instances would be labeled as negative if all the instances in that bag are negative. Otherwise, a bag containing at least one positive instance is regarded as positive. The automatic lung cancer diagnosis is an example of MIL, as in Figure 2.2. A patient has cancer if at least one malignant nodule exists, otherwise, this patient is regarded as non-cancer.

In [37], the best MIL solution is based on leaky noisy-or method [62]. The authors assumed that the nodules are independent causes of lung cancer. Also, consider the situation if malignant nodules are missed by the detection phase, the detected benign nodule would attribute the cause of cancer. [37] add a dummy nodule (as P_d) to address this challenge. Each nodule feature is fed to a sub-network to achieve a probability P_i, so the final cancer probability as:

$$P = 1 - (1 - P_d) \prod_i (1 - P_i) \tag{2.10}$$

Then, a cross entropy loss is used on the final probability and scan label. Another widely used MIL technique is called attention-based multiple instance learning (AMIL) [63]. The authors [63] proposed a permutation-invariant aggregation operator corresponding to an attention mechanism [64] based on a deep neural network. Some other ad-hoc solutions, such as feature combination, max-pooling, can also be used for MIL, which has been introduced in [37].

2.5 Sequential Learning

Recurrent Neural Networks (RNN) have been widely used in natural language processing (e.g., [65]) and speech recognition (e.g., [66]) to understand sequence data. The most popular variants of RNN included Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU [67]). The common LSTM unit is composed of a cell and three gates (forget gate, input gate and output gate), which is designed to be capable of learning long term dependencies. Recently, the RNN (especially the LSTM) has been introduced in spatiotemporal tasks (known as convolutional RNN) for precipitation nowcasting [68], pattern recognition [69, 70], image classification [71, 72], medical image analysis [61, 62], et cetera. In addition, the RNN structure helped to bridge multiple modal data, for example, text and image in visual question answering system [73, 74]. Some special topics, like [72] multi-label recognition, can also utilize the CNN-RNN structure.

The rationale of using convolutional RNN is to utilize both spatial and temporal information.

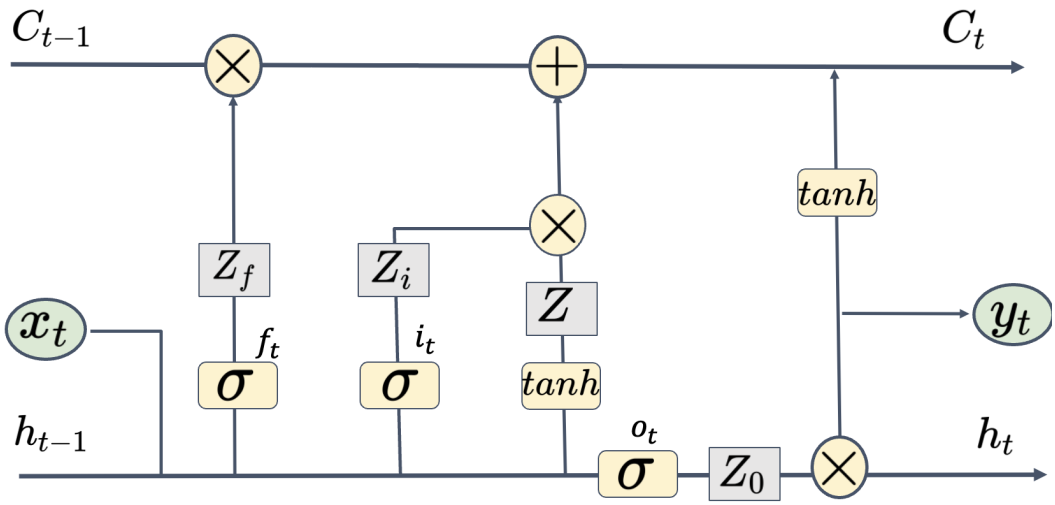


Figure 2.4: The illustration of LSTM. LSTM contains of three gates (i.e., forget gate f_t, input gate i_t and output gate o_t) and two middle states (hidden state h_t and cell state C_t). x_t and h_t is the input and output at timepoint t.

The RNNs are mainly designed with ordered sequence data, and some of works with RNN are designed to explore inner connection within one sample image or spatial-connected images. For example, [75] constructed a sequence for RNN by ranking the risk of different patches of whole slide image. The bidirectional convolutional LSTM was used in hyperspectral image for spectral-spatial feature extraction [76]. In lung cancer diagnosis, patients may have acquired a series of CT images, rather than one single scan. In screening programs, patients who meet the inclusion criteria are recommended to take annual scans [4, 77]. For incidentally discovered nodules, doctors may also recommend a re-visit for patients to see if the nodule is stable.

Recurrent neural networks (RNNs), leading methods to model temporal information, have been widely applied in multiple domains including natural language processing [65], speech recognition [66], computer vision [78], and healthcare [61]. The Long Short-Term Memory (LSTM) [79] and its variants are the most widely used RNNs. The LSTM was proposed to address the incapability of learning long-term dependency of conventional RNN. As in Figure 2.4, A LSTM cell is composed of three gates (forget gate, input gate, and output gate) and two middle states (hidden state and cell state). We introduce the details as follows (W,b are the learnable coefficients and bias of each component), where some descriptions are motivated by [80].

Recurrent networks, including LSTM, store previous memory (e.g., hidden state h_t, the bottom

line in Figure 2.4) and make predictions based on memory and current observation. Different from conventional recurrent networks, LSTM introduce a new middle state called cell state C_t (the upper horizontal line in Figure 2.4). Unlike the hidden state which includes the non-linear transformation of sigmoid function σ, the cell state runs through the timeline with only minor linear interactions. The forget gate f_t decides what information we are going the throw from the cell state. The forget gate depends on the previous hidden state $h_(t-1)$ and current input x_t, which ranges (0, 1) and is achieved by a sigmoid function σ. The forget gate is computed by

$$f_t = \sigma(W_f \cdot [h_(t-1), x_t] + b_f) \tag{2.11}$$

Then, the LSTM model controls what information is added to the cell state. First, the input gate i_t is obtained with the previous hidden state $h_(t-1)$, current input x_t, and related learned parameters. Second, the new candidate cell state $\tilde{C}_t$ is achieved in a similar way. Those can be expressed as

$$i_t = \sigma(W_i \cdot [h_(t-1), x_t] + b_i) \tag{2.12}$$

$$\tilde{C}_t = \tanh(W_c \cdot [h_(t-1), x_t] + b_c) \tag{2.13}$$

The new cell state is updated by the forget gate f_t, previous cell state C_t, input gate i_t, and new candidate cell state $\tilde{C}_t$ as follows:

$$C_t = f_t * C_(t-1) + i_t * \tilde{C}_t \tag{2.14}$$

Afterward, the output at this t point (i.e., h_t) is based on an output gate o_t and cell state C_t. Similar to forget gate and input gate, the output gate is obtained with a sigmoid function feeding with previous hidden state $h_(t-1)$ and current input x_t along with corresponding parameters. The output h_t is achieved by multiplying the output gate and the tanh transformation (ranges from -1 to 1) of cell state C_t. The equations are as follow:

$$o_t = \sigma(W_o \cdot [h_(t-1), x_t] + b_o) \tag{2.15}$$

$$h_t = o_t * tanh(C_t) \tag{2.16}$$

Generally, the LSTM cell executes several times to get a final output. There are lots of LSTM variants proposed to address specific targets. Peephole LSTM added a connection to let the gates look at the cell state [81]. Gated Recurrent Unit (GRU) simplifies the LSTM by integrating the forget and input gates as one [67]. Phased LSTM [82] included a new gate with three different phases to address event-based sequential data. MxDRNN [83] generalized the ordered "RNN+CNN" design to multiple dummy orders by permutating multiple images. Distanced LSTM [84, 85] introduced a global-time-based temporal emphasis model to forget and input gates to address lung cancer detection with sequential CTs.

2.6 Missing Data

Missing data is common in healthcare, which can be caused by many reasons including but not limited to, acquisition error, human mistakes, questions refused by patients. In general diagnosis of lung cancer, patients may have different levels of diagnosis data (e.g., demographics $->$ CT $->$ repeated CT $->$ radiographic report $->$ blood test $->$ biopsy). In practice, patients may skip some levels or stop at a certain level, due to specific conditions and the doctor's recommendation. As the diagnosis data is in different formats and may store at different platforms/sites, missing data may exist if collaborating is not effective enough.

The research of missing data can date back to 1970s [86], and is received new attention with development of machine learning [87, 88, 89]. We describe the problem definition of missing data motivated by [86, 88, 90] as follow.

Considering $x \in R^n$ is a data vector in a complete form and $m \in \{0,1\}^n$ is the binary mask that determines the missing location:

$$x \in p_\theta(x), m \in p_\phi(m|x) \tag{2.17}$$

Let x_0 denotes the observed elements of x, and x_m is the missing parts according to m. θ and ϕ are the parameters of data distribution and mask distribution, respectively. With maximum likelihood estimation, θ and ϕ are estimated by the following marginal likelihood, integrating over

the unknown missing data:

$$p(x_0, x_m) = \int p_\theta(x_o, x_m) p_\phi(m|x_o, x_m) dx_m \tag{2.18}$$

Three different categories of missing data are introduced by [86, 90] and relumed by [88, 91]: (1) missing completely at random (MCAR); (2) missing at random (MAR); (3) missing not at random (MNAR).

MCAR: the probability of data missing is the same and unrelated to the data. Take the example from [91], some of the weighting data are missing just because of bad luck when a weighing scale that ran out of batteries. The missing data mechanism can be written as:

$$p_\phi(m|x_o, x_m) = p_\phi(m) \tag{2.19}$$

MAR: the probability of being missing is the same only within groups defined by the observed data. Similar example from [91], the weighing scale would produce more missing data on a soft surface than on a hard surface. MAR has a broader domain than MCAR, which is more general and realistic. The missing data mechanism can be written as:

$$p_\phi(m|x_o, x_m) = p_\phi(m|x_o) \tag{2.20}$$

MNAR: the probability of being missing varies for reasons but not known, which is mutual exclusion with MCAR and MAR. Same example of the weight scale, the scale might produce more missing values because of the weather, position, or the weight of the object. Those are difficult to recognize and measure.

The distinction of missing data categories is essential to under the adaption range of algorithms, and then make valid inferences. The most simple and executable case is MCAR, while its assumption will introduce bias if it is not the situation in practice. Most works of missing data are under the assumption of MCAR or MAR, where the $p(x_0, m)$ can be factorized into $p_\theta(x_0) p_\phi(m|x_0)$. In clinical practice, the missingness resulting from data entry errors usually belongs to MCAR, while the data transform problem and patient drop can lead to MNAR. The solutions for missing data can be broadly divided into two categories: deletion and imputation. The deletion category refers to

learning with only observed data, and imputation represents replacing missing values with substituted data. Deletion methods include the ad-hoc solutions such as listwise deletion and pairwise deletion that organized by [91], pattern submodel [92], and deep learning-based methods such as M3Net [93]. Missing imputation refers to "making up" the missing values based on observed data and any other mechanisms, which is the most prevalent way to deal with missing data. Tensive imputation approaches have been studies across multiple domains. The mean imputation, comprehensively introduced in [91] with details, is a simple and widely-used method to fill the missing value with the population average of corresponding observed items. The last observation carried forward (LOCF) [91] imputes the missing data with the last observation in a sequential context, which has been used in clinical longitudinal trials. Soft-imputer [94] belongs to the matrix completion imputation category, provides a convex algorithm for minimizing the reconstruction error corresponding to a bound on the nuclear norm. MICE [95] and MissForest [96] are representative discriminative imputation models that do not need to model joint distribution of observed and missing data. With the success of deep learning, multiple modern machine learning techniques, especially two types of generative models: variational autoencoder (VAE) [97] / generative adversarial net (GAN) [98] and their variants, have been applied to missing imputation. [87] proposed the heterogeneous imputation VAEs to handle various types of tabular data (e.g., real-valued vs. categorical). MisGAN [88] was proposed to learn a complete data generator and a mask generator that models the missing data distribution. The partial bi-directional GAN (PBiGAN) [89] is an improved version of MisGAN and is based on an encoder-decoder structure, which has been validated with state-of-the-art imputation performances in various tasks. In brief, missing data imputations have been well studied in multiple domains such as image reconstruction, tabular missing imputation, and irregularly sampled sequences. However, most existing methods have limited imputation within a single modality, which can lead to two challenges in multimodal context: 1) it is hard to integrate data spanning across heterogeneous modalities (e.g., image vs. non-image) into a single-modal imputation framework, 2) recovering discriminative information is unattainable when data are largely missing in target modality (e.g., only the background of an image is observed).

Single Imputation and Multiple Imputation. Single imputation represents that the imputation for the missing data is a single value (i.e., one-to-one), as shown in Figure 2.5(1). multiple imputation creates more than one imputed value for missing data (i.e., one-to-many). Multiple imputation [99]

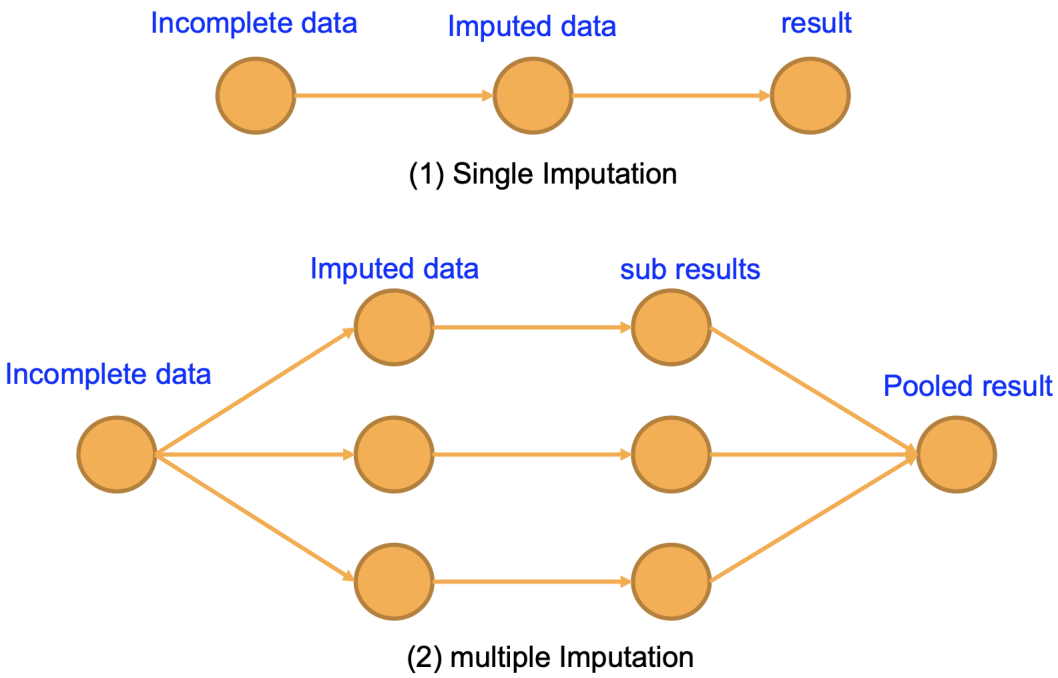


Figure 2.5: Illustration of single imputation and multiple imputation

reduces the problem of too small standard errors. Generally, multiple imputation follows three steps, as in Figure 2.5(2). Multiple times of single imputation. The imputed values are obtained m time follow a distribution. Analysis. Analyze the m imputed values separately, get sub-results. Pooling. Combining the sub-results with specific strategies.

CHAPTER 3

A Quality Assessment Tool for Machine Learning with Clinical CT

3.1 Introduction

Medical imaging plays an important role in medicine. It is widely used in clinical diagnosis and for screening programs such as National Lung Screening Trial (NLST) [4] and Vanderbilt Lung Screening Program (VLSP [77]). Moreover, medical imaging analyses using artificial intelligence has advanced dramatically in recent decades, especially with the development of deep learning [38, 84, 100, 101, 102].

In practice, factors such as network issues, accelerated acquisitions, motion artifacts, and imaging protocol design can impede interpretation of collections of two-dimensional images, especially under the context collaboration of multiple teams [22]. For example, the NLST dataset integrates CT images from over 30 sites across over 20 manufacturer and model combinations, which may trigger data issues when combined. Even within a single institution, such as the Vanderbilt University Institute for Imaging Science Center for Computational Imaging (VUIIS CCI) [23] database may encounter issues with incomplete transfer when the workload is heavy. Moreover, artifacts such as ringing artifacts, motion artifacts and metal artifacts are commonly encountered in clinical CT [103]. Potential issues with data quality can be exacerbated when the data collection pipeline includes human-based workflows while use limited views of the data at each step. In medical image de-identification, minor encoding errors or manipulation of DICOM file data could render these files useless or inaccessible. Data quality issues may obscure digital artifacts that can drive inference toward erroneous decisions. An intuitive example shown in Figure 3.1, when analyzing the lung cancer through CT image, the artifacts on primary pulmonary nodule may cause the misleading decision for doctors and encode biased information for machine learning algorithm.

Image quality assessment (IQA) can be divided into two categories: (1) subjective evaluation (assessed by a human reviewer) and (2) objective assessment (computed by algorithms) including machine learning based methods [104, 105]. The subjective assessment is usually regarded as the gold standard for IQA. Objective assessment of low contrast detectability in CT images was pro-

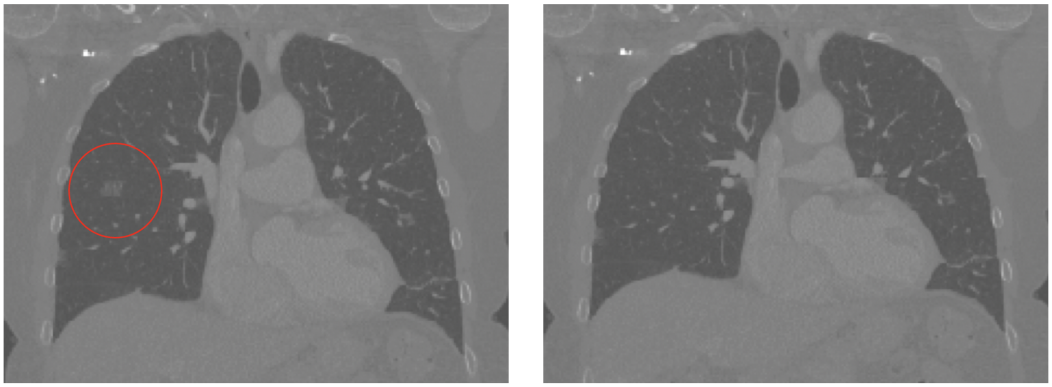


Figure 3.1: Potentially misleading interpretation of a low-quality image. (a) The complete image with a pulmonary nodule highlighted within the red circle. (b) The same image with a few slices missing. The image from (b) can be misinterpreted as lacking a pulmonary nodule to human reviewers and AI algorithms alike.

posed in [106]. Medical imaging quality noises from acquisition of full reference based IQA and no reference IQA are discussed in [106]. Deep learning has also been applied to CT image quality assessment [107]. The subjective assessment is reliable but is time and human-resource intensive. The objective assessment may be more efficient, but accuracy cannot be assured. Moreover, the IQA is especially unstable for cases not seen in the training as a common drawback of machine learning algorithms. In practice, there are multiple reasons result in that the image quality concerns. Some of cases can be fixed by dealing with the image itself and some cannot; this further increases the efforts of subjective assessment and decreases the robustness of objective assessment. A IQA pipeline that integrates the advantages of subjective and objective assessment for effectiveness and robustness is needed.

Previous research has aimed to improve the quality of the medical images, either during acquisition (e.g., [103, 108]) or post-processing (e.g., [109]). [108] used an updated credence cartridge radiomics phantom to detect subtle artifacts. The DTIPrep tool provides a pipeline with several quality control steps with a detailed protocoling and reporting facility for diffusion weighted images [109]. While the quality of images still needs to assess before clinical or research use [110].

As the largest publicly available chest CT dataset, NLST has been widely used in numerous medical imaging research studies. The reporting on IQA with the NLST has been varied: [38] excluded the volumes that failed to download or had unparseable DICOM and removed the volumes

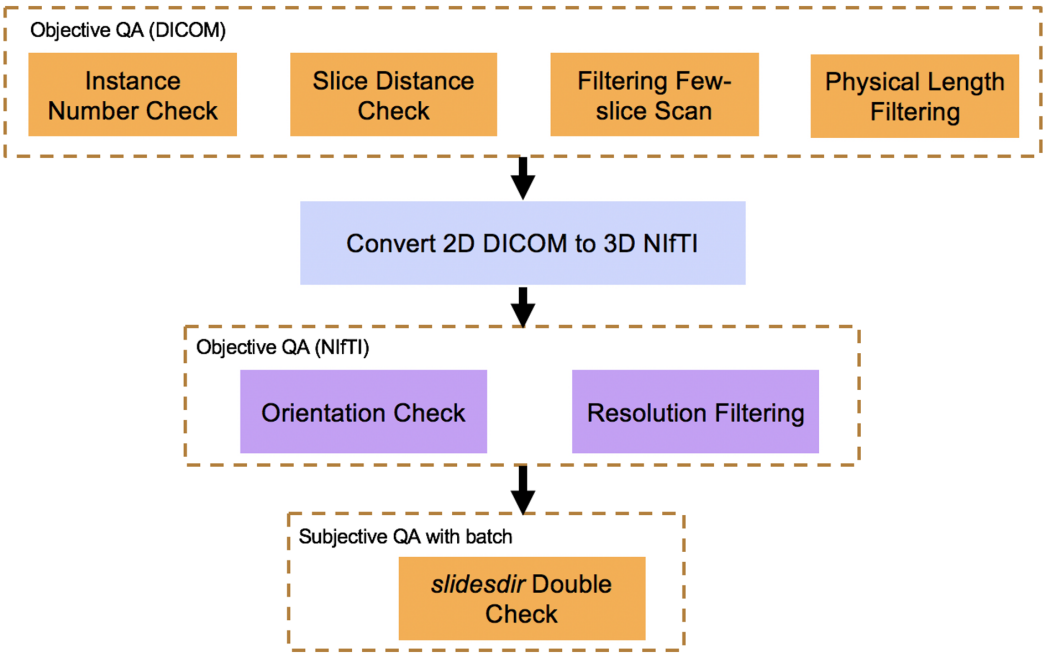


Figure 3.2: The components in our PIQA tool pipeline. Objective assessment on DICOM and NIfTI, subjective assessment with batch is included.

with (1) fewer than 50 slices, (2) slice spacing $\geqslant$ 5.0 mm, and (3) inconsistent pixel spacing. [111] skipped the CT images of corrupted images or incomplete images. However, [38, 111] do not detail how they performed IQA. Researches based on NLST are from multiple tasks including lung cancer risk prediction [112, 113, 114, 115], lung cancer incidence or mortality analysis [116, 117], vertebra segmentation [118], coronary artery calcium scoring [119], but no description of QA can be found in these references. The VUIIS CCI [23] database was developed to provide a new framework for algorithm development allowing large scale batch processing of images and scalable project management. IQA is a vital process for VUIIS CCI since it is unable to exclude the potential for noise or errors during medical imaging acquisition and/or data transfer. As more medical imaging datasets become available, researchers must be cognizant of the potential for faulty imaging studies being present. Furthermore, for the purposes of replicability, a standardized QA method is necessary.

To address IQA for clinical research and our own model development, we developed a series of automated data quality checks for medical imaging data, as shown in Figure 3.2. Our IQA tool integrates the advantages of subjective and objective assessments. The IQA tool can identify the issues from slice missing, low axial resolution, out of Region of Interest (ROI), and provides

a comprehensive report. First, our tool performs objective QA on DICOM images. Then, the DICOMs are converted to NIfTI and another objective QA is performed. The QA report can be automatically generated for each objective check, which is more efficient than human-only review. Finally, the NIfTI files are checked by subjective QA in batches using the *slicesdir* tool for a more focused manual review. Our IQA tool is flexible, allowing users to select specific QA steps as desired.

Additionally, our tool not only provide steps to identify certain errors, but also provide options to correct the image if it is possible with its own. We demonstrate its usage on NLST and two clinical imaging studies on VUIIS CCI. The code and tutorials are publicly available at https://github.com/MASILab/QA_tool.

3.2 Related Public Tools used in our Pipeline

dcm2niix [120]: dcm2niix is an open-source tool that is designed to convert neuroimaging data from the DICOM to NIfTI format.

slicesdir [121]: slicesdir is tool that takes in a list of images, and for each one, runs slicer to produce the same 9 default slices and combine them into a single GIF picture. All the GIF pictures can be viewed through one webpage. slicesdir belong to the family of FSL. FSL is a comprehensive library of analysis tools for medical imaging.

fslreorient2std [121]:this is a tool to reorient an image to match the orientation of the standard template images (MNI152) [122] so that they appear "the same way around" in FSLView.

3.3 Detailed Steps

3.3.1 Instance Number checking

Instance Number (IN) is a number that identifies this image in DICOM files, which was named Image Number in earlier versions. Generally, the INs of DICOM images in a single scan should be a series of consecutive integers. Our priority is to check if there are any missing INs of the CT scan. The rationale behind this step is that slices are missing when the number of slices is fewer than the number of the maximum IN minus the minimum IN plus one. We check the IN using

$$C_1 = max\{in_i\} - min\{in_i\} + 1 - size\{in_i\} \tag{3.1}$$

$$C_2 = \sum_{(} i, j, i < j) \mathbf{1}\{in_i == in_j\} \tag{3.2}$$

where $\{in_i\}$ is the list of INs of the DICOM files, and $size\{in_i\}$ indicates how many DICOMs in the scan. C_1 and C_2 represent how many slices are missed in the scan and how many slice-pairs with the same IN, respectively. The CT scan pass the Instance Number checking if $C_1 = 0$ and $C_2 = 0$.

3.3.2 Slice Distance checking

Slice Location is defined as the relative position of the image plane expressed in mm. This information is relative to an unspecified implementation specific reference point [123]. Slice Distance (SD) is defined as subtraction of Slice Location of the two consecutive DICOM image files. This step is introduced to check the SD between all consecutive DICOM files in their ordered sequence. The SD of all consecutive pairs should be the same in a scan, while in practice, the system error might result in very small and almost negligible difference. Herein, we introduce a self-defined threshold to specify our tolerance for slice distance variation.

The SD checking is

$$C_3 = \sum_i \mathbf{1}\{sd_i < \varepsilon\} \tag{3.3}$$

where $\{sd_i\}$ is the list of SDs of DICOM files, ε is the self-defined tolerance threshold, which should be related to the resolution. C_3 represents how many places which slice distance error. The CT scan pass the Slice Distance Check if $C_3 = 0$.

3.3.3 Filtering scans with few slices

Some scans with an unreasonably limited number of slices may pass the Instance Number and Slice Distance checks. For example, if a chest scan only with three DICOM slices, it is unlikely to be usable. This step filters few-slice scans. Filtering such scans can be done with a user-defined threshold on $size\{in_i\}$.

$$C_4 = \mathbf{1}\{size\{in_i\} < \delta\} \tag{3.4}$$

where δ is the self-defined threshold for filtering scans with a limited number of slices. The CT scan pass the filtering checking if $C_4 = 0$.

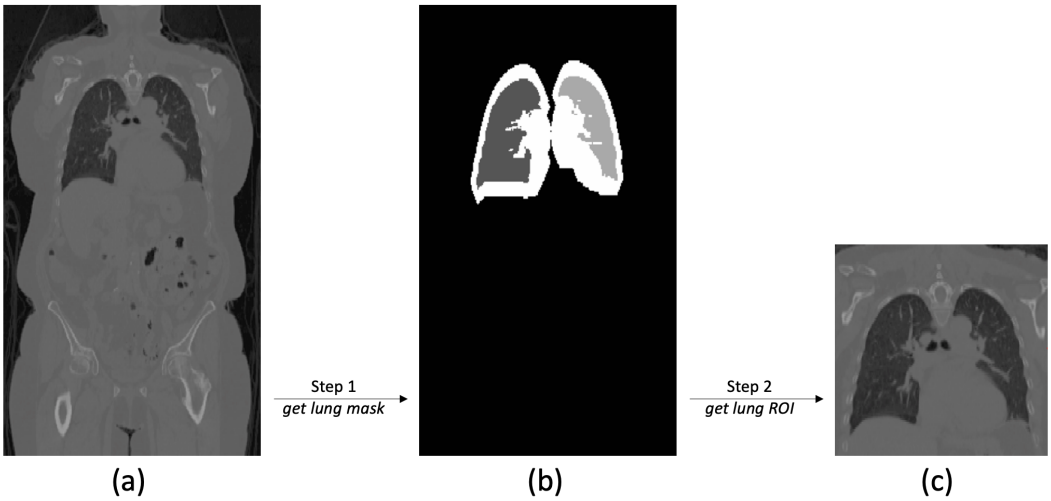


Figure 3.3: The two steps to segment scan that out of ROI. The Step 1 (getting lung mask) is adapted from Liao et al., which is based on thresholding, dilation. The Step 2 is cropping the image based on the lung mask with user-defined margin extension (e.g., 10%).

3.3.4 Physical Length Filtering

Some scans may extend outside of the region of interest (ROI). For example, the target scan is chest CT while is given a whole-body scan may have been provided. The Physical Length Filtering is designed for selecting those CTs with problematic physical body length for further processing or removal. Two thresholds are needed (i.e., low bound and upper bound of physical body length) for filtering

$$C_5 = \mathbf{1}\{\eta_1 < PL < \eta_2\} \tag{3.5}$$

where PL is the physical length computed from Slice Location of DICOMs, $\eta 1$ and $\eta 2$ are the self-defined lower bound and upper bound. The CT scan passes the filtering checking if $C_5 = 1$. We provide two additional steps to segment the ROI around the lung, as shown in Figure 3.3. Step 1. Create the lung mask based on the preprocessing of [37]. Step 2. Segment the lung ROI with lung mask.

3.3.5 NIfTI orientation Check and Resolution filtering

Image orientation is an important issue but may appear confusing [121]. This step is introduced to check and re-orient the image (if necessary).

After converting DICOM files to the NIfTI file format using an open-source tool dcm2niix, we

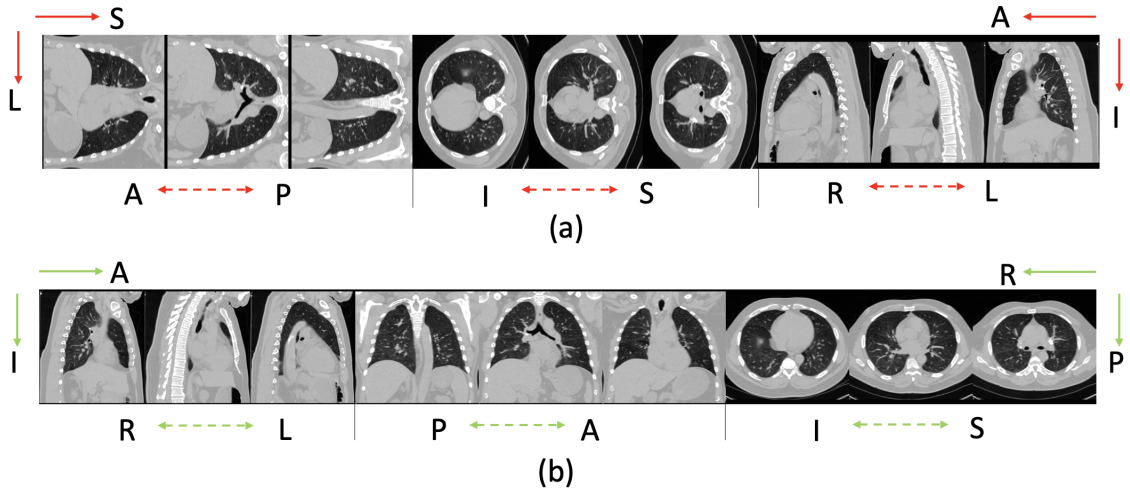


Figure 3.4: The nonstandard orientation case re-oriented by fslreorient2std tool, the scans are displayed by slicesdir. (a) nonstandard orientation case, (b) standard orientation case. The dash lines represent the direction of across slices. The full lines represent the CT direction of one single slice. The red and green lines represent the direction of nonstandard and standard cases, respectively. The Anterior (P), Posterior (P), Inferior (I), Superior (S), Right (R), Left (L) annotations follow the guides of https://itk.org/Wiki/Proposals:Orientation.

check the CT orientation and resolution in the affine matrix A, whose size is 4×4. We have

$$C_6 = \mathbf{1}\{-A_{11} == A_{22} == A_{33} > 0\} \tag{3.6}$$

$$C_7 = \sum_{(} i = 1)^3 \mathbf{1}\{|A_{ii}| > \Phi_i\} \tag{3.7}$$

where ϕ is the thresholds for each of the three dimensions. The CT scan is not with standard orientation if $C_6 = 0$, and we use the open source *fslreorient2std* [121] to convert CT to standard orientation, as Figure 3.4 shows. $C_7 > 0$ suggests that the CT fails to match resolution requirements.

3.3.6 Double check with visualized slices

The subjective assessment is regarded as the gold standard for IQA. In the final step, we apply the subjective assessment for validation using a batched approach, which is much faster than conventional subjective assessment. We use the *slicesdir* to visualize scans. As described in Section 3.2, the *slicesdir* tool can create a webpage for manual QA of a large batch of images.

	Objective QA				Objective QA		Subjective QA
	INC	SDC	FC	AFF	OC	RF	
NLST	0.4%	3.9%	0.2%	0.25%	0.14%	0.99%	0.6%
In-House	8.4%	10.3%	4.5%	4.8%	0.3%	8.5%	0.3%

Table 3.1: The IQA results of different steps

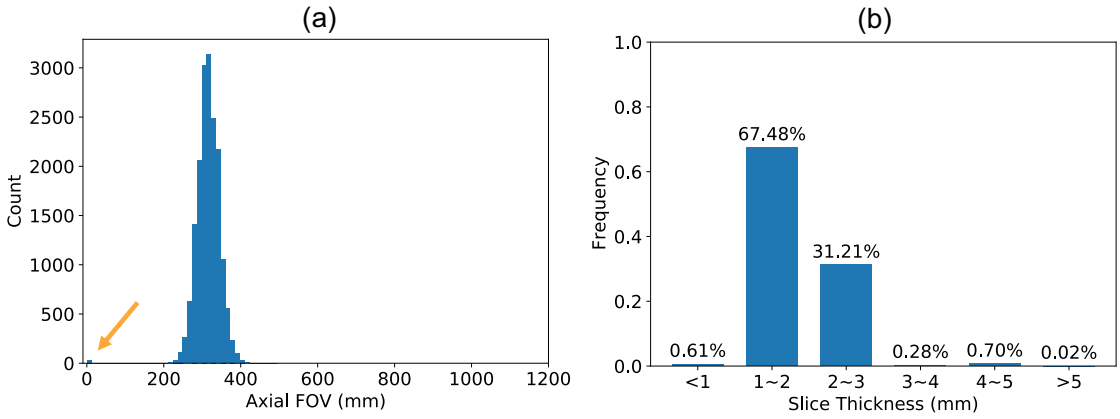


Figure 3.5: (a) the distribution of axial FOV in NLST. Some cases (see arrow) have very small axial FOV because the CT scans only have few slices (e.g., 1 or 2). (b) The distribution of axial resolution in NLST.

3.4 Experiments and Results

3.4.1 A case study with NLST

Data. National Lung Screening Trial (NLST) [4] is a randomized controlled clinical trial of screening tests for lung cancer, which is the largest publicly available chest CT dataset. Approximately 54,000 participants were enrolled in the study, and about 6,300 subjects with follow-up confirmed diagnosis and chest CTs. The goal of the NLST study was to assess if the chest CT screening can reduce the lung cancer mortality relative to chest radiography among high risk people [124].

The NLST dataset was downloaded from the official website (https://cdas.cancer.gov/nlst/) using the provided download instructions. We downloaded the JNLP file from the CDAS of Project NLST-7 and then used the Java Web Start Software to download the DICOM files. These DICOM images were stored in our local file system.

Results. QA results from the subsets of NLST dataset are shown in Table 3.1, with the results included only those CTs for which a diagnosis was ascertained, a total of 17392 scans. Note that (1) some scans may fail in multiple checks, so the total failed cases would be lower than summation of numbers in all checks, (2) some QA-failed cases may only indicate potential warnings and some

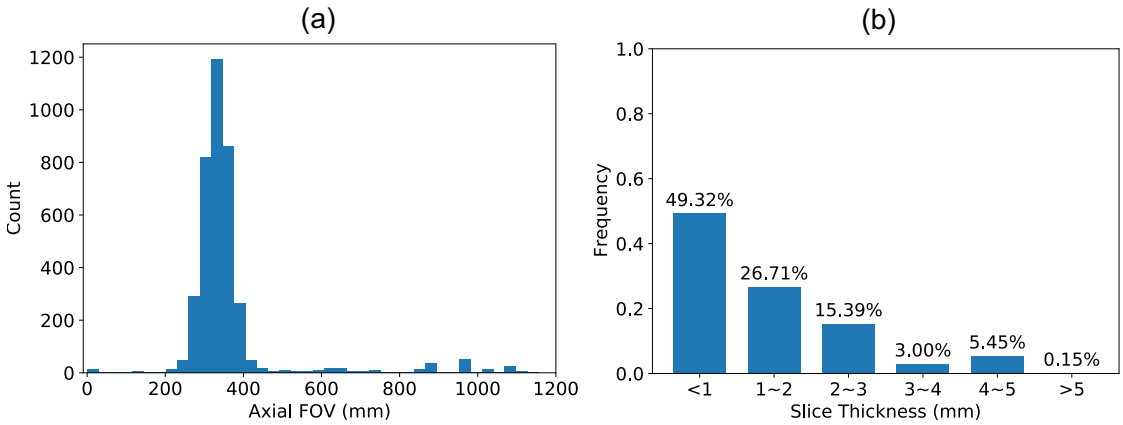


Figure 3.6: The distributions of axial FOV (a) and slice thickness (b) of in-house datasets.

still can be used in machine learning algorithms or clinical readers. For example, some scans with duplicated DICOMs. Figure 3.5 shows the axial length and resolution distributions in NLST. Figure 3.5(a) shows that most of the axial length are located in the range of 250 mm – 400 mm. The distribution of resolutions is primarily in the range of 1 mm – 3 mm, as expected.

3.4.2 A case study with the in-house datasets

Data. We consider two in-house clinical lung CT datasets in the evaluation: (1) The Molecular and Cellular Characterization of Screen-Detected Lesions (MCL [28]) and (2) Vanderbilt Lung Screening Program (VLSP, [27]), in total 3029 subjects with 5274 scans. These datasets are stored in the VUIIS CCI database. To illustrate the need for QA, we present analyses prior to our regular QA processes.

Results. Table 3.1 also shows the QA results from in-house clinical lung CT datasets, in total 5274 scans. As with the NLST dataset, (1) some scans may fail in multiple checks and (2) some QA-failed cases still can be used in AI algorithm or clinical usage which can be validated by subjective assessment. Most failing cases in "Instance Number Check", "Slice Distance Check" and "Filtering Few-slice scan" have been fixed by re-transferring the data through XNAT. We show the original number here to illustrate that several errors can occur in practice.

Figure 3.6 shows the axial length and resolution distributions of in-house datasets. We use the threshold $>$ 200 mm and $>$ 500 mm for axial length as outliers, threshold $>$ 3 mm for resolution as outliers, respectively, when reporting numbers in Table 3.1. Compared with our evaluation on

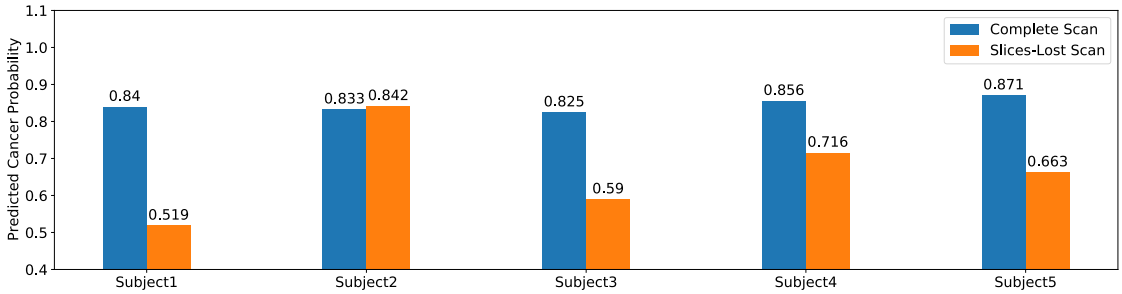


Figure 3.7: The comparison of predicted cancer probability of complete scan and slices-lost scan. The scans are all from cancer patients. The slices-lost is caused by data transfer problems, and re-transfer make the scan complete.

NLST, we find that the in-house datasets are with larger variations. The potential reasons are (1) the NLST dataset is collected with more strict inclusion/exclusion criteria, (2) the in-houses datasets combine multi-site data and cross large range of collecting years ($\leq$ 5 years).

Mini Study on impact with machine learning. As we mentioned, we find data slices can be missed during data transfer. To demonstrate the effect on learning from erroneous scans, we selected 5 chest CT scans of patients with known lung cancer from MCL. We compare the predicted cancer probability between the complete scans (which completed by re-transfer) and scans where some slices are missing. The predicted cancer probabilities are computed by the state-of-the-art method: [37] with their public pre-trained model. As Figure 3.7 shows, the predicted cancer probabilities change dramatically when the scan loses slices. This indicates that images that fail QA may adversely affect the machine learning process.

