Deep Learning Based Computer-aided Detection and Diagnosis Systems for Medical Imaging

A Thesis Submitted to the
College of Graduate and Postdoctoral Studies
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy
in the Department of Electrical and Computer Engineering
University of Saskatchewan
Saskatoon, Saskatchewan, Canada

By
Yi Wang

Permission to Use

In presenting this thesis in partial fulfilment of the requirements for a Postgraduate degree from the University of Saskatchewan, I agree that the Libraries of this University may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis work or, in their absence, by the Head of the Department or the Dean of the College in which my thesis work was done. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my thesis.

Requests for permission to copy or to make other use of material in this thesis in whole or part should be addressed to:

Head of the Department of Electrical and Computer Engineering
3B48 Engineering Building
University of Saskatchewan
57 Campus Drive
Saskatoon, Saskatchewan S7N 5A9
Canada

Or

Dean of College of Graduate and Postdoctoral Studies
116 Thorvaldson Building
University of Saskatchewan
110 Science Place
Saskatoon, Saskatchewan S7N 5C9
Canada

Abstract

Cancer is a major burden on public health worldwide and is the first leading cause of death. Early detection and diagnosis is the single most important way to improve cancer survival rates. Due to the high sensitivity to detect cancer, medical imaging has become a widespread screening modality for cancer diagnoses. With the advancement of technologies, modern medical imaging allows radiologists to perform interpretation in a 3D space thereby diagnostic performances of radiologists are improved. However, reviewing medical imaging is still a time-consuming task. For each patient case, radiologists need to review hundreds of images. With the continuous increment in new cancer patients daily, the limited number of radiologists can potentially delay treatment. To remedy this issue, computer-aided detection and diagnosis (CAD) system has been developed. The CAD system is designed to analyze medical imaging and provide a second objective opinion that supports radiologists in medical imaging interpretation and diagnostic decision-making. Nevertheless, there are currently not many uses for CAD systems applied in real clinical situations due to the high false-positive rates to detect lesions and limited performance to classify correct types of lesions. In recent years, deep learning has achieved great success in natural image classification, object detection, semantic segmentation and language processing. In many of these fields, deep learning can surpass or achieve near-human performance. The convolutional neural network (CNN), one of the most common deep learning architectures, shows a remarkable ability to extract discriminative features from the raw image input and perform prediction without any manual intervention. Thus, it is expected to improve the performance of existing CAD systems. In this thesis, several CNN architectures are developed to improve existing CAD systems, practicuarlly for lung cancer and anterior mediastinal nodular lesion diagnosis on computed tomography (CT) imaging and breast cancer diagnosis on automated breast ultrasound (ABUS) imaging. The content of this thesis can be divided into three parts: (1) A CNN-based computer-aided diagnosis (CADx) system for pulmonary nodule classification on CT imaging is proposed in this thesis. (2) A CNN-based CADx system for breast lesion classification on ABUS imaging is proposed in this thesis. (3) A fully convolutional neural network (FCN)-based computer-aided detection (CADe) system for anterior mediastinal

nodular lesion segmentation on chest CT imaging is proposed in this thesis.

CT imaging is widely used for lung cancer diagnosis. However, it is still not an easy task for experienced radiologists to differentiate pulmonary nodules from malignant to benign. In this thesis, a novel CNN architecture is proposed to predict the malignancy of pulmonary nodules on CT imaging. Compared to conventional CNN architecture, the proposed CNN adopts multi-scale convolutional layers and a multi-path feature extraction scheme. The proposed multi-scale convolutional layer utilizes a set of convolutional kernels to perform feature extraction that covers more effective local regions than the conventional convolution operation. In addition, the proposed multi-path feature extraction scheme combines features that are extracted from different layers. Hence, features that describe the global structures of the nodules are retained, including shapes, sizes and contours of nodules.

Handheld ultrasound (HHUS) imaging is commonly used for breast cancer diagnosis. However, HHUS imaging is highly operator depended since the viewing angle of acquired HHUS imaging is controlled by the operator. Recently, ABUS imaging has been introduced to construct the whole breast in an automatic fashion. As a single ABUS scan composites hundreds of slice images, it is a time-consuming task for radiologists to review all slice images. In this thesis, a CNN-based CADx system is proposed to predict the malignancy of breast lesions on ABUS imaging. The presence of a breast lesion on ABUS imaging can be visualized from different viewing planes, such as transverse view and coronal view. Instead of extracting features within a fixed one-dimensional plane, two multiview learning strategies are proposed that allows the proposed CNN to extract multiview features from both transverse and coronal view.

Anterior meidastinal nodular lesions (AMLs) are abnormal growths found in the anterior mediastinum. Chest CT imaging is the most common screening modality used for AML diagnosis. However, it is still challenging for radiologists to accurately detect AMLs from chest CT imagining due to the low contrast of reconstructed images, wide variations in intensity within an AML and similarity of AMLs to other tissues. Therefore, an automated detection procedure for AMLs has significance to save screening time. To improve the reading effort for radiologists, a FCN-based CADe system is proposed to realize automated AML segmentation from 2D chest CT slice images. For the proposed network, the self-attention

mechanism is utilized to improve the network performance by modelling global semantic correlations. Besides, an image-grid attention mechanism is employed to allow the proposed network to selectively focus on learning the characteristics of AMLs and disregard others in the image.