3.4.3 Analyses of each step in proposed IQA tool

There are multiple steps for checking quality in our tool and each step can give a different aspect of understanding of data. This process can help to identify the reason for failing QA and determine to fix the image or discard the image if cannot be fixed. The Instance Number Check aims to find how many slices are missed by checking the DICOM headers. Slice Distance Check can filter out the scans with duplicated chunks. The acquisition error resulting in few slices for one scan can be detected by Filtering Few-Slice Scan. The extending ROI case can be detected by the Physical Length Filtering, and further steps (e.g., segment the lung) can be processed accordingly. The inconsistent orientations problem may be harmful to machine learning since the data matrices are

not standardly oriented, which can be detected and corrected by the Orientation Check step. As the CT scans can be acquired by different settings, the Resolution Filtering can select the user-defined resolution range. Finally, a quick batch-based subjective QA process is achieved by slicesdir Double Check.

3.5 Conclusion

Image Quality Assessment (IQA) is crucial in machine learning since the model are driven by training data. Even though the subjective review is considered the gold standard for QA, it is difficult with large-scale clinical datasets. Also, missing slices may falsely appear correct on visual inspection which can deceive reviewers in a busy practice. Given the unknownness, heterogeneity, and variability in sources of error, objective and automated assessment is not without limitation. Our tool combines the advantages of subjective and objective assessments to provide an efficient QA method.

NLST is a well-known large-scale dataset with chest CTs. The previous QA tools on NLST either applying few objective criteria to filter obvious errors, or fully subjective QA requires substantial human efforts. In addition, compared with NLST which is quite standardized and large human efforts have spent on the collecting and assessment, in-house clinical datasets have more issues on CT image quality, especially when managing data from multiple sites. Our tool seeks a balance between the efficiency and complexity by several objective checks and subjective check in batches.

Our study has some limitations. First, our tool mainly focuses on the CT images already obtained, so we do not analyze the parameters of scanner. Second, our tool is not based on automatic image learning context, thus, the artifacts such as ring artifacts, motion artifacts and metal artifacts only can be found with subjective QA step. Third, the intent of our tool is for research studies, the tool cannot intervene in the typical clinical workflows and DICOM standards. In summary, we introduce a QA tool for CT images with multiple steps, including objective and subjective assessment, to address the data management gap between raw clinical data and machine learning inputs. We have made our tool publicly available https://github.com/MASILab/QA_tool.

CHAPTER 4

Recurrent Neural Networks for Collaborative Image Classification

4.1 Introduction

Convolutional neural networks (CNN) have been widely applied to extracting features for classification tasks (e.g., natural images, robotics, medical images etc.) and achieved the state-of-the-art performance with leading network infrastructures (e.g., VGGNet [125], ResNet [126], DenseNet [127], SENet [128], etc.) and novel loss functions (e.g., TripletLoss [129], CenterLoss [130], A-Softmax [131], etc.). One of the essential targets of feature extraction is to keep the discriminability for the class label and mitigate class-irrelevant noises. The "ideal" learning outcomes of a classification network should provide identical features for the images from the same class, but this is seldom achievable in practice even for state-of-the-art methods. Intra-class variation reduction could be the most intuitive way to address this problem. For example, the images from a same person can be varied across large range of attributes. The attributes like expression, age and face pose could be complicating factors for face classification task. Therefore, learning discriminative (usually attribute-irrelevant) features for multiple intra-class images (e.g., different photos of the same person) should be beneficial to leverage the classification performance. Customarily, there are two directions to address this target.

One direction is reducing intra-class variation under conventional CNN contexts. Traditionally, the training images were sampled independently from the entire training population. To improve the classification performance, in recent years, the researchers have started to control the learning strategies and add regularization on loss function by intentionally learning from pairs [132], triplets [129], clusters [130] of the training data within a batch. The idea behind such strategies is to take advantage of the intrinsic correlations between training samples by modeling the relationship rather than training them independently. Such methods target intra-class variation reduction at the batch-level.

Another direction is to learn more discriminative features by changing the conventional CNN structure and utilizing the order of multiple inputs. For example, the multi-view CNN [133] took

view-ordered images from the same subject by concatenating multi-path CNN features for 3D shape recognition. Another important contribution of learning from multiple images is the convolutional recurrent neural network (convolutional RNN [68]), which combined the advantages of both CNN and RNN to learn features from sequenced spatial data. Some methods took longitudinal data [61, 68] or encoded the different spatial patches of an image as a sequence with order information [75, 76] feeding to the RNN. In practice, no clear order or the order information cannot be obtained for many tasks.

The feature learning of classification with multiple attribute-ordered images can be interpreted as boosting the class discriminability by utilizing the order and mitigating the noise of attribute. For instance, five photos of a person across different ages as an ordered sequence should be better recognized than randomly sampled one photo from a large age range. Herein, we try to achieve a similar target with unorder images. We propose that different "dummy order" permutations can be introduced to learn attribute-irrelevant discriminative features. For instance, dummy orders "$a->b->c$", "$c->a->b$", "$b->c->a$" can be obtained from {a, b, c}. An intuitive idea to model different orders is to aggregate the information from different paths adaptively in a multi-path network (details in Figure 4.2). Motivated by keeping "memory" of sequence in the text and speech domain, we apply the widely used RNN structure for keeping class-discriminability within intra-class image sequence. In this case, multiple RNN paths can be employed, where each path learns one permutation of multi-image. The model is expected to be robust to the confounding attributes (e.g., age, pose in face images) while keeping the discriminability of class, since only the class label is distinctly included in the loss function (commonly, cross-entropy loss). Recent studies have taken different spatial patches of an image as sequence feeding to the network (e.g., [75, 76]). However, to the best of our knowledge, very limited (if any) previous methods have explored the convolutional RNN co-learning by sequencing independent unordered images.

Herein, we propose the multi-path x-D RNN (MxDRNN) to learn discriminative attribute-irrelevant features from multiple images of the same class. Briefly, we concatenate multiple RNN paths to collaboratively learn features from multiple images. Each path corresponds to a particular "dummy order" of the input images. By concatenating those "dummy orders", the proposed network structure can see multiple images (of the same class) from different "views". Except for belonging to the same class, we do not need any further restrictions (like attribute-ordered) of the

co-learning images. Unordered images are commonly available across different tasks. To verify the generalizability of our method, we conduct experiments on eight datasets of five different image domains.

Lung cancer detection is an example in the medical image, which with actual ordered scans. Among the prevalent lung cancer detection methods, a single scan is usually used for one subject. Better classification performance can be achieved when adopting our xDRNN method to the longitudinal CT data (multiple ordered CTs per subject). Furthermore, by adding extra "dummy orders", the multi-path version (MxDRNN) achieves higher performance.

In summary, the contributions of this work are three folds:

1. The proposed RNN+CNN strategy improves classification performance over leading methods by permutating intra-class unordered images. Results show that our method can learn the feature robust to category-irrelevant attributes (e.g., age, pose).

2. The proposed MxDRNN is a flexible structure, which can be used as (1) an end-to-end learning method by itself, or (2) a post component for existing networks.

3. The proposed MxDRNN is generalizable, which can be applied to (1) 1-D, 2-D, and 3-D learning scenarios, and (2) different domains (e.g., natural image, medical image). In addition, we introduce the multi-channel CNNs for fair comparisons, which take the same inputs of xDRNN and MxDRNN in the experiments.

4.2 Methods

4.2.1 Intuition

In an ideal classification-based feature learning, the feature vectors of two face images should be infinitely close when they are from same class. Unfortunately, this is nearly unachievable, even for the state-of-the-art methods (e.g., lightCNN9 [134], Light-CNN29 [134]). Including a man and a woman as examples, the normalized intra-class variances of each feature vector dimension are shown in Figure 4.1. The largest variation dimension is visualized by computing the average faces of images with high values ("high" face in Figure 4.1) and low values ("low" face in Figure 4.1) in that dimension. In the described ideal situation, the "high" face and "low" face should be nearly the same, and the variance of each dimension should close to zero. However, as an example, the intra-class feature learned by LightCNN9 [134] is not consistent across all the dimensions. The

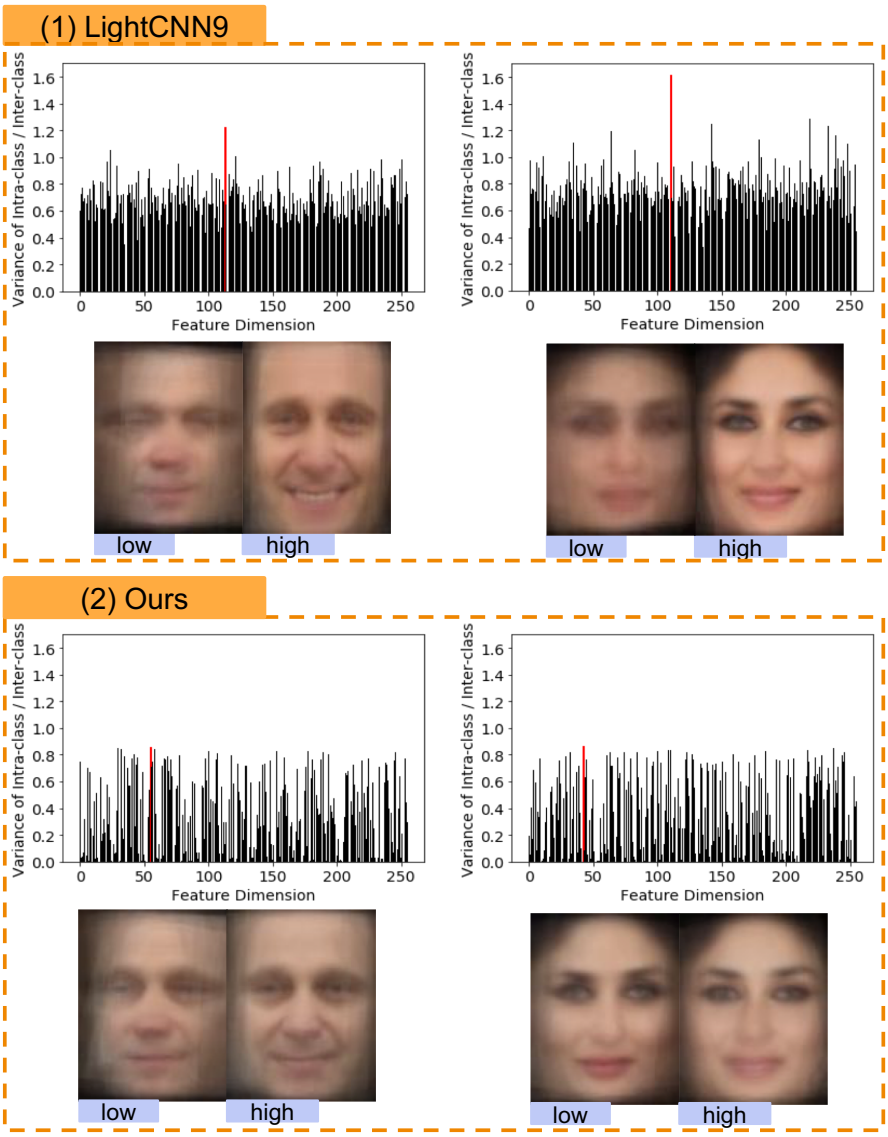


Figure 4.1: Examples of face images. We visualize the variations within about 700 images per person by computing the variance of intra-class over the variance of inter-class for each feature dimension (i.e., variance of intra-class/inter-class versus feature dimension in the plots). We select the maximum variance dimension (max-dimension highlighted in red, which indicates large intra-class variations) and compute the average of faces those with top 60 highest value in max-dimension as "high" face, and top 60 lowest value in max-dimension as "low" face to visualize the clear difference. Box (1) shows the images with the baseline method LightCNN9 and Box (2) show the images combining LightCNN9 with our method.

dimension with the largest variance distinguishes attributes like expression or pose, rather than referring to class-discriminative meanings (as the clear difference between "high" and "low" faces). This is a common limitation, indicating the non-discriminative attributes (e.g., expression, age, pose) have been encoded in the deep features. By contrast, using the proposed method (bottom in Figure 4.1), the general variations for deep features are reduced and the corresponding average faces at the largest variance dimension are more uniform.

"How to learn a feature representation closer to the ideal state and can the achieved feature representation leverage classification performance?" are the main focuses of this chapter.

4.2.2 Encoder for a Single Path

In some applications, we have more than one image per class in both training and test sets while without knowing attributes relation across multi-images. To collaboratively learning from multiple same-class images, we utilize the convolution LSTM framework. We model the unordered data as "dummy ordered" (longitudinal) input to x-D RNN.

Motivated by [68], we generalize the LSTM to x-D (i.e., 1-D, 2-D and 3-D) versions and unordered data in this chapter. Since our proposed method is generalizable for naive RNN and its variations like LSTM, we keep the "RNN" in our proposed algorithm's name. And we mainly experiment with LSTM. Our x-D RNN module can be formulized as:

$$
\begin{aligned}
i_t &= \sigma(W_{xi} \star \chi_k + W_{hi} \star h_(t-1) + W_{ci} \circ C_{(t-1)} + b_i) \\
f_t &= \sigma(W_{xf} \star \chi_k + W_{hf} \star h_{(t-1)} + W_{cf} \circ C_{(t-1)} + b_f) \\
C_t &= f_t \circ C_{(t-1)} + i_t \circ \tanh(W_{xf} \star \chi_k + W_{hf} \star h_{(t-1)} + b_f) \\
o_t &= \sigma(W_{xo} \star \chi_k + W_{ho} \star h_{(t-1)} + W_{co} \circ C_t + b_o) \\
h_t &= o_t \circ \tanh(C_t)
\end{aligned}
\tag{4.1}
$$

where "$\star$" is convolutional (1-D, 2-D, 3-D) operation. $\chi_k \in \{\chi_1, \cdots, \chi_T\}$ but not necessary in order. T is the number of images feeding to x-D RNN module, which also represents the number of co-learning samples each time (e.g., 2 or 3, and we call T "steps" in the following). χ_k is the x-D input data. Briefly, the main differences between xDRNN and convolutional LSTM are that the

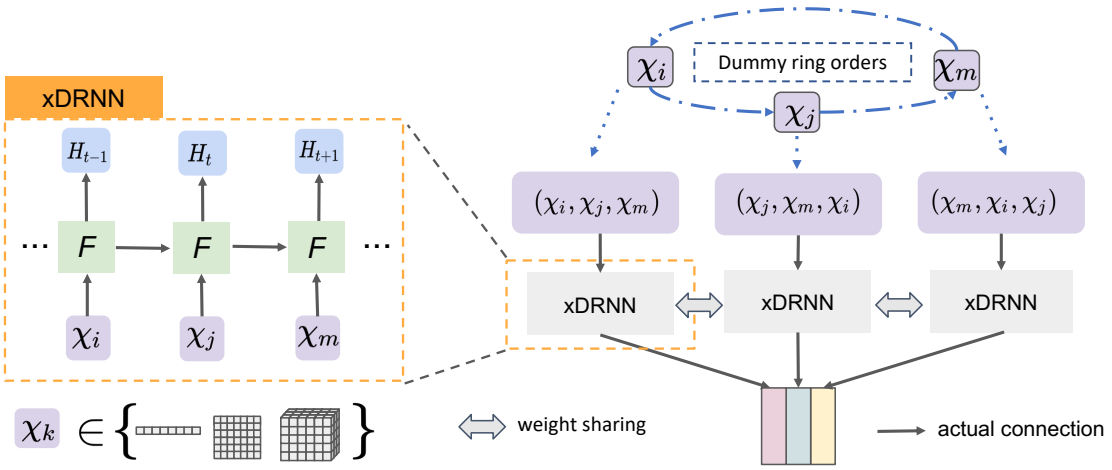


Figure 4.2: The framework of MxDRNN is presented with "three steps" (T = 3) as an example. The left panel shows the x-D RNN module, F represents the recurrent component. Different from the canonical RNN, and the input χ_k can be 1-D, 2-D and 3-D data. $\{\chi_i, \chi_j, \chi_m\}$ indicates the multi-image group input to MxDRNN. The length of $\{\chi_i, \chi_j, \chi_m\}$ equals to both the "steps" and the number of "dummy ring orders". We concatenate the output feature (at the channel dimension) of xDRNNs from all "dummy ring orders" to achieve the final classification. The output of xDRNN is final step of RNN, which is $H_{(t+1)}$ in this figure. Solid arrows indicate "actual connection" in the network, and dotted lines are only used as explanation.

input data χ_k is generalizable to 1-D, 2-D, 3-D and is not necessarily related to the order information.

4.2.3 Multi-path with Dummy Ring Order

The single path of x-D RNN can take advantage of information across images, but does not make full use of it. Multiple images can use different orders, which provide additional information to boosting performance. To balance the model concision and number of orders, we introduce the "dummy ring orders" rather than using all combination of multiple images. And the learning weights of different paths are shared to avoid overfitting. The framework of our method with "dummy ring orders" (DROs) is shown in Figure 4.2. DROs generate T (e.g., 2 or 3) dummy orders that starting with each image, respectively. For the dataset that the order is not externally defined, we randomly initialize the multi-image with a dummy order. For the dataset with an actual order (e.g., longitudinal data in Lung CTs), instead of randomly initializing the order of multi-image, the actual order is included in one of the DROs. Using examples for detail, when dealing 2 steps data, the MxDRNN could be described as

$$O = M(R(\chi_i, \chi_j), R(\chi_j, \chi_i)) \tag{4.2}$$

and if the step T is set to 3, the MxDRNN is:

$$O = M(R(\chi_i, \chi_j, \chi_m), R(\chi_m, \chi_i, \chi_j), R(\chi_j, \chi_m, \chi_i)) \tag{4.3}$$

where $R(\chi_i, \chi_j, \chi_m)$ is the x-D RNN operator (shown in Figure 4.2 as F), and O is the output of MxRNN, and M is the strategy combining multiple paths. Note that we will not change the number of inputs of training and test set when using MxDRNN and will keep the training and test set completely disjoint. For example, if the original data set is $\{\chi_1, \chi_2, \cdots \chi_n\}$. When applying the proposed MxRNN with "3 steps", the inputs of DROs are

$$\{\{\chi_{(n-1)}, \chi_n, \chi_1\}, \{\chi_n, \chi_1, \chi_2\}, \cdots, \{\chi_{(n-2)}, \chi_{(n-1)}, \chi_n\}\} \tag{4.4}$$

We see that the number of original samples equals to the number of inputs to MxDRNN. A single input $\{\chi_{(n-1)}, \chi_n, \chi_1\}$ for MxDRNN represents three paths $\chi_{(n-1)} -> \chi_n -> \chi_1, \chi_1 -> \chi_{(n-1)} -> \chi_n, \chi_n -> \chi_1 -> \chi_{(n-1)}$ that can be computed.

Briefly, multiple images from the same class or same subject are collaboratively learned in one single forward. The multi-path version MxDRNN with different paths of the multi-image further learns the discriminability of class and is robust to class-irrelevant attributes. Indicated by Figure 4.1, the learned feature from our method is less sensitive to variations.

4.3 Experiments and Results

Figure 4.3 illustrates the experiment design. In brief, we evaluate the performance of the proposed method on MNIST [135], 3D MNIST (https://www.kaggle.com/daavoo/3d-mnist), CIFAR10 [136], CIFAR100 [136], VGGFace2 [137] and lung screening computed tomography (CT) imaging (NLST [4] and non-public lung imaging data). For each dataset, we select a leading deep network on that application as a "base network". For our method, both xDRNN and MxDRNN, also termed as (M)xDRNN, are evaluated using the recurrent ideas.

To provide a fair comparison with multiple images consideration (e.g., seeing more than one image of an unknown class at once), we additionally implement the multi-channel versions. Different steps of the base network have been compared in MNIST, 3D-MNIST and CIFAR10. The MultiChannel-ToyNet concatenates multiple images as multiple input channels (MC-ToyNet in re-

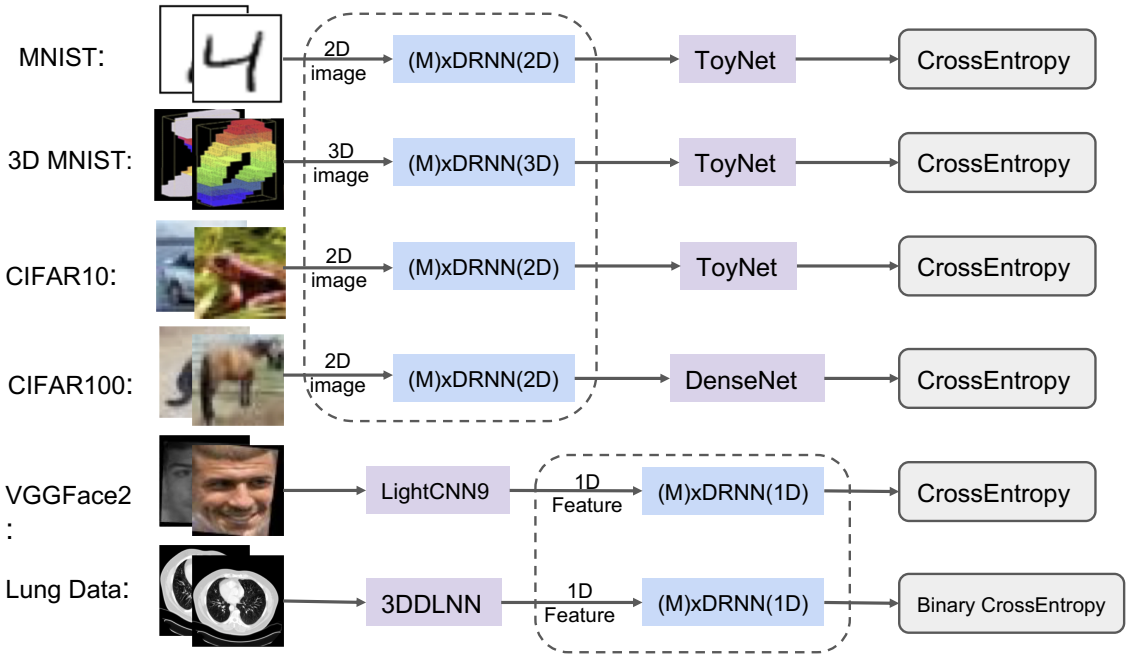


Figure 4.3: The proposed MxDRNN algorithms are presented in dotted line boxes. We test 1-D, 2-D and 3-D versions with different networks and different loss functions. The 3DDLNN net is also named as "Kaggle Top 1" method as the winner of the competition.

Datasets	Initial LR	Decreased Epochs	Decreased Ratio	Max Epoch	Batch Size	Optimizer	Weight Decay
MNIST	0.001	N/A	N/A	200	64	Adam	0
MNIST-3D	0.001	N/A	N/A	200	64	Adam	0
CIFAR10	0.001	N/A	N/A	200	60	Adam	0
CIFAR100	0.1	[150, 210]	0.1	300	60	SGD	5e-4
MNIST	0.0005	[20,30,40]	0.4	100	128	Adam	0
MNIST	0.01	[50,70,80]	0.4	100	128	Adam	0

Table 4.1: Hyper-parameters across different datasets. Initial LR represents initial learning rate. The learning rate would multiply the Decreased ratio at the Decreased Epochs. Our method is a post-network of pre-train model in VGGFace2 and Lung CTs.

sult tables). xDRNN-ToyNet and MxDRNN-ToyNet are xDRNN and MxDRNN, based upon the ToyNet core. "ToyNet" is replaced by "DenseNet" in the experiments of CIFAR100 and is replaced by "CNN" in the experiments of VGGFace2 and lung datasets. Note that the "CNN" is a simple 1-D convolutional layer to fairly compare with the 1-D convolutional "RNN" component in our method.

In the applications of MNIST, 3D MNIST, CIFAR10, CIFAR100 and VGGFace2, we test our algorithm with training/validation/testing splits. For lung datasets, five-fold cross-validation is performed to address the limited number of medical images available for testing. The hyper-parameters on different datasets are illustrated in Table 4.1. Our default Optimizer is Adam [138], but we follow the settings of open-source code in CIFAR100 for fair comparison. The hyper-parameters are varied

Network	2 Steps	3 Steps
ToyNet	99.15 / 0.031 (steps N. A.)	
MC-ToyNet	96.69 / 0.102	99.73 / 0.016
xDRNN-Toynet	99.73 / 0.013	99.87 / 7.87e-3
MxDRNN-Toynet	99.78 / 9.25e-3	99.90 / 1.35e-3

Table 4.2: Test accuracies (%) / test losses on MNIST

among different datasets, which are tuned based on validation set across all compared methods (not bias to our method).

4.3.1 MNIST

MNIST is a dataset of 10 classes of handwritten digits with the size of 32×32. Its training set with 60,000 examples, along with a test set with 10,000 examples. In this study, we split the training/validation/testing size as 54K/6K/10K.

The base network structure from the MNIST example of official PyTorch 0.41 repository (named as "ToyNet") is used for MNIST. It only contains two convolutional layers and one dropout layer, followed by two fully connected layers. Note that the same network structure is used in our experiments with 3D MNIST and CIFAR10. "# steps" in Table 4.2 represents the number of images input feeding to x-D RNN module each time is "#". We also use the "# steps" notion in the following experiments.

As seen in Table 4.2, the classification performance of xDRNN is superior compared with baseline methods, while the MxDRNN is further improved and outperforms the multichannel ToyNet. "3 steps" works better than "2 steps".

An explanatory experiment is performed to visualize the feature spaces of MNIST using the LeNet++. Briefly, we plot the test set of MNIST in Figure 4.4. We visualize the features from the testing set using the trained model at epoch=80. In the CNN method (LeNet++), the features are less discriminative in terms of intra-class variations and inter-class similarities. With our xDRNN, the classification surface is more discriminative, while the multi-path version brings further improvements. Rather than modifying the loss function like CenterFace [130], ArcFace [139], we only use the Cross-Entropy loss in the training.

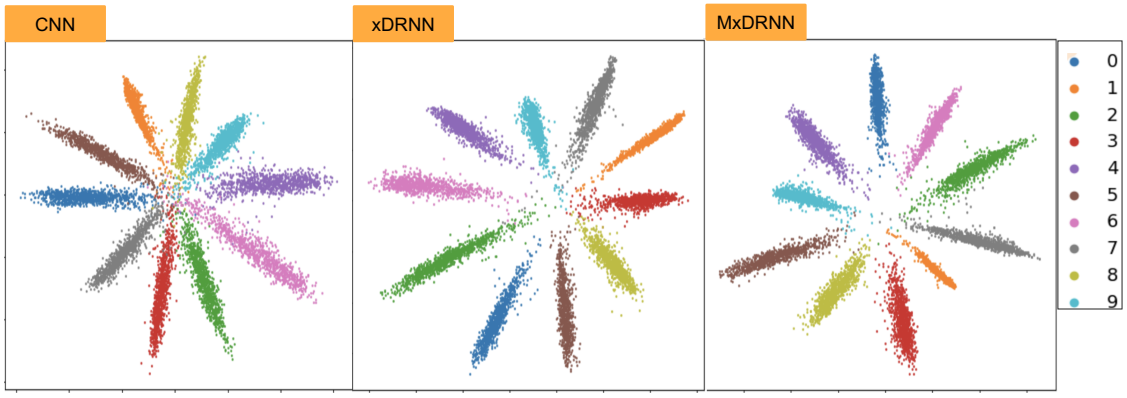


Figure 4.4: Visualization on feature space of MNIST on the test set. The left panel is the feature distribution map from the CNN. The middle panel is the feature distribution map from the proposed x-D RNN component, while the right is panel of the proposed MxDRNN version.

Network	2 Steps	3 Steps
ToyNet	92.44 / 0.271(steps N.A.)	
MC-ToyNet	96.27 / 0.160	97.98 / 0.055
xDRNN-Toynet	97.37 / 0.116	99.60 / 0.0293
MxDRNN-Toynet	97.88 / 0.0769	99.60 / 0.0290

Table 4.3: Test accuracies (%) / test losses on 3D-MNIST

4.3.2 3D-MNIST

3D MNIST is the 3D generalization of partial 2D MNIST from Kaggle with 12,000 16×16×16 volumes. The training/validation/testing splits are 9K/1K/2K. The same ToyNet design is extended from 2D to 3D for 3D MNIST. As seen in Table 4.3, xDRNN and MxDRNN achieve better validation accuracies in both "2 steps" and "3 steps" versions.

4.3.3 CIFAR10

The CIFAR10 dataset consists of 60K natural images of 10 classes with the size of 32×32. We split the training/validation/testing size as 45K/5K/10K. The same ToyNet structure as 2D MNIST experiment is applied to CIFAR 10.

As seen in Table 4.4, the proposed xDRNN and MxDRNN methods improve the performance with a large margin (e.g., accuracies from 60.19% to 78.07% and from 60.19% to 86.00%, respectively). The proposed methods achieve better accuracy compared with the MultiChannel-ToyNet (MC-ToyNet). As with the prior datasets, "3 steps" works better than "2 steps".

Network	2 Steps	3 Steps
ToyNet	60.19 / 1.02 (steps N.A.)	
MC-ToyNet	67.17 / 0.970	73.53 / 0.776
xDRNN-Toynet	75.16 / 0.718	85.06 / 0.439
MxDRNN-Toynet	78.06 / 0.621	86.00 / 0.378

Table 4.4: Test accuracies (%) / test losses on CIFAR10

4.3.4 CIFAR100

CIFAR100 is similar to CIFAR10 but has 100 classes containing 600 images each. There are 500 training images and 100 testing images per class. And the image size is 32×32. The training/validation/testing splits are 45K/5K/10K. To compare with GitHub methods (https://github.com/bearpaw/pytorch-classification) with the exactly same settings, the results from 50k/10K training/validation splits are also provided with the highest validation accuracies.

On CIFAR100 dataset, we compare our results with the state-of-the-art network structures. We take the DenseNet as base-net and include AlexNet [140], VGG19 [125], ResNet [126], DenseNet [127], WRN [138], ResNeXt [141], ShuffleNet [142], NasNet [143], Se-ResNet for comparison. xDRNN-DensNet(100, 12) and MxDRNN-DensNet(100, 12) are the proposed methods upon DensNet(100, 12). In Table 4.5, DenseNet represents DenseNet (100, 12), which means the Depth of network is 100 and Growth Rate is 12.

"3 steps" is applied in CIFAR100. The results are shown in Table 4.5, and the results with an asterisk are picked from the GitHub. The proposed method's accuracy on the test set is higher than all baseline methods. Note our experiments are trained on CIFAR100, so the number of report parameters are different from those reported in the GitHub (https://github.com/bearpaw/pytorch-classification), which were computed based on CIFAR10.

4.3.5 VGGFACE2

VGGFace2 dataset [137] has over 8000 identities in training set and 500 identities in test set. The identities in training and test sets are disjoint. VGGFace2 has large variations in pose, age, illumination, ethnicity and profession. To train our lightweight network ((M)xDRNN + fully connected layer) for face feature extracted by the existing model, we use 2000 identities (about 700,000 images) of the training set. Multi-Channel CNN + fully connected layer is introduced for fair com-

Network	Params (10^6)	Accuracy	Loss
AlexNet	2.50	43.87*	3.10*
VGG19-BN	20.09	71.95*	1.50*
ResNet-110	1.73	71.14*	1.04*
PreResNet-110	1.73	76.35*	1.02*
WRN28-10	36.54	81.86*	0.757*
ResNeXT29, 8×64	34.52	82.66*	0.740*
ResNeXT29, 16×64	68.25	82.70*	0.691*
DenseNet	0.800	77.12*	-
DenseNet(190, 40)	25.82	82.83*	0.751*
ShuffleNet	1.000	70.06**	-
NasNet	5.200	79.34**	-
SE-ResNet152	66.2	77.29**	-
DenseNet+	0.800	74.63	1.18
xDRNN-DenseNet	0.804	85.83	0.507
MxDRNN-DenseNet	0.808	87.76	0.450
xDRNN-DenseNet+	0.804	85.88	0.498
MxDRNN-DenseNet+	0.808	87.70	0.452

Table 4.5: Test accuracies and losses and on CIFAR100(%). The algorithms with "+" are with training/validation/testing splits and test accuracies are reported, and the rest are with training/validation splits on training/test sets of CIFAR100 and maximum validation accuracies are reported. The results with "*" are picked from GitHub (https://github.com/bearpaw/pytorch-classification). The results with "**" are gotten the code GitHub (https://github.com/weiaicunzai/pytorch-cifar100) "DenseNet" represents DenseNet (100, 12) in this table, which indicates the depth of DenseNet backbone is 100 and growth Rate is 12.

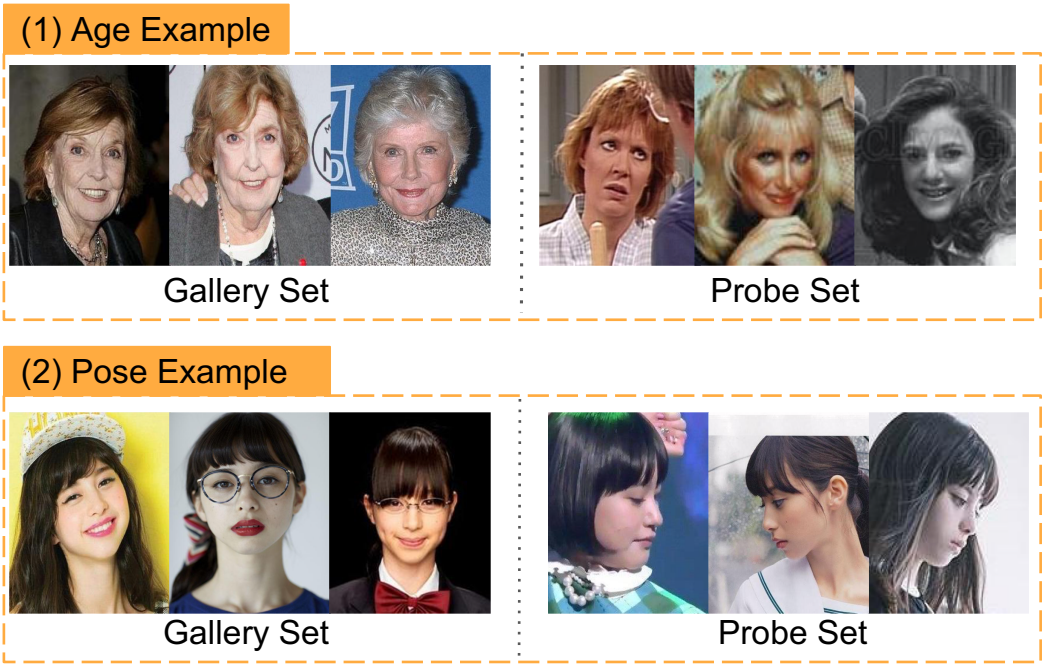


Figure 4.5: Example samples in face recognition tasks. The gallery set and probe set with great variations. The age examples in gallery set are mature, and those in probe set are young. The pose examples in gallery set and probe set are for front view and profile view, respectively.

parison. With these 2000 identities, we split 12 images per person for validation set and the rest for training. Faces are detected by MTCNN [144] and resized to 128 ×128.

There are two test subsets with pose and age variations, with examples shown in Figure 4.5. We have 100 identities with large age variations and 368 identities with large pose variations in the VGGFace2 test dataset. To make the task more challenging, we use the mature images as the gallery set for age set and the frontal images as the gallery set for pose, while for the probe set, we use images with the larger variations (i.e., with the larger age gap and pose angle between the gallery and probe sets).

We adapt (M)xDRNN as a network component upon the pre-trained state-of-the-art networks (LightCNN9). Briefly, the features from each individual image from the pre-trained networks were integrated to the final outputs using our light-weighted network (MxDRNN + fully connected layer, shown in Figure 4.3 as "(M)xDRNN(1D)").

"2 steps" is applied in VGGFace2, and the result is shown in Table 4.6. The proposed xDRNN and MxDRNN methods lead to significant improvements upon baseline methods (e.g., from 46.40%

Network	Age	Pose
Random Guess	1.00	0.30
LightCNN9	48.00	46.40
LightCNN29v2	70.94	71.97
ArcFace	52.30	54.40
SENet50	59.44	68.20
LightCNN9-MC-CNN	56.49	50.72
LightCNN9-xDRNN	71.70	76.13
LightCNN-MxDRNN	72.72	76.54

Table 4.6: Classification accuracies on VGGFace2 test set (%). The LightCNN9, LightCNN29v2 pre-train models are from https://github.com/AlfredXiangWu/LightCNN. Note the LightCNN29v2 model is even higher than the best performance reported in their paper. The ArcFace pre-train model is from https://github.com/ronghuaiyang/arcface-pytorch. The SENet50 pre-train model is from https://github.com/ox-vgg/vgg_face2.

Lung Data Source	NLST	MCL	VLSP
Total Subjects	1794	567	853
Longitudinal Subjects	1794	105	370
Cancer Frequency (%) 2.00	40.35	68.57	2.00
Gender (male, %)	59.59	58.92	54.87

Table 4.7: Demographic distribution in our experiments

to 76.54% with MxDRNN), which also improve upon multi-channel learning with LightCNN9 feature and the state-of-the-art networks (e.g., LightCNN29v2, and SENet50 with VGGFace2).

A qualitative analysis is illustrated in Figure 4.6. Five challenging cases with large pose are all failed with the baseline. Utilizing multi-image with our method, four of them are successfully recognized. The examples indicate our method is robust to the pose attribute, even both the training and testing are without specific pose information. The case with large domain variation compares to gallery set (i.e., image 5) also failed in our method.

4.3.6 Lung CT imaging

CT scans from 1794 subjects are employed from the National Lung Screening Trial (NLST), which is a large-scale lung screening study with CT screening exams public available. 1420 in-house clinically acquired subjects from Molecular Characterization Laboratories (MCL) and Vanderbilt Lung Screening Program (VLSP) are also used in evaluation (Table 4.7), which are used in de-identified form under institutional review board supervision.

Our preprocessing follows [37]. First, we resample the 3-D volume to $1 \times 1 \times 1$ mm isotropic

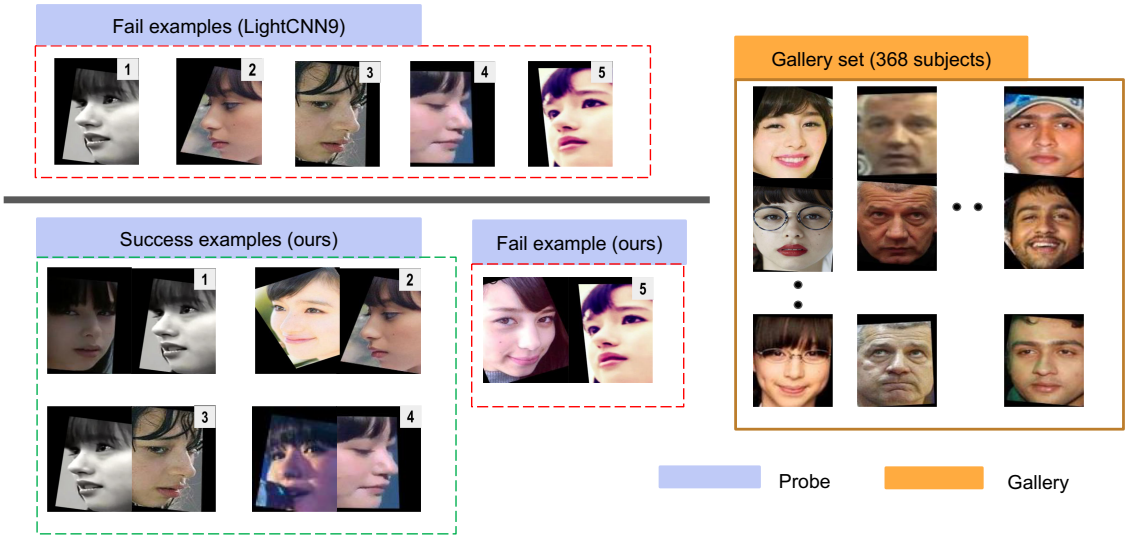


Figure 4.6: Examples of pose set in VGGFace2. The left panel comes from probe set (indicated by blue) and the right panel is the gallery set (indicated by yellow). Five challenging cases from same identity are shown with the baseline and our methods. These five are all failed in LightCNN9. Our method corrects four of them and the image from different domain still fails.

resolution, and, second, the lung is segmented using (https://github.com/lfz/DSB2017) from the original CT volume and the non-lung regions are zero-padded to Hounsfield unit score of 170. We use an existing CNN model to extract the CT image feature. Our algorithm is compatible with end-to-end training, or as a sub-network to process the features from existing models. In this section, we use the (M)xDRNN as a post network to process the features acquired from Liao et al. As shown in Figure 4.3, the (M)xDRNN is a network with 1-D convolutional and then followed by fully connected layer to the loss function. The feature dimension extracted from Liao et al. is 64 for each high-risk region. Five high-risk regions (possible nodules) of each scan are concatenated to 5×64 input as a scan-level feature. "Ori CNN" in Table 4.8 represents the "original" results obtained by the trained model of [37]. "MC-CNN" in this section represents multi-channel CNN (1D), which concatenates features from multi-scans in the "channel" dimension. Note the input of MC-CNN is the same as (M)xDRNN for a fair comparison. Since the number of available CT scans with three time points per subject is limited, we limit consideration to the "2 steps" design in the lung datasets.

Table 4.8 is the five-fold cross-validation results on NLST and our clinical datasets. In each fold, the training/validation/test ratio is around 3: 1: 1. The average with the standard deviation of test

Method	Accuracy	AUC	F1	Recall	Precision
Test results on NLST dataset					
Ori CNN	71.94(2.07)	74.18(2.11)	52.18(2.83)	38.07(2.63)	83.24(4.24)
MC-CNN	73.26(3.10)	77.96(0.98)	59.39(3.70)	47.91(4.87)	78.62(3.09)
xDRNNg	77.60(0.83)	79.55(1.33)	67.17(1.56)	57.88(2.34)	80.73(7.04)
xDRNN	77.05(1.46)	80.84(1.20)	67.85(2.41)	59.92(4.43)	78.68(3.32)
MxDRNNg	77.62(2.79)	80.38(1.42)	69.11(1.61)	62.90(2.59)	77.39(6.99)
MxDRNN	78.16(1.59)	81.62(1.27)	70.33(1.56)	63.46(1.65)	79.16(5.06)
Test results on our in-house datasets (MCL and VLSP)					
Ori CNN	84.80(2.43)	89.00(1.65)	70.29(4.26)	63.46(3.51)	78.83(5.50)
MC-CNN	84.51(1.29)	90.85(1.13)	70.55(1.29)	62.85(1.53)	78.83(5.50)
xDRNNg	85.72(2.31)	90.75(1.17)	73.20(3.57)	67.13(2.99)	80.76(6.58)
xDRNN	86.27(1.29)	92.27(1.15	74.17(2.47)	69.73(2.62)	79.56(5.69)
MxDRNNg	85.99(0.87)	90.35(1.25)	76.51(2.69)	74.97(3.14)	78.38(5.05)
MxDRNN	86.75(1.59)	90.68(1.32)	75.88(2.90)	72.95(3.59)	79.13(3.59)

Table 4.8: Experiments on Lung datasets. xDRNN and MxDRNN are with the backbone of LSTM. xDRNNg and MxDRNNg are with the backbone of GRU

results are reported. The upper part of Table 4.8 shows the lung cancer detection performance on the NLST cohort, where we only use the longitudinal data for training and testing. We evaluate the proposed method on the clinically acquired data (bottom of Table 4.8), where we use both longitudinal data and cross-sectional data in training and validation. The cross-sectional scans are duplicated to 2 steps and use longitudinal scans. The ratio of longitudinal scans and cross-sectional scans are the same in each fold. Different backbones (i.e., LSTM and GRU) in our method are compared, and the results are basically comparable (LSTM is a little bit better overall). The experiments with GRU and LSTM backbones indicate our method can easily transfer to other RNN structures.

4.3.7 Discussion

Our goal is to answer the question in Section 4.2: "How to learn a feature representation closer to the ideal state, and can the achieved feature representation leverage classification performance?" The (M)xDRNN method with "dummy ring orders" (DROs) gives a positive answer. The input of multi-image tuple may have variations at the image-level, while belong to the same class. Motivated by the widely use of RNN in the text and speech domains, which is designed to keep the "memory" of the sequence, we use the RNN path in our work is to keep the "memory" of the class to obtain class-discriminability and tolerant the intra-class variations. Ideally, the extracted feature should be more discriminative for classification. The multi-path strategy seeks more potential reasonable ways

to encode the multi-image especially when no specific order is known, which can be regarded as data augmentation. For a fair comparison with multi-image and validate the effectiveness of RNN-based structure, we also introduced the experiments that concatenate the multi-image at the channel dimension. Based our best understanding and empirical experiments validation, the network firstly collaboratively learns more discriminative feature with multiple images then a single image, since multiple images provide additive information for the same identity. In addition, the strategy of training with multi-image and multi-path increases the robustness of variation of test images.

Beyond the superior performance, we dig into the deeper level to visualize the samples using the proposed algorithm in Figure 4.4. Meanwhile, the normalized variances of the intra-class features are reduced, and variations (like pose, expression) are suppressed (Figure 4.1), which supports that our method is more robust to category-irrelevant attributes.

There are several limitations of the proposed method. First, although the performance of our method is superior to most existing methods, our approach requires multiple images within the same class. Second, the proposed methods introduce more parameters for the training models. Fortunately, the increased number of parameters of MxDRNN is relatively small. Take CIFAR100 as an example, we only increase the parameters from 0.800M (DenseNet (100, 12)) to 0.808M (MxDRNN + DenseNet (100, 12)), and the performance increased from 74.63% to 87.70% (Table 4.5).

4.4 Conclusion

In this chapter, we propose the generalizable MxDRNN method to leverage classification performance using more than one image per identity. It works for 1-D, 2-D and 3-D data across eight different datasets of five different tasks, which indicates the generalization ability of our method. The proposed MxDRNN brings large improvements in both end-to-end training or post-processing of deep features. Additionally, as shown in the face image example, the learned features from our method are robust to category-irrelevant attributes (Figure 4.1) and achieve much higher performance (Table 4.6).

CHAPTER 5

Time-Distanced Gates in Long Short-Term Memory Networks

5.1 Introduction

Longitudinal lung screening CT scans contain temporal relevant diagnostic information for lung cancer, and its effectiveness has been explored (e.g., [38, 61]). As lung screening is becoming more common, longitudinal lung CT scans are also becoming readily available for decision making in clinical practice. The guidelines for lung screening indicate annual imaging for high-risk patients (https://medlineplus.gov/lungcancer.html). However, in general practice, clinical screening is rarely precisely annual since patients may miss visits or may have less frequent scans due to competing factors. Additionally, if clinical concerns arise, more frequent scans may be possible. In our experiments, longitudinal CT scans are best modeled as irregularly sampled, which indicates the time interval between CT scans varies substantially. As the example shown in Figure 5.1, a benign nodule can exhibit substantial variations if the time interval is large, while the malignant nodule may vary little within a short time. Hence, careful consideration of the time interval is necessary to provide the context of the different signal between scans. This confounding factor challenges most of learning models (including canonical sequential models) that do not consider the precise timing of scans.

Recurrent Neural Networks (RNNs) are leading methods to apply deep learning to longitudinal data (e.g., natural language processing [65], speech recognition [66], computer vision [145], and medical imaging [61, 83, 146]). Some works [147, 148, 149] collaborate the RNNs with generative models. The Long Short-Term Memory (LSTM) [79] network is a RNN approach that captures both long-term and short-term dependencies within sequential data by introducing the cell state and three gates (i.e., input gate, forget gate and output gate). The LSTM is widely applied to multiple fields including temporal action recognition [150] and pulmonary nodule detection with 3D CNN [151]. Many variants of LSTM have been proposed. Peephole LSTM [81] adds a "peephole connection" that allows the gate layers have wider receptive field. The Gated Recurrent Unit [67] combined the input gates and forget gates as a single "gate". Phased LSTM [82] included a new gate with

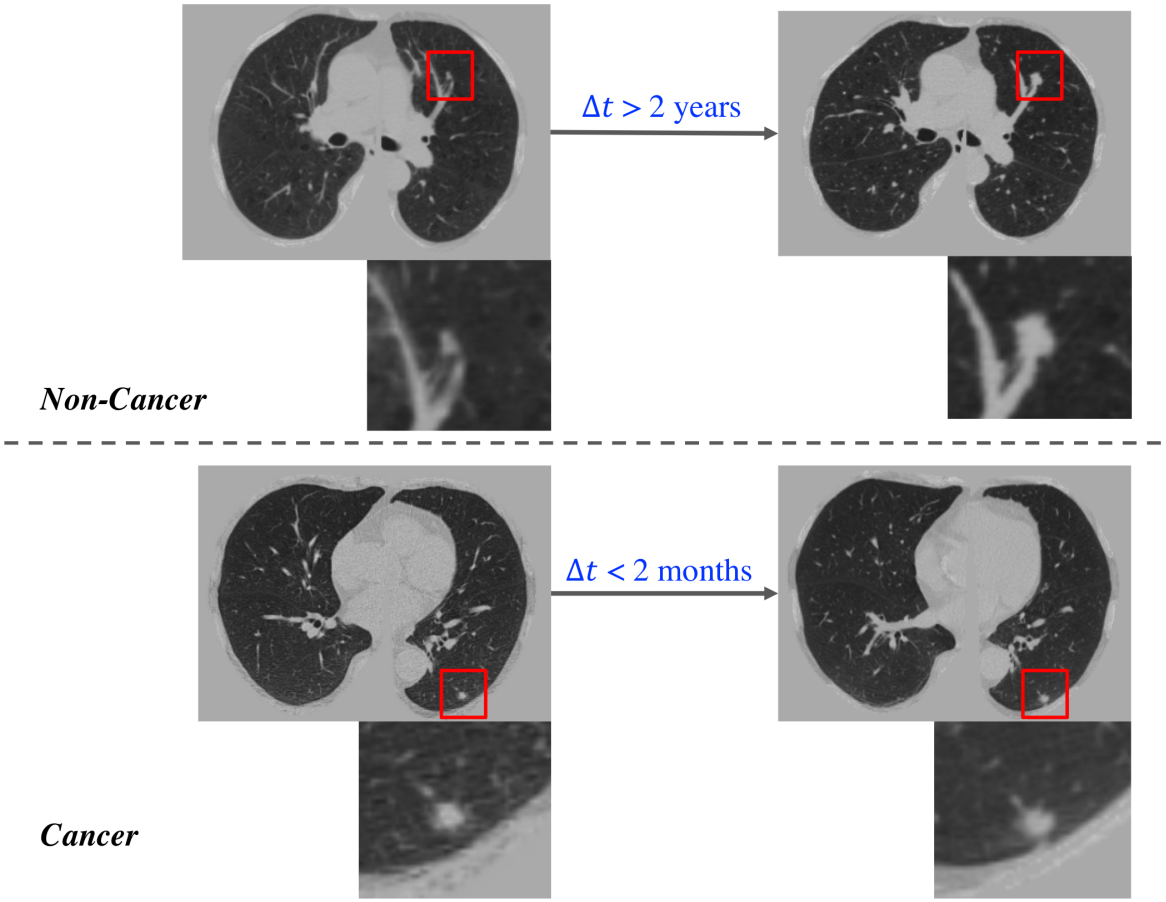


Figure 5.1: Challenging examples for conventional LSTM. One high-risk region per image is enlarged. The upper CT images are from a cancer-free patient, where the clear changes can be seen in nodule over 2 years. The lower CT images come from a cancer patient, where a clear difference is hard to be visualized within a short time interval.

three different phases to address event-based sequential data. Recently, the temporal intervals have been modeled in LSTM (e.g., recommendation system in finance [152] and abnormality detection with 2D chest X-ray [146]). However, no previous studies have been conducted to model global temporal variations and emphases to the best of our knowledge. The previous methods [146, 152] have modeled the relative intervals between consecutive scans. However, these methods did not include the global time information.

In this chapter, the Distanced LSTM (DLSTM) is proposed to perform lung cancer diagnosis using the longitudinal imaging features from lung CT scans. The novelty of this approach arises from a new Temporal Emphasis Model (TEM) to capture the global time distance from the previous time stamp scan to the last time point. The TEM is aggregated with forget gate and input gate to emphasize more recent scans.

Experiments on simulated data and CT datasets are included to evaluate our method. First, the toy dataset is simulated and termed as Tumor-CIFAR, which is generated by adding dummy nodules to CIFAR10 according to [151] that the malignant nodules grow 3 times faster than benign ones. The performance on Tumor-CIFAR highly supports that our DLSTM can capture the time stamp information in sequences. Second, we include three empirical lung screening CT scan datasets (the National Lung Screening Trial (NLST) [4], the Vanderbilt Lung Screening Program (VLSP) (https://www.vumc.org/radiology/lung) and the Molecular Characterization Laboratories (MCL) (https://mcl.nci.nih.gov)) including regular and irregular sampled scans. The MCL and VLSP are combined as our in-house dataset.

This chapter is the extension of the conference version [84]. Specifically, we (1) generalize and evaluate the DLSTM [84] with four temporal emphasis models and (2) include more comprehensive baseline methods and deeper analyses. In summary, the contributions of this manuscript are:

(1) The proposed DLSTM models the global temporal variations and emphasizes newer scan.

(2) The TEM model is proposed to encode temporal information with the forget gate and input gate in LSTM families.

(3) The evaluations of simulated datasets and three empirical datasets (including cross-validation and external-validation) are provided.

5.2 Theory

5.2.1 Task Description and Intuition

Given a set of patients $P = \{p_1, p_2, \cdots, p_n\}$ with longitudinal CT scans, the aim of the network is to predict a label for each patient to indicate whether the patient has cancer or not. For simplification, the following definitions are provided. Each patient p_i has m+1 longitudinal scans with data features $\{X_0, \cdots, X_m\}$ from scan acquisition times $\{T_0, \cdots, T_m\}$. The time intervals between scans $\{l_0, \cdots, l_{m-1}\}$ are computed by $l_t = T_{t+1} - T_t$. The time interval to last scan $\{d_0, \cdots, d_m\}$ is computed by $d_t = max\{T\} - T_t$. In this scenario design, $d_m = 0$.

The motivating idea is to model longitudinal data in the context of LSTM. While long term patterns are often of high importance in natural longitudinal learning (e.g., on natural language, voice, videos), we observe that recent scans may detect an event on onset in medical imaging. Two concerns should be addressed for the longitudinal CT scans for diagnosis (1) newer data usually bring more information for diagnosis and (2) timestamp interval information should be included (as shown in Figure 5.1). Therefore, the time distances of scans are introduced as $\{d_t\}$, which allow emphasis on the more recent data and encode the time interval information, in the proposed method (as shown in Figure 5.2).

5.2.2 Convolutional LSTM

Convolutional LSTM [68] was proposed to integrate LSTM with computer vision tasks [78, 145]. More details can be found in Chapter 2.5. Zhu et al. [150] proposed the Time-LSTM, which added a time-interval based gate to LSTM with better modeling user behavior in a recommender system. Santeramo et al. [153] included an additive term with time interval information in LSTM equations, which was proposed for abnormalities detection of chest X-ray. We employ [150, 153]as the benchmark methods in this study since those are the most representative time modulated algorithms. Note that for clarity in the remainder of this manuscript, LSTM implicitly refers to convolutional LSTM.

5.2.3 Distanced LSTM

LSTM family is the most widely used RNNs in classification/prediction with sequential data. The input gate i_t and forget gate f_t are designed to control the information to be stored and forgotten at step T_t and before step T_t, respectively. In classical LSTM, the time points are treated as uniform

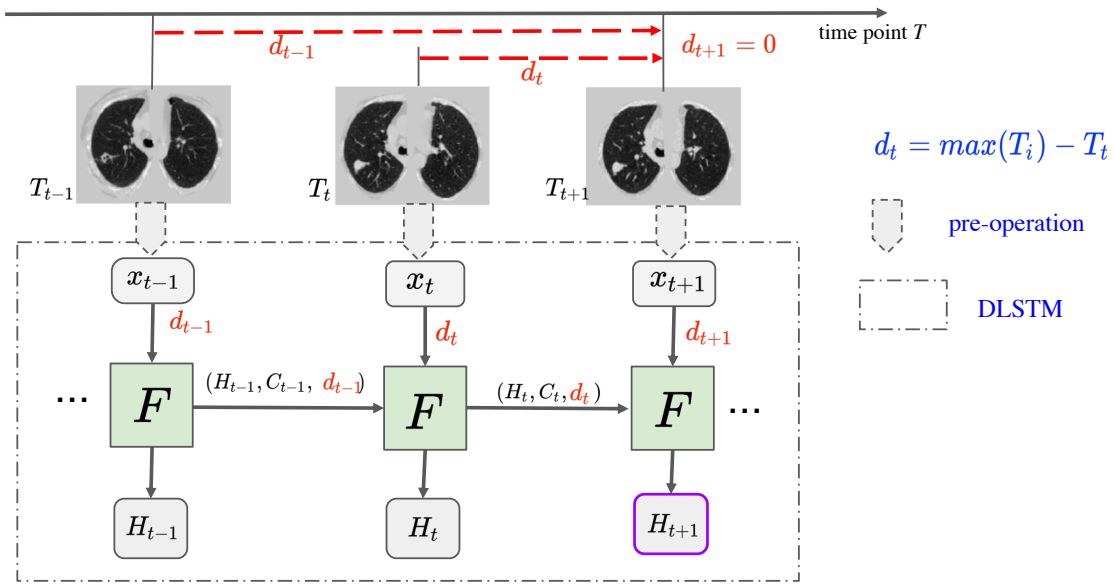


Figure 5.2: The framework of DLSTM (three "steps" in the example). The pre-operation can be image preprocessing or a feature extraction network. x_t is the input data at time point t, and d_t is the time distance from the time point t to the latest time point. "F" represents the learnable DLSTM component (convolutional version in this chapter). H_t and C_t are the hidden state and cell state, respectively. The input data, x_t, could be 1D, 2D, or 3D. The last step's output (e.g., H_{t+1}) is the output of DLSTM.

distribution without modeling the time intervals.

Herein, a focusing term is introduced into the DLSTM method by proposing a Temporal Emphasis Model (TEM). The TEM encodes the time distance d_t with a parameter learnable mathematical function. In this chapter, four different variations of the DLSTM are introduced as:

$$D_1(d_t, a, c) = a \cdot e^{-c \cdot d_t} \tag{5.1}$$

$$D_2(d_t, a, c) = a \cdot max\{1 - c \cdot d_t, \varepsilon\} \tag{5.2}$$

$$D_3(d_t, a, c) = a \cdot max\{1 - c \cdot d_t^2, \varepsilon\} \tag{5.3}$$

$$D_4(d_t, a, c) = a \cdot \log(1 + c * e^{d_t}) \tag{5.4}$$

where a and c are positive learnable parameters. ε is a small positive value without prior knowledge. $D(d_t, a, c)$ represents the TEM in the following.

The proposed TEM model is multiplied with the input gate and forget gate in LSTM. Here, we

follow the format of LSTM to form the DLSTM:

$$
\begin{aligned}
i_t &= D(d_t,a,c)\sigma(W_{xi} * X_t + W_{hi} * H_{t-1} + W_{ci} \circ C_{t-1} + b_i) \\
f_t &= D(d_{t-1},a,c)\sigma(W_{xf} * \chi_k + W_{hf} * h_{t-1} + W_{cf} \circ C_{t-1} + b_f) \\
C_t &= f_t \circ C_{t-1} + i_t \circ \tanh(W_{xf} * \chi_k + W_{hf} * h_{t-1} + b_f) \\
o_t &= \sigma(W_{xo} * \chi_k + W_{ho} * h_{t-1} + W_{co} \circ C_t + b_o) \\
h_t &= o_t \circ \tanh(C_t)
\end{aligned}
\tag{5.5}
$$

Since the input gate handles the current step t, the TEM model encodes current time distance d_t into $D(d_t,a,c)$ to form the time-distanced input gate i_t. The forget gate multiplies the TEM model $D(d_{t-1},a,c)$ at time $t-1$ because the forget gate f_t addresses the "previous" information.

5.3 Method

5.3.1 Simulation: Tumor-CIFAR

First, we examine the asymptotic performance of the temporal learning models as the datasets become large. Here, simulations provide both scalability and certain ground truth.

5.3.1.1 Dataset

The public CIFAR10 dataset [136] contains 60,000 natural images with size of 32×32, across highly heterogenous classes. It is widely used to evaluate methods while requiring minimal efforts on preprocessing and computing (given the small image size). In our simulation, each image in CIFAR10 has been extended to five sequential images with two gradually growing nodules, as shown in Figure 5.3. The nodule size s_i is computed by $s_i = t_i \times g$, where t_i is the time stamp from the beginning point, g is the growth rate, i is the sequential index. The difference between malignant and benign nodules is the growth rate g. We follow the finding of [151] that the growth rate of malignant pulmonary nodules is approximately three times as the benign one. The growth rate g of simulated nodules is

$$
g = \frac{s_i}{t_i} \sim \begin{cases} |N(3,1.8)| & malignant \\ |N(1,0.2)| & benign \end{cases}
\tag{5.6}
$$

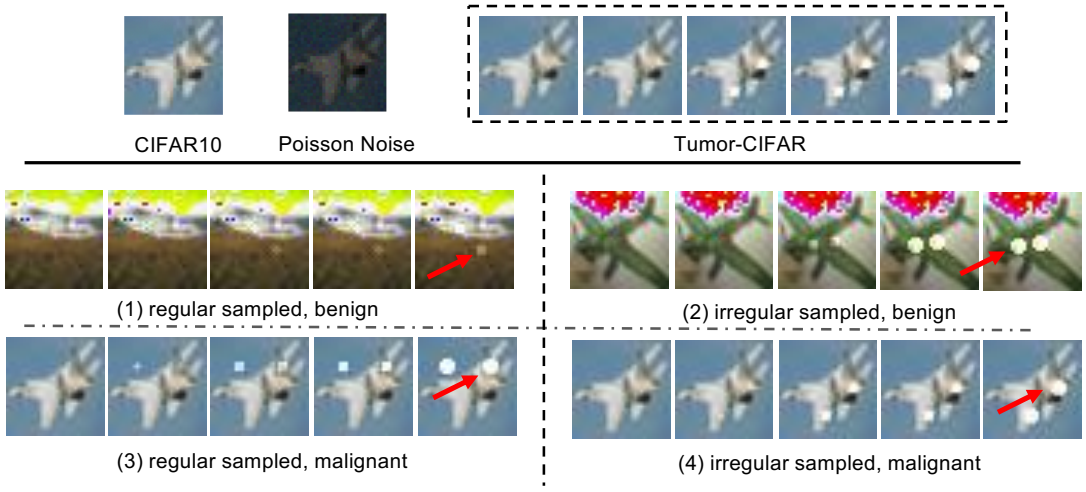


Figure 5.3: Illustration of the Tumor-CIFAR. The upper panel shows the differences between CIFAR10 and Tumor-CIFAR. Each image in CIFAR10 will be transformed into a five-step longitudinal sample by adding growing nodules and Poisson noise (the intensities of noise map in the figure are magnified ten times for better visualization). The bottom panel show more examples in the two version datasets we simulated (e.g., nodules are added to "airplane"). The bottom-left panel is from version 1, which has the same time interval distribution, different nodule sizes between benign and malignant. The bottom-right panel is from version 2, which has the same nodule size distribution, different time intervals between benign and malignant. the dummy nodules are shown as white blobs (some are indicated by red arrows).

where $N(\mu, \sigma^2)$ represents the Gaussian distribution with μ as mean and as variance. Thus, the classification of malignant and benign nodules is transferred to classification of growth rate g, which is computed by nodule size s_i and time information t_i. The simulation code, detail generating descriptions and more image examples are publicly available at https://github.com/MASILab/tumor-cifar. Motivated by [103] that Poisson noise is one of the prevalent noises in CT imaging, Poisson noise (intensities of noise map are linearly normalized to 0-10) is added to the Tumor-CIFAR. Another implementation of adding salt and pepper noise can be found in the public GitHub repository.

We study two applications of the DLSTM model here with two versions of the Tumor-CIFAR.

Regular Sampled (version 1): the image samples have the same "interval distribution" but different nodule "size distribution" between benign and malignant (bottom left panel of Figure 5.3).

Irregular Sampled (version 2): the same nodule "size distribution" but different "interval distribution" between benign and malignant (bottom right panel of Figure 5.3).

The Regular Sampled version is designed to verify if the emphasis on different scans will be

Initial Learning Rate	Decreased Epochs	Decreased Ratio	Max Epoch	Optimizer	Weight Decay
0.01	[50, 70, 80]	0.4	100	Adam	0

Table 5.1: Training parameters in Tumor-CIFAR and CT datasets

helpful for classification when the time interval is under the same distribution for benign and malignant (rough regularly sampled). The Irregular Sampled version, with extremely irregularly sampled (e.g., the time interval can differ more than 2 times across subjects) data, measures if our method can capture the time distance difference between the malignant and benign samples.

5.3.1.2 Experimental Design

There are 50,000 training samples and 10,000 test samples in Tumor-CIFAR. Each sample has 5 sequential images as longitudinal data. The simulated malignant prevalence is 50% in both training and test sets, and the training set is further randomly split into training and validation as 4:1.

The base network structure (CNN in results Figure 5.7), termed as "ToyNet", is borrowed from the official example for MNIST of PyTorch 0.41 [154]. The ToyNet contains two convolutional layers (a 2D dropout after the second) and followed by two fully connected layers along with a 1D dropout layer in the middle. The methods "LSTM" and "DLSTM" in Figure 5.7 represent the 2D convolutional LSTM and 2D convolutional DLSTM stacked in the beginning of the ToyNet, respectively. Training parameters are illustrated in Table 5.1. The initial learning rate is set as 0.01 and is multiplied by 0.4 at 50th, 70th and 80th epochs.

5.3.2 Empirical Chest CTs

Three different evaluation settings are conducted on empirical chest computed tomography (CT) datasets. (1) cross-validation on longitudinal national lung screening trial (NLST) datasets (which are rough regularly sampled), (2) cross-validation on clinical cohort (including cross-sectional and longitudinal, and largely irregularly sampled data), (3) trained on NLST and test on clinical longitudinal data as external-validation.

5.3.2.1 Dataset

We include three lung screening CT datasets in this chapter: National Lung Screening Trial (NLST), Molecular Characterization Laboratories (MCL) and Vanderbilt Lung Screening Program (VLSP).

Lung Data Source	NLST	MCL	VLSP
Total Subject	1794	567	853
Longitudinal Subject	1794	105	370
Cancer Frequency (%)	40.35	68.57	2.31
Gender (male, %)	59.59	58.92	54.87

Table 5.2: Demographic distribution in our experiments

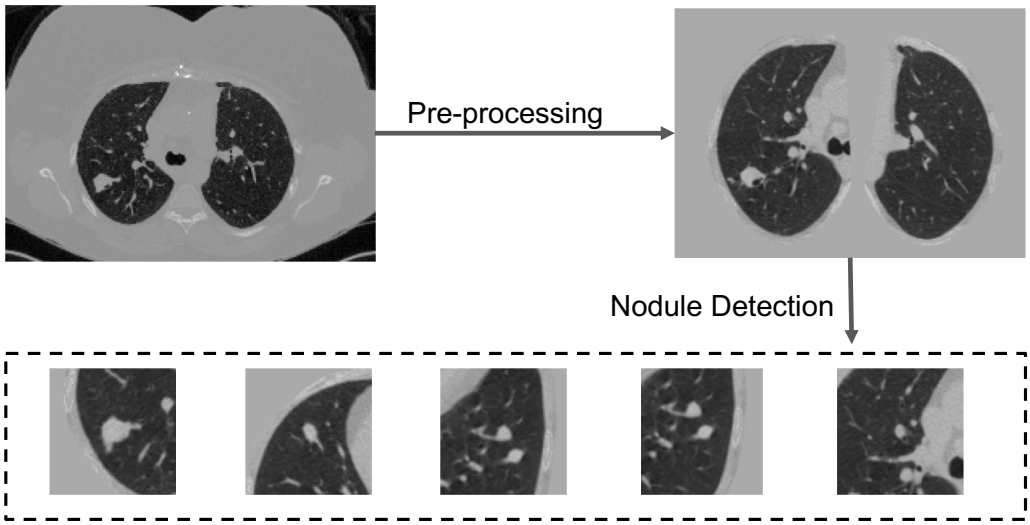


Figure 5.4: Preprocessing and nodule detection. Both steps follow the open-source code of Liao et al. Briefly, the preprocessing segments the lung and get rid of the background in chest CT, and nodule detection detects five highest risk regions. If the number of detected nodules is less than five, patches of all zeros are added to create the five patches.

The demographics of each are shown in Table 5.2. NLST [4] is a large-scale randomized controlled trial for early diagnosis of lung cancer with low-dose CT screening exams. From the machine learning perspective, NLST is unbalanced since cancer patients are less frequent than non-cancer patients. We obtain a subset (1794 longitudinal subjects) from NLST, termed as "NLST" in Table 5.2, which includes all the longitudinal subjects with the label "follow-up confirmed lung cancer" (the ground truth is 1) and a random subset of "follow-up confirmed not lung cancer" longitudinal scans (the ground truth is 0). The in-house datasets VLSP and MCL are combined as the clinical dataset cohort. These data are used in the de-identified form under internal review board supervision.

5.3.2.2 Data Preprocessing and Nodule Detection

In terms of lung CT datasets, we follow the data preprocess and nodule detection of Liao et al. [37]. The CT scans are resampled to $1 \times 1 \times 1 mm^3$ isotropic resolution, and then the scan is segmented by

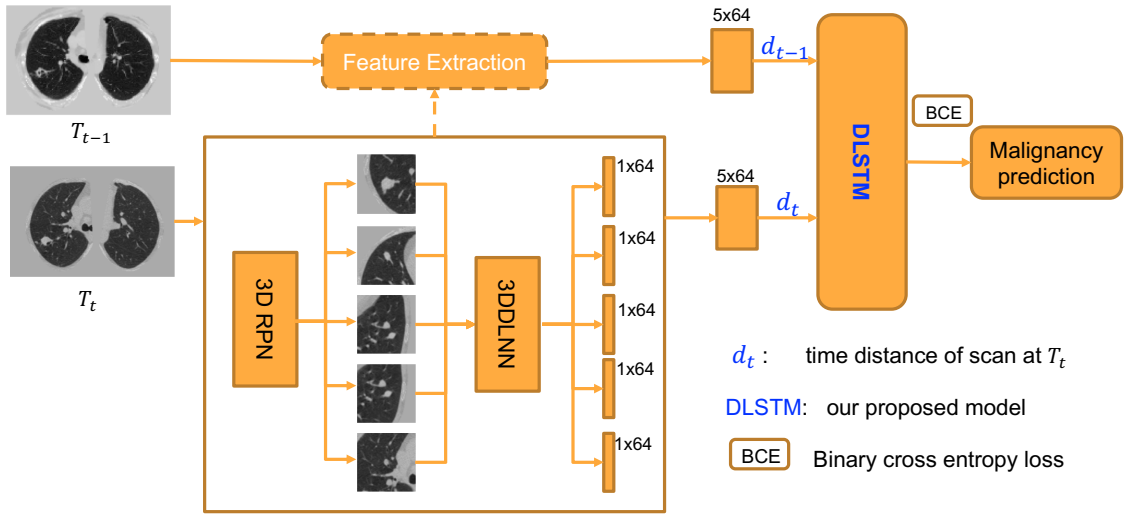


Figure 5.5: The pipeline for chest CTs. The serial CT images are from the same person at T_{t-1} and T_t. The 3D RPN and 3DDLNN are the CNNs borrowed from Liao et al. to extract scan-level feature. The details of DLSTM and time distance definition d_t are illustrated in Figure 5.2

the open source code (https://github.com/lfz/DSB2017). Briefly, the CT images are first converted to Hounsfield Unit (HU) and the image volumes are normalized by a window of [-1200, 600]. The lung masks from [37] are used to remove the context outside the lung. The $128 \times 128 \times 128$ volume patches are put into 3D RPN [49] to locate the pulmonary nodules as [37]. The top five highest confidence regions are selected, as shown in Figure 5.4, for classifying the whole scan.

5.3.2.3 Experimental Design

We follow the image preprocessing and nodule detection pipeline of [37]. Our network can be trained end-to-end or considered as a lightweight post-processing component. In this section, we evaluate the effectiveness of the proposed method as the post-processing network.

The pipeline including the CNNs and DLSTM components for chest CTs is shown in Figure 5.5. The five highest risk regions (possible nodules) for each CT scan are selected by 3D RPN. After feeding into the pre-trained 3DDLNN model of [37], each region is modeled as a 1D feature (1×64 vector). The scan-level feature is achieved by concatenating region features into a 5×64 matrix.

Figure 5.6 presents the experimental design. For a fair comparison, the same features are feed to the Multi-channel CNN (MC-CNN), LSTM, Time-LSTM, tLSTM, and DLSTM networks. MC-

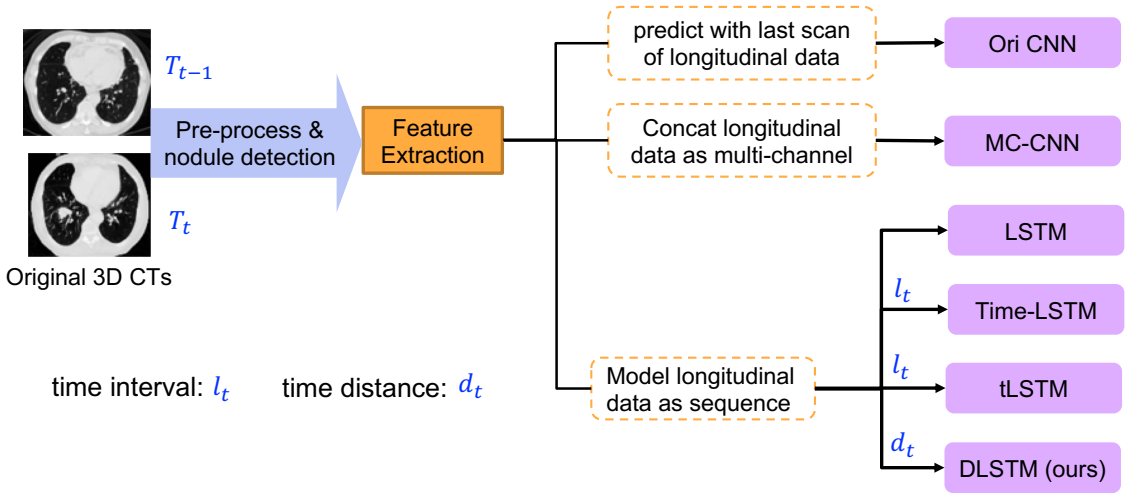


Figure 5.6: The experimental design of CT images. The 3DDLNN is the network structure from Liao et al. Six different methods are compared in our experiments, including two newly time-modeled LSTM algorithms (Time-LSTM and tLSTM). Those two integrate the time interval l_t in the model, while our method introduces the new concept of time distance d_t.

CNN concatenates multi-scan features in the "channel" dimension, which is motivated by the strategy in [38]. The LSTM-based methods model the longitudinal data as the sequence. The training parameters are shown in Table 5.1. Metrics of Accuracy, AUC, F1 score, Recall, Precision are compared. McNemar test has performed combining predicting of five folds. Since CT image data is precious and most lengths of longitudinal steps are less than three, the sequence length in this section is set to two and the last two longitudinal scans are selected.

The "Ori CNN" in Table 5.3, 5.4, 5.4 represents the results obtained by open source code and

Method Accuracy	AUC	F1	Recall	Precision	p-value	
Ori CNN	84.80(2.43)	89.00(1.65)	70.29(4.26)	63.46(3.51)	78.83(5.70)	<0.05
MC-CNN	84.51(1.29)	90.85(1.13)	70.55(1.29)	62.85(1.53)	80.84(4.42)	< 0.05
LSTM	86.27(1.29)	90.27(1.15)	74.17(2.47)	69.73(2.62)	79.56(5.69)	0.08
Time-LSTM	85.79(2.37)	90.81(1.57)	74.57(3.81)	71.08(3.56)	78.71(6.48)	0.42
tLSTM	86.42(1.48)	91.06(1.48)	74.36(1.99)	68.55(1.55)	**81.49(5.28)**	0.40
DLSTM1	**86.97(1.45)**	91.17(1.53)	**76.11(2.68)**	**72.71(2.38)**	80.04(5.18)	*(base)
DLSTM2	86.98(1.20)	**91.41(1.51)**	75.54(1.67)	71.24(5.01)	81.22(6.11)	–
DLSTM3	85.99(1.13)	91.10(1.69)	74.68(2.89)	70.51(6.07)	80.23(5.55)	–
DLSTM4	86.91(1.37)	91.07(1.28)	75.85(1.94)	72.39(3.65)	80.21(6.34)	–

Table 5.3: Experimental results on clinical datasets (%, average (std) of cross-validation). The average and standard deviation (std) of five-fold test results are reported. The best average results are shown in bold. The $p < 0.05$ indicates our method significantly improve the compared method (McNemar test).

Method Accuracy	AUC	F1	Recall	Precision	p-value	
Ori CNN	71.94(2.07)	74.18(2.11)	52.18(2.83)	38.07(2.63)	83.24(4.24)	<0.05
MC-CNN	73.26(3.10)	77.96(0.98)	59.39(3.70)	47.91(4.87)	78.62(3.09)	<0.05
LSTM	77.05(1.46)	80.84(1.20)	67.85(2.41)	59.92(4.43)	78.68(3.32)	<0.05
Time-LSTM	77.91(2.18)	81.41(0.45)	69.01(2.85)	61.16(3.71)	79.60(4.68)	<0.05
tLSTM	77.37(2.97)	80.80(1.45)	67.47(2.46)	58.65(5.12)	79.81(3.34)	<0.05
DLSTM1	**78.96(1.57)**	**82.55(1.31)**	**70.85(1.82)**	**61.61(2.01)**	83.38(4.34)	*(base)
DLSTM2	78.63(1.45)	81.51(1.11)	68.35(2.03)	57.49(3.87)	84.88(4.56)	–
DLSTM3	78.68(1.51)	81.54(0.94)	68.76(1.78)	57.76(3.25)	**85.40(4.06)**	–
DLSTM4	78.05(2.01)	82.09(1.38)	68.90(2.52)	59.84(3.48)	81.44(3.43)	–

Table 5.4: Experimental results on NLST dataset (%, average (std) of cross-validation). The average and standard deviation (std) of five-fold test results are reported. The best average results are shown in bold. The $p < 0.05$ indicates our method significantly improve the compared method (McNemar test).

Method	Accuracy	AUC	F1	Recall	Precision	p-value
Train and Test both on longitudinal subjects						
OriCNN (all scans)	83.42	83.50	52.53	45.77	62.66	<0.05
Ori CNN	87.58	85.10	59.31	55.13	64.18	<0.05
MC-CNN	85.89	76.54	56.21	55.13	57.33	<0.05
LSTM	85.89	83.80	57.32	57.69	56.92	<0.05
Time-LSTM	88.00	87.82	65.87	68.75	63.22	<0.05
tLSTM	86.73	88.69	66.31	79.49	56.88	<0.05
DLSTM1	88.63	89.05	68.24	74.36	63.04	*(base)
DLSTM2	89.47	88.62	69.14	70.00	68.29	–
DLSTM3	89.47	87.39	67.11	63.75	70.83	–
DLSTM4	89.26	88.42	69.46	72.50	66.67	–

Table 5.5: Experimental results on cross-dataset test (%, external-validation). The best results are shown in bold. The $p < 0.05$ indicates our method significantly improve the compared method (McNemar test).

trained model of [37]. The results are reported at patient-level rather than scan-level. The "Ori CNN" reports the performance of the last scan for each patient. Briefly, we have the following three experimental settings:

Cross-validation on NLST longitudinal scans (setting 1). We only include the patients with longitudinal scans in NLST and perform cross-validation as shown in Table 5.4 for 1794 subjects.

Cross-validation on combining cross-sectional and longitudinal scans (setting 2). As shown in the dataset demographic table (Table 5.2), more than half of the patients only have a single CT scan (cross-sectional) from the clinical in-house cohort. We duplicate the cross-sectional scans to the dummy "two steps" longitudinal scans. For the time information involved methods (i.e., Time-LSTM, tLSTM, and DLSTM), we set both the time interval and time distance of dummy longitudinal scans to zero.

External-validation on longitudinal scans across data cohorts (setting 3). To test the generalization ability of our model, we train the model on NLST and test the model on longitudinal clinical data as external validation. The parameter is tuned within the NLST dataset, and then directly applied to the model in longitudinal clinical subjects. Note that the longitudinal data are rough regularly sampled in NLST while the clinical dataset is largely irregular acquired subjects. The NLST dataset is split into five folds (as in setting 1) for training, and the final predicted cancer probability for each subject is the average of five models trained on five folds of NLST when calculating the five metrics.

5.4 Experimental Results

5.4.1 Simulation: Tumor-CIFAR

The results of simulation are shown in Figure 5.7. From Figure 5.7(1), our DLSTM achieves better results (AUC 0.989) than the baseline methods CNN (AUC 0.937) and LSTM (AUC 0.958) on the regularly sampled (version 1) Tumor-CIFAR. Figure 5.7(2) shows the experimental results on Version 2 Tumor-CIFAR. The irregularly sampled (version 2) Tumor-CIFAR is an extremely irregularly sampled dataset, whose nodule size distributions are the same between benign and malignant samples. The DLSTM is extremely predictive (AUC 0.996), while the algorithms without time information (CNN and LSTM) achieve minimal discrimination between malignant and benign samples. Our method significantly improves the LSTM and CNN in both version 1 and version 2 (p

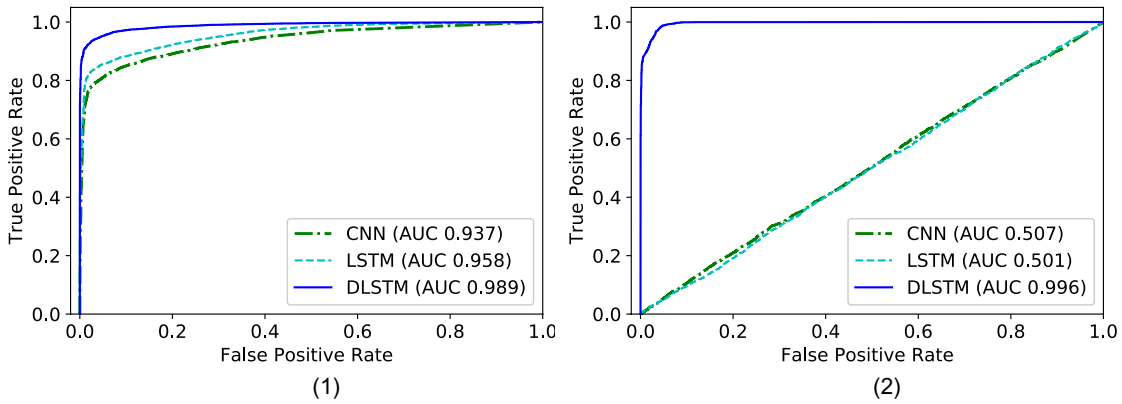


Figure 5.7: The receiver operating characteristic (ROC) curves of the results on Tumor-CIFAR. The right bottom of the figures shows the Area Under the Curve (AUC) values of different methods. (1) version 1: rough regularly sampled data. The CNN and LSTM achieve reasonable performance, and the proposed DLSTM performs better. (2) version 2: extremely irregularly sampled data. The CNN and LSTM achieve minimal learning while the proposed DLSTM achieve high performance. (best view in color).

< 0.05, McNemar test).

5.4.2 Empirical Chest CTs

The experimental results of setting 1 are shown in Table 5.4. Our methods achieve the highest performances on all five evaluation metrics across the compared methods. Table 5.3 illustrates the five-fold cross-validation of 1420 clinical subjects (setting 2), and our DLSTM shows competitive results. Table 5.5 shows the results of external-validation on longitudinal scans (setting 3). The "Ori CNN (all scans)" in Table 5.5 represents the results computed by all scans of longitudinal subjects independently, and the "Ori CNN" only includes the last scan for each subject.