Acknowledgements

I would first like to express my deep sense of thanks and gratitude to my supervisor Dr. Seokbum Ko of the Department of Electrical and Computer Engineering at the University of Saskatchewan. I was extremely lucky to have a supervisor who always willing to help me in all the time of research and writing of this thesis. Without his patience and knowledge, I would not have been accomplished the works presented in the thesis.

I am truly grateful to have the opportunities to work with Dr. Hao Zhang, Dr. Kum Ju Chae, Dr. Eun Jung Choi, Dr. Won Gi Jeong, Dr. Younhee Choi and Dr. Gong Yong Jin. Their valuable suggestions and advice make all work possible to be presented in this thesis.

I would like to express my sincere thanks to my committee members, Dr. Li Chen, Dr. Khan Wahid and Dr. Fang X. Wu for their time to review this thesis and their contributions to improve the quality of this thesis.

Finally, I would like to thank my parents for their continuous supports and encouragements throughout the period of writing this thesis.

Contents

List of Tables

List of Figures

List of Abbreviations

ABUS	Automated Breast Ultrasound
AML	Anterior Mediastinal Nodular Lesion
ANN	Artificial Neural Network
AUC	Area Under Receiver Operating Characteristic Curve
CAD	Computer-aided Detection and Diagnosis
CADe	Computer-aided Detection
CADx	Computer-aided Diagnosis
CBAM	Convolutional Block Attention module
CNN	Convolutional Neural Network
Conv	Convolutional
CT	Computed Tomography
DICOM	Digital Imaging and Communication in Medicine
DL	Deep Learning
DSC	Dice Similarity Coefficient
FC	Fully-connected
FCN	Fully Convolutional Neural Network
GAP	Global Average Pooling
HOG	Histogram of Oriented Gradients
HU	Houndsfield Unit
IoU	Intersection over Union
KNN	K-nearest Neighbor
LBP	Local Binary Pattern
LDA	Linear Discriminant Analysis
MHA	Multi-Head Attention
ML	Machine Learning
MLP	Multilayer Perceptron
PCA	Principal Component Analysis

ReLU	Rectified Linear Unit
ROC	Receiver Operating Characteristic Curve
SGD	Stochastic Gradient Descent
SVM	Support Vector Machines
US	Ultrasound

Chapter 1
Introduction

This chapter presents the usefulness of incorporating computer-aided detection and diagnosis systems with medical imaging to assist radiologists with cancer diagnoses. Due to the high false-positive rate and limited classification performance to categorize different types of lesions, improving existing computer-aided detection and diagnosis (CAD) systems is desired. As the most advanced technique for implementing deep learning, the convolutional neural network (CNN) is getting more and more attention in recent years. CNNs achieve near or even better than human-level performance in various computer vision tasks, including image classification, object detection and semantic segmentation. Thus, it is expected to be effective for the design of CAD systems. This motivates the works presented in this thesis to explore the feasibilities of applying CNNs to improve existing CAD systems, particularly for lung cancer, breast cancer and anterior mediastinal nodular lesion diagnosis. Section 1.1 describes the use of medical imaging and CAD systems. Besides, the advantage of CNN is presented in Section 1.1. The motivations for the works presented in this thesis are discussed in Section 1.2. The objectives of this thesis is presented in Section 1.3. The contributions for the works described in thesis are described in Section 1.4. The overview of this thesis is presented in Section 1.5.

1.1 Cancer Diagnoses with Computer-aided Detection and Diagnosis Systems

The foundation of the human body is made up of trillions of cells. Cells are essential for structuring the human body, absorbing nutrients from food, and transforming nutrients into

energy. Cells can grow or split into new cells based on human body needs. When cells are too old or damaged, to sustain the functionality of the human body, new cells are generated to replace the old or damaged cells. Cancer is a disease that interferes with the normal process of cells. Affected cells grow uncontrollably and have the potential to spread throughout the body, which disrupts organ function and causes death. Nowadays, with the rapid improvement of health technology, cancer remains a major burden on public health worldwide. According to Canadian Cancer Statics reported in 2022 [1], cancer is the first leading cause of death in Canada. 43% of all Canadian population is expected to arise cancer during their lifetime. However, the cancer survival rate is low. About 1 in 4 patients diagnosed with cancer will eventually die from cancer. Hence, early detection and diagnosis is an important way to reduce the cancer mortality rate.

Although cancer has been discovered in early 400 BC, there was no efficient way to diagnose cancer without performing surgery. Until the later 18th century, medical imaging has been introduced with the invention of the X-ray. By employing ionizing radiation through the body, constructed X-ray imaging visualizes the interior of the body such as organs or tissues. X-ray imaging becomes an efficient screening modality to diagnose cancer as X-ray imaging allows radiologists to seek internal structures of the body and determine abnormal formations. In the early 20th century, computed tomography (CT) has been developed. By using multiple X-ray sources with various angles, created CT imaging produces a 3D view of the internal body. Compared to X-ray imaging, CT imaging delivers better visual representations to describe the position, shape and size of internal body structures. CT imaging has become a gold-standard screening modality for pulmonary cancer diagnosis nowadays. In addition, CT imaging has been widely used for the diagnosis of stomach cancer, liver cancer, pancreatic cancer, etc. Instead of using ionizing radiation for medical imaging reconstruction, lots of efforts have been attempted to apply ultrasound technologies to the medical imaging field. Ultrasound (US) imaging uses ultrasound sound waves with high frequency to construct the internal structures of the body. Since the US has no ionizing radiation that may cause additional harm to the examiner's health, US imaging has become a popular screening modality and commonly utilized for breast cancer diagnosis. However, US imaging is highly operator dependent. For instance, interpreting breast cancer relies on the operator's

experience as reconstructed US imaging visualizes the entire breast that is limited within one orientation. To minimize operator dependence, automated breast ultrasound (ABUS) imaging has been introduced to visualize the whole breast in a 3D view. In contrast to the US, the screening time and breast cancer detection rate have been improved by using ABUS imaging. With the advancement in medical imaging technologies, radiologists are allowed to interpret imaging in a 3D view thereby diagnostic performances are improved. Nevertheless, screening CT and ABUS imaging are time-consuming tasks even for experienced radiologists. For example, a single patient's CT generates hundreds of slice images used to reconstruct the 3D volume. To perform diagnosis, radiologists need to review hundreds of slice images for each patient. With the dramatic increase in cancer diagnoses nowadays, the limited number of radiologists can lead to the patient not being offered effective treatment on time.

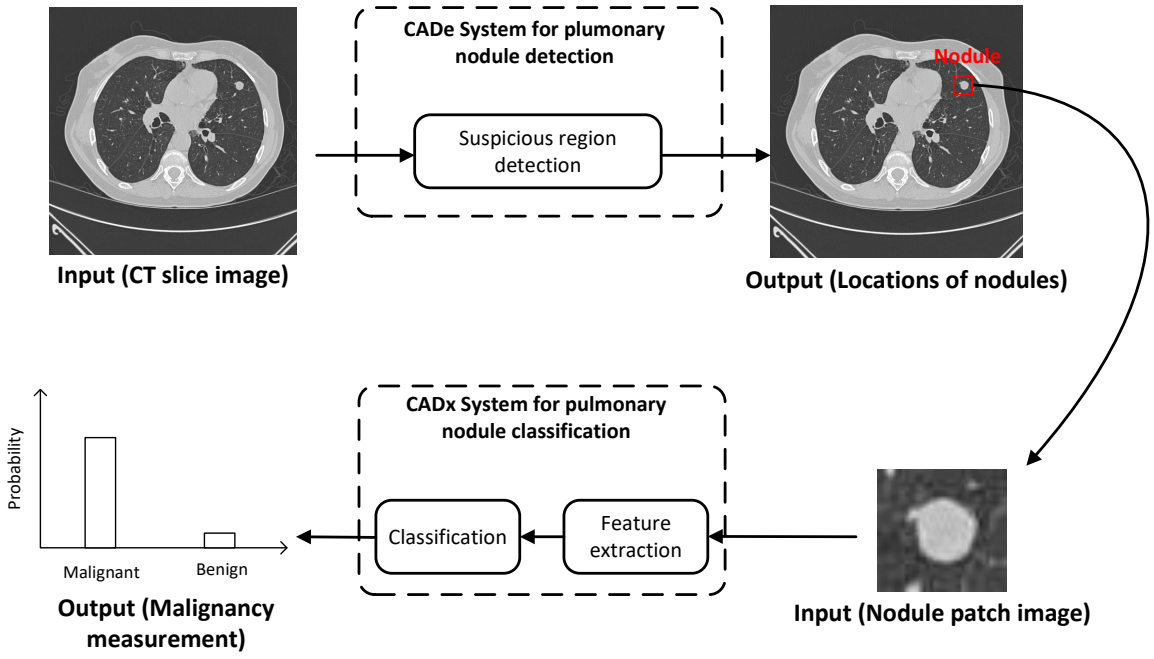


Figure 1.1: A general workflow of a CAD system for lung cancer diagnosis on CT imaging.

To reduce the time in medical imaging interpretation, the CAD system has been developed. The CAD system is a computerized procedure designed to assist radiologists by providing a second diagnostic opinion. A general CAD system consists of two sub-systems, which are a computer-aided detection (CADe) system and a computer-aided diagnosis (CADx) sys-

tem. An example of CAD system designed for lung cancer diagnosis on CT imaging is shown in Figure 1.1. Specifically, a CADe system is designed to mark suspicious regions inside imaging that help radiologists to identify the locations of abnormalities. Besides, a CADx system is utilized to measure the malignancies of suspicious regions by extracting features from these regions. Many previous studies have demonstrated the usefulness of employing CAD systems to support radiologists for improved cancer diagnosis. An experiment was conducted to compare the diagnostic performances of radiologists with the CADe system used for breast lesion detection in ABUS imaging [2]. Their experimental results concluded that employing the CADe system improved the detection rate of radiologists while screening time is saved. Chae et. al [3] designed a CADx system to assist radiologists as the second reviewer for pulmonary nodule characterization in chest CT imaging. During their observation performance test, with the aid of the CADx system, physicians showed significant improvement in classifying pulmonary nodules between benign and malignant. Although it is shown that CAD systems improve radiologists' diagnostic performance and reading efficiency, there are currently not many practical clinical usages of CAD systems due to high false-positive rates and difficulty of characterizing lesions with a wide variation in appearances [4]. Hence, improving existing CAD systems to fit in clinical practices is desired.