We show the qualitative results (Figure 5.8) in response to the challenge examples in Figure 5.1. The MC-CNN and LSTM, which do not include temporal information, fail in the challenging case. While when time information is included, the algorithms perform better, and our DLSTM achieves superior results.

In both cross-validation and external-validation, our method achieves competitive results across all five metrics including accuracy, AUC, F1 score, recall and precision. In this cross-validation, our method is empirically evaluated to be effective in the longitudinal subjects set from NLST (setting 1) and clinical datasets combining cross-sectional and longitudinal subjects (setting 2). For example,

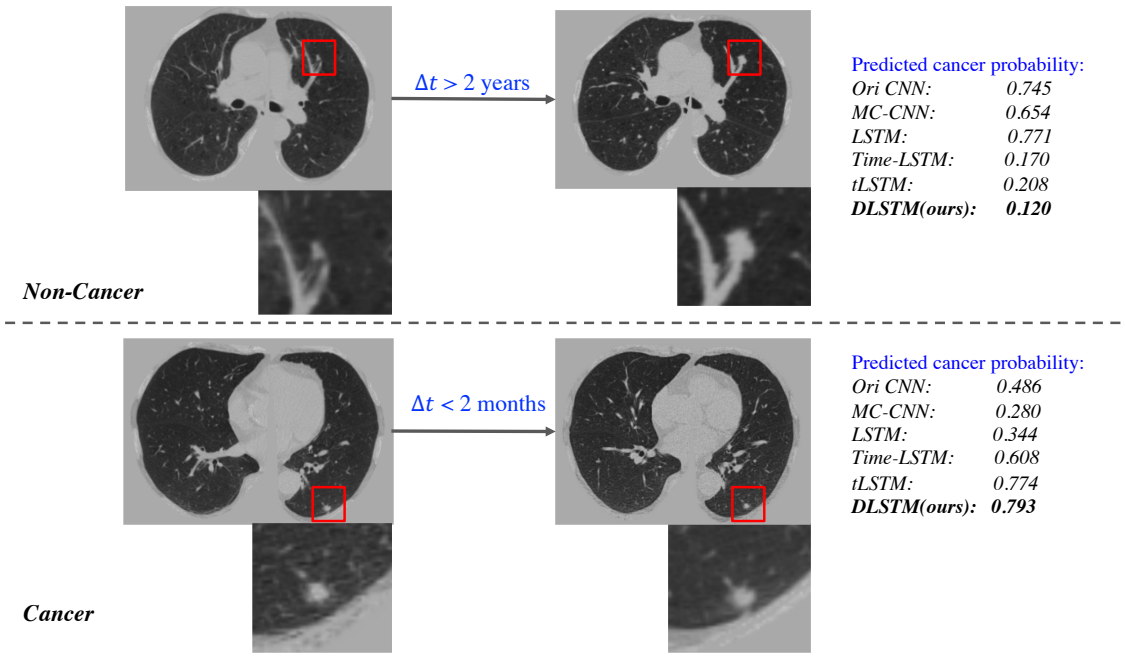


Figure 5.8: Qualitative results related to Figure 5.1. The upper part is from a non-cancer patient, which with large time interval between two scans. The bottom part is from a cancer, and the two scans is close at time distance. The DLSTM is the exponential version.

our proposed DLSTM improves the conventional LSTM on F1 score from 0.6785 to 0.7085 (Table 5.4, NLST dataset), and from 0.7417 to 0.7611 (Table 5.3, clinical cohort).

The experiments of external-validation (setting 3) support that the two concerns should be addressed: (1) the latest scans achieve higher performance compared with all scans (as Table 5.3), which supports that the emphasis on last scan is meaningful. (2) The method without time information could be worse than the "Ori CNN" performance, indicating the time-information exclusion may ruin the model in practice. Additionally, our model addresses the two concerns effectively and outperforms the existing time-information included models comprehensively.

5.5 Discussion

Experiments on Tumor-CIFAR. The proposed DLSTM (AUC 0.984) outperforms the baselines (i.e., CNN (AUC 0.917) and LSTM (AUC 0.953)) in Tumor-CIFAR-v1, which indicates that the DLSTM can work better on regularly sampled longitudinal data. The results of version 2 indicate that the classification is very challenging for time-free models (i.e., not include the time information and only feed the image) if the longitudinal data is extremely irregularly sampled, while the

proposed DLSTM captures the time dependence effectively.

We provide the analytical equation for prediction (malignant vs. benign) using the lesion size and time stamp for simulation. In theory, the analytical equation can be adopted to assess the growth rate of the pulmonary nodule and then the growth rate can be further applied to predict the nodule malignancy. However, in practice, the exact pulmonary nodule size is usually not available. Also, the indicators of lung cancer are complicated which may not only be evidenced by the nodule size or growth rate. More indicators (such as nodule shape, nodule intensity, tissue around the nodule, nodule location) could also play essential roles. These factors can be indicated by the CT images. Thus, it is hard to directly use an analytical equation in practice that with only considers the pulmonary nodule size and time stamps. However, we believe the size and time can be included as important indicators if the data is available.

Experiments on Lung CT Cohorts. The traditional CNN network (e.g., [37]) only takes one scan per patient for the lung cancer diagnosis, ignoring the additional variations encoded in longitudinal scans. The multi-channel CNN strategy (similar to [38]) concentrates the longitudinal scans at the channel dimension, which does not highlight the time and order information. The LSTM utilizes the order of sequence while overlooking the timestamp of scans. The found time-included methods (e.g., Time-LSTM and tLSTM) can model the time intervals between consecutive scans, but neglect the global information that newer scans are typically more informative.

Our DLSTM is motivated by the explanation of forget gate and input gate in LSTM. The temporal emphasis model (TEM) in DLSTM is the decrement function taking time distance, indicating the longer distance scan receives less emphasis. The time distances of longitudinal scans include global information, and the local differences (similar as the time interval in [146, 152]) are encoded by two adjacent time distances. Briefly, our method introduces the explainable time-distanced gates without changing the LSTM structure.

Our proposed method shows significant improvement ($p < 0.05$) over compared methods under the contexts of longitudinal imaging (setting 1 and 3). Under the setting 2 of cross-validation on combining cross-sectional (single scan per subject) and longitudinal scans, the overall improvements can be indicated by the five metrics, while the p-values indicate that improvement from our method on the LSTM-based methods is not significant. The potential reason is the sequential methods (ours and the compared LSTM-based methods) can be biased by the large ratio ($\sim 66\%$,

indicated by Table 5.2) of cross-sectional scans.

Another interesting finding is the comparison of the four different backbone functions in the DLSTM. Overall, those models achieve similar performance, indicating that the DLSTM approach is compatible with families of linear, quadratic, exponential, and log-exponential temporal models. In refined comparison, the quadratic version (DLSTM3) achieves the "least satisfying" comprehensive performances, and it is the only concave function among compared backbones. In practice, we would recommend using the convex function as the backbone since the DLSTM1 achieves the most robust performances across different settings and metrics.

Summary. We propose the novel Distanced LSTM (DLSTM) along with time-distanced gates to model the global temporal intervals between longitudinal CT scans for lung cancer diagnosis. The experiments on the simulated datasets (Tumor-CIFARs) and empirical CT datasets with five metrics (including 1794 NLST and 1420 in-house subjects) demonstrate the effectiveness of DLSTM. Our method is generally superior to baseline methods and the representative existing time-information included methods (i.e., Time-LSTM and tLSTM) under the cross-validation and external-validation settings. The core of DLSTM should be generalizable, indicating that the concept of "time distance" is easy to be extended with other temporal dependence without increasing the model complexity.

CHAPTER 6

Deep Multi-path Network Integrating Incomplete Biomarker and Chest CT Data for Evaluating Lung Cancer Risk

6.1 Introduction

Clinical data elements (CDEs) and biomarkers have also been widely used for cancer risk estimation in research and practice . The National Lung Screening Trial (NLST) selected subjects according to CDEs (e.g., age >=55, pack-years >= 30) [4]. The Mayo team introduced model that identifies malignancy of nodules with clinical data and radiological characteristics of pulmonary nodules [30]. Kammer et al. [35] empirically validated the blood marker hs-CYFRA 21-1 for improving the diagnosis of lung cancer. Several methods (e.g., [155]) integrated the CT reader information (e.g., nodule size) and biomarkers to predict the lung cancer risk. With flourishing of deep learning in computer vision fields, the deep convolutional neural network (CNN) has been widely adopted to extract medical image features. Several subsequent methods extended Liao's framework to serial CT images [83, 85] and multi-task networks [101, 102]. As validated, the CDEs/biomarkers and CT images are helpful for lung cancer detection. Yet, not all the subjects have the complete multi-modality data of CDEs, biomarkers and chest CT, which brings the challenge to learn a deep network with missing data. Note that in this chapter, CDEs and biomarkers are regarded as one modality and CT image as another. For clarity, the CDEs/biomarkers will be termed as biomarkers in the following.

One drawback of majority existed imputation models is that they are based on the same type of data (e.g., impute images based on observed images). In practice, the missing data are usually heterogenous and may include image and non-image types. The ad-hoc solutions of listwise deletion and pairwise deletion can be extended to the field of machine learning. A straightforward operation is to delete the samples with missing component when training a machine learning algorithm, which is the listwise deletion extension. However, the imperfect data may also contribute, e.g., Yang et al. [102] ignored the backpropagation of one task when its related label is missing when training a multi-label network for chest CT.

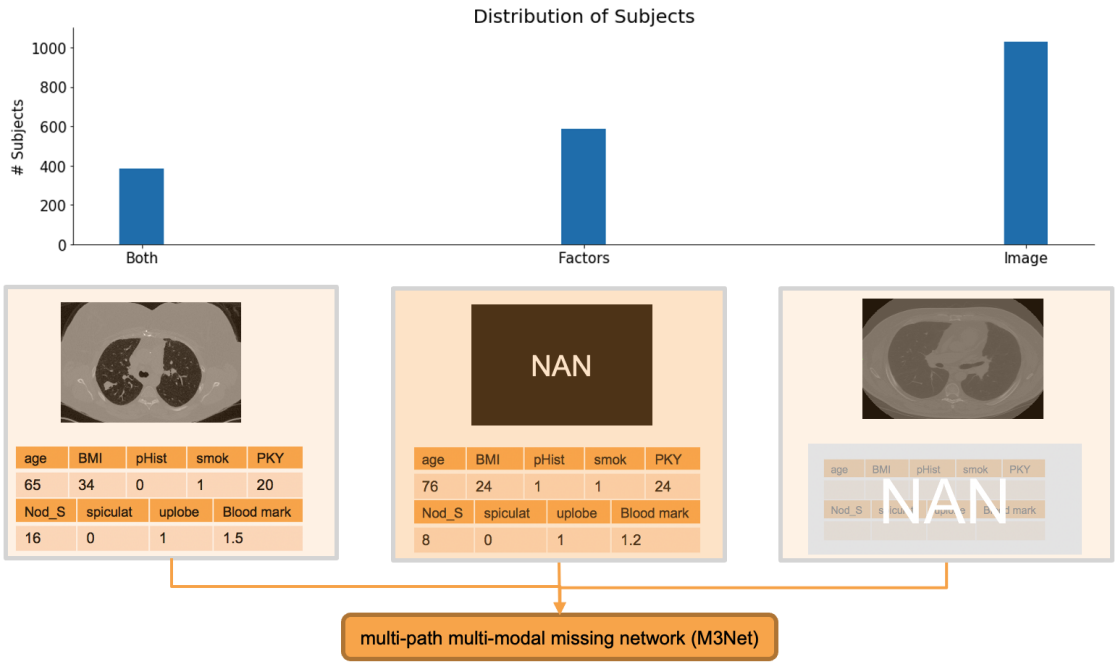


Figure 6.1: The intuition of the proposed M3Net. In practice, not all subjects have both clinical variables, biomarker and CT images. Our network takes the available data, including complete data and data with missing modality, to train a uniform deep network. During the test phase, our model can predict lung cancer risk with incomplete data.

In contrast to previous research (Chapter 2.6) with missing values [87, 156], image inpainting [88], we focus on the missing modality across image and non-image data in this chapter. We propose a new network, termed as multi-path multi-modal missing network (M3Net), to cope with missing data in lung cancer risk estimation, as shown in Figure 6.1. Generally, the number of subjects with one modality (either biomarkers or CT) is relatively high, while the fewer subjects with both complete modalities of biomarker and CT. The M3Net integrates (1) multi-modality data (i.e., CT images and biomarkers) and (2) data with missing modality in an end-to-end training network with multi-path. The data with missing modality can be included in both training and testing sets. Our datasets come from three sites: Vanderbilt University Medical Center (VUMC), University of Colorado Denver (UCD), the Detection of Early Cancer Among Military Personnel (DECAMP) and University of Pittsburgh Medical Center (UPMC). Our network is evaluated by cross-validation with available subset of VUMC + UCD + DECAMP (termed as VDD in the following), and external validated with the UPMC cohort. Our M3Net shows superior performance over the single modality predictions and learning without incomplete data.

Cohorts	VDD	UPMC
Total Subjects	1232	99
With biomarkers	585	99
With images	1030	99
With image and biomarkers	383	99

Table 6.1: The number of subjects in the cohorts

6.2 Methods

6.2.1 Data

The data used in our experiments is subset of the project the Consortium for Molecular and Cellular Characterization of Screen-Detected Lesions (MCL). We have collected the subjects from VUMC, UCD, DECAMP and UPMC under Institutional Review Board (IRB) supervision. The subjects from VUMC, UCD and DECAMP are combined as the cohort VDD. The five-fold cross-validation is performed in the VDD cohort, and UPMC cohort is used as external validation. The data distribution is illustrated in Table 6.1. The UPMC cohort is composed of patients with indeterminate pulmonary nodules, whose nodule size is between 6 and 30 mm.

6.2.2 Multi-path Multi-modal Missing Network

We call the proposed model as multi-path multi-modal missing network (M3Net) as the model has multiple paths and is designed to handle missing modality of biomarkers and image. The framework is shown in Figure 6.2. The network path used for learning from CT image (the image-path) has two main parts. The Convolutional Neural Network (CNN) includes the nodule detection network and the feature extraction network from Liao et al. [37]. The output of the CNN is the five 128-dimension features respect to five nodules in this chapter, which have the top five confidence scores. The AMIL represents Attention-based Multi-Instance Learning adapted from [63], which is a sub-net in our framework that feed by the five nodule-features. The input of the biomarker-path is a 10-dimension vector that contains the nine available biomarkers as Figure 6.1 and a higher-level factor Mayo risk. Two dense layers are applied to extract feature from biomarkers. The outputs from image-path and biomarker-path are concentrated with two other factors, blood and Mayo risk, which have been empirically validated as helpful to estimating cancer risk [30, 35]. The combined-path is composed of two dense layers and its output is the final cancer risk prediction. The reported

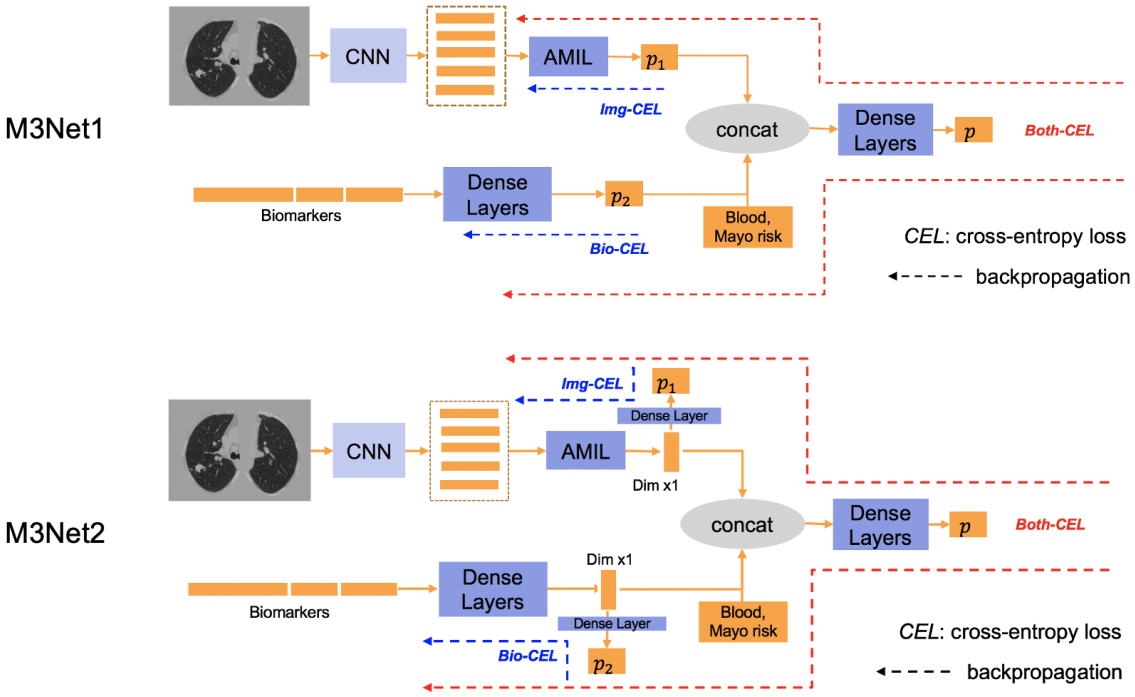


Figure 6.2: The framework of the proposed M3Net, including two versions M3Net1 (upper) and M3Net2 (bottom). The M3Net includes three paths, one for CT image and another for biomarkers, and one for combining multi-modalities. The cross-entropy loss (CEL) may be included in three positions. "Dim" in M3Net2 represents the dimension of the feature with which will be used in the concatenation. The sub-paths in M3Net1 provide the intermediate estimated risks while sub-paths in M3Net2 provide high level features for concatenation.

performance is based on the final prediction of combined-path. The cross-entropy loss (CEL) is deployed in three positions in the pipeline for image-path, biomarker path and combined path, respectively. The image-path and biomarker-path can be trained and learned independently, as the motivation is to take advantage of the subjects with missing modality. There are three situations of the framework: (1) subject with image only: the M3Net will only include the Img-CEL and only the image-path will be trained; (2) subject with biomarkers only: the M3Net will only include the Bio-CEL and only the biomarker-path will be trained; (3) subject with both image and biomarkers: the three CELs will be included and all learnable parameters will be trained.

We provide two versions of M3Net, as shown in Figure 6.2, termed as M3Net1 and M3Net2. In M3Net1, intermediate estimated risks (i.e., p_1,p_2) which are supervised by CEL from image-path and biomarker-path have feed to the combine-path. Instead of intermediate estimated risk, the feature vector has been feed to the combine-path in M3Net2.

6.3 Experiments and Results

6.3.1 Experimental Settings

Two experimental validation settings are applied in this chapter: (1) cross-validation in VDD, and (2) external-validation in UPMC. The VDD cohort has been randomly split into five folds. In the cross-validation, each fold data has been held out from training as the test set, and remaining four folds are split as 3:1 for training and validation sets. In external validation, UPMC cohort is the held-out test set, four folds in VDD is the training set, and one-fold is the validation set. If the biomarkers acquisition date is missing, it would be matched with the last CT scan date by default.

Our experiments are implemented in Python 3.7 with PyTorch 1.5 using the GTX Titan X. The max-training-epoch is set to 100. The initial learning rate is 0.01 and is multiplied by 0.2 at the 40th, 60th, 80th epochs. The optimizer used in the training is stochastic gradient descent (SGD).

6.3.2 Experimental Results

The test-set results of cross-validation and external-validation are shown in Table 6.2. The results are reported on the subjects with both biomarkers and CT in Table 6.1. The compared benchmarks including (1) the methods Mayo Model [30] and Liao et al. [37] pre-trained model, which does not need training in this chapter, (2) single modality prediction of image only and biomarkers only.

Methods	VDD	UPMC1 (mean ± std)	UPMC2 (CI)
Mayo Model^{+}	0.682*	0.858*	0.858(0.776-0.926)*
Liao et al. pretrain model^{+}	0.663*	0.810*	0.810 (0.717-0.897)*
Image only	0.712 ± 0.051 *	0.863 ± 0.004 *	0.863 (0.784-0.935)*
Biomarkers only	0.757 ± 0.059 *	0.826 ±0.040 *	0.852 (0.770-0.919)*
Learning without imperfect data	0.793 ± 0.027 *	0.886 ± 0.007 *	0.889 (0.815-0.950)*
M3Net1 (ours)	0.816 ± 0.048	**0.913 ± 0.005**	**0.917 (0.857-0.966)**
M3Net2 (Dim=5) (ours)	**0.848 ± 0.052**	0.910± 0.011	0.916 (0.851-0.968)

Table 6.2: The AUC of test set in cross-validation VDD and external-validation UPMC.

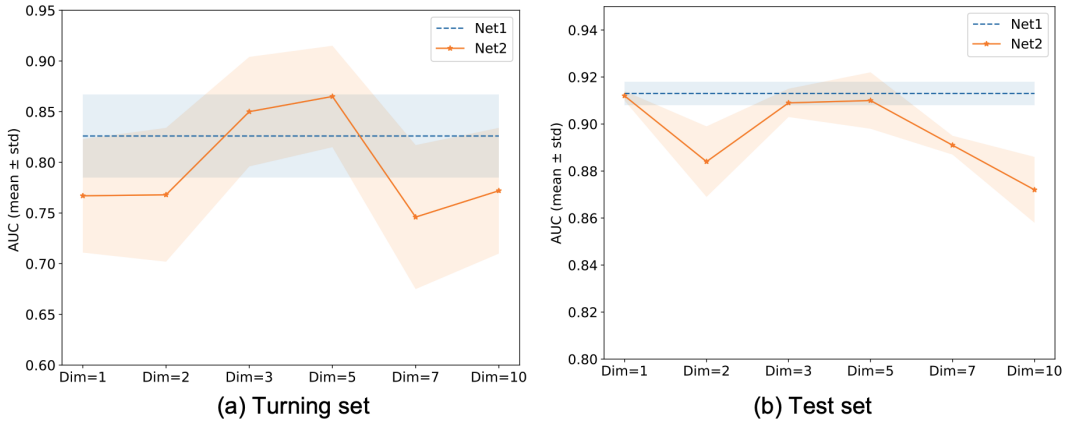


Figure 6.3: The comparison of M3Net1 and M3Net2 in the validation set and test set in the external validation setting, where AUC (mean ± std) of five folds is shown: (a) The performance on validation set (VDD). (b) The performance on test set (UPMC).

The single modality prediction has the same training procedure and training data of the proposed method, but the evaluation is based on the p_1 and p_2 of M3Net1 in Figure 6.2, respectively. (3) Learning while excluding the subjects that with missing modality which is termed as Learning without imperfect data in Table 6.2. This baseline has the same network structure as M3Net1, the only difference being that the training data only includes those subjects with both CT image and biomarkers.

Our method significantly improves the compared benchmarks ($p < 0.05$, bootstrap two-tailed test). The computing of p-value and 95% confidential interval (CI) is adapted from: https://github.com/mateuszbuda/ml-stat-util.

Figure 6.3 shows the comparison of M3Net1 and M3Net2 in the validation set and test set in the external validation setting. In M3Net2, we compare the settings of different "Dim" parameters. In general, the performance of M3Net1 is comparable with the best performance of M3Net2 under our settings and datasets. The performance decreases when "Dim" arises, which probably results from

that the network being unable to learn effective large-dimension feature given the limited data size.

6.4 Discussion

In this chapter, we propose a new framework M3Net to estimate lung cancer risk by integrating multiply modalities with missing data. The proposed network shows superior performance compared with single modality predicting and learning without imperfect data, which indicates the combining multi-modalities and utilizing more subjects even with missing data can increase the performance of lung cancer diagnosis. We compare two versions of our model with their performance on validation set and test set. A limitation of this chapter is the data size is not as large as general classification tasks in computer vision field, therefore, the generalization across different patient cohorts requires care.

CHAPTER 7

Deep Multi-modal Prediction with Incomplete Data

Risk factors, including clinical data elements (e.g., age, cancer history, and smoking status) and radiomics features (such as nodule size), are usually used as a form of tabular data. These factors have been widely used in machine learning and established clinical models [29, 30, 31, 114] for lung cancer risk estimation (LCRE). With the success of deep learning, high-level features can be automatically extracted from high-dimensional images (such as CTs). For example, [40] analyzed lung cancer risk with manually detected nodules in CTs (nodule-level). [37] developed a detection-classification system with multi-instance learning for LCRE at the scan-level. [85] utilized the repeated scans of the same patient (patient-level) to capture the longitudinal information, either in a simple concatenation way or refined sequential modeling. Furthermore, previous studies have shown that data from multiple modalities provide complementary information for prediction, not only in general computer vision tasks [157] but also in LCRE (with CT image and risk factors [93]).

In clinical practice, missing data is common in LCRE especially considering data from multiple modalities. Generally, CT images and clinical risk factors can be collected from lung cancer screening programs, as shown in Figure 7.1. Clinical data elements (e.g., age and smoking status) are usually collected from electronic medical records and questionnaires from clinical visits. Based on the collected clinical information, doctors will determine if a CT screening is needed. Further, radiology reports including radiomic features (e.g., nodule size and spiculation) can be created based on the CT image. According to clinical guidelines, patients maybe be suggested to undergo repeated CT scans. Thus, ideally, clinical tabular data and repeated CT scans can be available for lung cancer diagnosis. However, in practice, data can be missing due to intricate reasons including data entry, data exchange, or loss of follow-up (Figure 7.1). Furthermore, missing data can be more severe in the multi-site context, given potential inconsistent protocols and high heterogeneities of data collection.

As illustrated in Figure 7.1, missing data may exist in either/both modalities. General imputation methods that only use information from one modality (we call it target modality), we posit

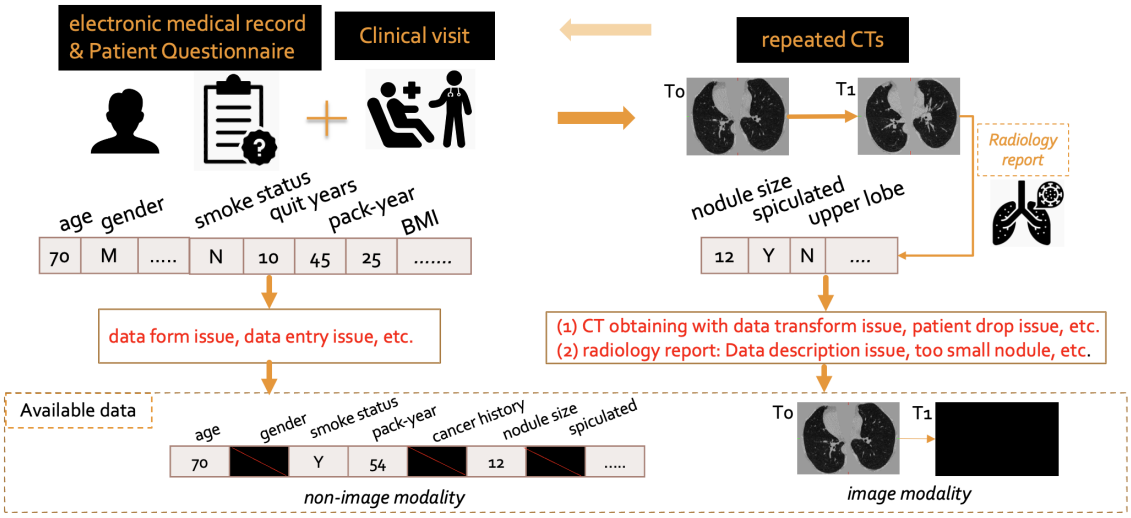


Figure 7.1: Missing data in multiple modalities. The upper panel shows a general lung screening process. In clinical practice, missing data can happen at different phases (as red text). The lower panel shows that patient may miss risk factors or/and miss follow-up CT scans.

information from another modality (termed as conditional modality) may provide essential contribution to recover its class discrimination when imputing. For example, a digit image with only the background can recover the number with the help of another modal data (as in our illustration example), which is impossible only with the digit modality. In LCRE, a CT image can help to impute missed radiographic factors (e.g., nodule size), and nodules can be easier reconstructed when informed by risk factors.

In this study, we propose a new model, Conditional PBiGAN (C-PBiGAN), based on the encoder-decoder structure as introduced in PBiGAN, to capture the joint distribution for imputation across modalities. Specifically, we introduce 1) a conditional latent space in multi-modal missing imputation context; 2) a class regularization loss to capture discriminative information during imputation. For evaluation, we start with an illustration example with MNIST and Fashion-MNIST (with 70,000 data samples) as two modalities. MNIST and Fashion-MNIST have the same number of classes/samples, which is convenient to simulate as classification on samples with two-modality. Further, we focus on the real clinical problem that deals with the missing data for lung cancer risk estimation, where risk factors (non-image) and serial CT scans (image) are two important modalities for rendering clinical decisions.

In summary, our contributions are three-fold: (1) To the best of our knowledge, the proposed

C-PBiGAN is the first deep imputation of missing data by capturing the joint distribution of multi-modal data with adversarial training.

(2) Our model can both recovers realistic data (e.g., lung nodules in CTs and handwritten digits) and recover the class-discriminative information, even when data are largely missing.

(3) Evaluated with two types of multi-modal data (image + image, image + tabular), our model achieves superior downstream predicting performance compared with benchmark imputation methods.

7.1 Theory

7.1.1 Task Description and Intuition

In this section, we describe the task of multi-modal imputation, given a set of samples $1, 2, \cdots, n$ with the data source from two modalities A and B. In the illustration example, each sample has two paired images from MNIST and Fashion-MNIST as two modalities. In the task of LCRE, each patient has CT images and tabular clinical data as two modalities. X^A and X^B are the complete data space from modalities A and B. In practice, as Figure 7.1, we may not be able to observe complete data. Take modality A for example, $x^A \in R^n$ denotes completely observed data, and $m \in \{0,1\}^n$ is a missing indicator with the same dimension of x^A that determines which entries in x^A are missing (i.e., 1 for observed, 0 for missing). So, $(x^A, m) \in X_o^A$ represents the observed data in modality A. Conventional imputation methods "make up" missing data with the knowledge of a single modality (e.g., complete (x^A, m) based on the knowledge of X_o^A). However, the data from modality B (either fully observed or not) may help the imputation in A. Our motivation is imputing the missing data by extending the knowledge from modality A to B with a newly designed model. When the data X_o^A and X_o^B are imputed, i.e., become complete data $\tilde{X}^A$ and $\tilde{X}^B$, which can be directly used for downstream tasks with a unified model.

7.1.2 Partial Bi-directional GAN framework

Generative adversarial networks (GANs) [98] have shown great success in generating realistic samples, which have also been adapted to the field of missing data imputation [88, 89, 156]. GANs have a clever way of training with two sub-models: a generator and a discriminator, where the generator is trained to generate fake samples to fool the discriminator and the discriminator is designed for

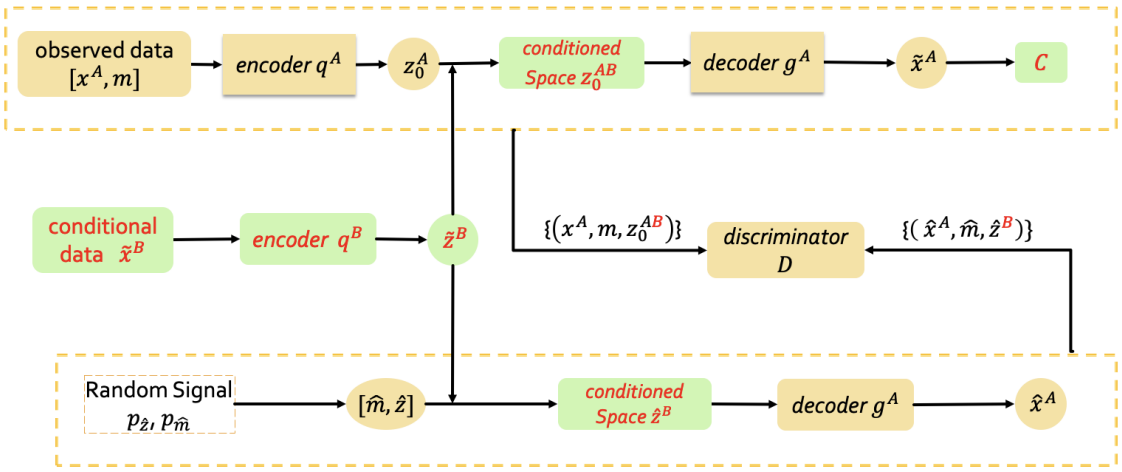


Figure 7.2: Structure of the proposed C-PBiGAN. The orange and green characters highlight our contributions compared with PBiGAN. m is the missing index of target modality A and z is the corresponding latent space. $\tilde{x}^B$ is the complete data of conditional modality B, which can be fully observed or imputed. $\tilde{x}^A$ is the imputed data of A based on observed data $[x^A, m]$ and $\tilde{x}^B$. $\hat{x}^A$ is the generated data of A based on $\tilde{x}^B$ and noise distributions of $p_{\hat{z}}$ and $p_{\hat{m}}$. C is a classifying module along with cross-entropy loss regularizing the generator for keeping the identities of imputed data.

separating real and fake samples. PBiGAN [89] is a recently proposed imputation method with an encoder-decoder framework, based on bi-directional GAN (BiGAN) [158], has been validated with state-of-the-art performance in several tasks.

In this chapter, the proposed Conditional PBiGAN (C-PBiGAN) is based on the structure of PBiGAN. We show C-PBiGAN and PBiGAN in Figure 7.2, where the "black text" components are from original PBiGAN and our contributions are highlighted with red text. Note that the original PBiGAN is designed for a single modality (i.e., modality A in Figure 7.2).

The PBiGAN is composed of an encoder, decoder and discriminator. As in Figure 7.2, the decoder g^A transforms a latent code z into a complete data space X^A, where z can be a feature space (e.g., z_o^A) produced by the encoder, or sampled from a simple distribution (e.g., Gaussian). The encoder $q^A(z_o^A|x^A, m)$, denoted as q_A for simplification, maps the missing distribution p_m of an incomplete data (x^A, m) into a latent vector z_o^A, where $x^A \in \mathbb{R}^n$ denotes complete data, and $m \in \{0,1\}^n$ is a missing indicator with same dimension of x^A that determines which entries in x^A are missing (i.e., 1 for observed, 0 for missing).

The discriminator D of PBiGAN takes the observed data (x^A, m) and its corresponding latent code z_o^A as the "real" tuple in adversarial training. Similarly, the fake sample $(\hat{x}^A, \hat{m}, \hat{z}))$ is comprised

of 1) a random latent code $\hat{z}$ sampled from a simple distribution $p_{\hat{z}}$ (e.g., Gaussian), 2) missing indices $\hat{m}$,$p_{\hat{m}}$, and 3) the generated data $\hat{x}^A$ based on random latent code $\hat{z}$.

The loss of PBiGAN is defined as follows, which is minimax optimized as general GAN:

$$\begin{aligned} L\left(D, g^A, q^A\right) =& \mathbb{E}_{(x^A,m)\sim p_m} \mathbb{E}_{z_o^A \sim q^A(z_o^A|x^A,m)} [\log D(x^A, m, z_o^A)] \\ &+ \mathbb{E}_{(\cdot,\hat{m})\sim p_{\hat{m}}} \mathbb{E}_{\hat{z}\sim p_{\hat{z}}} [\log(1 - D(g^A(\hat{z},\hat{m}),\hat{m},\hat{z}))] \end{aligned} \tag{7.1}$$

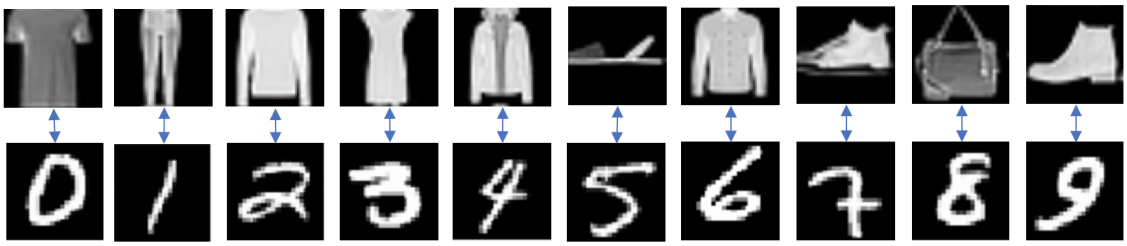
7.1.3 The Proposed Conditional PBiGAN

One drawback of PBiGAN [89] is the imputation is finished within single modality, which does not take advantage of complementary information from multiple modalities. The information from complementary modalities (i.e., a conditional modality) can be essential, especially when the missing rate in the target modality is high. In this work, we propose a new algorithm, called conditional PBiGAN (C-PBiGAN), that includes (1) conditional latent space with knowledge of both target and conditional modalities, and (2) a classification regularization loss is optimized during generator training to effectively preserve discrimination for imputed data.

As shown in Figure 7.2, there are three main differences between the proposed C-PBiGAN and PBiGAN. First, another encoder (i.e., q^B) is introduced to extract the feature $\tilde{z}^B$ for data $\tilde{x}^B$ of the conditional modality B. $\tilde{x}^B$ is completed data, either fully observed or has been imputed. Second, the latent space (e.g., z_o^A) for reconstructing data has been replaced with the conditional space z_o^{AB}, which combines the information from modality $A(z_o^A)$ and modality $B(\tilde{z}^B)$. Third, a classification regularization loss (i.e., cross-entropy loss) has been added to reconstruction data with another feature extraction net (i.e., the red text C in Figure 7.2). The generative adversarial loss of C-PBiGAN can be written as:

$$\begin{aligned} L_{D,g^A,q_A,q_B} =& \mathbb{E}_{(x^A,m)\sim p_m} \mathbb{E}_{z_o^{AB} \sim [q^A(z_o^A|x^A,m), q^B(\tilde{z}^B|\tilde{x}^B)]} [\log D(x^A, m, z_o^{AB})] \\ &+ \mathbb{E}_{(\cdot,\hat{m})\sim p_{\hat{m}}} \mathbb{E}_{\hat{z}^B \sim [p_{\hat{z}}, q^B(\hat{z}^B|\tilde{x}^B)]} [\log(1 - D(g^A(\hat{z}^B,\hat{m}),\hat{m},\hat{z}^B))] \end{aligned} \tag{7.2}$$

] where (x^A, m) is observed data in target modality A, z_o^{AB} and $\tilde{z}^B$ are the conditional latent space of the "true" and "fake" sample in adversarial training. To enforce the imputed $\tilde{x}^A$ having the same identity with x^A even when data are largely missing, we further introduce a feature extraction net C



Fashion-M	T-shirt	Trouser	Pullover	Dress	Coat	Sandal	Shirt	Sneaker	Bag	Ankle boot
MNIST	0	1	2	3	4	5	6	7	8	9

Figure 7.3: The illustration of MM-MNIST. A sample in MM-MNIST contains two paired samples from Fashion-MNIST and MNIST. The paired samples have the sample class label index.

along with a cross-entropy loss when training the generator. Specifically, C-PBiGAN is designed to solve the following minimax optimization problem:

$$\min_{g^A,q^A}(\max_{D}(L_G(D,g^A,q^A,q^B)) - \mathbb{E}_{\tilde{x}^A\sim g^A(\cdot)}[\log p(y|C(\tilde{x}^A))]) \tag{7.3}$$

where y is the class label. q_B, C can be pretrained or trained with g_A,q_A simultaneously. The proposed C-PBiGAN is also different from conditional GAN [159] in two ways: 1) our model can utilize the partially observed data in the imputation context, and 2) a classifier C along with binary cross-entropy loss is introduced to highlight identity preservation of imputed data.

A limiting case of C-PBiGAN is to impute data that is completely missing (i.e., $m = \mathbf{0}$). In this case, complete data for training (i.e., $m = \mathbf{1}$) are needed, and it is the generated $\hat{x}^A$, rather than $\tilde{x}^A$ as in Fig. 7.2, that used for downstream task. In Eq. (3), the $\tilde{x}^A$ is replaced with $\hat{x}^A$. One of our tasks imputing a nodule image belongs to this limiting case, as Figure 7.6(b).

7.2 Experiments on Illustration Dataset

7.2.1 Dataset Introduction

To visualize the effectiveness of C-PBiGAN, we select two widely used datasets with the same number of data samples and classes in computer vision: MNIST [135] and Fashion-MNIST (F-MNIST) [160] to simulate a multi-modal dataset. Frechet inception distance (FID) [161] is used to evaluating the quality of generated images.

MNIST and Fashion-MNIST contain 10 classes of digits and wearings, respectively, which both

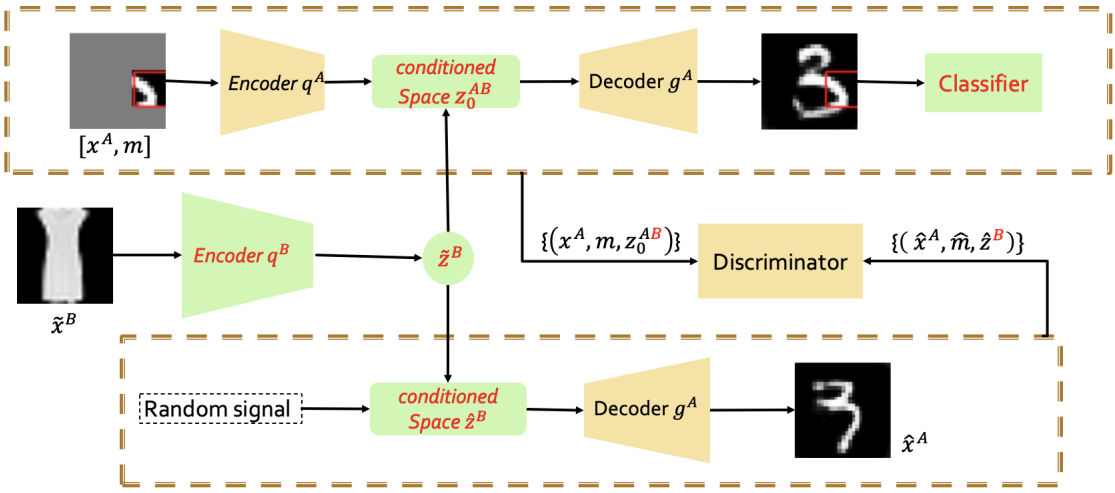


Figure 7.4: The C-PBiGAN instantiation for MM-MNIST dataset. The Fashion-MNIST are treated as conditional modality (i.e., modality B) and MNIST samples are from target modality (i.e., modality A).

have a 60000/10000 of train/test split. In this work, we further split the 60000 training samples as 50000/10000 of train/validation. Paired samples from MNIST and Fashion-MNIST are simulated as two modalities. Two modalities (one from MNIST and one from Fashion-MNIST) are both images but with huge distinction. We term this simulated multi-modal MNIST dataset as MM-MNIST (Figure 7.3).