In 1943, the concept of machine learning (ML) was coined by Arthur Samuel. ML is a subset of artificial intelligence, that offers the ability of computers to learn input data and use learned patterns to solve problems or make decisions. As ML allows computers to predict results without explicitly designing a complex procedure, ML has brought significant interest in medical imaging analysis. Numerous studies have made successful attempts to incorporate ML algorithms in CAD systems. In 1994, Wolberg et al. [5] developed an ML-based CAD system to diagnose breast cancer from fine needle aspirate slides. In 1996, ML algorithms were utilized to diagnose gastric and oesophageal cancer [6]. However, due to limited computation power around that time, ML algorithms were designed on small scale and thereby performances of early developed ML-based CAD systems were limited. In the early 2000s, with the rapid growth in computation power, it makes possible to model large-scale ML algorithms, known as deep learning (DL). During that time, the considerable improvement in the ability to gather large-scale databases also stimulated the development

of DL algorithms. In 2012, AlexNet [7], The CNN derived from DL algorithms, achieved significant improvement over its predecessors in natural image classification. Thereafter, many research efforts have been made to improve the performances of CNNs. In 2016, two novel CNN architectures, Inception-v3 [8] and ResNet [9], have been developed and achieved remarkable image classification accuracies even better than humans. In addition, CNNs have come to exceed other proceeding methods on other computer vision tasks such as image segmentation [10, 11] or object detction [12, 13]. As a subset of ML, CNNs have become a proper candidate to replace conventional ML algorithms due to two reasons as follows: (1) A conventional ML algorithm requires several manual steps to select a feature extractor and a classifier. In contrast, without any manual intervention, CNN is designed as an end-to-end system that performs feature extraction and classification in an automatic manner. (2) As shown in Figure 1.2, when data is small scale, the conventional ML algorithms outperform CNNs. However, the performances of conventional ML algorithms are saturated as the size of data increases.

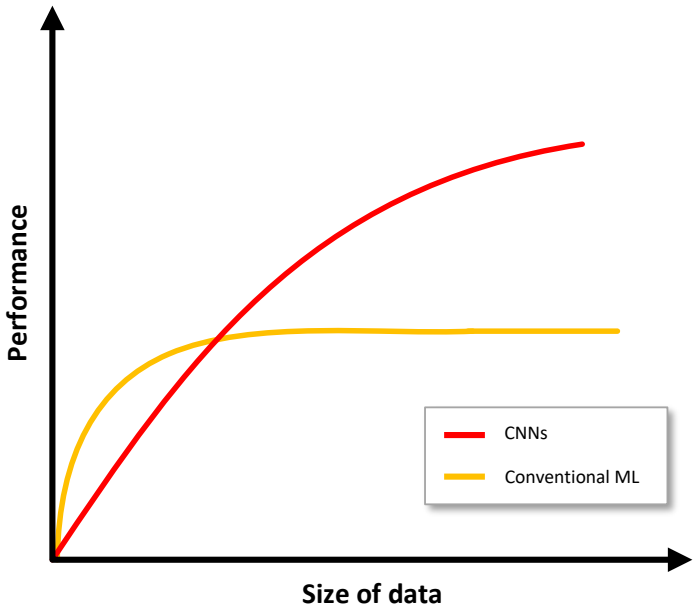


Figure 1.2: Performances of convolutional neural network and conventional machine learning algorithm regarding different data size [14].

In recent years, the availability of accessing large-scale medical imaging databases has

become possible. Thus, CNNs are expected to be helpful to improve the diagnostic performances of existing CAD systems. In this thesis, the focus is put on the design of deep learning-based CAD systems to assist radiologists in medical imaging interpretation. Particularly Particularly, different novel CNN architectures tailored for medical imaging tasks are presented in this thesis.

1.2 Motivations

As the most frequently diagnosed cancer, lung cancer remains the first leading cause of cancer death worldwide [15]. According to the statistics of the American Lung Association [16], the five-year survival rate for lung cancer patients is 18.6%, which is significantly lower than other commonly diagnosed cancer such as breast cancer or prostate cancer. Therefore, early diagnosis is important to reduce the lung cancer mortality rate. CT imaging is the most common screening modality for lung cancer diagnosis because of its high sensitivity to detect pulmonary nodules. Nevertheless, it is still a challenging task for radiologists to differentiate malignant nodules from benign ones because nodules have a wide variation in shapes and sizes while malignant nodules may have similar visual representation as benign nodules. Early studies attempted to utilize non-textual features to detect nodules from CT imaging [17, 18]. Nevertheless, non-textual features are limited to describe variations in shapes, sizes or intensities. Non-textual features may not be generic to categorize pulmonary nodules between malignant and benign. Thus, a CADx system that can provide effective pulmonary nodule classification is desired.

More recent studies have demonstrated the usefulness of employing conventional ML algorithms for pulmonary nodule classification [19–21]. Compared to non-textual feature extractors, these methods are effective to extract high-level features having better generalization ability and discriminative power [20]. Compared to conventional ML approaches, CNNs take raw data as input and output predicting outcomes without explicitly designing a feature extractor and a classifier. To realize an ML-based CADx system, the optimal feature extractor and classifier are manually selected with numerous trials and errors. Besides, many large lung image databases become publically available, which allows CNN to take the

advantage of large data size. By considering these factors, the usefulness of utilizing CNN for pulmonary nodule classification should be investigated.

Early attempts have been made to incorporate CNNs into CADx systems for pulmonary nodule classification. Specifically, two design strategies are suggested to realize a CNN-based CADx system, including customized CNN and transfer learning. Previous studies have demonstrated the effectiveness of employing customized CNNs for pulmonary nodule classificaiton [22, 23]. These self-designed CNNs commonly consist of fewer layers than the state-of-the-art deep CNNs tailored for natural image classification. Compared to the natural image databases, the sizes of available lung image databases are still relatively small. Therefore, training a deep CNN using current lung image databases is difficult to get a trained model having robust generalization. To remedy this problem, transfer learning is commonly utilized [24, 25]. With transfer learning, the deep CNN, previously trained on a large-scale natural image database, is fine-tuned with a lung image database. However, a question is raised to select customized CNN or transfer learning for pulmonary nodule classification since both approaches showed promising results. Thus, the effectiveness of customized CNN tailored for pulmonary nodule classification should be explored and compared to transfer learning approaches.

Breast cancer is the most commonly diagnosed cancer among women and is the second leading cause of cancer death worldwide. To provide an opportune treatment and increase the survival rate, diagnosing breast cancer at an early stage is vital. ABUS imaging is an advanced screening modality that allows radiologists to perform screening in a 3D volume. Compared to conventional US imaging, screening ABUS imaging is less time-consuming and better reproducibility [26]. However, diagnosing breast cancer is still a difficult task. The reasons are as follows: (1) breast lesions have notable intra-class appearance variations to speckled artifacts and shadowing artifacts. (2) fat breast tissues and contact artifacts have similar visual representations as breast lesions. (3) breast lesions have a broad range of sizes, shapes and contours. By considering these issues, a CADx system is expected to assist radiologists for the classification of breast lesions as benign or malignant.

While numerous researches focused on developing CADx systems, particularly for US imaging, there have been few studies focusing on ABUS imaging. Particularly, ML algorithms

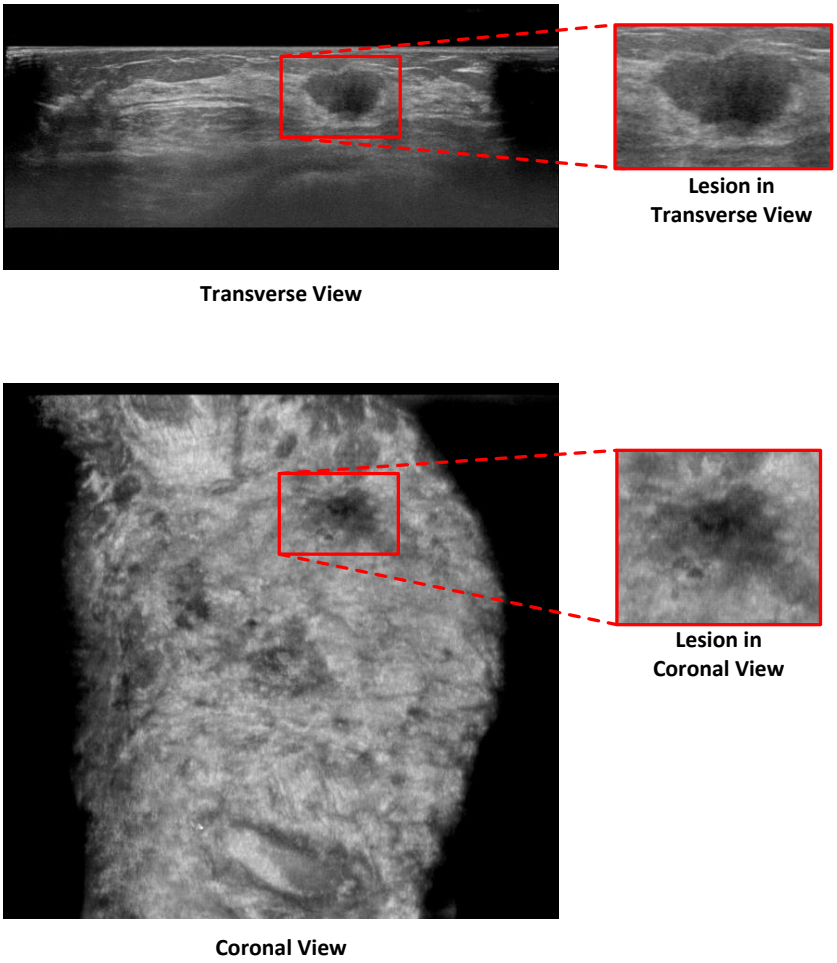


Figure 1.3: Different visual representations of a malignant breast lesion from different views.

are the dominant trend in opinion. However, these methods are designed to take processed lesion image data as input, which lesions need to be segmented [27, 28]. Therefore, the performance of these methods could be affected by segmentation algorithms. Since CNN is designed to take raw image data as input, it is expected to reduce intermediate steps that may downgrade overall system performance.