7.2.2 Method

The theory of C-PBiGAN (Figure 7.2) is instantiated as Figure 7.4 for MM-MNIST dataset. We follow the net backbones and experimental setup of PBiGAN [89] in MNIST of its GitHub [162].

The encoder, containing three convolutional layers and two fully connected layers, is fed with 28×28 size images. The decoder is composed of one fully connected layer and three deconvolutional layers, which is fed with latent feature vectors. The discriminator (i.e., critic model) is fed with tuples of latent feature vector and image and its final output is scores of real vs. fake. The max training epochs is 1000. The MNIST is the target modality with 88% pixels missing, and Fashion-MNIST is the fully observed conditional modality. We include predictions of raw missing image, PBiGAN imputed image, and fully observed image for comparison. The same net structure is used for obtaining the final prediction accuracies.

Method	Random guess	Raw missing	PBiGAN	C-PBiGAN	complete
Fashion-MNIST	10.00	-	-	-	-
Imputed MNIST	10.00	28.14	52.09	95.08	98.49
MM-MNIST	10.00	81.28	90.99	96.41	99.83
FID (test set)	-	139.8	51.8	21.3	10.2

Table 7.1: Predicting test accuracies (%) of MM-MNIST and its single modality (Fashion-MNIST is fully observed)

7.2.3 Results and Analysis

We compare four methods qualitatively and quantitatively. Table 7.1 shows predicted accuracies of multi-modal prediction and the single modality respectively. The digits MNIST modality is only partially observed (i.e., only a box of the image is observed, this setting is motivated by PBiGAN). As in Table 7.1, our C-PBiGAN outperforms the compared baselines with a large gap on the imputed MNIST modality (e.g., C-PBiGAN 95.1% vs. PBiGAN 52.1%), indicating the information from the conditional modality does help to recover the identity information. In the multi-modal prediction context (i.e., "MM-MNIST" in Table 7.1), the proposed C-PBiGAN can still outperforms baselines (C-PBiGAN 96.4% vs. PBiGAN 91.0%).

Figure 7.4 visualizes an example of how the proposed approach works. The "dress" and "3" are two modalities of same class from Fashion-MNIST and MNIST, respectively. When C-PBiGAN conditioned on Fashion-MNIST, the data modality of MNIST is not only realistically imputed but also preserving the class information, even only a small part of original data observed. By comparison, PBiGAN can obtain a realistic output, while fails to keep the class identity. Our model C-PBiGAN achieves the lowest FID (i.e., indicating the most realistic) among the settings with missing data. The PBiGAN is designed to generate samples that match original data distribution while regardless of its identity, which may lead to generated samples in the "middle" states of multiple identities. For example, as shown in Figure 7.5, the generated sample from PBiGAN is not a typical handwritten digit (even though it looks like an "8" at a glance, while it is not a typical handwritten "8"). Thus, it may decrease the reality of generated samples and lead to higher FID. Instead, our model introduces the "identity" information from another modality with the conditional latent space. It can avoid generating "middle" states of difference identities (e.g., the "middle" states of a "0" image and a "2" image are probably not images of realistic number) at some degree and then

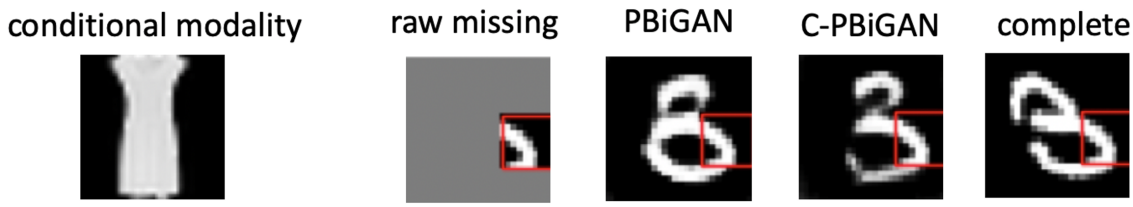


Figure 7.5: Qualitive results of MM-MNIST. Only a similar box of the digit is observed. The PBiGAN can impute the image to a "smooth" completed image. However, it is more like a "8" rather than a "3". By comparison, the C-PBiGAN preserves the identity of the imputed image.

lead to more realistic generations.

7.3 Experiments on Empirical Lung Cancer Datasets

7.3.1 Dataset Introduction

Chest CTs and tabular risk factors are readily available source of data for lung cancer risk estimation. We consider two longitudinal CTs as the complete data for image modality, where time point 0 (TP0) indicates the previous CT and time point 1 (TP1) indicates the current CT. The non-image modality (i.e., risk factors) includes the following 14 risk factors: age, sex, education, body mass index (BMI), race, quit smoke time, smoke status, pack-year, chronic obstructive pulmonary disease (COPD), personal cancer history, family lung cancer history, nodule size, spiculation, upper lobe of nodule. The first 11 factors do not need the effort of the radiologists' reading of CTs because they are derived from electronic medical records, questionnaires, clinical visits. The last 3 factors are from radiology report that requires radiologists' manual efforts (as Figure 7.1). The selection of the risk factors in this chapter are highly motivated by the clinical models, i.e., Mayo [30], PLCOm2012 [29], Brock models [31].

Two cohorts of lung cancer datasets are studied in this work, 1) the national lung screening trail (NLST) [4] and 2) an in-house screening cohort from the Vanderbilt Lung Screening Program (VLSP), a research study under our Institutional Review Board's supervision. Patients in the NLST are included if 1) they have 14 selected risk factors available, 2) have a tissue-based diagnosis, and 3) the diagnosis happened within 2 years of the last scan for cancer cases. Note that included subjects are all high-risk patients (all received biopsies), the distinction between cancer / non-cancer in this subset is harder than the whole NLST population. In total, we have 3889 subjects from the NLST of whom 601 were diagnosed with cancer. In VLSP, data from 404 subjects have data (including

missing data) for evaluated, in which 45 were diagnosed with lung cancer. Due to issues as Figure 7.1, the available factors of VLSP have an average of 32% missing rate, and 60% of patients do not have complete longitudinal scans.

7.3.2 Method

C-PBiGAN has been instantiated to impute risk factors and longitudinal images. Risk factor imputation follows the general C-PBiGAN (Figure 7.2) since the factors can be partially observed even when some data are missing. In this case, we only need to replace modality A with partially observed risk factors and modality B with CT (imputed or observed). The C-PBiGAN instantiation for factor imputation is shown in 7.6(a). Image imputation is under the limiting case of C-PBiGAN (instantiated as Figure 7.6(b)), since partially observed "nodule" in CT is not a practical setting. We follow the C-PBiGAN theory for image imputation, and we also utilize information from longitudinal context in practice. We assume the background of a nodule would not substantially change between TP0 and TP1. Thus, motivated by masking strategies in [44, 163], nodule background is borrowed from observed CT (i.e., TP0 image) of the same patient by masking its center when generating the target time point (i.e., TP1 image), see "TP0 background" in Figure 7.6(b) and more examples in Figure 7.8.

Given a CT scan, we follow Liao' pipeline [37] of preprocessing and detect the top five confidential nodule regions for downstream work. Rather than imputing a whole 3D CT scan, we focus on imputing the nodule areas of interest in 2D context with axial/coronal/sagittal directions as 3 channels. Considering 1) the radiographic reports regarding TP0 are rarely available, and 2) TP1 plays a more important role in lung cancer risk estimation [85], we focus on the imputation on TP1 of image modality in this study. The TP0 image is imputed as a copy of the TP1 image when TP1 is observed and TP0 is missing.

7.3.3 Network Structure

The structures of encoder, decoder, and discriminator are 1) adapted from face example in PBiGAN for image modality, and 2) separately comprised of four dense layers for non-image modality. A unified multi-modal longitudinal model (MLM), which contains an image path and a non-image path, is used for lung cancer risk estimation to evaluate the effectiveness of imputations. The image

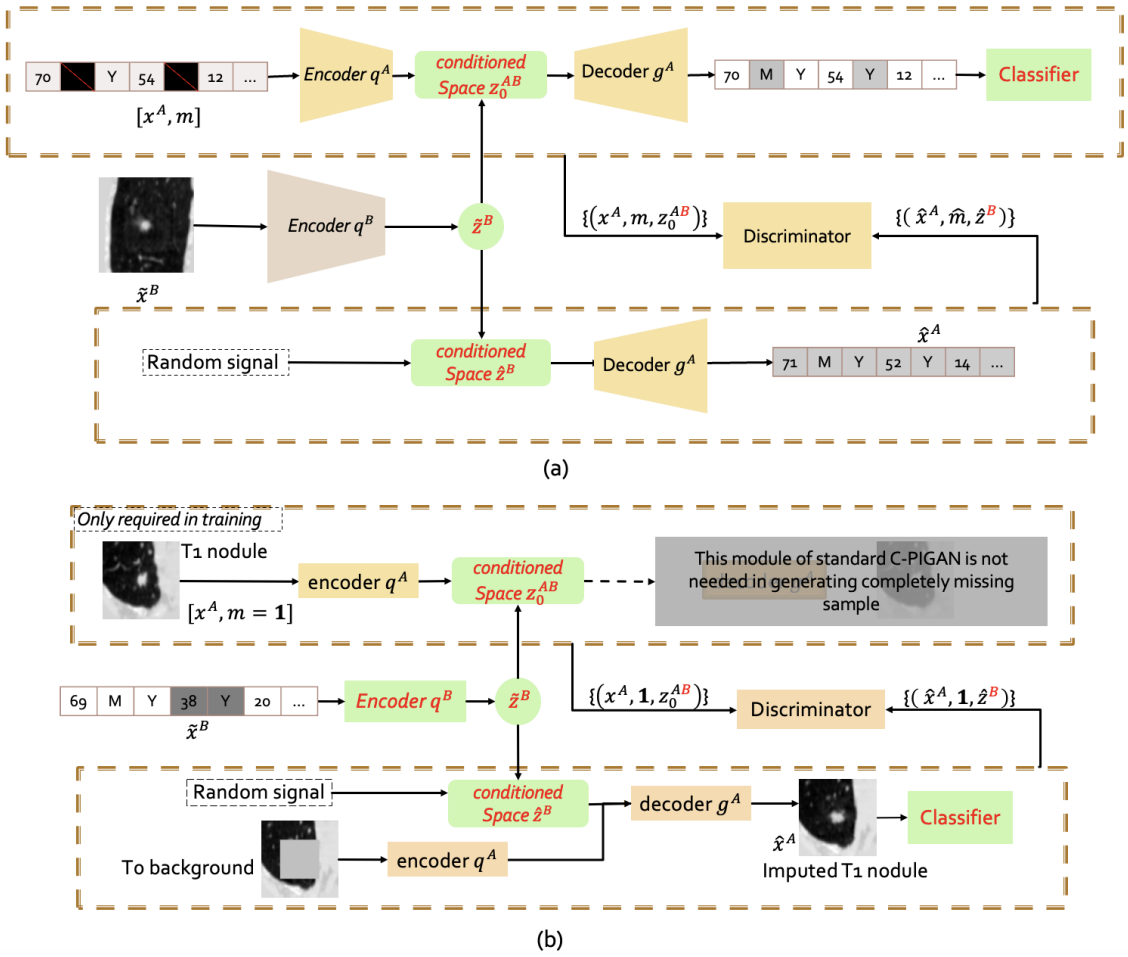


Figure 7.6: (a) C-PBiGAN instantiation for clinical factors imputation. (b) An instantiation of C-PBiGAN limiting case for CTs imputation.

Method	Image-only	Mean imputation	Soft-imputer	PBiGAN	C-PBiGAN	fully-observed factors
factor-only	N/A	79.73	79.46	79.14	83.04	86.24
LOCF	75.45	83.76	83.80	83.79	84.00	86.21
PBiGAN	76.54	83.02	83.82	83.29	83.51	85.90
C-PBiGAN$^{\#}$	82.70	85.00	85.62	85.17	85.87	86.72
C-PBiGAN	84.15	85.72	85.90	85.91	86.20	88.27
fully-observed images	87.48	88.23	88.40	88.44	88.46	89.57

Table 7.2: AUC results (%) of test set on NLST. Generally, each row or each column represents an imputation option for image-missing or risk-factor-missing, respectively. "Image-only" or "Factor-only" represents predictions only using imputed longitudinal-images or factors, respectively.

Method	Image-only	Mean imputation	Soft-imputer	PBiGAN	C-PBiGAN
factor-only	N/A	75.17	83.46	84.40	86.56
LOCF	75.52	82.83	87.11	86.99	87.63
PBiGAN	73.44	80.85	84.43	84.88	85.65
C-PBiGAN$^{\#}$	80.59	83.87	86.57	87.19	87.69
C-PBiGAN	82.61	85.29	88.11	88.49	89.19

Table 7.3: AUC results (%) of external in-house set. Generally, each row or each column represents an imputation option for image-missing or risk-factor-missing, respectively. "Image-only" or "Factor-only" represents predictions only using imputed longitudinal-images or factors, respectively.

path includes a backbone of ResNet18 to extract image features and a LSTM to integrate longitudinal image features (from TP0 and TP1). The risk factor features are extracted by a subnet with four dense layers. The image path and non-image path in the MLM are validated to be effective by comparing with representative models (i.e., AUC in NLST: image-path model (0.875) vs. Liao et al. (0.872) with image data only, non-image path model (0.883) vs. Mayo clinical model (0.829)). The image and non-image features are combined for the final prediction.

7.3.4 Experimental Settings and Evaluations

The NLST is randomly split into train / validation / test sets with 2340 / 758 / 791 subjects. The in-house dataset of 404 subjects is externally tested when training is finished in NLST. We follow the experimental setup of PBiGAN opensource code [89] when training C-PBiGAN, e.g., using Adam optimizer with a learning rate of 1e-4. The maximum number of training epochs is set to 200. Our experiments are based on Python 3.7 and PyTorch 1.5 on GTX Titan X. The area under the receiver operating characteristic (AUC) for lung cancer risk estimation is used to quantitatively evaluate the effectiveness of imputations and FID to evaluate the image qualities. We also show imputed nodules of benign and malignant cases as qualitative results.

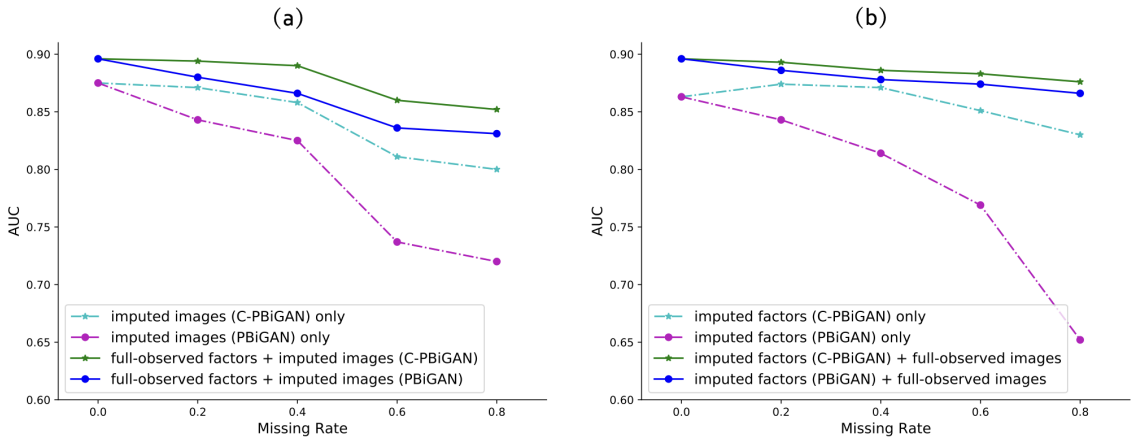


Figure 7.7: (a) AUCs of various TP1-image missing rates when factors are fully observed in NLST, (b) AUCs of various missing rates of factors when images are fully observed in NLST. The left start point is under condition that data is not missing (i.e., missing rate = 0.0).

Representative imputations in image and non-image modalities are combined for comparison as in Table 7.2 and Table 7.3. Non-image imputation baselines include mean imputation, soft-imputer and PBiGAN. LOCF and PBiGAN are compared for image modality imputation. As a comparable way utilizing observed TP0 when imputing TP1, C-PBiGAN$^{\#}$ denotes feeding of TP0 nodule image without masking the central "nodule", rather than "TP0 background" in Figure 7.6.

7.3.5 Results and Analyses

Table 7.2 and Table 7.3 show 1) test results of NLST with 30% missing risk factors and 50% missing in longitudinal TP1 image (upper), and 2) external tests of in-house data, respectively. Figure 7.7 compares the proposed C-PBiGAN with the baseline PBiGAN [89] in terms of lung cancer risk estimation performance in NLST under (a) various TP1 missing rates when factors are fully observed, (b) various factor missing rates when images are fully observed. A qualitative result on image imputation is shown in Figure 7.8.

In Table 7.2, The C-PBiGAN combination (bold) significantly improves all method combinations without C-PBiGAN across the image and non-image modalities, in both MCAR (a hard subset of NLST) and MNAR (clinical dataset with real missing) contexts ($p < 0.05$, bootstrap two-tailed test (n=2000)), indicating that our model effectively imputes the multi-modal data for lung cancer risk estimation.

As shown in Figure 7.7, our model outperforms PBiGAN in the image-missing and factor-

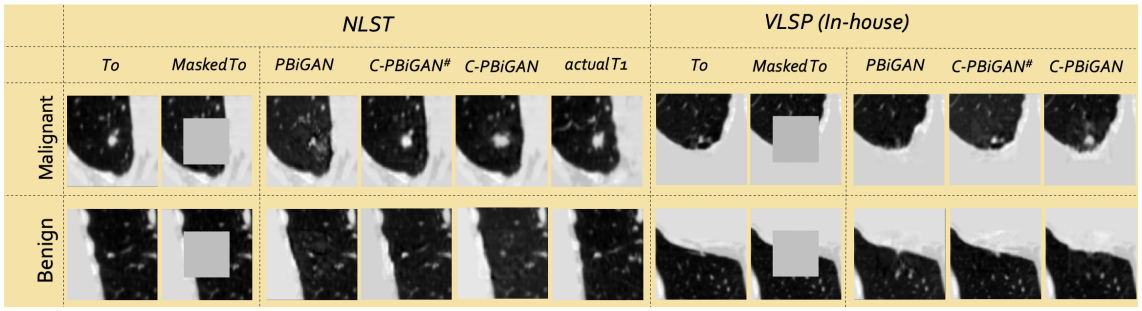


Figure 7.8: Qualitative results of imputed longitudinal images (upper: malignant cases, bottom: benign cases). "PBiGAN" and "C-PBiGAN" impute the TP1 by feeding the masked TP0 as "TP0 background" in Figure 7.3, and C-PBiGAN$^{\#}$ feeds TP0 including center rather than TP0 background.

Method	Masked TP0	PBiGAN	C-PBiGAN$^{\#}$	C-PBiGAN
FID	95.4	34.7	32.1	35.0

Table 7.4: FID comparison of different methods in the NLST

missing contexts of different rates. A more obvious superiority can be found when only using the imputed modality for prediction (e.g., C-PBiGAN: 0.830 vs. PBiGAN: 0.652 when risk factors have 0.8 missing rate in Figure 7.7(b), $p < 0.05$), indicating the knowledge from conditional modality in C-PBiGAN does help impute data for lung cancer prediction. Even at some missing rates, the imputed factors conditioned on images can even achieve higher AUC than the fully observed factors (see Figure 7.7(b)).

Figure 7.8 shows malignant and benign cases in NLST and the in-house dataset, respectively. Both PBiGAN and proposed C-PBiGAN can reconstruct visually realistic images, while the malignant and benign cases from PBiGAN are harder to distinguish. Table 7.4 shows that FID values of imputation settings with the original data. We find that the FID of PBiGAN and two settings of C-PBiGAN (i.e., C-PBiGAN and C-PBiGAN$^{\#}$) is close, which is lower than the setting with 64×64 center masked within 128×128 images.

When comparing C-PBiGAN$^{\#}$ to C-PBiGAN, the latter is more effective where the center of fed TP0 image is masked, given the current pipeline. We find that when feeding TP0 without masking center to provide nodule background (i.e., C-PBiGAN$^{\#}$), the central nodule region of imputed TP1 can be fit to the center of TP0, just like the nodule background of imputed TP1 is designed to fit TP0 nodule background. This limits the discrimination of imputed TP1, as the generated malignant

nodule in Figure 7.8 demonstrates. It appears that separating the "background" from the "nodule" region during learning conveys important data structure, since we want the "background" of imputed TP1 to be close to observed TP0 while the "nodule" of imputed TP1 should mainly be conditioned on tabular data. Motivated by masking strategies in [44, 163], our C-PBiGAN is fed with the TP0 background masking the center when imputing the TP1 image (Figure 7.6(b)). Different from ours, conditional GAN [159] used in [44, 163] only targeted generating nodules while ignoring their malignancy.

Table 7.4 shows FIDs of different settings (its qualitative results are shown in Figure 7.8). We find (1) PBiGAN and C-PBiGAN achieve similar FID and (2) C-PBiGAN$^{\#}$ achieve a little lower FID. Different from the MM-MNIST use case, the "middle" states of lung nodules still can be realistic. For example, the middle state of digits "2" and "3" is not realistic, while the middle state of a benign nodule (say its nodule size is 1mm) and a malignant nodule (e.g., with 20mm nodule size) can be a realistic nodule with a size around 10mm. This explains that (1) PBiGAN and C-PBiGAN achieve similar FID. Second, C-PBiGAN$^{\#}$ is fed with a whole TP0 nodule including more image contexts, resulting to a more realistic generation. However, as we have discussed, the generated TP1 from C-PBiGAN$^{\#}$ is easier to fit into TP0 and less discriminative to the following prediction task.

7.4 Conclusion

In this work, we explore the missing imputation across multi-modality data structures by proposing a new GAN-based algorithm to modeling joint distribution of available data. We hypothesize that missing imputation can be benefited from modeling the joint distribution of multi-modal data and validate the hypothesis with two types of multi-modal data (image + image, image + tabular).

Starting with an illustration dataset (combining MNIST and FashionMNIST as MM-MNIST) with large-scale data (70,000 samples), we visualize how our model support the hypothesis. Then, we validate our model with clinical datasets for lung cancer risk estimation, including internal validation on the NLST and external validation within a local population of missing in practice. We have successfully imputed a set of sequential lung CT scans informed by existing tabular data as well as missing tabular data informed by existing CT scans.

In summary, we propose a novel deep learning based missing imputation methodology for

multi-modal data. We validate our method on 1) multi-class classification of the MM-MNIST with incomplete image, and 2) lung cancer risk estimation on large-scale NLST dataset with simulated missing as well as another independent external validation cohort with real missing data. Our model achieves significantly better results compared with benchmarks, including the state-of-the-art model (PBiGAN).

CHAPTER 8

A Comparative Study of Confidence Calibration in Prediction Models

8.1 Introduction

Confidence calibration [164], a.k.a. model calibration or predicting calibration, measures the agreement between the predicted probability and the true correctness likelihood. Due to the availability of large-scale datasets and powerful computing resources, the recent development of deep learning models has dramatically improved the discrimination of prediction models. However, improvements in model calibration are comparatively less impressive.

Confidence calibration is important across domains, especially healthcare [165], for trustworthy prediction. Even though deep learning is more accurate with larger scale training data and more learnable parameters, deep models can be poorly calibrated. It is believed that hyperparameters like model depth, weight decay and batch normalization influence final model calibration [164]. A predictive model should know when its prediction is inaccurate. For example, in computer-aided diagnosis system, the model should inform its users, doctors, of the uncertainty around its prediction. Another obstacle to achieving desirable calibration is the unbalanced nature of available data. The available training data for clinical diagnosis are usually highly unbalanced. For example, the national lung screening trial (NLST) has included over 23,000 patients in the chest CT trail, but only about 1000 patients were ultimately diagnosed with lung cancer [4]. Even though great progress has been made in deep learning with publicly available well-balanced large-scale datasets such as ImageNet [166], CelebA [167], COCO [168], model discrimination and calibration can be challenging in the imbalanced training [169, 170, 171, 172]. As shown in Figure 8.1, the imbalanced training damage the class discrimination and confidence calibration. Though studies have been conducted to improve the discrimination in a re-sampling [173, 174] or re-weighting contexts [166, 175, 176, 177, 178], the calibration analyses on imbalance training is relatively lacked especially in clinical contexts.

Confidence calibration is commonly performed in one- or two- stages. One-stage refers to the model is developed end-to-end for discrimination and calibration simultaneously which tries

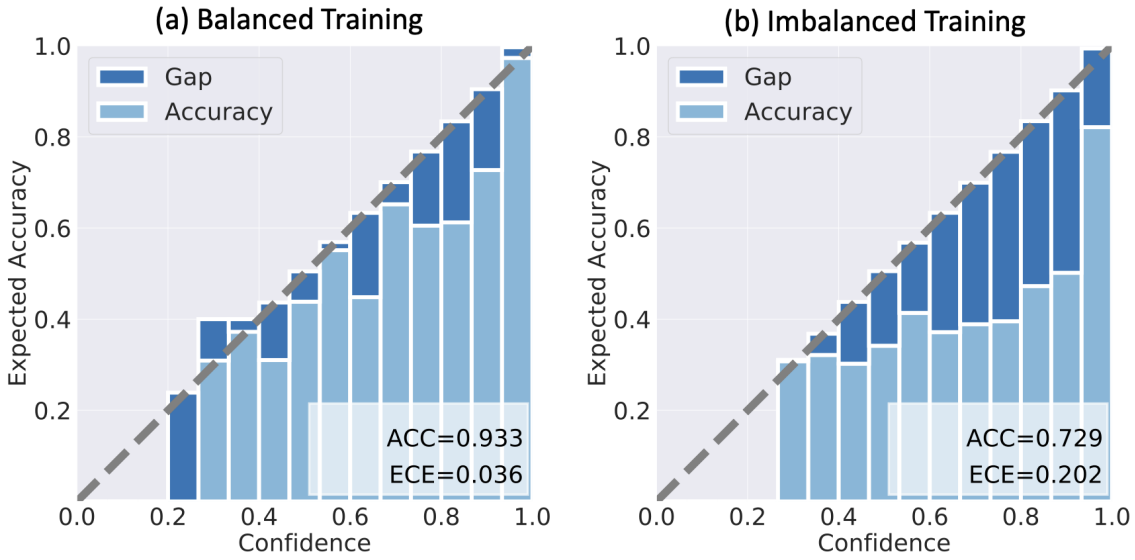


Figure 8.1: The reliability diagram between the balance training and imbalanced training on the same test set (more details can be found in experiments section). Both the performances of class discrimination reflected by accuracy (ACC, larger is better) and confidence calibration reflected by expected calibration error (ECE, smaller is better) are decreased under the imbalanced training.

to produce calibrated predictions to begin with. Two stage models are developed with a separate calibration step that tries to take the original predictions and make them more calibrated. One stage model mixup [179] is a regularization for a deep learning model by interpolating the input and label. Thulasidasan et al. [180] find that training with mixup can significantly improve the model calibration. Another branch to improve model calibration is under uncertainty estimation contexts. A well-calibrated model should be accurate when it is certain (with high confidence) and be uncertain (with low confidence) when it is inaccurate. Krishnan et al. [181] introduce differentiable accuracy versus uncertainty calibration (AvUC) loss function that allows models to have well-calibrated uncertainties. There are also calibration models in the two-stage framework. One of the earliest works on calibrating deep learning models is discussed in [164], which offers calibration insights into neural network learning and provides a simple but effective way (i.e., temperature scaling) to calibrate models. Label smoothing [182] is another widely used regularization technique to avoid overfitting during training. Muller et al. [183] introduce a comprehensive study on label smoothing which can increase confidence calibration. Zhong et al. [169] incorporate a label-aware smoothing in the calibration model for long-tail learning. Lukasik et al. [184] show label smoothing can mitigate label noise.

Even though promising results have been demonstrated in previous arts, most of them are under standardized or simulated classification task and long-tail challenges especially in clinical contexts are not well explored. These are approaches proposed recently to calibrate deep prediction, while there are no studies found to demonstrate how these representative models work in different challenging contexts. In this work, we conduct a comparative study of four leading calibration models across one- and two-stage (12 combinations in total including baseline CEL for comparison). Started from a balanced multi-class classification with large-scale training set, we then simulate an imbalanced training but still with large-scale set. Furthermore, we extend our analyses to a clinical prediction task (i.e., lung cancer diagnosis) with CT image. Our study can help readers to understand big picture of confidence calibration especially the four leading calibration models and understand their gap across different domains.

8.2 Method

In this section, we introduce the used representative approaches for calibration in this paper. Our motivation to choose the approaches is (1) well-known and widely used, (2) cover multiple categories such as one- vs. two- stages. With the 4 selected approaches and the baseline, we have 12 combinations in total for comparison.

Focal loss [178]. Focal loss was first proposed for dense object detection to deal with large class imbalance problems. To avoid easily classified negatives that comprise most of the loss and dominate the gradient, focal loss includes a modulating factor $(1-p_t)^{\gamma}$ to cross-entropy loss. In practice, the focal loss sometimes is adopted in the α-balanced form:

$$FL(p_t) = -\alpha_t(1-p_t)^{\gamma}\log(p_t) \tag{8.1}$$

where p_t is the predicted probability to match the correct class.

Note: Focal loss was proposed in ICCV 2017 (top-tier, H5 Index: 172) [178] and won the best paper, then was studied for calibration in NeurIPS 2020 (top-tier, H5 Index: 192) [185]. Until March, 2022, It has gotten 12,000+ citations.

Label-Aware Smoothing [183]. The standard cross-entropy loss is defined as: $H(y,p) = \sum_{k=1}^{K} -y_k \log(p_k)$, where y_k is "1" for the correct class and "0" for the rest. For label smooth-

ing, $y_k^{LS} = y_k(1-\alpha) + \alpha/K$. To improve calibration for long-tailed recognition, Zhong et al. [169] propose the Label-aware Smoothing (LAS) as below:

$$y_k^{LAS} = y_k(1 - f(N_y)) + \frac{f(N_y)}{k}, \tag{8.2}$$

large N_y has a larger smoothing factor.

Note: Label-smoothing was proposed and studied in CVPR 2016 (top-tier, H5 Index: 301) [182] and specially studied in NeurIPS 2019 (top-tier, H5 Index: 192) [183] and ICML 2021 (top-tier, H5 Index 151) [184], then was studied for calibration in CVPR 2021 (top-tier, H5 Index: 301) [169]. Until March, 2022, it has gotten 19,000+ citations.

Mixup [179]. Mixup strategy is motivated by the Vicinal Risk Minimization [186]. As in [179], the vicinal points $(\tilde{x}, \tilde{y})$ are generated according to the following rules:

$$\begin{aligned} x &= \lambda x_i + (1-\lambda)x_j \\ y &= \lambda y_i + (1-\lambda)y_j \end{aligned} \tag{8.3}$$

where x_i and x_j are randomly selected sample pairs. y_i and y_j are related one-hot labels of x_i and x_j. $\lambda \in [0,1]$ and $\lambda \sim beta(\alpha, \alpha)$ for $\alpha \sim (0, \infty)$. The mixup strategy incorporates the prior knowledge that linear interpolation of inputs should linear interpolation of outputs.

Note: mixup was proposed in ICLR 2017 (top-tier, H5 Index: 166) [179], then was studied for calibration in NeurIPS 2019 (top-tier, H5 Index: 192) [180]. Until March, 2022, it has gotten 3400+ citations.

Temperature Scaling [164]. Platt scaling is a parametric approach for model calibration, which learns two parameters a,b to achieve a new prediction $q_i = \sigma(a \cdot z_i + b)$ given the original prediction z_i. Temperature scaling is a simple and straightforward extension of Platt scaling, which only has one learnable parameter. In the multi-class classification context, given the logit vector z_i, the new prediction is obtained as follows:

$$\hat{q}_i = \max_k \sigma_{SM}(\frac{z_i}{T})^{(k)} \tag{8.4}$$

where the parameter T is called temperature and σ_{SM} is the Softmax function. Since both platt scaling and temperature scaling do not change the ranking of predictions, so they do not affect the discrimination of the overall model.

Note: temperature scaling was studied for calibration in ICML 2017 (top-tier, H5 Index: 192) [164]. Until March, 2022, it has gotten 2200+ citations.

8.3 Evaluation Metrics

Expected Calibration Error (ECE). ECE [187] is one of the most popular metrics to measure calibration performance. There are two main steps to compute ECE: (1) dividing the prediction value space into equal-space bins, (2) calculating the weighted average of the difference of accuracy and confidence. The ECE is defined as:

$$ECE = \sum_{m=1}^{M} \frac{|B_m|}{n} |acc(B_m) - conf(B_m)| \tag{8.5}$$

where the n is the number of samples, B_m is the m-th bin.

Accuracy (ACC). Accuracy is a metric that measures the fraction of model predictions get right.

$$ACC = \sum_{i=1}^{N} \mathbf{1}\{\hat{y}_i == y_i\}/N \tag{8.6}$$

where $\hat{y}_i$ and y_i is the predicted label and ground truth for sample i, respectively. N is the number of samples.

Area Under the ROC Curve (AUC). ROC curve (receiver operating characteristic curve) is a figure showing the performance of a prediction/classification model at all thresholds. The curve plots true positive rate as y-axis and false positive rate as the x-axis. The AUC represents the area of the two-dimensional region under the whole ROC curve from (0,0) to (1,1).

Reliability Diagrams. Reliability diagrams [42] are the figures of the predicted probabilities (confidence) and the observed frequency (expected frequency), which shows the frequency of predicted probabilities that happened in practice (that is, calibration). Examples can be found in Figure 8.2 and Figure 8.3.

8.4 Experiments on CIFAR10

8.4.1 Data Introduction

The CIFAR10 consists of 60,000 tiny images evenly from 10 classes. There are originally 50,000 samples in the training set and 10,000 samples in the test set. We further split the original training set into training/validation splits with the ratio of 9:1. The validation set is used for tuning hyper-parameters and selecting epoch numbers.

The original CIFAR10 [136] dataset is balanced, and we term it as CIFAR10-Ori in this chapter. Following the setting of [169], we create a variant of CIFAR10-Ori with long-tail distribution called CIFAR10-LT with the imbalance factor (IF) 0.01 for the training set. The number of samples in class i is defined as $N_i = N_{ori} \cdot IF^{i/9}$, where N_{ori} is the sample number in CIFAR10-Ori. The validation and test sets of CIFAR10-LT are kept the same as CIFAR10-Ori.

8.4.2 Implementation

Our backbone model is highly motivated by [169]. We utilize the ResNet-32 as our backbone network. The training process has up to 2 stages. In Stage 1, the feature extraction model ResNet-32 is trained from scratch with Focal loss or cross-entropy loss (CEL), and we also have the comparison of with mixup and without mixup. For Stage 2, the feature extraction model is fixed and only trains the classifier (e.g., last layer) or scales the logits. We include the LAS [169] and Temperature Scaling for comparison. The experiments are conducted on both CIFAR10-Ori and CIFAR10-LT.

We use the SGD optimizer with a momentum of 0.9 and a base learning rate (BLR) of 0.1. In the first five epochs, the learning rate is $BLR * epoch_{index}/5$. The learning rate is divided by 10 at 150th and 250th epoch. The max number of epochs is 300. The batch size is set to 128 and the weight decay is 5e-4. Our experiments are conducted with PyTorch 1.6. The hyper-parameter α in mixup is set to 1 for CIFAR10-LT and 0.2 for CIFAR10-Ori. The hyper-parameters in focal loss are $\alpha = 1, \gamma = 2$. The number of bins when computing the ECE is 15.

8.4.3 Results and Analyses

Herein, we start with an easier setting of balanced training and then dive into a harder one with imbalanced training.

Balanced CIFAR10-Ori Results.

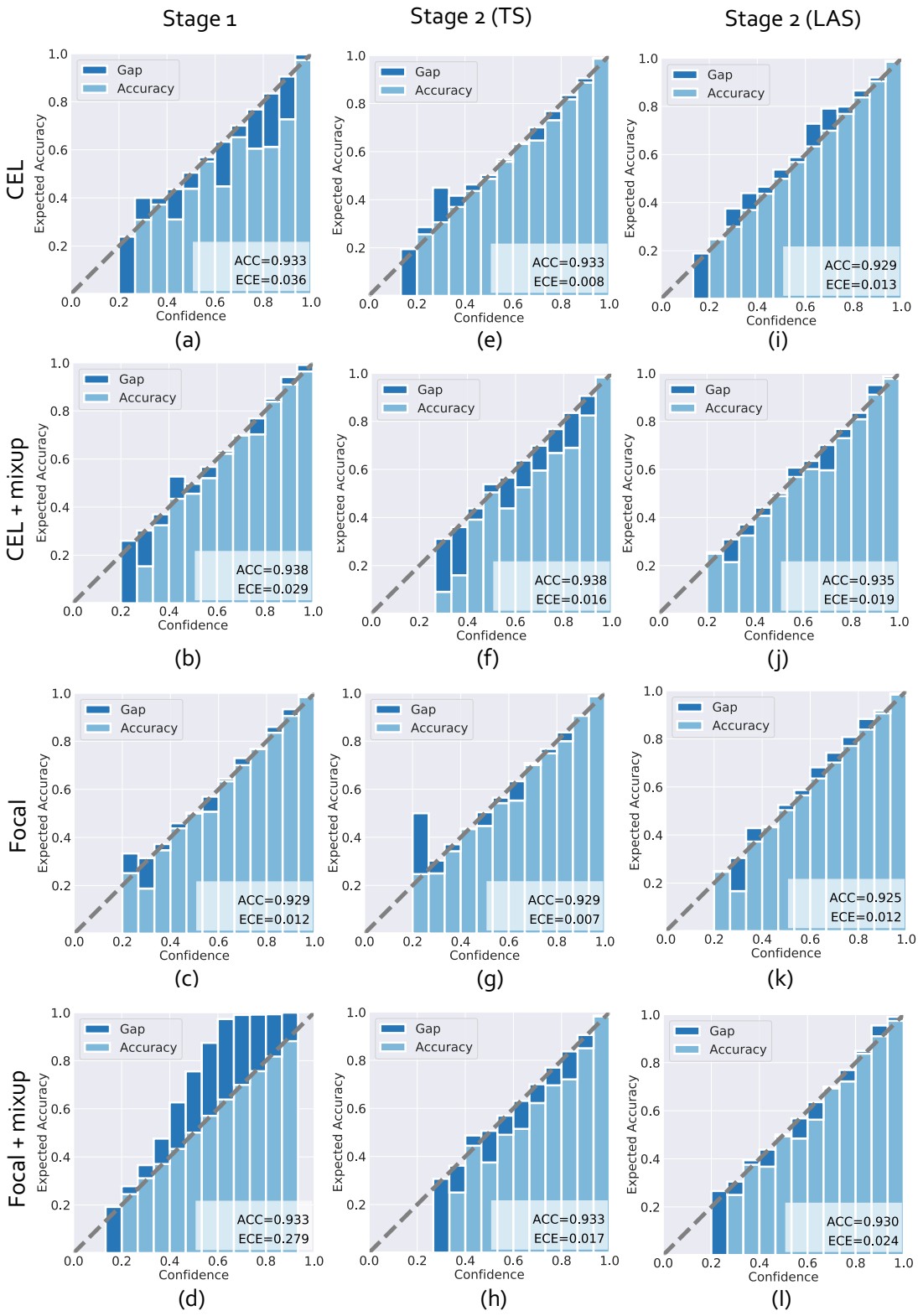


Figure 8.2: Reliability Diagrams of experiments on CIFAR10-Ori. This figure corresponds to the results in Table 8.1.