As ABUS imaging provides a way to visualize lesions in volumetric views, it is feasible to gather various contextual information of a lesion from different views. For example, as shown in Figure 1.3, the visual representation of a malignant lesion in the transverse view is different from the coronal view. By taking this finding, CNN can be designed to perform feature extraction from multiple views. With multiview feature extraction, more useful features are

extracted and expected to enhance CNN for improved lesion classification.

Anterior mediastinal nodular lesions (AMLs) often begin in the area of the chest that separates the lungs and the sternum. Although the prevalence of incidental AMNs is extremely rare [29], it is still an urgent medical concern due to a wide variety of diseases that are associated with AMLs, including thymoma, teratoma, thyroid goitre and lymphoma [30]. Hence, early diagnosis is important to prevent further cancer development. To detect AMLs, chest CT imaging is the most common screening modality for the initial assessment [31]. CT imaging provides a volumetric chest view, which is more effective than X-ray imaging. However, AMLs are characterized by a broad range of shapes and sizes while adjacent tissue or organs have similar appearances compared to the AMLs. By considering those facts, it is still a challenging task for radiologists to detect AMLs by interpreting CT imaging due to the high chances of missing or misreading AMLs. In addition, to verify lesion findings, the American College of Radiology recommends performing an additional magnetic resonance (MR) imaging scan [32]. Nevertheless, the workload of radiologists could be significantly increased due to the additional examination. A possible direction to alleviate those problems is to realize an automated CADe system that improves reading effort for radiologists.

Semantic segmentation is the process of dividing an image into multiple distinct regions that correspond to different objects. In natural image processing, fully convolutional neural networks (FCNs), a variant of CNNs, are widely used for semantic segmentation due to several reasons as follows: 1) FCNs are robust to handle large amounts of variations in the appearance of objects, such as size, shape and orientation. 2) Unlike conventional segmentation methods (e.g., region growing, graph cut and contour-based methods), FCNs perform the entire segmentation task without requiring any manual feature engineering. By taking those findings, CNN-based segmentation methods can be adopted to realize an automated CADe system for detecting and segmenting AMLs from chest CT imaging. In addition, the contour of AML is one of the most important diagnostic indexes to recognize the types of lesions in-between benign and malignant [33]. Therefore, a CADe system that provides a way to segment AMLs from chest CT imaging does not only help radiologists to identify the locations of the AMLs but also assists radiologists in determining the type of the detected AMLs without any manual intervention.

While numerous studies have demonstrated the usefulness of applying deep learning techniques to segment pulmonary nodules or lymph nodes from chest CT imaging [34–37], limited studies have explored the feasibility of adopting deep learning techniques for automated detection of AMLs. Huang et al. [38] suggested using a two-staged 3D ResUNet to segment AMLs from chest CT imaging. However, the network is likely to be overfitted and under-segmented due to many parameters. To overcome those issues, a pre-processing stage was suggested to remove irrelevant anatomical structures prior to employing the network. Still, the performance of the network was highly correlated with the pre-processing stage. Therefore, a novel FCN architecture should be considered, which does not require any pre-processing or post-processing.

1.3 Objectives of Thesis

By considering all challenges and limitations discussed in Section 1.2, the objectives of this thesis are as follows:

- A CNN-based CADx system is developed to assist radiologists for lung cancer diagnosis on CT imaging. The developed system provides an automated workflow by taking a lung nodule patch image as input and estimating the malignancy rate of the input patch image.

- A CNN-based CADx system is designed to assist radiologists for breast cancer diagnosis on ABUS imaging. By taking a breast lesion patch image as input, the designed system automatically classifies the input patch image in-between malignant and benign.

- To assist radiologist for AML detection, a FCN-based CADe system is developed to segment AMLs from CT imaging. Without any manual intervention, the developed system identifies the locations of AMLs from the CT slice image and predicts the contours of detected AMLs.

1.4 Contributions of Research Works

In this thesis, the feasibility of applying deep learning techniques for cancer diagnoses is investigated by designing new network architectures and novel feature extraction strategies.

For pulmonary nodule classification on CT imaging, a novel CNN architecture is proposed by introducing multi-scale convolutional layers and the multipath feature extraction method. The proposed CNN is evaluated on a public chest CT imaging database. Compared to existing ML-based approaches, the proposed CNN shows better classification performances. Besides, the proposed CNN outperforms the state-of-the-art CNNs that employ customized architectures and transfer learning methods. In addition, the usefulness of the proposed CNN for clinical usage is verified. With the aid of the proposed CNN, radiologists show improved diagnostic performance when interpreting the malignancy of the nodules.

In addition, a multiview CNN is proposed for breast lesion classification on ABUS imaging. To the best of our knowledge, this was the first CNN-based CADx system approach tailored for breast lesion classification on ABUS imaging. One novel aspect of the proposed CNN is that the proposed multiview learning strategy allows the CNN to perform feature extraction from different views. Compared to conventional CNN approaches relying on single-view features, the proposed CNN shows better classification performance. Besides, compared to previous ML approaches, the proposed CNN has no need to design or select a proper feature extractor explicitly. The proposed CNN still outperforms previous ML approaches. The clinical usefulness of applying the proposed CNN for breast cancer diagnosis is explored. By referring to the predicting outcomes of the proposed CNN, human reviewers show significant improvement on their diagnostic performances.