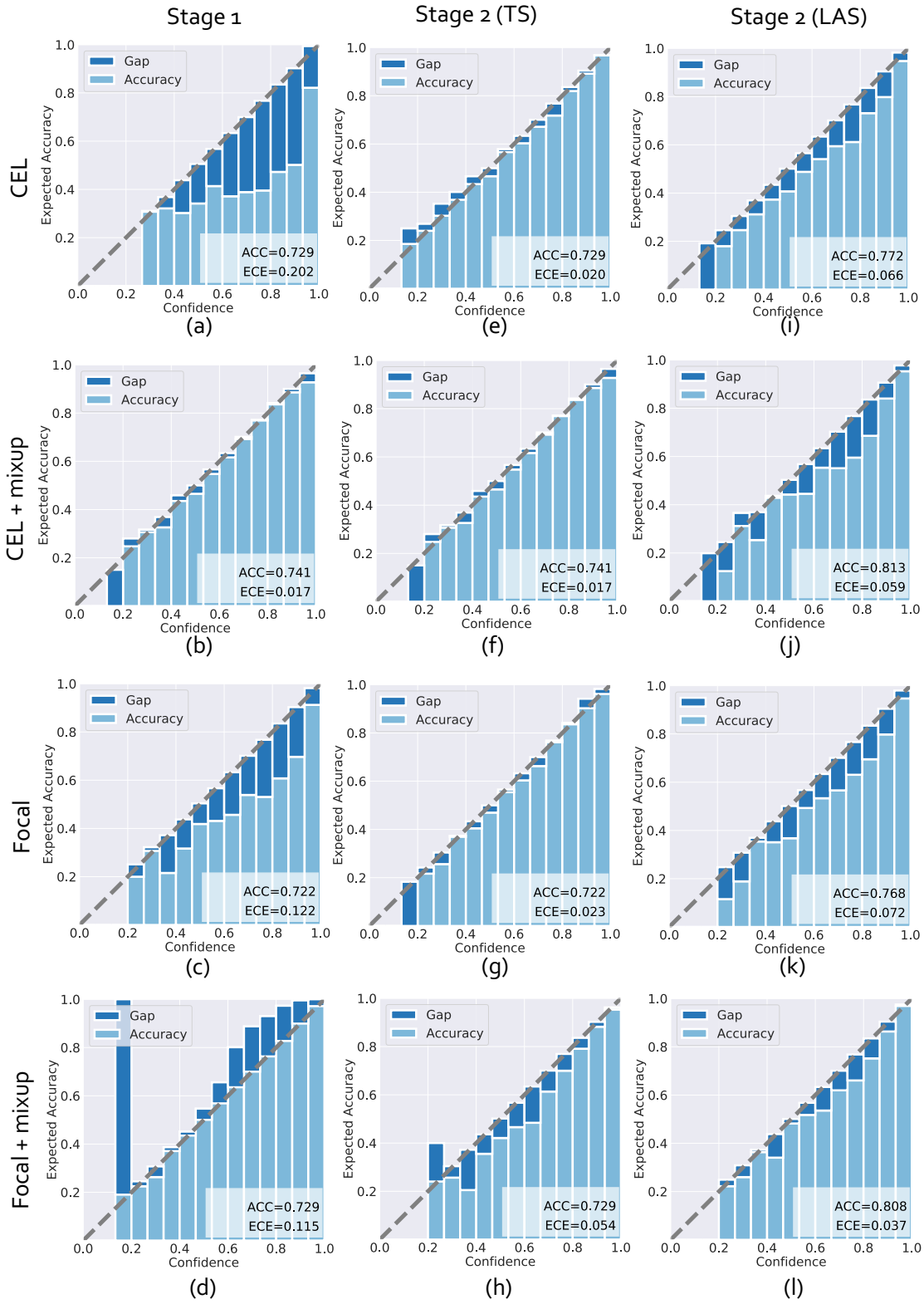


Figure 8.3: Reliability Diagrams of experiments on CIFAR10-LT. This figure corresponds to the results in Table 8.2.

Stage 1 Model	Stage 1		Stage 2 (TS)		Stage 2 (LAS)	
	ACC	ECE	ACC	ECE	ACC	ECE
CEL	93.32	3.60	93.32	0.75	92.86	1.32
CEL + Mixup	93.83	2.86	93.83	1.57	93.54	1.87
Focal	92.95	1.22	92.95	0.65	92.55	1.23
Focal + Mixup	93.26	27.92	93.26	1.66	93.00	2.40

Table 8.1: Test results of training with CIFAR10-Ori (%). TS represents temperature scaling and LAS represents label-aware smoothing.

Stage 1 Model	Stage 1		Stage 2 (TS)		Stage 2 (LAS)	
	ACC	ECE	ACC	ECE	ACC	ECE
CEL	72.90	20.20	72.90	2.02	77.19	6.64
CEL + Mixup	74.10	1.72	74.10	1.72	81.27	5.88
Focal	72.18	12.22	72.18	2.32	76.83	7.19
Focal + Mixup	72.93	11.48	72.93	5.37	80.82	3.69

Table 8.2: Test results of training with CIFAR10-LT (%). TS represents temperature scaling and LAS represents label-aware smoothing.

Table 8.1 and Figure 8.2 show the experiments on CIFAR10-Ori. From the results of Stage 1, the discrimination performance (ACC) is similar across four settings. Compared with the baseline CEL, the strategy of mixup slightly improves the ACC while improve calibration performance with a clear margin (ECE: $3.60\% - > 2.86\%$). Contrastively, replacing the CEL with Focal loss slightly decreases the ACC and improves the ECE with a large margin (ECE: $3.60\% - > 1.22\%$). Interestingly, when adding both the mixup and Focal loss, the calibration performance is surprisingly poor.

When applying the temperature scaling (TS) in Stage 2 for recalibration, the class discrimination performance does not change, and the confidence calibration are improved with a large margin across all the four settings. When using the label aware smoothing (LAS), the ACCs are only slightly changed. Generally, LAS improves calibration performance in Stage 2 but it is less impressive compared with TS.

Balanced CIFAR10-Ori Analyses.

In Stage 1 comparison, when trained by CEL, the predicted risks are vastly overconfident. Both the mixup and focal loss can relief the overconfidence of CEL, which can result to better calibration. However, when applying both mixup and focal loss, the prediction will be underconfident, which lead to worse calibration performance. As for class discrimination, there is no clear impact found

(ACC changes $< 1\%$) from mixup and focal loss.

Temperature scaling does not change discrimination as it only scales the logits and will not change the logits ranking of different classes, which is consistent through all experiments. As the second stage post-processing methods, both LAS and temperature scaling make a contribution in confidence re-calibration under this balanced training setting.

Imbalanced CIFAR10-LT Results.

The experiments of training with CIFAR10-LT are shown in Table 8.2 and Figure 8.3. Compared with CIFAR10-Ori, the discrimination performance is much lower ($\sim$20% ACC decreased).

In Stage 1, mixup can improve the calibration performance (ECE) greatly and improve the discrimination performance slightly (ACC). The Focal loss can improve the baseline model CEL in terms of ECE, while increase the ECE on the top of the mixup (i.e., ECE: CEL + mixup (1.72%) vs. Focal Loss + mixup (11.48%)).

The LAS can improve both the calibration and discrimination in general (except calibration in the setting of CEL + mixup).

Imbalanced CIFAR10-LT Analyses.

As the imbalanced training nature, the patterns of tailed classes have limited diversity in the training set, which weakens the model's generalization ability and results to lower discrimination performance. Also, the difference of class frequency in training can lead model overfit to head classes (i.e., the classes with more samples).

Like the analyses of CIFAR10-Ori, mixup, Focal loss and even the second stage model LAS are approaches relief the overconfidence, which can improve the calibration performance. However, combining mixup and Focal loss should be careful.

8.5 Experiments on Lung Cancer Datasets

8.5.1 Data Introduction

We have included three datasets in this experiment: national lung screening trial (NLST) [4], two cohorts UPMC, and UCD from MCL [188].

NLST. The details of the whole NLST are in Chapter 1. Here, we use the subjects if 1) have a tissue-based diagnosis, and 2) the diagnosis happened within 1 year of the last scan if it is a cancer case, 3) passed image quality check. There are in total 606 cancer subjects out of 5344 subjects.

The whole NLST dataset is randomly split into five even folds. Four folds are used for training and one-fold for validation.

UPMC and UCD. Different NLST which is a screening cohort and has a low cancer rate (0.1) after our inclusion criteria, UPMC and UCD cohorts are dominated with incidental cases and have much a higher cancer rate ($\sim$ 0.5). There are 78 cancer cases out of 155 patients in UPMC and 52 cancer cases out of 96 patients in UCD.

8.5.2 Implementation

We have included two methods (i.e., CEL and Focal loss) in Stage 1 and two re-calibration methods (i.e., TS and LAS) in Stage 2 for comparison. Mixup strategy is designed for single image classification, while in the lung cancer diagnosis experiment, we use multi-instance learning by feed multiple nodule proposals. We drop the mixup technique here as the multi-instance learning context is not directly match the mixup intuition.

The neural network backbone is motivated by the image branch of our previous work [189]. We first apply the preprocessing steps and nodule detection model of [37] to raw CT image data. As suggested by [37], the top five confidence nodule proposals are enough to cover all the nodules. To speed the training and test process, we use 2D images rather than 3D. Axial/coronal/sagittal directions of the nodule proposal are formulated as 3 channel data with the dimension of ($3 \times 128 \times 128$). We use a ResNet-18 backbone to extract the image feature of each nodule proposal. Then, the image features of the five nodule proposals are transferred to a single image feature vector with an attention-based multi-instance learning layer.

We use the SGD optimizer with a momentum of 0.9 and a learning rate of 0.005. The max number of epochs is 100 and the final model is used for testing as we empirically notice that the training is converged after the 100th epoch in our setting. The batch size is set to 128 and the weight decay is 1e-4. Our experiments are conducted with PyTorch 1.6. The hyper-parameters in focal loss are $\alpha = 1, \gamma = 2$. The number of bins when computing the ECE is 10.

8.5.3 Results and Analyses

The results on NLST, UCD, and UPMC are shown in Table 8.3, Table 8.4, and Table 8.5, respectively. The Focal loss has improved discrimination over CEL, which has been observed across all

Stage 1 Model	Stage 1		Stage 2 (TS)		Stage 2 (LAS)	
	AUC	ECE	AUC	ECE	AUC	ECE
CEL	90.71	1.72	90.71	1.72	90.79	7.43
Focal	91.53	16.16	91.53	1.54	91.54	7.39

Table 8.3: Internal-validation results on NLST (%). TS represents temperature scaling and LAS represents label-aware smoothing.

Stage 1 Model	Stage 1		Stage 2 (TS)		Stage 2 (LAS)	
	AUC	ECE	AUC	ECE	AUC	ECE
CEL	74.50	10.10	74.50	6.73	74.78	31.03
Focal	75.20	11.13	75.20	13.50	75.05	28.37

Table 8.4: External validation results on UCD. TS represents temperature scaling and LAS represents label-aware smoothing.

three datasets (e.g., 91.53% vs. 90.71% in NLST, 87.20% vs. 83.50% in UPMC). The temperature scaling generally improves the calibration performance across the losses of CEL and Focal loss. The temperature scaling does not change the discrimination as it only scales the logits and does not change the ranking. In the major situations, the LAS does not improve the calibration performance but can improve the discrimination minorly.

8.6 Discussion

In this chapter, we systematically conduct a comparative study on confidence calibration and class discrimination with four representative and widely used calibration approaches. Those four approaches include one-stage and two-stage methods, and we have 12 method combinations for comparison. Started with an easy setting with balanced training with large-scale training set on natural image classification (i.e., CIFAR10-Ori), then we extend the analyses to imbalanced training setting but still with large-scale training set (i.e., CIFAR10-LT). Furthermore, as the clinical practice task, lung cancer diagnosis is studied in the context cross- and external- validations with three data

Stage 1 Model	Stage 1		Stage 2 (TS)		Stage 2 (LAS)	
	AUC	ECE	AUC	ECE	AUC	ECE
CEL	83.50	12.60	83.50	6.53	83.70	24.23
Focal	87.20	10.50	87.20	7.97	87.20	19.89

Table 8.5: External validation results on UPMC. TS represents temperature scaling and LAS represents label-aware smoothing.

cohorts. In summary, our contributions are as follows:

- We strengthened some conclusions draw by previous studies with more general and difficult settings. The temperature scaling maintains the discrimination improves confidence calibration in general including balanced- vs. imbalanced, nature image vs. medical image et cetera. Focal loss and label-aware smoothing improves the calibration of natural image classification in both balanced and imbalanced setting (large-scale training set). The label-aware smoothing even improves the discrimination with a large margin in CIFAR10-LT.

- We have interesting new findings that are not well-studied in previous studies. For example, in balanced training with CEL, the prediction is overconfident and lead to large calibration error. Both the mixup and focal loss improves the calibration performance by relieving the overconfidence of the prediction. However, combining focal loss and mixup will lead to under confidence and then damage the calibration.

- We have some contradictive findings when comparing the results of nature image and medical imaging. The calibration performance of both Focal loss and label-aware smoothing are contradictive between our medical diagnosis and nature image classification, which highlight that transfer knowledge from computer vision domain to medical imaging requires care considering the application contexts are usually different. As a simple but robust approach, temperature scaling improves the calibration performance in general, not only nature image classification but also the internal- and external- validation in the lung cancer diagnosis, which reminder us to put a high priority on "simpler" approaches when transferring knowledge from general computer vision to medical imaging.

CHAPTER 9

Cancer Risk Estimation Combining Lung Screening CT with Clinical Data Elements

9.1 Introduction

Manually human-curated imaging semantic features (e.g., nodule size) and CDEs have been integrated for cancer risk estimation [31, 114, 190]. The Brock University PanCan study (called the Brock model) [114] is a logistic regression model incorporating imaging semantic features and CDEs for cancer risk estimation after nodule discovery. However, the acquisition and management of imaging semantic features require manual efforts from radiologists. In addition, most individuals participating in lung cancer screening programs do not have lung nodules large enough to be documented. The potential for missing information makes it difficult to compute the lung cancer risk for these patients using methods that rely on imaging semantic features (e.g., Brock model [31]), and may therefore miss early-stage cancer. Several lung cancer CT screening studies, including the NLST and European NELSON trial, was used to derive positivity criteria for Lung-Reporting and Data System (Lung-RADS) [191, 192]. In the NLST population, the cancer rate in Lung-RADS 1 and Lung-RADS 2 is lower than 1%, while these rates were greater than 1% for Lung-RADS 3 and greater than 10% for Lung-RADS 4.

Deep learning techniques are transforming the medical imaging field [100] as the result of the success in general computer vision [49, 98, 193, 194].

We hypothesized that CT and CDEs provide complementary information for lung cancer risk estimation. In this study, we developed a model to integrate CT image feature and CDEs in a unified machine learning framework. Specifically, we adopted deep learning techniques to extract quantitative imaging features and to train a co-learning deep learning model end-to-end by inputting CT imaging features and CDEs. The preprocessing and feature extraction of CT images were adapted from Liao's work [37] and the CDE selection by the PLCOm2012 model [29]. We evaluated our method using cross-validation on the NLST dataset and performed external testing using data from the Vanderbilt Lung Screening Program (VLSP).

Screening Program	NLST	VLSP
No. patients	23505	147
No. patients with cancer	722 (3.2%)	21 (14.2%)
No. scans	64898	220
No. scans with cancer	1037 (1.6%)	40 (18.1%)
Age, years	62 ± 5	65 ± 5
No. men/women	13838/9667	82/65
BMI, kg/m2	28.07 ± 5.01	28.28 ± 5.68
COPD	1196 (5.1%)	41 (27.9%)
Personal cancer history	972 (4.1%)	30 (20.4%)
Family lung cancer history	5103 (21.7%)	38 (25.9%)
Tobacco use status (former, current)		
Former	12321	60
Current	11184	87
Pack years (mean, standard deviation)	55.58 ± 23.12	48.94 ± 19.82
Tobacco use quit time, years	4.66 ± 5.62	3.33 ± 6.22
Education		
Less than high school	6904 (29.4%)	6 (4.1%)
High school graduate or GED	3309 (14.1%)	29 (19.7%)
Post-high school training, excluding college	5450 (23.2%)	5 (3.4%)
Associate's degree	4002 (17.0%)	37 (25.2%)
Bachelor's degree	3386 (14.4%)	35 (23.8%)
Graduate	425 (1.8%)	35 (23.8%)
Race		
White	21801 (92.8%)	134 (91.2%)
Black	1030 (4.4%)	12 (8.2%)
Asian	518 (2.2%)	0
Pacific	82 (0.3%)	0
Latino	0	1 (0.7%)
Indian	74 (0.3%)	0

Table 9.1: Demographics of NLST and VLSP used in this study (scan-level). Values shown as either n (percent) or mean ± standard deviation. COPD = chronic obstructive pulmonary disease

9.2 Method

9.2.1 Patient Selection

The in-house program with existing de-identified data that was performed in accordance with the Health Insurance Portability and Accountability Act (HIPAA) with approval fron our Institutional Review Board (IRB, #181279). Two screening cohorts were used in this study. The NLST dataset is a large-scale randomized controlled trial for early diagnosis of lung cancer, which was conducted in the United States wherein approximately 54,000 participants were enrolled between August 2002 and August 2004. Our Vanderbilt Lung Screening Program (VLSP) is a comprehensive program

Screening Program	NLST	VLSP
Age	55-74 years old	55-80 years old
Tobacco use status	Current or former tobacco use	Current or former tobacco use
Pack years	≤ 30	≤ 30
Quit Smoking Time	Quit tobacco use ≤ 15 years	Quit tobacco use ≤ 15 years
Specific exclusion criteria	Prior lung cancer Chest CT within 18 months Hemoptysis Unexplained weight loss ≥ 15 lbs prior year	Cancer diagnosis within past 5 years current surveillance with chest CT Signs or symptoms of lung cancer pneumonia at the time of screening

Table 9.2: Inclusion and Exclusion Criteria in NLST and VLSP

that offers annual lung screening CT and management by the Department of Radiology of our University Medical Center. Patients included in the study from VLSP were enrolled between 2015 and 2018 and written informed consent from patients is waived by IRB. The demographics of NLST and VLSP included in this study are shown in Table 9.1. The NLST and VLSP have comparable screening eligibility criteria (Table 9.2).

Motivated by the PLCOm2012 model [29], CDEs included in our model are those data elements defined in PLCOm2012: age, education, body mass index (BMI), personal cancer history, family lung cancer history, tobacco-use status, tobacco-use quit time, and pack-years.

In addition to limiting inclusion to patients who meet the eligibility criteria defined in Table 9.2, we included samples (64,898 scans; Table 9.1) for which we were able to successfully obtain source data and where imaging data met the following criteria: (a) meeting quality standards (https://github.com/MASILab/QA_tool) [195], (b) successful preprocessing as described by Liao et al. [37], and (c) meeting the criteria of defining positive or negative case. A positive case was defined as a biopsy-confirmed diagnosis of lung cancer within two years of the imaging date. A negative case was defined as a biopsy that was not consistent with a lung cancer diagnosis or stable radiographic findings for two or more years.

In our in-house dataset, we excluded 740 patients (with CT) who do not have definitive confirmation of cancer status or with data missing (many of these are expected to be negative cases). Thus, the cancer rate of in-house is different from general screening cohorts.

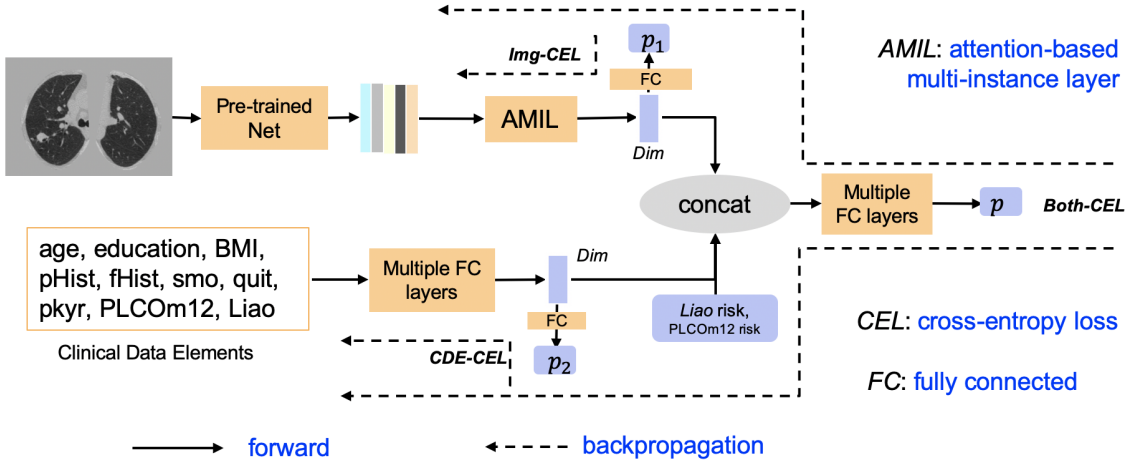


Figure 9.1: Our proposed co-learning framework. The "Pre-trained Net" is illustrated in Figure 9.1. We apply an attention-based multi-instance layer to combine the features from top five nodule proposal. Finally, the clinical factors are concatenated with the image feature and followed with fully connected layers for final prediction.

9.2.2 CT Acquisition

The NLST data information can be found in [4]. For CTs in VLSP, the slice thickness (0018, 0050) is 1.0, the kilovoltage peak (0018, 0060) is 120.0, data collection diameter (0018, 0090) is 500.0, the reconstruction diameter (0018, 1100) is 447.0, the acquisition type (0018, 9302): 'SPIRAL', the manufacturer (0008, 0070) is 'Philips', the scan options (0018, 0022) is 'HELIX'.

9.2.3 Algorithm Design

The framework of our co-learning model was adapted from the previous work [93], as shown in Figure 9.1. The image preprocessing and pulmonary nodule detection are shown in Chapter 2. The input of the framework is a CT image and associated CDEs from a patient, and the output is a predicted lung cancer risk. As described in the previous section, the five proposals with top confidence score were obtained with a nodule detection model. Each nodule proposal was further converted to a 1×128 dimension feature vector. Patients were considered to have lung cancer if any nodule was malignant; therefore, the lung cancer prediction was form as a multiple instance learning problem where nodule proposals are instances [37]. In a multiple instance learning model, the input is multiple instances from a sample and the output is the prediction of the sample. The sample is labeled as positive if any of the instances are positive, otherwise, the sample is negative when all the instances are negative.

Methods	AUC	AUPRC
No Skill+	0.50	0.016
PLCOm2012	0.69 ± 0.02	0.038 ± 0.006
Brock Model	0.84 ± 0.01	0.27 ± 0.03
Liao Model	0.86 ± 0.02	0.32 ± 0.02
Ours	0.88 ± 0.02	0.34 ± 0.02

Table 9.3: AUC and AUPRC on the NLST Validation Dataset. the AUC of Brock model is computed by padding the default values (nodule size: 2 mm, speculation: no, upper lobe: no, nodule type: nonsolid) when factors are not available. Note that only nodule size smaller than 4 mm are missing in the NLST dataset. No Skill represents predicting without any knowledge, equivalent to random guessing

9.2.4 Model Comparisons

We compared our co-learning model with the popular CDE-only method PLCOm2012 [29] and the image-only method Liao et al [37]. The PLCOm2012 model uses 11 different epidemiological CDEs for risk prediction. The Brock model was also included for comparison with appropriate imputation on those patients who miss imaging semantic features. The Liao model takes only the CT image as the input. We applied the pre-trained model of [37] that won the Kaggle challenge [196] for comparison (termed as Liao model in the following). In NLST, the nodule annotations are missing when the nodule size is less than 4 mm. The VLSP dataset did not contain nodule annotations when the lung scan was categorized as Lung-RADS 1 or 2.

9.2.5 Data Imputation for Brock Model

We imputed the missing values according to the criteria of Lung-RADS [197] and clinical experiences. We imputed the scan in Lung-RADS 1 as {nodule size: 0 mm, spiculated: no, nodule type: solid}, and Lung-RADS 2 as {nodule size: 3 mm, spiculated: no, nodule type: solid}. The missed "upper lobe" variable was achieved by a logistic regression trained with the available ones in NLST. To test if the model prediction was sensitive to imputed values, we included different imputation combinations for the Brock model. The nodule count was set as 1 when computing the Brock model. Note that the imputed values were only applied when computing the Brock model, as other compared methods (including ours) did not need human-curated radiomic features.

Imputed Nodule Size [Lungrads 1, Lungrads 2]	Brock model	Ours
$[0mm, 2mm]$	0.80 (0.71, 0.89)	0.91 (0.85, 0.95)
$[0mm, 3mm]$	0.78 (0.68, 0.88)	0.91 (0.85, 0.95)
$[1mm, 3mm]$	0.78 (0.68, 0.88)	0.91 (0.85, 0.95)
$[1mm, 4mm]$	0.78 (0.67, 0.88)	0.91 (0.85, 0.95)

Table 9.4: Comparison of Imputed Values for the Developed Model Compared to the Brock Model on the External Testing VLSP Dataset. Values are shown with 95% CIs.

9.2.6 Cross-validation on NLST

We randomly split the NLST cohort (in Table 9.1) into five folds. Five-fold cross-validation (details are in Appendix E1 (supplement)) was applied in NLST dataset, as shown in Table 9.3. In each fold of the evaluation, 20% of the cohort (i.e., one-fold) were held out from training as the test set, and the remainder of the data were split as 3:1 for training and tuning. The splitting was random. The model selection was based on the tuning set.

9.2.7 External Testing on the VLSP

To test the generalizability of our model, external validation was performed (i.e., the model was trained with NLST and tested with VLSP). In external testing, we applied the same models developed during the cross-validation training on NLST without fine-tuning on VLSP. The prediction on VLSP was based on the average of the predictions of the five-fold models.

In the VLSP dataset, nodule annotations were not reported for nodules smaller than 6 mm in diameter. We compared performance when imputing other nodule diameter values in Table 9.4. We included four more combinations of imputed nodule size values for Lung-RADS 1 and Lung-RADS 2 and the imputed value combinations were based on the definition of Lung-RADS. As with the NLST, the nodule count was set as 1 when computing the Brock model. The nodule size, spiculation, location of the primary nodule (upper lobe) is imputed with last CT records of a patient if values from the current CT were missing and values from the last CT are available, which may happen because some patients from Lung-RADS 3 or 4 can change to Lung-RADS 2 if the nodule was stable. The ROC and PRC curves and their AUC values were computed with averaging predicting of five models from different folds.

9.2.8 Statistical Analysis

The predicted performance was evaluated by the area under the receiver operating characteristic curve (AUC) and the area under the precision-recall curve (AUPRC) and their corresponding 95% CI. We show both bootstrapped 95% CIs and P-value for indicating a statistical difference. The bootstrapped two-tailed test and DeLong test [198] were used to compare the performance between the different models. No different conclusions were observed between those two tests (ie reported $P < .05$ values were verified with both bootstrapped two-tailed test and DeLong test). The computation of bootstrapped two-tailed test and 95% CIs were adapted from: https://github.com/mateuszbuda/ml-stat-util.

9.3 Results

9.3.1 Patient Overview

Two datasets are studied in this work: the NLST and VLSP. A total number of 23505 patients with 64898 CT scans are included in the NLST (average age: 62, men/women: 13838/9667). The VLSP contains a total number of 147 patients with 220 scans (average age: 65, men/women: 82/65).

9.3.2 Model Performance

The model AUCs and AUPRCs are reported in Table 9.3. Our method had a higher performance (AUC, 0.878) than the Liao model (image-only; AUC, 0.864; $P < .05$) and other clinical models including PLCOm2012 (AUC, 0.692; $P < .05$) and Brock model (AUC, 0.845; $P < .05$).

9.3.3 External Testing on VLSP

The model AUCs and AUPRCs are shown in Figure 9.2 and Table 9.4. We found that our model had a higher performance (AUC, 0.905) compared to Liao models (AUC, 0.881; $P < .05$) and Brock model (AUC, 0.782; $P < .05$). These results were found even as our model used different imputed values where the nodule characteristic data elements were missing. In addition, two examples (one cancer and one non-cancer) are shown in Figure 9.3 with predicted cancer probabilities of different methods.

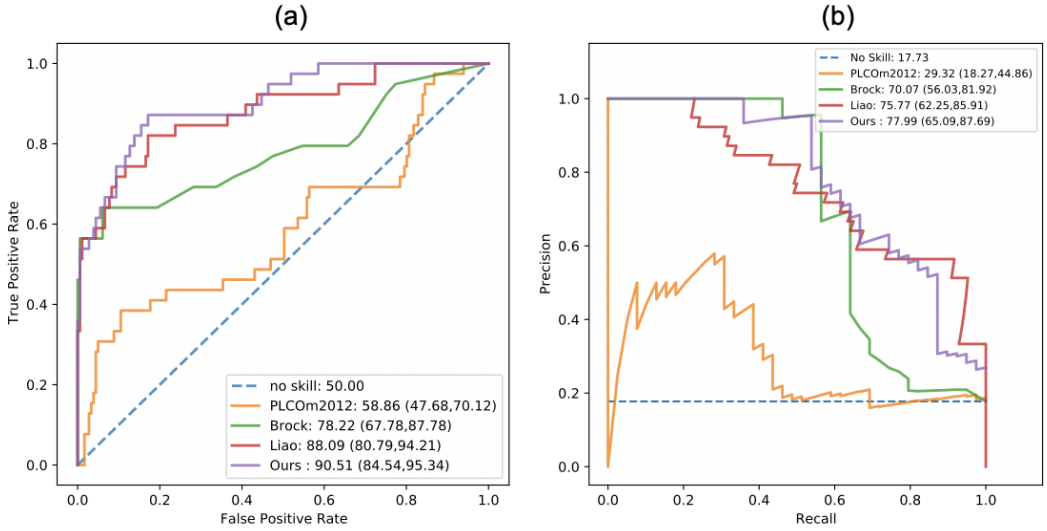


Figure 9.2: (a) Area under the receiver operating characteristic curve (AUC) and (b) area under the precision-recall curve (AUPRC) on the external VLSP test dataset. (b) The AUC and AUPRC of Brock model were computed by imputing the default values (nodule size: 0 mm for Lung-RADS 1 and 3 mm for Lung-RADS 2, spiculation: no, type: solid, upperlobe achieved by logistic regression) when data from the patients were not available.

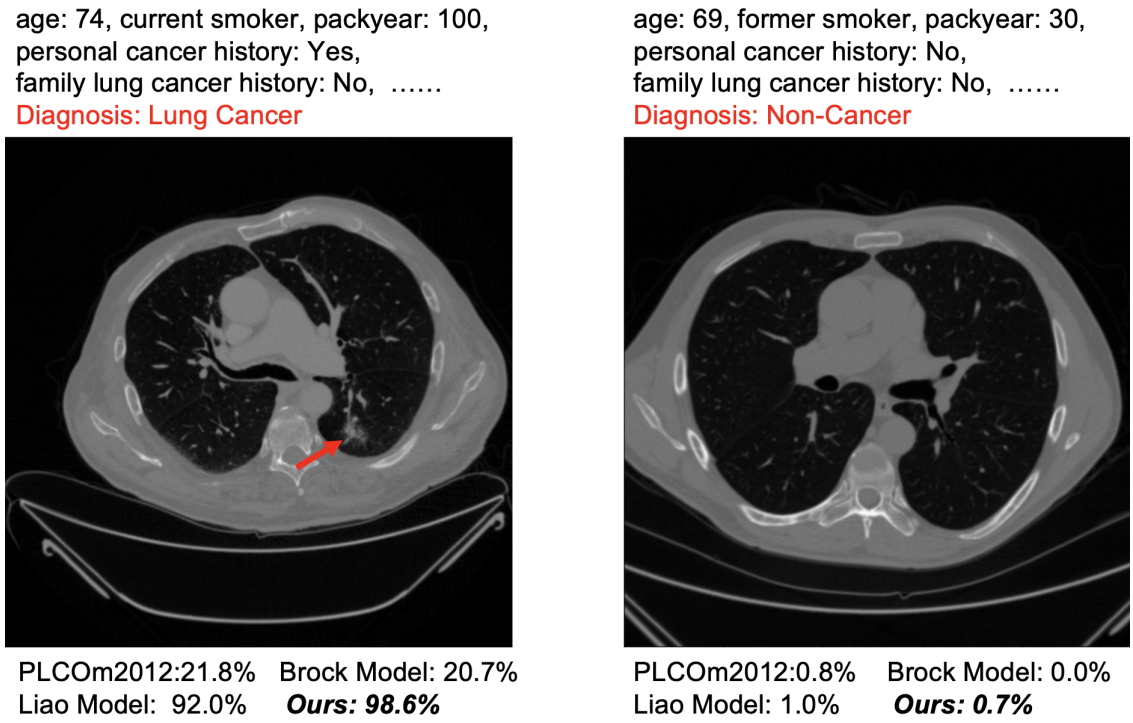


Figure 9.3: The Left: a cancer case, the Right: a non-cancer case. The clinical data are shown in the upper of each case and the predicted cancer risks are shown in the bottom. Our prediction is calibrated with a sigmoid function which does not change AUC value.

9.4 Discussion

Risk prediction modeling can help clinicians make more informed decisions regarding invasive procedures and diagnostic testing. We hypothesize that providing additional information (e.g., predicted risk from artificial intelligence) may support subsequent evaluation with physiologic imaging (i.e., PET/CT or tissue sampling with biopsy) while a lower risk would support short-term follow-up with CT imaging. All three of these courses are currently acceptable for patients with suspicious pulmonary nodules on screening (Lung-RADS 4). Current American College of Chest Physicians guidelines suggest further testing when the risk for lung cancer is 5%-65% and referral for a tissue diagnosis above 65% cancer risk [32].

The benefit of lung screening is in part due to annual follow up which is recommended for patients in both the Lung-RADS 1 and 2 categories. These categories are sometimes interchangeably used, particularly when patients have small nodules that remain stable on subsequent exams. Though small and with low potential of malignancy, these nodules still can be tested and used in developing machine learning models.

Our study demonstrates that a deep learning framework integrating CDEs, and CT imaging features can be helpful in lung cancer risk estimation. Risk estimation amongst lung screening participants will become even more important with the impending expansion of screening guidelines to include those patients who are considered lower risk only based on age and history of tobacco use. However, these patients may be identified as high-risk by using machine learning imaging algorithms of low-dose CT scans. When pulmonologists or thoracic surgeons care for patients with a positive CT scan, CDEs might be helpful to determine the best management for each patient. Motivated by the synergy between CT and CDEs, we included the CDEs in a multi-modal context with CT images for better risk estimation. Additionally, our model is flexible to extend by adding more elements (e.g., nodule size), if available. Thus, we believe a potential future direction for improving current standard Lung-RADS recommendations is considering predicted risk from a multi-modal deep learning model with standard input of CT images and clinical data.

In this work, we applied established machine learning techniques (e.g., multiple instance learning, multi-modality fusion) to improve computer-aided diagnosis. Our approach is automatic to extract high-risk regions (ie, nodule proposals) from CT without radiologist participation and integrate image features with CDEs by deep learning techniques. The false-positive nodules detected

are handled by multi-instance learning techniques. Different from radiomics models (e.g., Brock model) and Lung-RADS-based evaluation, the need for manual nodule segmentation and nodule characteristics (e.g., nodule size) extraction is not necessary for our pipeline. CDEs are usually collected from clinical decision-making visits and/or questionnaires, which also takes manual effort to collect and input them into systems. The role of radiologist is still irreplaceable in terms of looking for and reporting clinically significant findings (emphysema, pulmonary fibrosis, atelectasis, etc.).

Our model was evaluated by cross-validation with data from the NLST and external testing with an in-house dataset (VLSP). The results show our model higher performance compared to the image-only model and established risk calculators ($P < .05$ for both). The proposed model achieved +2.4% AUC values over the Liao model (image-only model) in external testing using the VLSP dataset, the expected number of patients would benefit from a better estimate would be 24 out of every 1000 CT scans. Improved discrimination performance for cancer versus non-cancer helps manage the patients for future treatment. We also find that adding more patients for fine-tuning an image-based model (see the Appendix E3 (supplement)) can increase prediction performance, which motivates us to use as much data as possible for our multi-modal model.

Some other methods, like the PLCOm2012, which uses CDEs as inputs for cancer risk estimation and selection criteria for screening, cannot distinguish between cancer and non-cancer effectively among the high-risk population. The potential reason is that CDEs only reflect the general "long-term" risk and PLCOm2012 do not have a data source (e.g., CT image) reflecting current symptoms.

Image-only methods are well developed because of the success of feature representation in machine learning and large acquisition of medical image data. In lung cancer risk estimation, deep learning methods are especially popular due to large publicly available datasets (e.g., LIDC-IDRI [24] and NLST [4]) and the promotion of challenges (e.g., Kaggle Data Science Bowl [199], LUNA16 [26]). DeepSEED [200] developed a 3D convolutional neural network with encoder-decoder structure, which achieves the state-of-the-art nodule detection performance on LUNA16 and LIDC-IDRI. Liao et al [37] won the Kaggle challenge with a 3D deep neural network that predicting cancer risk by inputting a whole CT image. Ardila et al [38] proposed that a deep learning algorithm with screening CT images can outperform radiologists. The image-based methods are automatic which may be important in busy imaging practice. While these methods ignore the

complementary information from CDEs even they are routinely obtained by lung cancer screening programs. Our model integrates the CDEs and CT image features in the end-to-end machine learning framework.

The Brock model incorporates the "current" clinical status of a patient as gleaned from curated lung nodule radiomic features along with the CDEs. Huang et al [114] encoded the human-curated radiographic feature and CDEs into a multiple layer perceptron network. The Brock model is widely used in lung cancer risk estimation (e.g., McWilliams et al [31]), yet one of the challenges is that it requires additional human effort to reliably annotate and report nodule features in a structured manner to be used in routine practice. Our model similarly combines the patient-level clinical features (CDEs and radiomic information), but extracts radiomic features directly from CT images in an efficient and reliably automated manner using a deep learning model.

Our study has several important limitations. We only used a single time point from each patient to predict lung cancer risk; prior CT records may be helpful for prediction. Integrating longitudinal information in this multi-modal context will be an interesting topic to explore in the next step. Second, we evaluated the "discrimination" of prediction models with the metrics AUC and AUPRC. We have not introduced another aspect, called "confidence calibration [164]", in clinical decision making, which refers to the agreement between observed probability and predicted risk. One of our future works is integrating uncertainty calibration in our framework, which would make the machine learning system more reliable. In addition, the currently available data in VLSP is limited compared with popular screening programs, such as the NLST. The potential impact on patient management and ultimately improvement in survival would be better evaluated with larger data sets and including more available clinical data elements, which we hope to address with future research.

In conclusion, we combined CDEs and CT images for lung cancer risk estimation in a unified machine learning model. Our risk prediction model had a higher performance compared with established risk calculators (PLCOm2012 and Brock model) and the image-based method (Liao et al.) on the NLST and VLSP datasets. Our model can be extended when more data elements are available.

Please refer to [201] for supplementary materials.

CHAPTER 10

Reducing Uncertainty in Cancer Risk Estimation for Indeterminate Pulmonary Nodules

10.1 Introduction

Lung cancer has the leading cancer death rate among all types of cancers [2]. The management of indeterminate pulmonary nodules (IPNs) suffers from wasting resources from benign disease and delaying the treatment for malignant cases. In the United States, there is an estimation of 1.5 million IPNs detected annually [6].

Currently, the management of IPN patients is mainly based on risks from clinical models such as the Mayo model [30]. High-risk patients can be recommended for biopsy, surgery, or other related treatment. Low-risk patients can be recommended to follow the general screening guidelines. The most challenging routine relies on the indeterminate groups which can be overtreated or have diagnosis errors [202]. To have improved care for IPN patients, reducing the indeterminate rate is essential.

There are works that have been done target for IPNs. For example, Mayo model is an established risk calculator for small radiologically indeterminate modules taking six clinical acquired features. Kammer et al. [203] proposed a combined biomarker for the management of indeterminate pulmonary nodules.

In this work, we extend our framework [93] for IPN management. Our model can integrate data from image and non-image modalities for a better risk estimation. We validate that the improved risk estimation with this deep learning model can used for IPN management without extra radiation exposure and delay treatment, as illustrated in Figure 10.1.

10.2 Method

10.2.1 Patient Selection

Four cohorts from MCL are used in our experiments. The data Vanderbilt University Medical Center (VUMC) are used as the learning cohort. Independent cohorts from the University of Colorado Denver (UCD), the University of Pittsburgh Medical Center (UPMC) and the Detection of Early

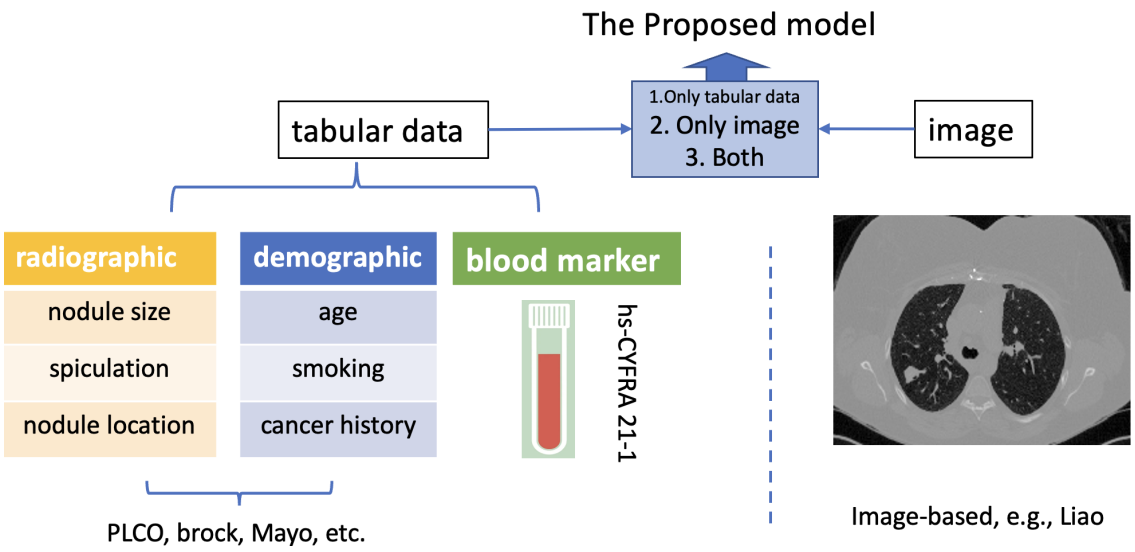


Figure 10.1: Flamework of our study. Some high impact established risk calculators (e.g., PLCOm2012, Brock, Mayo models) are based on radiographic and demographic risk factors. Image-based methods (e.g., Liao et al.) are feed with CT images using deep learning techniques. We combine the radiographic, demographic and blood risk factors as the non-image modality and propose a new machine learning framework integrating the image and non-image modality in end-to-end manner. Our model can predict lung cancer risk when 1) only tabular data are available, 2) only image data are available, or 3) both are available.

	Total Subject	With Image	With Clinical Data	With Both
# Subject	1284	1033	655	404
# Cancer	813	695	387	269

Table 10.1: The training cohort population from VUMC

Cancer Among Military Personnel (DECAMP) are served as external validation. For cancer patients, the final diagnosis was biopsy-proven. Non-cancer patients confirmed with biopsy-proven or with no evidence of growth based on at least two-year longitudinal imaging follow-up. The subjects from the external validation sets were incidentally found to have a nodule, and IPN represents the nodule that has the largest axial diameter between 6- and 30-mm. The data distribution of the learning cohort is shown in Table 10.1.

There are two modalities included: tabular clinical data and CT images. The tabular clinical data include radiographic features (nodule size, spiculation, nodule location), demographic elements (age, smoking status, body mass index, pack-year), and a blood marker (hs-CYFRA 21-1). For the learning cohort VUMC, we include the patient in the training set if at least one modality is available. For validation cohorts, we only include the patients with two modalities.

10.2.2 Data Preprocessing, Pulmonary Nodule Detection, and Network Structure

The network structure follows our proposed technique chapter [93], which has described in Chapter 6. Also, same as [93], our data preprocessing steps follow [37].

10.2.3 Evaluation Metrics

We use the Area Under the ROC Curve (AUC) to evaluate the discrimination of models, and bias-corrected clinical net reclassification improvement (cNRI) to compute the reclassification performance . We also show the reclassification confusion matrix with our method. We use cross-validation from the learning cohort VUMC. The cohort is randomly split into five folds. For the evaluation of each fold, four folds are used for training, and one-fold is used for validation. The other three cohorts are evaluated as external cohorts.

10.2.4 Model Comparison

We included three other models for comparison. Mayo model, Brock model and Liao model. Mayo model computes the probability of malignancy using a logistic regression of 3 clinical and 3 radiographic variables (age, smoking status, personal cancer history, nodule size, spiculation, upper lobe). Brock model estimates the cancer probability using a logistic regression of the following elements: age, sex, family history of lung cancer, emphysema, nodule size, nodule type, upper lobe, nodule count, spiculation. Liao model is an image-based deep learning model that contains two modules: nodule detection and malignancy evaluation. The nodule detection module is a 3-D version of the region proposal network [49] with a modified UNet structure trained by LUNA16 and private annotated nodules from DSB data [199]. The malignancy evaluation module is built on the nodule detection module and trained by DSB data.

10.2.5 Experiment Settings

We use the VUMC set as the learning cohort. As our previous studies [93, 201], five-fold cross-validation has been applied. We further externally validate our model with 3 cohorts from other sites. For each cohort, we show the results of all the patients and the patients with IPNs when reporting the AUC. The thresholds for low-, moderate-, and high- separation is 0.1 and 0.7 for the Mayo model based on the British Thoracic Society (BTS) guideline [204]. The thresholds for our

	VUMC (val)	UPMC		DECAMP		UCD	
		All	IPN only	All	IPN only	All	IPN only
Mayo	70.7	86.7	85.9	57.6	57.4	67.9	66.3
Brock	71.9	88.6	87.6	64.0	63.7	71.4	71.4
Liao	70.9	82.8	79.7	70.9	70.2	74.6	74.3
Ours	78.7	91.8	89.7	71.1	71.3	84.7	83.2
#Patient	404	155	134	136	130	96	81
#Cancer	269	78	67	67	64	52	43

Table 10.2: The discrimination performance (AUC) and the evaluated patient population of related cohort

model are obtained from the validation set. The benign and malignant groups are separated when computing the cNRI, which follows [203]. The final prediction are the median of the predictions from five models of five-fold cross-validation.

10.3 Experimental Results

10.3.1 Discrimination

IPN represents indeterminate pulmonary nodule, whose nodule size ranges from 6 to 30 mm. The population of VUMC are shown in Table 10.1. The VUMC are randomly split into 5 folds. The model trained using 5-fold cross validation with VUMC. Specifically, the network is trained five times. In each time, 4 folds are used as training set and one-fold is used as validation set. The external sets are tested without model tuning out of VUMC set.

Table 10.2 shows the AUC values of compared methods and cohorts reflect the discrimination of models.

The results of VUMC are reported on the validation sets. The AUCs of UPMC, DECAMP, UCD are external validated results. The patient number and number of cancer cases are shown at the bottom of Table 10.2. Compare with the represented clinical models (Mayo and Brock) and image-only model (Liao), our model achieves superior performance across all the settings (e.g., AUC: ours (0.92) vs. Mayo (0.87) vs. Brock (0.89) vs. Liao (0.83)).

10.3.2 Re-classification

The reclassification performance is illustrated in Figure 10.2. The samples are considered reclassified if our model puts them in different risk groups than Mayo model risks. The frequency tables

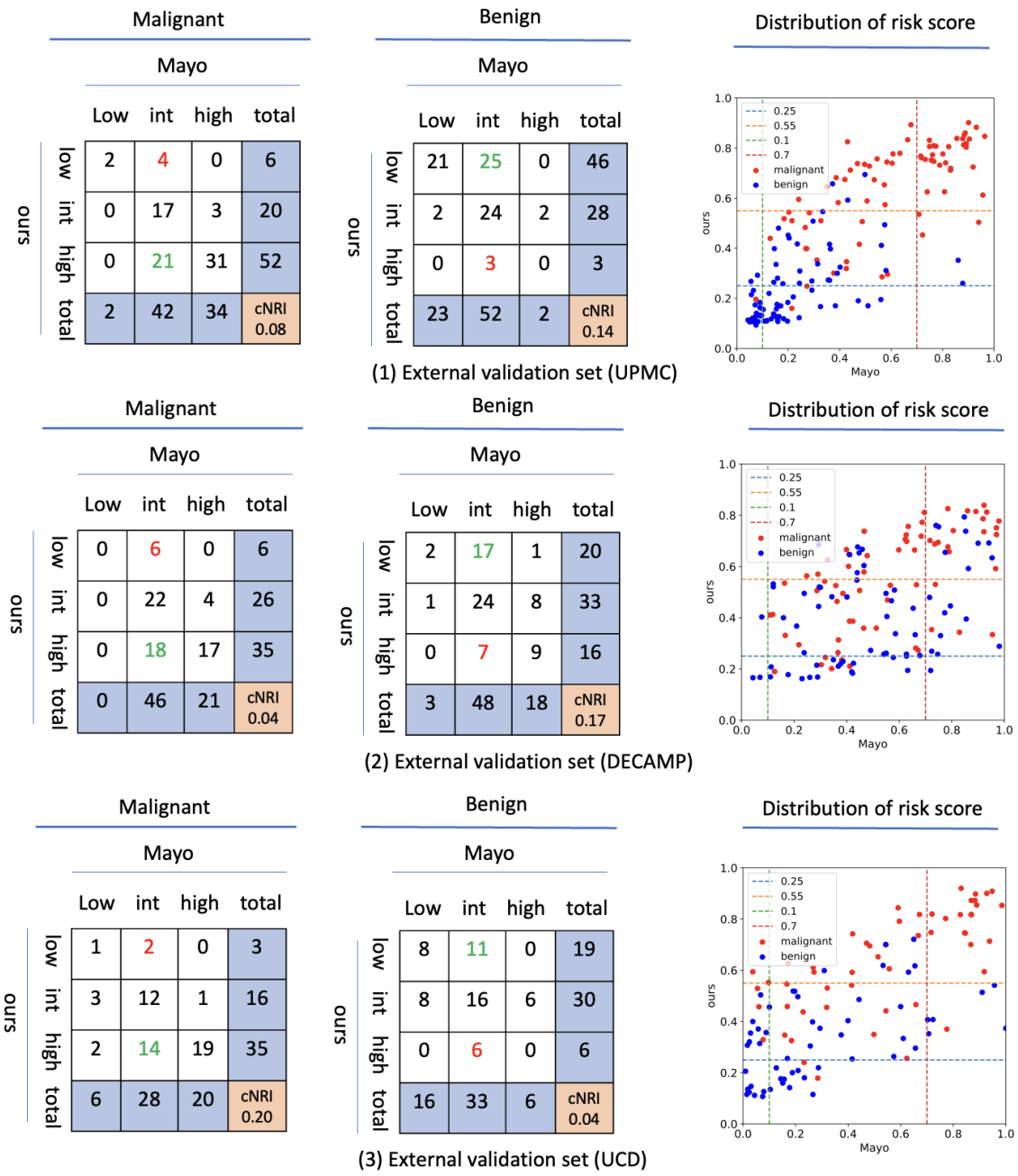


Figure 10.2: Reclassification of three external cohorts. For each cohort, we have reclassification matrices (frequency tables) of malignant and benign, and distribution of risk scores. For the baseline Mayo model, the thresholds for low-, intermediate-, and high- risk are 0.1 and 0.7. The thresholds of our models are achieved from validation set with maximum value of $cNRI_{benign}$+ $cNRI_{malignant}$. cNRI > 0 represents our model has better reclassification performance than Mayo.

and distributions of risk scores of UPMC, DECAMP, and UCD are shown in 10.2(1), 10.2(2), and 10.2(3), respectively. Among all three external validation sets, our model has better reclassification than the baseline mayo model (i.e., cNRI > 0).

10.4 Discussion

To our best knowledge, this is the first clinical validation work integrating CT image, clinical data elements, and the blood marker for the management of IPNs within a deep learning model, which is an extension of our previous technique chapter [93] (Chapter 6) and screening cohort validation [201] (Chapter 9). Our model only requires CT images, standard radiographic features, clinical elements, and inexpensive blood tests. The CT image, standard radiographic feature, and clinical element are usually available in previous standard managements. The CYFAR 21-1 blood biomarker is motivated by a clinical study [203]. Our image features are automatically extracted by a deep learning model from CT images, which does not need human reading effort or external software.

We hypothesize that data from different resources can provide complementary information and can be helpful for lung cancer diagnosis and IPN management. We in total involve four data cohorts with a large amount of IPNs for evaluation. The VUMC cohort includes 1284 patients but not all of them have complete data and our model can handle the missing modality problem. Except for the cross-validation on VUMC, we further have three external cohorts for validation. Our experiments have supported our hypotheses.

This work has some limitations. The clinical data and CT images of patients with IPNs are limited resources and we do not have a huge number of patients to train and validate like the field of computer vision with ImageNet. So, the power of deep learning may not have fully explored in the task. We look forward to future work on IPN management with deep learning.

In summary, multi-modal prediction with deep learning is promising for disease diagnosis and IPN management. More clinical data if available can be integrated with our framework. IPN management decisions are made by multiple experts from different backgrounds, with consultation with the patients. Quantitative lung cancer risk plays an important role in IPN management. Our framework shows superior performance over compared models and has the potential for better results if more data are available.

CHAPTER 11

Conclusion and Future Works

11.1 Impact of the Dissertation

This dissertation focuses on lung cancer diagnosis to target the opportunities/challenges of the prediction model, including image quality issues, sequential data, irregularly sampling, multiple modalities, missing data, and confidence calibration.

We start with the challenges of obtaining clean data; then, we effectively utilize more data to achieve more accurate predictions. At the same time, we tackle the challenges of the imperfectness of data including irregular sampling and missing data. Moreover, we include different aspects to analyze prediction models and conduct real clinical validation studies. In summary, our work extends the current research of lung cancer diagnosis with deep learning. Moreover, multiple works can be generalizable to other disease diagnosis and prediction models.

Lung cancer diagnosis with machine learning models is well developed in recent years. Extensive studies have been conducted on either predicting using lung nodule / single CT scan or tabular clinical data. With the development of data acquisition, larger-scale datasets and data from multiple modalities can be available for machine learning models. Ideally, we can have sequential CT images and clinical tabular data from each patient, which brings new opportunities for more accurate predictions. However, not all the data are ideally obtained, which leads to the imperfectness of data.

Starting with the data acquisition process, we find the raw medical imaging data (de-identified) may have quality issues that may result from network issues, accelerated acquisitions, motion artifacts, and imaging protocol design, which may damage the effectiveness of machine learning models. We developed a quality assessment tool to rule out the cases that cannot meet the quality standard or correct some errors if possible [195]. The work has been presented in Chapter 3.

Multiple samples from the same subject/class can be available in many classification applications. We proposed a new recurrent model to utilize multi-sample information. Beginning with evaluation at general computer vision tasks where the multiple samples per class have no semantic order, we extended our model to evaluate the lung cancer diagnosis with sequential multiple CTs

with semantic order. Experiments show that our model has superior performance over the model using one sample per class/subject and conventional sequential models. We validated that multi-sample per subject/class can be more discriminative than one-sample per subject/class [83]. The details can be found in Chapter 4.

Traditional sequential models ignore the interval differences across subjects, meanwhile, chest CT images can be irregularly acquired in practice and such sampling patterns may be commingled with clinical usages. We proposed a new version of the prominent LSTM approach called distanced LSTM to explicitly model the time distance of each scan to the last scan. The effectiveness of our model is validated with both simulated data and real CT images from three datasets [84, 85]. We validated that the time information of each sequential point should be properly modeled in irregularly sampled problems. The details of this model can be found in Chapter 5.

Except for the CT image modality, several standardized clinical data elements can be available for lung cancer prediction. We proposed a new deep learning model that integrates the image and non-image data for lung cancer diagnosis and validate that multi-modal data provide complementary information for each other [93]. Moreover, considering some patients may not have complete two modalities, our model can train and predict with only one modality available. We validated that multi-modal data provide complementary information and more subjects (even only one modality available) can help to train a model. More details can be found in Chapter 6.

More generally, a patient record may not completely miss one whole modality data. Partial missing data can exist in clinical cohorts across multiple modalities. Considering existing missing imputation methods mainly limit the operation within one modality, we addressed the imputation of missing data by modeling the joint distribution of multi-modal data [189]. Our proposed C-PBiGAN model is evaluated with both illustration examples (i.e., MNIST and Fashion-MNIST) and real lung datasets (i.e., chest CT images and tabular clinical data). We validated that essential information missed in one modality can be maintained in another which helps missing data imputation and eventually improves prediction performance. We describe the details in Chapter 7.

Except for discrimination, confidence calibration is another important aspect of the prediction/-classification model. We conducted a contrastive analysis with recently representative calibration models. Starting with the computer vision dataset CIFAR10, we extended our analysis to the clinical problem of lung cancer diagnoses. Our work can provide guidance for choosing calibration

models. More can be found in Chapter 8.

Finally, to closely address clinical interests, we conducted two clinical validation studies on screening and incidental cohorts, respectively [201]. We validated that our previously developed model contributes to clinical communities in Chapter 9 and Chapter 10.

11.2 Beyond Lung Cancer Risk Estimation

The central goal of our research is to build intelligent models handling imperfect data in practice (especially healthcare). With the development of deep learning community, there are opportunities and challenges that coexist:

Opportunity 1: Repeated data. With the development of big data, repeated measurements of classes/subjects are commonly available. For example, patients may take repeated CT scans from lung cancer screening programs. A person may have multiple images available for face recognition. How to effectively utilize the abundant data is a key question.

Opportunity 2: Multi-modal data. Data from multiple modalities can be available for prediction. Take the lung cancer risk estimation, for example, image modality including CT images and non-image modalities such as clinical data elements and blood biomarkers can be incorporated for more accurate prediction.

Opportunity 3: Multi-site data. To obtain a larger dataset, data may be collected from multiple sites. For example, the NLST dataset CT images from over 30 sites across over 20 manufacturer and model combinations.

In addition to the opportunities, there are challenges that need to handle as well:

Challenge 1: Irregularly sampled data. Irregularly sampled series exist in multiple domains, especially in healthcare. They do not naturally produce a fixed-interval representation required by many standard machine learning models, which requires special care. For example, chest CT imaging acquisition can be irregularly sampled, and such sampling patterns may be commingled with clinical usages.

Challenge 2: Missing data. In clinical practice, missing data is common in lung cancer risk estimation especially considering data from multiple modalities. Data can be missing due to intricate reasons including data entry, data exchange, or loss of follow-up.

Challenge 3: Distribution shifting. As data may come from multiple sites, the training and

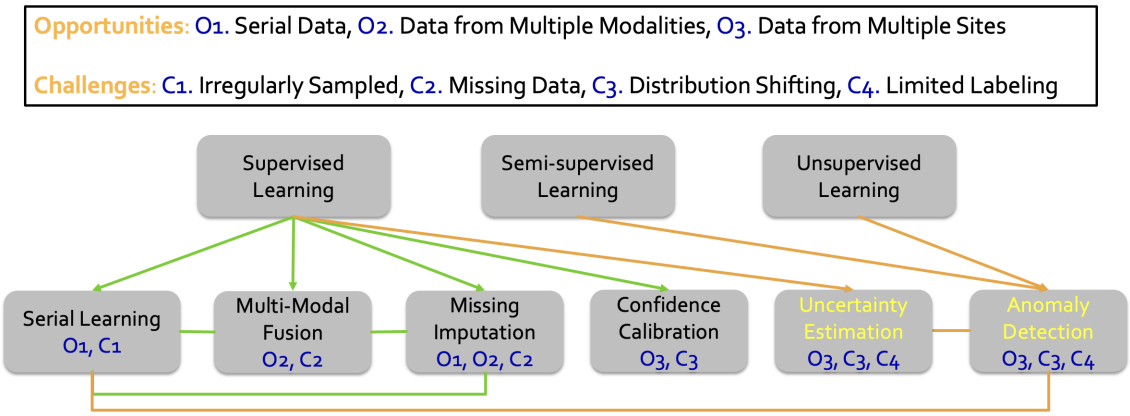


Figure 11.1: Overview of prediction models. We list three opportunities and four challenges with the development of deep learning community (upper box). The bottom flow chart shows potential categories/topics existed in prediction models. The green lines show the connections we have explored, and orange lines are potential future directions.

testing cohorts may have distribution shifting. For example, the screening cohort usually has more samples available for training, while the incidental cohort has different distribution than the screening cohort. Moreover, imbalance training is also popular in screening cohorts.

Challenge 4: Limited labeling. As we mentioned before, large-scale data can be collected with the development of deep learning. However, annotation, especially in healthcare, is always expensive and even cannot be obtained. Supervised learning has been well-developed, while semi-supervised learning or unsupervised learning with limited or no label available is still challenging.

Figure 11.1 shows an overview of prediction models from our perspective. Based on opportunities and challenges, we show our contribution and potential future works. We proposed serial learning models to utilize Opportunity 1 and tackle Challenge 1 (Chapter 4, 5). To address Opportunity 2 and Challenge 2, we proposed a multi-modal fusion model for lung cancer risk estimation (Chapter 6). The multi-modal missing imputation in Chapter 7 covers Opportunity 1,2 and Challenge 2. We provided a contrastive analysis of confidence calibration for prediction models in Chapter 8.

Our Model: a general prediction model perspective. We aim to develop a generalizable model that can integrate sequential multi-modal data for calibrated prediction even when under the challenges of irregularly sampled and data missing, which are motivated by the above opportunities and challenges. The key points and contributions of the overall model have been distributed in our previous chapters. When extending to other prediction tasks, our overall model should be easy to

be adapted. For example, using the brain MRI and clinical data (e.g., BraTS [205]), for disease prediction, we can adapt our model to integrate the multi-modal data (T1, T2, clinical elements, etc) and handle missing data, as described in Chapters 6 and 7. We can adapt our model using the idea of Chapter 5 when there are irregularly sampled sequences in the disease classification as MIMIC-III [206].

To extend our current model with more challenges that have not been addressed, we have following future works.

11.3 Future Works

In this dissertation, our contributions mainly lie in the framework of supervised learning. There are some directions that can be explored in the future work:

FW1. Uncertainty Estimation in the Supervised Learning framework. Obtaining accurate uncertainty quantification plays an important role in trustworthy prediction. Different from the confidence calibration discussed in Chapter 8, uncertainty estimation models usually can achieve an uncertainty score for each sample, rather than an overall calibration score for the whole dataset. As the lung cancer dataset may come from multiple sites and distribution may exist across training and test sets, it is important for models to know what the uncertainty is about their prediction.

FW2. Anomaly Detection with Unsupervised- and Semi-supervised- Learning. Anomaly detection (AD) refers to detecting uncommon samples out of inlier distribution. AD can be performed either in an unsupervised manner when only inliers are involved during the training process, or in a semi-supervised manner when few labeled anomalies are available (Challenge 4). The lung cancer screening program matches the setting of AD. In addition, anomaly localization techniques should be helpful for more accurate therapy.

FW3. Anomaly Detection in Time Series. Repeated CT scans is popular in lung cancer diagnosis. Fundamentally, repeated scanning is to find the abnormal timepoint when some normal timepoints are presented. As Challenge 4, even large-scale datasets can be collected, golden standard (e.g., biopsy) diagnosis of cancer is raw. This matches the setting of anomaly detection, and the adaption needed is to transfer from a single timepoint to sequential settings.

FW4. Bridge the Uncertainty Estimation and Anomaly Detection. The intuition is that anomalies should have higher uncertainty when a normal space is trained with only normal data. Some

works have been explored for general computer vision [181], while uncertainty estimation for anomaly detection has not been studied in lung cancer risk estimation.

To conclude, my dream is to develop "gentle-and-strict" models that are 1) easy to implement and user-friendly (gentle), and 2) motivated by practical challenges and theoretically solid (strict).

References

[1] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer Statistics, 2018," *CA: A Cancer Journal for Clinicians*, 2018.

[2] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2019," *CA: A Cancer Journal for Clinicians*, 1 2019.

[3] K. D. Miller, R. L. Siegel, C. C. Lin, A. B. Mariotto, J. L. Kramer, J. H. Rowland, K. D. Stein, R. Alteri, and A. Jemal, "Cancer treatment and survivorship statistics, 2016," *CA: A Cancer Journal for Clinicians*, vol. 66, pp. 271–289, 7 2016.

[4] N.L.S.T.R.T.J., "The national lung screening trial: Overview and study design," *Radiology*, vol. 258, pp. 243–253, 1 2011.

[5] L. L. Humphrey, M. Deffebach, M. Pappas, C. Baumann, K. Artis, J. P. Mitchell, B. Zakher, R. Fu, and C. G. Slatore, "Screening for lung cancer with low-dose computed tomography: A systematic review to update the U.S. preventive services task force recommendation," *Annals of Internal Medicine*, vol. 159, pp. 411–420, 9 2013.

[6] M. K. Gould, T. Tang, I. L. A. Liu, J. Lee, C. Zheng, K. N. Danforth, A. E. Kosco, J. L. Di Fiore, and D. E. Suh, "Recent trends in the identification of incidental pulmonary nodules," *American Journal of Respiratory and Critical Care Medicine*, vol. 192, pp. 1208–1214, 11 2015.

[7] T. Lokhandwala, M. A. Bittoni, R. A. Dann, A. O. D'Souza, M. Johnson, R. J. Nagy, R. B. Lanman, R. E. Merritt, and D. P. Carbone, "Costs of Diagnostic Assessment for Lung Cancer: A Medicare Claims Analysis," *Clinical Lung Cancer*, vol. 18, pp. e27–e34, 1 2017.

[8] S. L. Lee, A. Z. Kouzani, and E. J. Hu, "Automated detection of lung nodules in computed tomography images: A review," *Machine Vision and Applications*, vol. 23, pp. 151–163, 1 2012.

[9] R. L. Drake, W. Vogl, Mitchell, and A. W.M, "Gray's Anatomy for Students (3rd ed.)," *Edinburgh: Churchill Livingstone/Elsevier*, pp. 167–174, 9 2014.

[10] N. R. Standring, Susan. Borley, "Gray's Anatomy: the anatomical basis of clinical practice," *Edinburgh: Churchill Livingstone*, vol. 91-B, no. 7, pp. 992–1000, 2008.

[11] "Lung fissures — Radiology Reference Article — Radiopaedia.org." https://radiopaedia.org/articles/lung-fissures. accessed 2020/12/10.

[12] "Lung Carcinoma - Pulmonary Disorders - Merck Manuals Professional Edition." https://www.merckmanuals.com/professional/pulmonary-disorders/tumors-of-the-lungs/lung-carcinoma#sec05-ch062-ch062b-1405. accessed 2020/12/10.

[13] "Non-Small Cell Lung Cancer Treatment (PDQ®)–Health Professional Version - National Cancer Institute." https://www.cancer.gov/types/lung/hp/non-small-cell-lung-treatment-pdq#_470. accessed 2022/01/06.

[14] S. A. Kenfield, E. K. Wei, M. J. Stampfer, B. A. Rosner, and G. A. Colditz, "Comparison of aspects of smoking among the four histological types of lung cancer," *Tobacco Control*, vol. 17, pp. 198–204, 6 2008.

[15] "Small Cell Lung Cancer Treatment (PDQ®)–Health Professional Version - National Cancer Institute." https://www.cancer.gov/types/lung/hp/small-cell-lung-treatment-pdq. accessed 2022/01/06.

[16] F. C. Detterbeck, D. J. Boffa, A. W. Kim, and L. T. Tanoue, "The Eighth Edition Lung Cancer Stage Classification," 1 2017.

[17] "Lung Cancer - Small Cell: Stages — Cancer.Net." https://www.cancer.net/cancer-types/lung-cancer-small-cell/stages. accessed 2020/12/10.

[18] "Computed Tomography (CT)." https://www.nibib.nih.gov/science-education/science-topics/computed-tomography-ct. accessed 2020/12/09.

[19] J. R. Mayo, "CT evaluation of diffuse infiltrative lung disease: Dose considerations and optimal technique," in *Journal of Thoracic Imaging*, vol. 24, pp. 252–259, 11 2009.

[20] K. Yamashita, S. Matsunobe, T. Tsuda, T. Nemoto, K. Matsumoto, H. Miki, and J. Konishi, "Solitary pulmonary nodule: Preliminary study of evaluation with incremental dynamic CT," *Radiology*, vol. 194, no. 2, pp. 399–405, 1995.

[21] J. D. Hyer and G. Silvestri, "Diagnosis and Staging of Lung Cancer," *Clinics in Chest Medicine*, vol. 21, pp. 95–106, 3 2000.

[22] R. e. a. Botvinik-Nezer, "Variability in the analysis of a single neuroimaging dataset by many teams," *Nature*, vol. 582, pp. 84–88, 5 2020.

[23] R. L. Harrigan, B. C. Yvernault, B. D. Boyd, S. M. Damon, K. D. Gibney, B. N. Conrad, N. S. Phillips, B. P. Rogers, Y. Gao, and B. A. Landman, "Vanderbilt University Institute of Imaging Science Center for Computational Imaging XNAT: A multimodal data archive and processing environment," *NeuroImage*, vol. 124, pp. 1097–1101, 1 2016.

[24] S. G. Armato, G. McLennan, L. Bidaut, M. F. McNitt-Gray, C. R. Meyer, A. P. Reeves, B. Zhao, D. R. Aberle, C. I. Henschke, E. A. Hoffman, E. A. Kazerooni, H. MacMahon, E. J. Van Beek, D. Yankelevitz, A. M. Biancardi, P. H. Bland, M. S. Brown, R. M. Engelmann, G. E. Laderach, D. Max, R. C. Pais, D. P. Qing, R. Y. Roberts, A. R. Smith, A. Starkey, P. Batra, P. Caligiuri, A. Farooqi, G. W. Gladish, C. M. Jude, R. F. Munden, I. Petkovska, L. E. Quint, L. H. Schwartz, B. Sundaram, L. E. Dodd, C. Fenimore, D. Gur, N. Petrick, J. Freymann, J. Kirby, B. Hughes, A. Vande Casteele, S. Gupte, M. Sallam, M. D. Heath, M. H. Kuhn, E. Dharaiya, R. Burns, D. S. Fryd, M. Salganicoff, V. Anand, U. Shreter, S. Vastagh, B. Y. Croft, and L. P. Clarke, "The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): A completed reference database of lung nodules on CT scans," *Medical Physics*, vol. 38, no. 2, pp. 915–931, 2011.

[25] S. Wang, M. Zhou, Z. Liu, Z. Liu, D. Gu, Y. Zang, D. Dong, O. Gevaert, and J. Tian, "Central focused convolutional neural networks: Developing a data-driven model for lung nodule segmentation," *Medical Image Analysis*, vol. 40, pp. 172–183, 8 2017.

[26] A. A. A. Setio, A. Traverso, T. de Bel, M. S. Berens, C. v. d. Bogaard, P. Cerello, H. Chen, Q. Dou, M. E. Fantacci, B. Geurts, R. v. d. Gugten, P. A. Heng, B. Jansen, M. M. de Kaste, V. Kotov, J. Y. H. Lin, J. T. Manders, A. Sóñora-Mengana, J. C. García-Naranjo, E. Papavasileiou, M. Prokop, M. Saletta, C. M. Schaefer-Prokop, E. T. Scholten, L. Scholten, M. M. Snoeren, E. L. Torres, J. Vandemeulebroucke, N. Walasek, G. C. Zuidhof, B. v. Ginneken, and C. Jacobs, "Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: The LUNA16 challenge," *Medical Image Analysis*, vol. 42, pp. 1–13, 12 2017.

[27] "Vanderbilt Lung Screening Program — Department of Radiology." https://www.vumc.org/radiology/lung. accessed 2020/12/17.

[28] "Consortium for Molecular and Cellular Characterization of Screen-Detected Lesions." https://mcl.nci.nih.gov/. accessed 2020/12/16.

[29] M. C. Tammemägi, H. A. Katki, W. G. Hocking, T. R. Church, N. Caporaso, P. A. Kvale, A. K. Chaturvedi, G. A. Silvestri, T. L. Riley, J. Commins, and C. D. Berg, "Selection criteria for lung-cancer screening," *New England Journal of Medicine*, vol. 368, pp. 728–736, 2 2013.

[30] S. J. Swensen, "The Probability of Malignancy in Solitary Pulmonary Nodules," *Archives of Internal Medicine*, vol. 157, p. 849, 4 1997.

[31] A. e. a. McWilliams, "Probability of cancer in pulmonary nodules detected on first screening CT," *New England Journal of Medicine*, vol. 369, no. 10, pp. 910–919, 2013.

[32] P. Lambin, E. Rios-Velazquez, R. Leijenaar, S. Carvalho, R. G. Van Stiphout, P. Granton, C. M. Zegers, R. Gillies, R. Boellard, A. Dekker, and H. J. Aerts, "Radiomics: Extracting more information from medical images using advanced feature analysis," *European Journal of Cancer*, vol. 48, pp. 441–446, 3 2012.

[33] H. MacMahon, D. P. Naidich, J. M. Goo, K. S. Lee, A. N. Leung, J. R. Mayo, A. C. Mehta, Y. Ohno, C. A. Powell, M. Prokop, G. D. Rubin, C. M. Schaefer-Prokop, W. D. Travis, P. E. Van Schil, and A. A. Bankier, "Guidelines for management of incidental pulmonary nodules detected on CT images: From the Fleischner Society 2017," 7 2017.

[34] M. Takada, N. Masuda, E. Matsuura, Y. Kusunoki, K. Matui, K. Nakagawa, T. Yana, I. Tuyuguchi, I. Oohata, and M. Fukuoka, "Measurement of cytokeratin 19 fragments as a marker of lung cancer by CYFRA 21-1 enzyme immunoassay," *British Journal of Cancer*, vol. 71, no. 1, pp. 160–165, 1995.

[35] M. N. Kammer, A. K. Kussrow, R. L. Webster, H. Chen, M. Hoeksema, R. Christenson, P. P. Massion, and D. J. Bornhop, "Compensated Interferometry Measures of CYFRA 21-1 Improve Diagnosis of Lung Cancer," *ACS Combinatorial Science*, vol. 21, no. 6, 2019.

[36] D. Shen, G. Wu, and H. I. Suk, "Deep Learning in Medical Image Analysis," *Annual Review of Biomedical Engineering*, vol. 19, pp. 221–248, 6 2017.

[37] F. Liao, M. Liang, Z. Li, X. Hu, and S. Song, "Evaluate the Malignancy of Pulmonary Nodules Using the 3-D Deep Leaky Noisy-or Network," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–12, 2 2019.

[38] D. Ardila, A. P. Kiraly, S. Bharadwaj, B. Choi, J. J. Reicher, L. Peng, D. Tse, M. Etemadi, W. Ye, G. Corrado, D. P. Naidich, and S. Shetty, "End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography," 6 2019.

[39] P. P. Massion, S. Antic, S. Ather, C. Arteta, J. Brabec, H. Chen, J. Declerck, D. Dufek, W. Hickes, T. Kadir, J. Kunst, B. A. Landman, R. F. Munden, P. Novotny, H. Peschl, L. C. Pickup, C. Santos, G. T. Smith, A. Talwar, and F. Gleeson, "Assessing the Accuracy of a Deep Learning Method to Risk Stratify Indeterminate Pulmonary Nodules," *American journal of respiratory and critical care medicine*, vol. 202, pp. 241–249, 7 2020.

[40] L. Liu, Q. Dou, H. Chen, J. Qin, and P. A. Heng, "Multi-Task Deep Model with Margin Ranking Loss for Lung Nodule Analysis," *IEEE Transactions on Medical Imaging*, vol. 39, pp. 718–728, 3 2020.

[41] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, pp. 861–874, 6 2006.

[42] H. C. Hartmann, T. C. Pagano, S. Sorooshian, R. Bales, and A. . Hartmann, "Confidence builders: Evaluating seasonal climate forecasts from user perspectives," *Bulletin of the American Meteorological Society*, 2002.

[43] "Hounsfield scale - Wikipedia."

[44] D. Jin, Z. Xu, Y. Tang, A. P. Harrison, and D. J. Mollura, "CT-realistic lung nodule simulation from 3D conditional generative adversarial networks for robust lung segmentation," in *MICCAI*, vol. 11071 LNCS, pp. 732–740, Springer Verlag, 6 2018.

[45] T. Mitchell, *Machine Learning*. New York: McGraw Hill, 1997.

[46] C. M. Bishop, *Pattern recognition and machine learning*. 2006.

[47] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," 5 2015.

[48] S. Dasiopoulou, V. Mezaris, I. Kompatsiaris, V. K. Papastathis, and M. G. Strintzis, "Knowledge-assisted semantic video object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15, pp. 1210–1224, 10 2005.

[49] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, pp. 1137–1149, 6 2017.

[50] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, pp. 386–397, 2 2020.

[51] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in *Computer Vision and Pattern Recognition*, pp. 779–788, 2016.

[52] J. Redmon and A. Farhadi, "YOLO9000: Better, Faster, Stronger," in *Computer Vision and Pattern Recognition*, vol. 2017-January, Institute of Electrical and Electronics Engineers Inc., 12 2017.

[53] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal Speed and Accuracy of Object Detection," *arXiv*, 4 2020.

[54] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single Shot MultiBox Detector," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9905 LNCS, pp. 21–37, 12 2015.

[55] B. van Ginneken, S. G. Armato, B. de Hoop, S. van Amelsvoort-van de Vorst, T. Duindam, M. Niemeijer, K. Murphy, A. Schilham, A. Retico, M. E. Fantacci, N. Camarlinghi, F. Bagagli, I. Gori, T. Hara, H. Fujita, G. Gargano, R. Bellotti, S. Tangaro, L. Bolaos, F. D. Carlo, P. Cerello, S. Cristian Cheran, E. Lopez Torres, and M. Prokop, "Comparing and combining algorithms for computer-aided detection of pulmonary nodules in computed tomography scans: The ANODE09 study," *Medical Image Analysis*, vol. 14, pp. 707–722, 12 2010.

[56] A. A. A. Setio, F. Ciompi, G. Litjens, P. Gerke, C. Jacobs, S. J. Van Riel, M. M. W. Wille, M. Naqibullah, C. I. Sanchez, and B. Van Ginneken, "Pulmonary Nodule Detection in CT Images: False Positive Reduction Using Multi-View Convolutional Networks," *IEEE Transactions on Medical Imaging*, vol. 35, pp. 1160–1169, 5 2016.

[57] Q. Dou, H. Chen, L. Yu, J. Qin, and P. A. Heng, "Multilevel Contextual 3-D CNNs for False Positive Reduction in Pulmonary Nodule Detection," *IEEE Transactions on Biomedical Engineering*, vol. 64, pp. 1558–1567, 7 2017.

[58] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International conference on medical image computing and computer-assisted intervention*, vol. 9351, pp. 234–241, Springer Verlag, 5 2015.

[59] W. Shen, M. Zhou, F. Yang, C. Yang, and J. Tian, "Multi-scale convolutional neural networks for lung nodule classification," in *Information Processing in Medical Imaging*, vol. 9123, pp. 588–599, Springer Verlag, 2015.

[60] L. Liu, Q. Dou, H. Chen, I. E. Olatunji, J. Qin, and P. A. Heng, "Mtmr-net: Multi-task deep learning with margin ranking loss for lung nodule analysis," in *MICCAI*, vol. 11045 LNCS, pp. 74–82, Springer Verlag, 2018.

[61] Y. Xu, A. Hosny, R. Zeleznik, C. Parmar, T. Coroller, I. Franco, R. H. Mak, and H. J. Aerts, "Deep learning predicts lung cancer treatment response from serial medical imaging," *Clinical Cancer Research*, vol. 25, pp. 3266–3275, 6 2019.

[62] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.

[63] M. Ilse, J. M. Tomczak, C. Louizos, M. Welling, and M. W. Nl, "DIVA: Domain Invariant Variational Autoencoders," *MIDL2020*, pp. 1–27, 1 2020.

[64] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Transformer: Attention Is All You Need," *Behavioral and Brain Sciences*, no. Nips, pp. 1–101, 2017.

[65] T. H. Wen, M. Gašić, N. Mrkšić, P. H. Su, D. Vandyke, and S. Young, "Semantically conditioned lstm-based Natural language generation for spoken dialogue systems," in *EMNLP*, pp. 1711–1721, 2015.

[66] S. Han, J. Kang, H. Mao, Y. Hu, X. Li, Y. Li, D. Xie, H. Luo, S. Yao, Y. Wang, H. Yang, and W. J. Dally, "ESE: Efficient speech recognition engine with sparse LSTM on FPGA," in *FPGA*, pp. 75–84, 2 2017.

[67] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *EMNLP*, pp. 1724–1734, 2014.

[68] X. Shi, Z. Chen, H. Wang, D. Y. Yeung, W. K. Wong, and W. C. Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," in *Advances in Neural Information Processing Systems*, vol. 2015-Janua, pp. 802–810, 2015.

[69] B. Zhao, X. Li, X. Lu, and Z. Wang, "A CNN–RNN architecture for multi-label weather recognition," *Neurocomputing*, vol. 322, pp. 47–57, 12 2018.

[70] M. Lv, W. Xu, and T. Chen, "A hybrid deep convolutional and recurrent neural network for complex activity recognition using multimodal sensors," *Neurocomputing*, vol. 362, pp. 33–40, 10 2019.

[71] C. Shi and C. M. Pun, "Multi-scale hierarchical recurrent neural networks for hyperspectral image classification," *Neurocomputing*, vol. 294, pp. 82–93, 6 2018.

[72] J. Wang, Y. Yang, J. Mao, Z. Huang, C. Huang, and W. Xu, "CNN-RNN: A Unified Framework for Multi-label Image Classification," in *IEEE CVPR*, 2016.

[73] Z. Yu, J. Yu, C. Xiang, and J. Fan, "Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering," *IEEE transactions on neural networks and learning systems*, 2018.

[74] Z. Yu, J. Yu, Y. Cui, D. Tao, and Q. Tian, "Deep Modular Co-Attention Networks for Visual Question Answering," in *IEEE CVPR*, 2019.

[75] G. Campanella, M. G. Hanna, L. Geneslaw, A. Miraflor, V. Werneck Krauss Silva, K. J. Busam, E. Brogi, V. E. Reuter, D. S. Klimstra, and T. J. Fuchs, "Clinical-grade computational pathology using weakly supervised deep learning on whole slide images," *Nature Medicine*, vol. 25, pp. 1301–1309, 8 2019.

[76] Q. Liu, F. Zhou, R. Huang, and X. Yuan, "Bidirectional-Convolutional LSTM Based Spectral-Spatial Feature Learning for Hyperspectral Image Classification," *Remote Sensing*, vol. 9, p. 1330, 12 2017.

[77] "Vanderbilt Lung Screening Program — Department of Radiology."

[78] C. Finn, I. G. Openai, S. Levine, and G. Brain, "Unsupervised Learning for Physical Interaction through Video Prediction," in *Advances in neural information processing systems*, 2016.

[79] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, pp. 1735–1780, 11 1997.

[80] "Understanding LSTM Networks – colah's blog." https://colah.github.io/posts/2015-08-Understanding-LSTMs. accessed 2019/03/17.

[81] F. A. Gers and J. Schmidhuber, "Recurrent nets that time and count," in *International Joint Conference on Neural Networks*, vol. 3, pp. 189–194, IEEE, 2000.

[82] D. Neil, M. Pfeiffer, and S.-C. Liu, "Phased LSTM: Accelerating Recurrent Network Training for Long or Event-based Sequences," in *NIPS*, 2016.

[83] R. Gao, Y. Huo, S. Bao, Y. Tang, S. L. Antic, E. S. Epstein, S. Deppen, A. B. Paulson, K. L. Sandler, P. P. Massion, and B. A. Landman, "Multi-path x-D recurrent neural networks for collaborative image classification," *Neurocomputing*, vol. 397, pp. 48–59, 7 2020.

[84] R. Gao, Y. Huo, S. Bao, Y. Tang, S. L. Antic, E. S. Epstein, A. B. Balar, S. Deppen, A. B. Paulson, K. L. Sandler, P. P. Massion, and B. A. Landman, "Distanced LSTM: Time-Distanced Gates in Long Short-Term Memory Models for Lung Cancer Detection," in *International Workshop on Machine Learning in Medical Imaging*, vol. 11861 LNCS, pp. 310–318, 2019.