Moreover, to support radiologists in AML detection, a modified UNet, the most commonly used FCN architecture for medical imaging segmentation, is proposed to segment AMLs from chest CT imaging. To enhance the performance of the proposed network, the self-attention mechanism is utilized and allows the network to extract global semantic correlations, which helps the network to differentiate AMLs from the dense background. Besides, the proposed network adopts an image-grid attention mechanism, a convolutional block attention module, to extract more robust features by focusing on the regions of the AMLs. Compared to the

state-of-the-art FCN-based segmentation networks, the proposed network achieved superior segmentation performance.

Below is the list of accepted and submitted publications, arranged according to the order of appearance in this thesis:

- Chapter 3: *Convolutional Neural Network Based Computer-aided Diagnosis System for Pulmonary Nodule Classification on Computed Tomography*:

  - **Y. Wang**, H. Zhang, K. Chae, G. Jin and S. Ko, "Novel Convolutional Neural Network Architecture for Improved Pulmonary Nodule Classification on Computed Tomography," in *Multidimensional Systems and Signal Processing*, vol.31, pp. 1163-1183, Jan 2020.

- Chapter 4: *Convolutional Neural Network Based Computer-aided Diagnosis System for Breast Lesion Classification on Automated Breast Ultrasound*:

  - **Y. Wang**, E. Choi, Y. Choi, H. Zhang, G. Jin and S. Ko, "Breast Cancer Classification in Automated Breast Ultrasound using Multi-View CNN with Transfer Learning," in *Ultrasound in Medicine & Biology*, vol. 46, no. 5, pp. 1119-1132, May 2020.

- Chapter 5: *Fully Convolutional Neural Network Based Computer-aided Detection System for Anterior Mediastinal Nodular Lesion Segmentation from Chest Computed Tomography*:

  - **Y. Wang**, W. Jeong, H. Zhang, Y. Choi, G. Jin and S. Ko, "Anterior Mediastinal Nodular Lesion Segmentation from Chest Computed Tomography Imaging Using UNet based Neural Network with Attention Mechanisms," under review at *Academic Radiology*.

- Other publications that are not included in this thesis:

  - **Y. Wang,**, H. Zhang, K. Oh, J. Lee, S. Ko, "Energy efficient spiking neural network processing using approximate arithmetic units and variable precision weights," in *Journal of Parallel and Distributed Computing*, vol.158, pp. 164-175, Dec. 2021.

- K. Chae, G. Jin, S. Ko, **Y. Wang**, H. Zhang, E. Choi, H. Choi, "Deep Learning for Classification of A Small (≤2cm) Pulmonary Nodule on CT Imaging: A Preliminary Study," in *Academic Radiology*, vol. 27, no. 4, pp. e55-e63, Apr 2020.

- **Y. Wang**, K. Shahbazi, H. Zhang, K. Oh, J. Lee and S. Ko, "Efficient Spiking Neural Network Training and Inference with Reduced Precision Memory and Computing," in *IET Computers & Digital Techniques*, vol. 13, no. 5, pp. 397-404, Sep 2019.

1.5 Overview of Thesis

This thesis is organized into six chapters as follows:

- **Chapter 1**: This chapter includes three parts. First, the advantages and disadvantages of CAD systems that incorporate medical imaging for cancer diagnoses are presented. Second, by considering the limitations of CAD systems, the motivation of this thesis is explained. Third, the overview and contributions of the works presented in this thesis are provided.

- **Chapter 2**: This chapter contains three parts that provide background information of the works presented in this thesis. First, the fundamentals of CNN are explained including the uses of convolutional layers, activation layers, pooling layers and fully-connected layers. Second, the process of training a CNN is described, including the ways to define a loss function, select an optimizer and perform backpropagation. Third, to realize an ultra-deep CNN that can be utilized for medical imaging databases, the transfer learning method is presented.

- **Chapter 3**: This chapter presents a CNN-based CADx system tailored for pulmonary nodule classification on CT imaging. The proposed CNN is designed to take nodule patch images as input and predict the types of the nodule in-between benign and malignant. Compared to conventional CNN architecture, a multi-scale convolutional layer adopting various convolutional kernels with different kernel sizes is proposed to allow the CNN to capture more robust nodule features from the input. Inspired by

the ResNet [9], the proposed CNN is designed to have multiple feedforward paths to combine nodule features extracted from the convolutional layers at the beginning of the CNN and the end of the CNN. Combined features via the proposed multipath feature extraction method are effective to improve the classification performance of the proposed CNN.

- **Chapter 4**: A CNN-based CADx system designed for breast lesion classification on ABUS imaging is presented in this chapter. The proposed CNN takes lesion patch images as input and classifies them in-between benign and malignant. By adopting transfer learning, the proposed CNN follows the architecture of Inception-v3 [8] which is a high-performance ultra-deep CNN originally used for natural image classification. To improve the classification performance of the proposed CNN, two multiview learning strategies are proposed. With multiview learning strategies, the proposed CNN predicts the type of a lesion by extracting features over the corresponded patch images captured from different views.

- **Chapter 5**: This chapter presents a FCN-based CADe system utilized for AML segmentation from chest CT imaging. A modified UNet [10] architecture is proposed to take a 2D slice image as input and outputs the corresponding mark image that outlines the areas of AMLs at the pixel level. To enhance the robustness of feature extraction, the proposed network adopts the self-attention mechanism and the convolutional block attention module, an image-grid attention mechanism. In addition, to preserve both conovlutional features and self-attentive features, a multi-path feature extraction stage is proposed.