[85] R. Gao, Y. Tang, K. Xu, Y. Huo, S. Bao, S. L. Antic, E. S. Epstein, S. Deppen, A. B. Paulson, K. L. Sandler, P. P. Massion, and B. A. Landman, "Time-Distanced Gates in Long Short-Term Memory Networks," *Medical Image Analysis*, vol. 65, p. 101785, 10 2020.

[86] D. B. Rubin, "Inference and missing data," *Biometrika*, vol. 63, no. 3, pp. 581–592, 1976.

[87] A. Nazabal, P. M. Olmos, Z. Ghahramani, and I. Valera, "Handling Incomplete Heterogeneous Data using VAEs," *Pattern Recognition*, 7 2020.

[88] S. C. X. Li, B. M. Marlin, and B. Jiang, "Misgan: Learning from incomplete data with generative adversarial networks," *7th International Conference on Learning Representations, ICLR 2019*, pp. 1–20, 2019.

[89] S. C. X. Li and B. M. Marlin, "Learning from irregularly-sampled time series: A missing data perspective," in *37th International Conference on Machine Learning, ICML 2020*, vol. PartF16814, pp. 5893–5902, 2020.

[90] L. E. Richards, R. J. A. Little, and D. B. Rubin, "Statistical Analysis with Missing Data," *Journal of Marketing Research*, vol. 26, no. 3, p. 374, 1989.

[91] S. Van Buuren and K. Oudshoorn, "Flexible multivariate imputation by MICE," tech. rep., 1 1999.

[92] S. Fletcher Mercaldo and J. D. Blume, "Missing data and prediction: the pattern submodel," *Biostatistics (Oxford, England)*, vol. 21, pp. 236–252, 4 2020.

[93] R. Gao, Y. Tang, K. Xu, M. Kammer, S. Antic, S. Deppen, K. Sandler, P. Massion, Y. Huo, and B. A. Landman, "Deep multi-path network integrating incomplete biomarker and chest CT data for evaluating lung cancer risk," in *SPIE 2021*, p. 46, 10 2021.

[94] R. Mazumder, T. Hastie, H. Edu, R. Tibshirani, T. Edu, and T. Jaakkola, "Spectral Regularization Algorithms for Learning Large Incomplete Matrices," *Journal of Machine Learning Research*, vol. 11, pp. 2287–2322, 2010.

[95] M. J. Azur, E. A. Stuart, C. Frangakis, and P. J. Leaf, "Multiple imputation by chained equations: What is it and how does it work?," *International Journal of Methods in Psychiatric Research*, vol. 20, pp. 40–49, 3 2011.

[96] D. J. Stekhoven and P. Bühlmann, "Missforest-Non-parametric missing value imputation for mixed-type data," *Bioinformatics*, vol. 28, no. 1, pp. 112–118, 2012.

[97] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *International Conference on Learning Representations*, 12 2014.

[98] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *NIPS*, vol. 3, pp. 2672–2680, 2014.

[99] D. Rubin, *Multiple imputation for nonresponse in surveys*. John Wiley and Sons Ltd, 2004.

[100] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," 12 2017.

[101] R. Gao, L. Li, Y. Tang, S. L. Antic, A. Paulson, Y. Huo, K. Sandler, P. P. Massion, and B. A. Landman, "Deep multi-task prediction of lung cancer and cancer-free progression from censored heterogenous clinical imaging," in *Medical Imaging 2020: Image Processing*, p. 12, SPIE-Intl Soc Optical Eng, 3 2020.

[102] Y. Yang, R. Gao, Y. Tang, S. L. Antic, S. Deppen, Y. Huo, K. L. Sandler, P. P. Massion, and B. A. Landman, "Internal-transfer weighting of multi-task learning for lung cancer detection," in *Medical Imaging 2020: Image Processing*, vol. 11313, p. 74, SPIE, 3 2020.

[103] F. E. Boas and D. Fleischmann, "CT artifacts: Causes and reduction techniques," *Imaging in Medicine*, vol. 4, no. 2, pp. 229–240, 2012.

[104] R. Robinson, V. V. Valindria, W. Bai, H. Suzuki, P. M. Matthews, C. Page, D. Rueckert, and B. Glocker, "Automatic quality control of cardiac MRI segmentation in large-scale population imaging," in *MICCAI*, vol. 10433 LNCS, pp. 720–727, Springer Verlag, 9 2017.

[105] I. Oksuz, B. Ruijsink, E. Puyol-Antón, J. R. Clough, G. Cruz, A. Bustin, C. Prieto, R. Botnar, D. Rueckert, J. A. Schnabel, and A. P. King, "Automatic CNN-based detection of cardiac MR motion artefacts using k-space data augmentation and curriculum learning," *Medical Image Analysis*, vol. 55, pp. 136–147, 7 2019.

[106] D. Racine, A. H. Ba, J. G. Ott, F. O. Bochud, and F. R. Verdun, "Objective assessment of low contrast detectability in computed tomography with Channelized Hotelling Observer," *Physica Medica*, vol. 32, pp. 76–83, 1 2016.

[107] Q. Gao, S. Li, M. Zhu, D. Li, Z. Bian, Q. Lv, D. Zeng, and J. Ma, "Combined global and local information for blind CT image quality assessment via deep learning," in *Medical Imaging 2020: Image Perception, Observer Performance, and Technology Assessment* (F. W. Samuelson and S. Taylor-Phillips, eds.), vol. 11316, p. 39, SPIE, 3 2020.

[108] L. R. F. Branco, R. B. Ger, D. S. Mackin, S. Zhou, L. E. Court, and R. R. Layman, "Technical Note: Proof of concept for radiomics-based quality assurance for computed tomography," *Journal of Applied Clinical Medical Physics*, vol. 20, pp. 199–205, 11 2019.

[109] I. Oguz, M. Farzinfar, J. Matsui, F. Budin, Z. Liu, G. Gerig, H. J. Johnson, and M. Styner, "DTIPrep: Quality control of diffusion-weighted images," *Frontiers in Neuroinformatics*, vol. 8, 1 2014.

[110] L. S. Chow and R. Paramesran, "Review of medical image quality assessment," *Biomedical Signal Processing and Control*, vol. 27, no. May, pp. 145–154, 2016.

[111] A. Schreuder, C. Jacobs, L. Gallardo-Estrella, M. Prokop, C. M. Schaefer-Prokop, and B. van Ginneken, "Predicting all-cause and lung cancer mortality using emphysema score progression rate between baseline and follow-up chest CT images: A comparison of risk model performances," *PLoS ONE*, vol. 14, 2 2019.

[112] K. ten Haaf, J. Jeon, M. C. Tammemägi, S. S. Han, C. Y. Kong, S. K. Plevritis, E. J. Feuer, H. J. de Koning, E. W. Steyerberg, and R. Meza, "Risk prediction models for selection of lung cancer screening candidates: A retrospective validation study," *PLoS Medicine*, vol. 14, no. 4, 2017.

[113] M. C. Tammemägi, K. ten Haaf, I. Toumazis, C. Y. Kong, S. S. Han, J. Jeon, J. Commins, T. Riley, and R. Meza, "Development and Validation of a Multivariable Lung Cancer Risk Prediction Model That Includes Low-Dose Computed Tomography Screening Results," *JAMA Network Open*, vol. 2, p. e190204, 3 2019.

[114] P. e. a. Huang, "Prediction of lung cancer risk at follow-up screening with low-dose CT: a training and validation study of a deep learning method," *The Lancet Digital Health*, vol. 1, pp. e353–e362, 11 2019.

[115] A. Schreuder, C. M. Schaefer-Prokop, E. T. Scholten, C. Jacobs, M. Prokop, and B. Van Ginneken, "Lung cancer risk to personalise annual and biennial follow-up computed tomography screening," *Thorax*, vol. 73, pp. 626–633, 7 2018.

[116] E. F. Patz, E. Greco, C. Gatsonis, P. Pinsky, B. S. Kramer, and D. R. Aberle, "Lung cancer incidence and mortality in National Lung Screening Trial participants who underwent low-dose CT prevalence screening: A retrospective cohort analysis of a randomised, multicentre, diagnostic screening trial," *The Lancet Oncology*, vol. 17, pp. 590–599, 5 2016.

[117] R. Yip, D. F. Yankelevitz, M. Hu, K. Li, D. M. Xu, A. Jirapatnakul, and C. I. Henschke, "Lung cancer deaths in the national lung screening trial attributed to nonsolid nodules," 2016.

[118] N. Lessmann, B. van Ginneken, P. A. de Jong, and I. Išgum, "Iterative fully convolutional neural networks for automatic vertebra segmentation and identification," *Medical Image Analysis*, vol. 53, pp. 142–155, 4 2019.

[119] N. Lessmann, B. van Ginneken, M. Zreik, P. A. de Jong, B. D. de Vos, M. A. Viergever, and I. Išgum, "Automatic calcium scoring in low-dose chest CT using deep neural networks with dilated convolutions," *IEEE Transactions on Medical Imaging*, vol. 37, pp. 615–625, 11 2017.

[120] X. Li, P. S. Morgan, J. Ashburner, J. Smith, and C. Rorden, "The first step for neuroimaging data analysis: DICOM to NIfTI conversion," *Journal of Neuroscience Methods*, vol. 264, pp. 47–56, 5 2016.

[121] M. W. Woolrich, S. Jbabdi, B. Patenaude, M. Chappell, S. Makni, T. Behrens, C. Beckmann, M. Jenkinson, and S. M. Smith, "Bayesian analysis of neuroimaging data in FSL.," *NeuroImage*, vol. 45, no. 1 Suppl, 2009.

[122] G. Grabner, A. L. Janke, M. M. Budge, D. Smith, J. Pruessner, and D. L. Collins, "Symmetric atlasing and model based segmentation: An application to the hippocampus in older adults," in *MICCAI*, vol. 4191 LNCS, pp. 58–66, Springer Verlag, 2006.

[123] "DICOMLookup." http://dicomlookup.com/lookup.asp?sw=Ttable&q=C.7-9. accessed 2020/11/09.

[124] D. R. Aberle, A. M. Adams, C. D. Berg, W. C. Black, J. D. Clapp, R. M. Fagerstrom, I. F. Gareen, C. Gatsonis, P. M. Marcus, and J. R. D. Sicks, "Reduced lung-cancer mortality with low-dose computed tomographic screening," *New England Journal of Medicine*, vol. 365, pp. 395–409, 8 2011.

[125] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," 9 2014.

[126] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016-Decem, pp. 770–778, 2016.

[127] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," in *IEEE CVPR*, 2017.

[128] J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation Networks," in *IEEE CVPR*, 2018.

[129] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *IEEE CVPR*, vol. 07-12-June-2015, pp. 815–823, IEEE Computer Society, 10 2015.

[130] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *ECCV*, vol. 9911 LNCS, pp. 499–515, Springer Verlag, 2016.

[131] W. Liu, Y. Wen, Z. Yu, and M. Yang, "Large-Margin Softmax Loss for Convolutional Neural Networks," in *ICML*, 2016.

[132] Y. Sun, Y. Chen, X. Wang, and X. Tang, "Deep Learning Face Representation by Joint Identification-Verification," in *NIPS*, 2014.

[133] H. Su, S. Maji, E. Kalogerakis, and E. Learned-Miller, "Multi-view Convolutional Neural Networks for 3D Shape Recognition," in *ICCV*, 2015.

[134] X. Wu, R. He, S. Member, Z. Sun, and T. Tan, "A Light CNN for Deep Face Representation with Noisy Labels," *IEEE Transactions on Information Forensics and Security*, 2018.

[135] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[136] A. Krizhevsky and G. Hinton, "Learning Multiple Layers of Features from Tiny Images," tech. rep., University of Toronto, 2009.

[137] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "VGGFace2: A dataset for recognising faces across pose and age," in *IEEE international conference on automatic face & gesture recognition*, 2018.

[138] S. Zagoruyko and N. Komodakis, "Wide Residual Networks," in *British Machine Vision Conference*, vol. 2016-September, pp. 1–87, 2016.

[139] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive Angular Margin Loss for Deep Face Recognition," in *IEEE CVPR*, 2019.

[140] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *NIPS*, 2012.

[141] S. Xie, R. Girshick, P. Dollár, Z. Tu, K. He, and U. San Diego, "Aggregated Residual Transformations for Deep Neural Networks," in *IEEE CVPR*, 2017.

[142] X. Zhang, X. Zhou, and M. Lin, "ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices," in *IEEE CVPR*, 2018.

[143] B. Zoph, G. Brain, V. Vasudevan, J. Shlens, and Q. V. Le Google Brain, "Learning Transferable Architectures for Scalable Image Recognition," in *IEEE CVPR*, 2018.

[144] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint Face Detection and Alignment using Multitask Cascaded Convolutional Networks," *IEEE Signal Processing Letters*, vol. 23, pp. 1499–1503, 4 2016.

[145] K. Gregor, I. Danihelka, A. Graves, D. J. Rezende, and D. Wierstra, "DRAW: A recurrent neural network for image generation," in *International Conference on Machine Learning*, vol. 2, pp. 1462–1471, International Machine Learning Society (IMLS), 2015.

[146] J. Chung, K. Kastner, L. Dinh, K. Goel, A. Courville, and Y. Bengio, "A recurrent latent variable model for sequential data," in *Advances in Neural Information Processing Systems*, vol. 2015-January, pp. 2980–2988, 2015.

[147] J. Bayer and C. Osendorfer, "Learning Stochastic Recurrent Networks," in *NIPS 2014 Workshop on Advances in Variational Inference*, 11 2014.

[148] W. Zhu, C. Lan, J. Xing, W. Zeng, Y. Li, L. S. T. A. C. . . . , and u. 2016, "Co-occurrence feature learning for skeleton based action recognition using regularized deep LSTM networks," in *the AAAI conference on artificial intelligence*, 2016.

[149] W. Zhu, C. Liu, W. Fan, and X. Xie, "DeepLung: Deep 3D Dual Path Nets for Automated Pulmonary Nodule Detection and Classification," in *IEEE Winter Conference on Applications of Computer Vision*, 1 2018.

[150] Y. Zhu, H. Li, Y. Liao, B. Wang, Z. Guan, H. Liu, and D. Cai, "What to do next: Modeling user behaviors by Time-LSTM," in *IJCAI International Joint Conference on Artificial Intelligence*, pp. 3602–3608, International Joint Conferences on Artificial Intelligence, 2017.

[151] F. G. Duhaylongsod, V. J. Lowe, E. F. Patz, A. L. Vaughn, R. E. Coleman, and W. G. Wolfe, "Lung tumor growth correlates with glucose metabolism measured by fluoride-18 fluorodeoxyglucose positron emission tomography," *The Annals of Thoracic Surgery*, vol. 60, no. 5, pp. 1348–1352, 1995.

[152] J. Cai, L. Lu, Y. Xie, F. Xing, and L. Yang, "Improving Deep Pancreas Segmentation in CT and MRI Images via Recurrent Neural Contextual Learning and Direct Loss Function," tech. rep., 7 2017.

[153] R. Santeramo, S. Withey, and G. Montana, "Longitudinal detection of radiological abnormalities with time-modulated lstm," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support.*, vol. 11045 LNCS, pp. 326–333, Springer Verlag, 2018.

[154] A. e. a. Paszke, "PyTorch: An Imperative Style, High-Performance Deep Learning Library," in *NeurIPS*, 2019.

[155] Y. Shen and M. Gao, "Brain Tumor Segmentation on MRI with Missing Modalities," in *International Conference on Information Processing in Medical Imaging*, vol. 11492 LNCS, pp. 417–428, Springer Verlag, 2019.

[156] J. Yoon, J. Jordon, and M. Van Der Schaar, "GAIN: Missing data imputation using generative adversarial nets," in *International Conference on Machine Learning*, vol. 13, pp. 9042–9051, 6 2018.

[157] Y. Tian, J. Shi, B. Li, Z. Duan, and C. Xu, "Audio-Visual Event Localization in Unconstrained Videos," in *ECCV*, vol. 11206 LNCS, pp. 252–268, 2018.

[158] J. Donahue, T. Darrell, and P. Krähenbühl, "Adversarial feature learning," in *ICLR*, 5 2017.

[159] M. Mirza and S. Osindero, "Conditional Generative Adversarial Nets," *arXiv preprint arXiv:1411.1784*, 11 2014.

[160] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-mniST: A novel image dataset for benchmarking machine learning algorithms," in *arXiv:1708.07747*, 8 2017.

[161] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium," *Advances in Neural Information Processing Systems*, vol. 2017-December, pp. 6627–6638, 6 2017.

[162] S. C.-X. Li, "An encoder-decoder framework for learning from incomplete data," 2020.

[163] Y. Mirsky, T. Mahler, I. Shelef, and Y. Elovici, "CT-GAN: Malicious tampering of 3D medical imagery using deep learning," in *Proceedings of the 28th USENIX Security Symposium*, pp. 461–478, USENIX Association, 1 2019.

[164] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On Calibration of Modern Neural Networks," *34th International Conference on Machine Learning, ICML 2017*, vol. 3, pp. 2130–2143, 6 2017.

[165] P. C. Austin and E. W. Steyerberg, "The Integrated Calibration Index (ICI) and related metrics for quantifying the calibration of logistic regression models," *Statistics in Medicine*, vol. 38, pp. 4051–4065, 9 2019.

[166] J. Byrd and Z. C. Lipton, "What is the Effect of Importance Weighting in Deep Learning?," *ICML*, 12 2018.

[167] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep Learning Face Attributes in the Wild," in *CVPR*, 2015.

[168] T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common Objects in Context," in *ECCV*, vol. 8693 LNCS, pp. 740–755, Springer Verlag, 5 2014.

[169] Z. Zhong, J. Cui, S. Liu, and J. Jia, "Improving Calibration for Long-Tailed Recognition," in *CVPR*, 2021.

[170] J. Tan, C. Wang, B. Li, Q. Li, W. Ouyang, C. Yin, and J. Yan, "Equalization Loss for Long-Tailed Object Recognition," 2020.

[171] B. Kang, S. Xie, M. Rohrbach, Z. Yan, A. Gordo, J. Feng, and Y. Kalantidis, "Decoupling Representation and Classifier for Long-Tailed Recognition," 10 2019.

[172] A. G. Roy, J. Ren, S. Azizi, A. Loh, V. Natarajan, B. Mustafa, N. Pawlowski, J. Freyberg, Y. Liu, Z. Beaver, N. Vo, P. Bui, S. Winter, P. MacWilliams, G. S. Corrado, U. Telang, Y. Liu, T. Cemgil, A. Karthikesalingam, B. Lakshminarayanan, and J. Winkens, "Does Your Dermatology Classifier Know What It Doesn't Know? Detecting the Long-Tail of Unseen Conditions," 4 2021.

[173] M. Buda, A. Maki, and M. A. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," *Neural Networks*, vol. 106, pp. 249–259, 10 2017.

[174] N. Japkowicz and S. Stephen, "The class imbalance problem: A systematic study," *Intelligent Data Analysis*, vol. 6, pp. 429–449, 1 2002.

[175] Y.-X. Wang, D. Ramanan, and M. Hebert, "Learning to Model the Tail," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[176] C. Huang, Y. Li, C. C. Loy, and X. Tang, "Learning Deep Representation for Imbalanced Classification," in *IEEE CVPR*, 2016.

[177] Y. Cui, M. Jia, T. Y. Lin, Y. Song, and S. Belongie, "Class-Balanced Loss Based on Effective Number of Samples," *IEEE CVPR*, 1 2019.

[178] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal Loss for Dense Object Detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, pp. 318–327, 8 2020.

[179] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond Empirical Risk Minimization," in *ICLR*, 10 2017.

[180] S. Thulasidasan, G. Chennupati, J. Bilmes, T. Bhattacharya, and S. Michalak, "On Mixup Training: Improved Calibration and Predictive Uncertainty for Deep Neural Networks," *Advances in Neural Information Processing Systems*, vol. 32, 5 2019.

[181] R. Krishnan and O. Tickoo, "Improving model calibration with accuracy versus uncertainty optimization," in *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[182] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," *IEEE CVPR*, 12 2016.

[183] R. Müller, S. Kornblith, and G. Hinton, "When Does Label Smoothing Help?," in *Advances in Neural Information Processing Systems*, vol. 32, Neural information processing systems foundation, 6 2019.

[184] M. Lukasik, S. Bhojanapalli, A. K. Menon, and S. Kumar, "Does label smoothing mitigate label noise?," in *ICML*, vol. PartF168147-9, International Machine Learning Society (IMLS), 3 2020.

[185] J. Mukhoti, V. Kulharia, A. Sanyal, S. Golodetz, P. H. S. Torr, and P. K. Dokania, "Calibrating Deep Neural Networks using Focal Loss," *arXiv*, 2 2020.

[186] O. Chapelle, J. Weston, L. Bottou, and V. Vapnik, "Vicinal Risk Minimization," *Advances in Neural Information Processing Systems*, vol. 13, 2000.

[187] M. P. Naeini, G. Cooper, and M. Hauskrecht, "Obtaining Well Calibrated Probabilities Using Bayesian Binning," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 29, 2 2015.

[188] "Consortium for Molecular and Cellular Characterization of Screen-Detected Lesions."

[189] R. Gao, Y. Tang, K. Xu, H. H. Lee, S. Deppen, K. Sandler, P. Massion, T. A. Lasko, Y. Huo, and B. A. Landman, "Lung Cancer Risk Estimation with Incomplete Data: A Joint Missing Imputation Perspective," in *MICCAI*, 7 2021.

[190] M. C. Tammemagi, H. Schmidt, S. Martel, A. McWilliams, J. R. Goffin, M. R. Johnston, G. Nicholas, A. Tremblay, R. Bhatia, G. Liu, K. Soghrati, K. Yasufuku, D. M. Hwang, F. Laberge, M. Gingras, S. Pasian, C. Couture, J. R. Mayo, P. V. Nasute Fauerbach, S. Atkar-Khattra, S. J. Peacock, S. Cressman, D. Ionescu, J. C. English, R. J. Finley, J. Yee, S. Puksa, L. Stewart, S. Tsai, E. Haider, C. Boylan, J. C. Cutz, D. Manos, Z. Xu, G. D. Goss, J. M. Seely, K. Amjadi, H. S. Sekhon, P. Burrowes, P. MacEachern, S. Urbanski, D. D. Sin, W. C. Tan, N. B. Leighl, F. A. Shepherd, W. K. Evans, M. S. Tsao, and S. Lam, "Participant selection for lung cancer screening by risk modelling (the Pan-Canadian Early Detection of Lung Cancer [PanCan] study): a single-arm, prospective study," *The Lancet Oncology*, vol. 18, pp. 1523–1531, 11 2017.

[191] C. I. Henschke, R. Yip, D. F. Yankelevitz, and J. P. Smith, "Definition of a positive test result in computed tomography screening for lung cancer," *Annals of Internal Medicine*, vol. 158, pp. 246–252, 2 2013.

[192] N. Horeweg, C. M. v. d. Aalst, R. Vliegenthart, Y. Zhao, X. Xie, E. T. Scholten, W. Mali, E. Thunnissen, C. Weenink, H. J. Groen, J.-W. J. Lammers, K. Nackaerts, J. v. Rosmalen, M. Oudkerk, and H. J. d. Koning, "Volumetric computed tomography screening for lung cancer: three rounds of the NELSON trial," *European Respiratory Journal*, vol. 42, pp. 1659–1667, 12 2013.

[193] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "SphereFace: Deep hypersphere embedding for face recognition," in *IEEE CVPR*, vol. 2017-January, pp. 6738–6746, Institute of Electrical and Electronics Engineers Inc., 11 2017.

[194] R. Gao, F. Yang, W. Yang, and Q. Liao, "Margin Loss: Making Faces More Separable," *IEEE Signal Processing Letters*, vol. 25, no. 2, 2018.

[195] R. Gao, M. S. Khan, Y. Tang, K. Xu, S. Deppen, Y. Huo, K. L. Sandler, P. P. Massion, and B. A. Landman, "Technical Report: Quality Assessment Tool for Machine Learning with Clinical CT," 7 2021.

[196] "Data Science Bowl 2017 — Kaggle." https://www.kaggle.com/c/data-science-bowl-2017. accessed 2019/12/01.

[197] "Lung-RADS — Radiology Reference Article — Radiopaedia.org." https://radiopaedia.org/articles/lung-rads?lang=us. accessed 2020/05/22.

[198] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, "Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach," *Biometrics*, vol. 44, p. 837, 9 1988.

[199] "Data Science Bowl 2017 — Kaggle."

[200] Y. Li, H. Liu, and Y. Fan, "DeepSEED: 3D Squeeze-and-Excitation Encoder-Decoder ConvNets for Pulmonary Nodule Detection," in *IEEE International Symposium on Biomedical Imaging*, 4 2020.

[201] R. Gao, Y. Tang, M. S. Khan, K. Xu, A. B. Paulson, S. Sullivan, Y. Huo, S. Deppen, P. P. Massion, K. L. Sandler, and B. A. Landman, "Cancer Risk Estimation Combining Lung Screening CT with Clinical Data Elements," *Radiology: Artificial Intelligence*, 10 2021.

[202] D. E. Ost and M. K. Gould, "Decision making in patients with pulmonary nodules," *American journal of respiratory and critical care medicine*, vol. 185, pp. 363–372, 2 2012.

[203] M. N. Kammer, D. A. Lakhani, A. B. Balar, S. L. Antic, A. K. Kussrow, R. L. Webster, S. Mahapatra, U. Barad, C. Shah, T. Atwater, B. Diergaarde, J. Qian, A. Kaizer, M. New, E. Hirsch, W. J. Feser, J. Strong, M. Rioth, Y. E. Miller, Y. Balagurunathan, D. J. Rowe, S. Helmey, S.-C. Chen, J. Bauza, S. A. Deppen, K. Sandler, F. Maldonado, A. Spira, E. Billatos, M. B. Schabath, R. J. Gillies, D. O. Wilson, R. C. Walker, B. Landman, H. Chen, E. L. Grogan, A. E. Barón, D. J. Bornhop, and P. P. Massion, "Integrated Biomarkers for the Management of Indeterminate Pulmonary Nodules," *https://doi.org/10.1164/rccm.202012-4438OC*, vol. 204, pp. 1306–1316, 12 2021.

[204] D. R. Baldwin and M. E. Callister, "The British Thoracic Society guidelines on the investigation and management of pulmonary nodules," *Thorax*, vol. 70, pp. 794–798, 8 2015.

[205] S. e. a. Bakas, "Identifying the Best Machine Learning Algorithms for Brain Tumor Segmentation, Progression Assessment, and Overall Survival Prediction in the BRATS Challenge," in *arXiv preprint arXiv:1811.02629*, vol. 124, 11 2018.

[206] A. E. Johnson, T. J. Pollard, L. Shen, L. W. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, and R. G. Mark, "MIMIC-III, a freely accessible critical care database," *Scientific Data 2016 3:1*, vol. 3, pp. 1–9, 5 2016.

Appendix A

Copyright from Publishers

A.1 Copyright from arXiv

Our technique report (Chapter 3) is under the license of CC BY-NC-ND. I, Riqiang Gao, am the creator and holder, retain ownership of the manuscript. A screenshot of copyright/License information are shown in A.1. Our Chapter 8 and Chapter 10 are under preparation for submission, whose newer version may be appearing in arXiv.

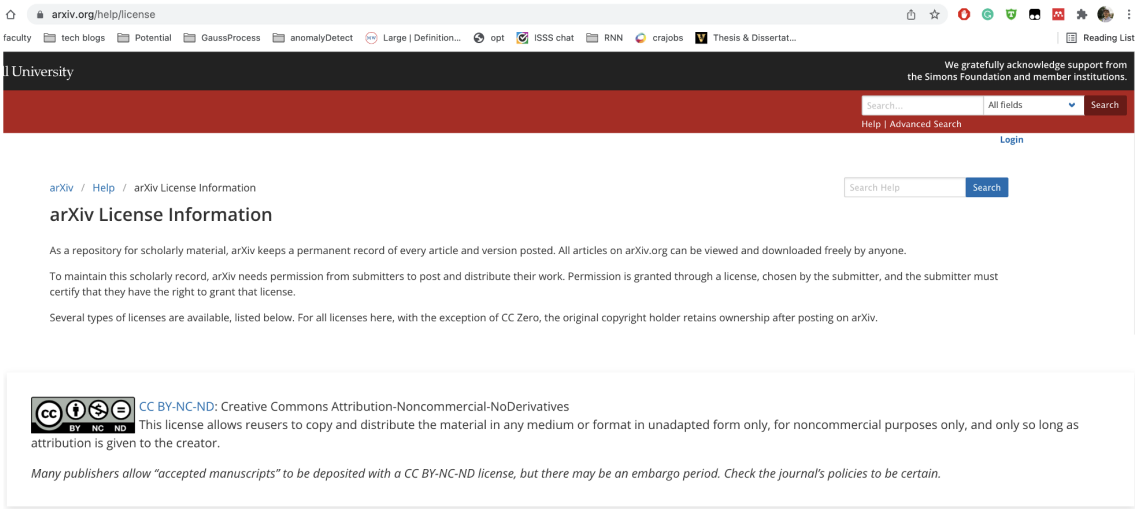
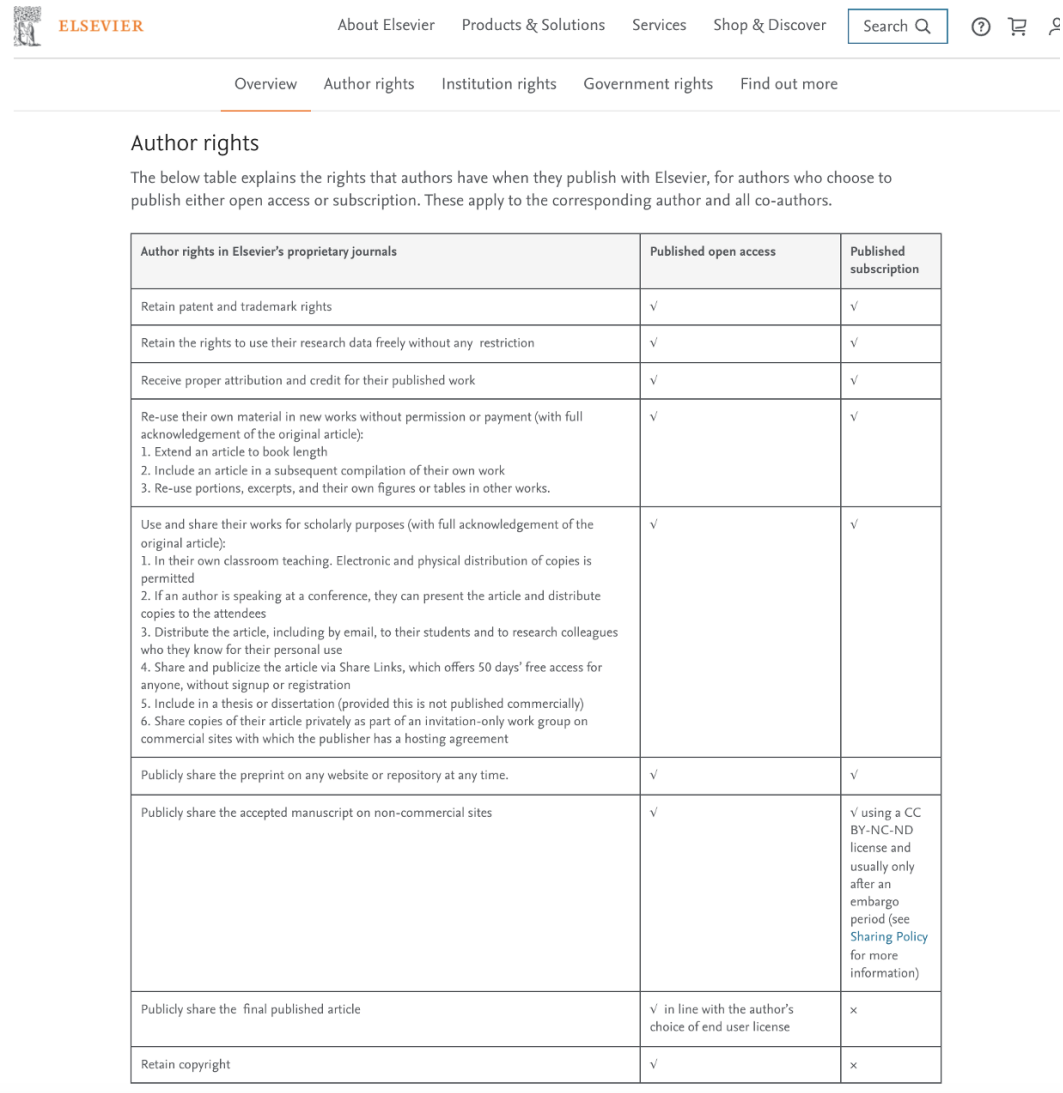


Figure A.1: Copyright from arXiv

A.2 Copyright from Elsevier

Author rights in Elsevier's proprietary journals (A.2) include re-use portions, excerpts, and their own figures or tables in other works. Our Neurocomputing paper (Chapter 4) and Medical Image Analysis paper Chapter 5 are under Elsevier.



ELSEVIER About Elsevier Products & Solutions Services Shop & Discover Search

Overview Author rights Institution rights Government rights Find out more

Author rights

The below table explains the rights that authors have when they publish with Elsevier, for authors who choose to publish either open access or subscription. These apply to the corresponding author and all co-authors.

Author rights in Elsevier's proprietary journals	Published open access	Published subscription
Retain patent and trademark rights	√	√
Retain the rights to use their research data freely without any restriction	√	√
Receive proper attribution and credit for their published work	√	√
Re-use their own material in new works without permission or payment (with full acknowledgement of the original article): 1. Extend an article to book length 2. Include an article in a subsequent compilation of their own work 3. Re-use portions, excerpts, and their own figures or tables in other works.	√	√
Use and share their works for scholarly purposes (with full acknowledgement of the original article): 1. In their own classroom teaching. Electronic and physical distribution of copies is permitted 2. If an author is speaking at a conference, they can present the article and distribute copies to the attendees 3. Distribute the article, including by email, to their students and to research colleagues who they know for their personal use 4. Share and publicize the article via Share Links, which offers 50 days' free access for anyone, without signup or registration 5. Include in a thesis or dissertation (provided this is not published commercially) 6. Share copies of their article privately as part of an invitation-only work group on commercial sites with which the publisher has a hosting agreement	√	√
Publicly share the preprint on any website or repository at any time.	√	√
Publicly share the accepted manuscript on non-commercial sites	√	√ using a CC BY-NC-ND license and usually only after an embargo period (see Sharing Policy for more information)
Publicly share the final published article	√ in line with the author's choice of end user license	×
Retain copyright	√	×

Figure A.2: Copyright from Elsevier

A.3 Copyright from LNCS

Authors retains the right to use the content for mon-commercial internal and educational purposes, etc. Our Chapter 7 is a extension of our MICCAI paper under Copyright from LNCS (A.3).

§ 2 Rights Retained by Author

Author retains, in addition to uses permitted by law, the right to communicate the content of the Contribution to other research colleagues, to share the Contribution with them in manuscript form, to perform or present the Contribution or to use the content for non-commercial internal and educational purposes, provided the original source of publication is cited according to the current citation standards in any printed or electronic materials. Author retains the right to republish the Contribution in any collection consisting solely of Author's own works without charge, subject to ensuring that the publication of the Publisher is properly credited and that the relevant copyright notice is repeated verbatim. Author may self-archive an author-created version of his/her Contribution on his/her own website and/or the repository of Author's department or faculty. Author may also deposit this version on his/her funder's or funder's designated repository at the funder's request or as a result of a legal obligation. He/she may not use the Publisher's PDF version, which is posted on the Publisher's platforms, for the purpose of self-archiving or deposit. Furthermore, Author may only post his/her own version, provided acknowledgment is given to the original source of publication and a link is inserted to the published article on the Publisher's website. The link must be provided by inserting the DOI number of the article in the following sentence: "The final authenticated version is available online at https://doi.org/[insert DOI]." The DOI (Digital Object Identifier) can be found at the bottom of the first page of the published paper.

Prior versions of the Contribution published on non-commercial pre-print servers like ArXiv/CoRR and HAL can remain on these servers and/or can be updated with Author's accepted version. The final published version (in pdf or html/xml format) cannot be used for this purpose. Acknowledgment needs to be given to the final publication and a link must be inserted to the published Contribution on the Publisher's website, by inserting the DOI number of the article in the following sentence: "The final authenticated publication is available online at https://doi.org/[insert DOI]".

Author retains the right to use his/her Contribution for his/her further scientific career by including the final published paper in his/her dissertation or doctoral thesis provided acknowledgment is given to the original source of publication. Author also retains the right to use, without having to pay a fee and without having to inform the Publisher, parts of the Contribution (e.g. illustrations) for inclusion in future work. Authors may publish an extended version of their proceedings paper as a journal article provided the following principles are adhered to: a) the extended version includes at least 30% new material, b) the original publication is cited, and c) it includes an explicit statement about the increment (e.g., new results, better description of materials, etc.).

Figure A.3: Copyright from LNCS

A.4 Copyright from SPIE

This is the permission obtained from SPIE. The email screenshot is shown in A.4.

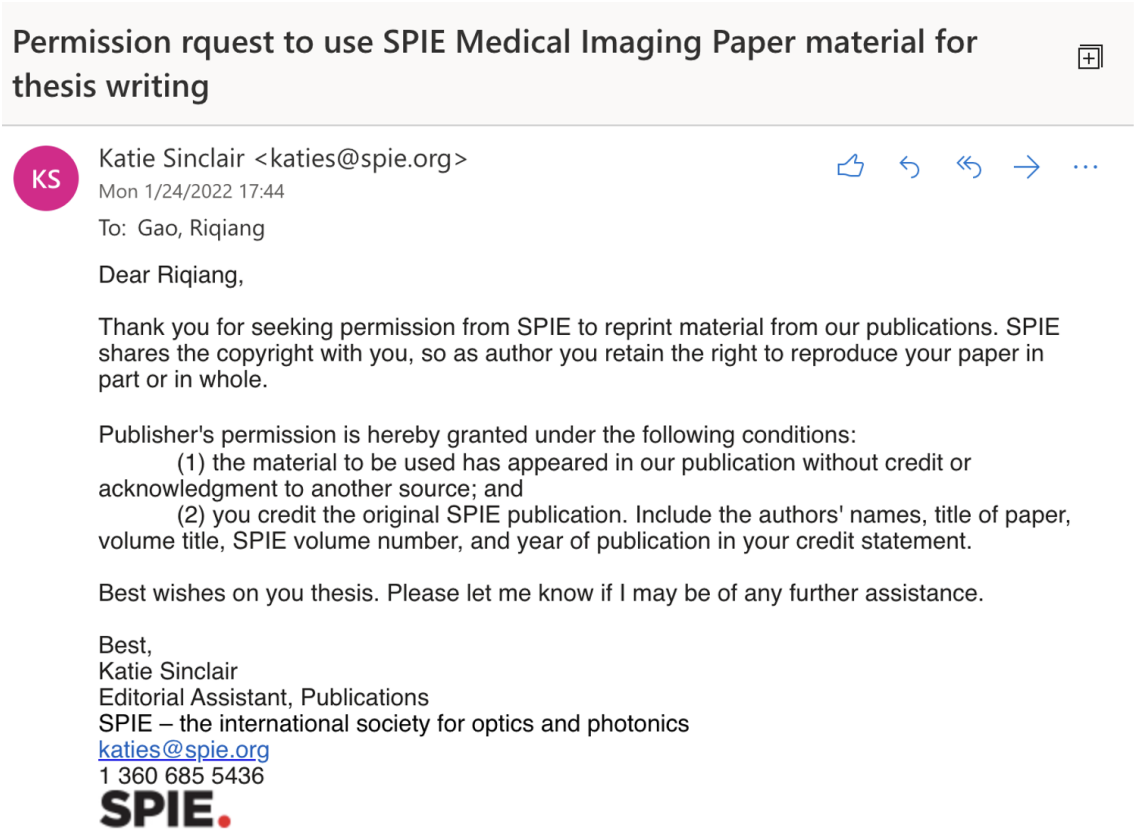


Figure A.4: Copyright from SPIE

A.5 Copyright from RSNA

This is the permission obtained from RSNA, Radiology AI (Chapter 9). The email screenshot is shown in A.5.

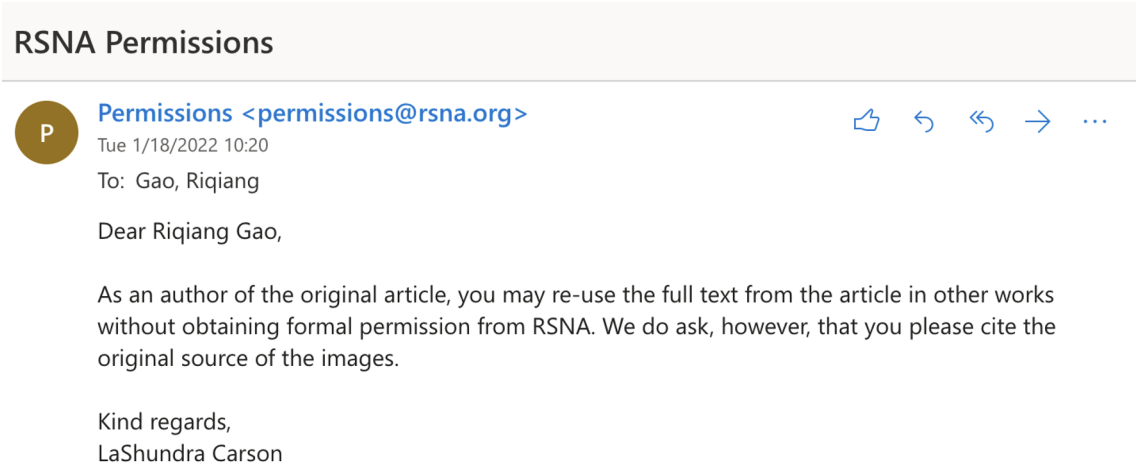


Figure A.5: Copyright from RSNA