- **Chapter 6:** This chapter summarizes the presented works of this thesis and the plans for future research.

Chapter 2

Background

This chapter provides the background information of the works that are presented in this thesis. In Section 2.1, the principles of convolutional neural network (CNN) are explained, including the fundamental building blocks of the CNN. Section 2.2 describes the process of training a CNN from scratch. The previous two sections focus on providing general steps to realize customized CNNs that are used in the proposed works. In addition, transfer learning is utilized in the proposed designs. Section 2.3 presents the concept of transfer learning and the methodology to apply transfer learning in medical imaging.

2.1 Fundamentals of Convolutional Neural Network

The CNN is the most commonly used deep learning model. Due to its comparable performance to human-level, CNN has been widely utilized in many different fields, including image classification [39], object detection [40], image segmentation [41], image reconstruction [42] and natural language processing [43]. As a subset of machine learning (ML), the biggest advantage of CNN compared to its preceding methods is that CNN is an end-to-end system that automatically extracts and identifies features without explicitly designing a feature extractor or any human interference [44]. CNNs are derived from the artificial neural network (ANN) that employs artificial neurons to simulate the functionality of the human visual cortex. The structure of a conventional ANN is a multilayer perceptron (MLP) that consists of multiple layers to create fully-connectivities. An example of a 3-layered ANN is shown in Figure 2.1. Instead of relying on fully-connected blocks, CNN builds complex local connections via shared weights of convolutional kernels. Local connectivities allow CNN to focus on subtle regions, which makes it possible to understand nuances of spatial arrangement. Besides, the weight

sharing feature significantly reduces the number of trainable parameters. Therefore, realizing a deep CNN is possible. For example, ResNet, a common CNN architecture, adopts 50 layers to perform image classification. However, it is impossible to train a 50-layered ANN with current computational power.

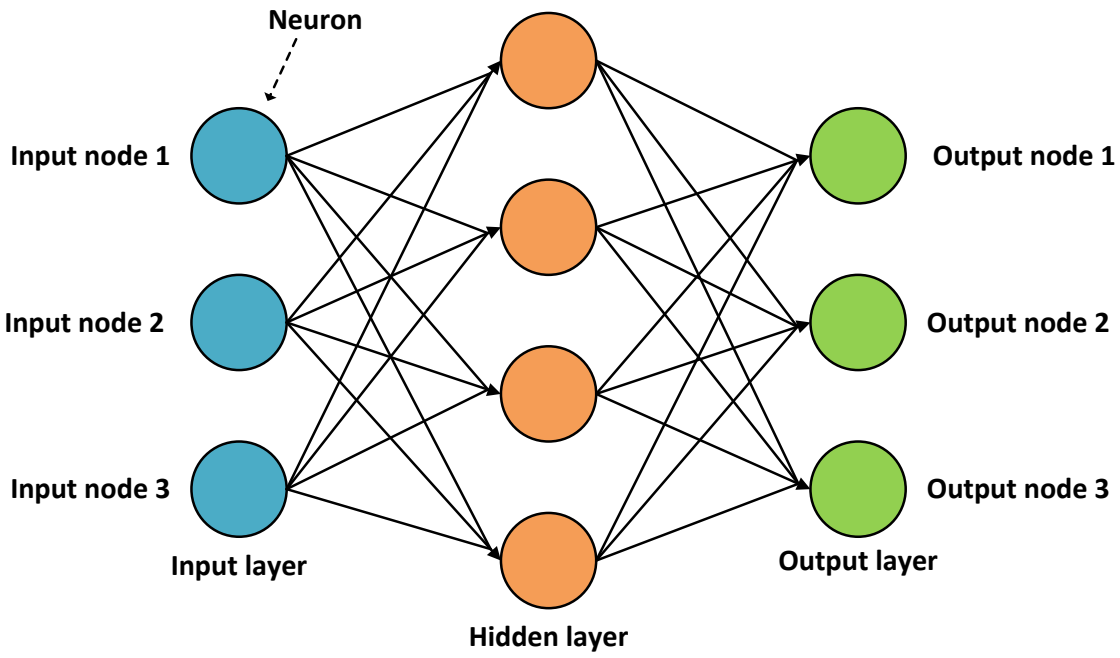


Figure 2.1: An example of a 3-layered artificial neural network employing fully connectivities.

A CNN comprises a sequence of layers to transform input images from pixel values to class scores. There are four common layers used to construct a CNN, including convolutional (Conv) layers, activation layers, pooling layers and fully-connected (FC) layers. An example of the LeNet [45] CNN architecture for digits image classification is shown in Figure 2.2.

In a CNN model, each layer is organized into several feature maps. Each feature map has a three-dimensional shape ($w \times h \times d$), where w and h are the width and height of each feature map, respectively. In practice, the height of the feature map is equal to its width. In addition, d refers to the depth of the feature maps. For example, the first Conv layer of the LeNet produces 6 feature maps to form its output with a size of $28 \times 28 \times 6$, where the depth (d) is equal to 6.

In a Conv layer, a group of trainable kernels is utilized to generate feature maps. The

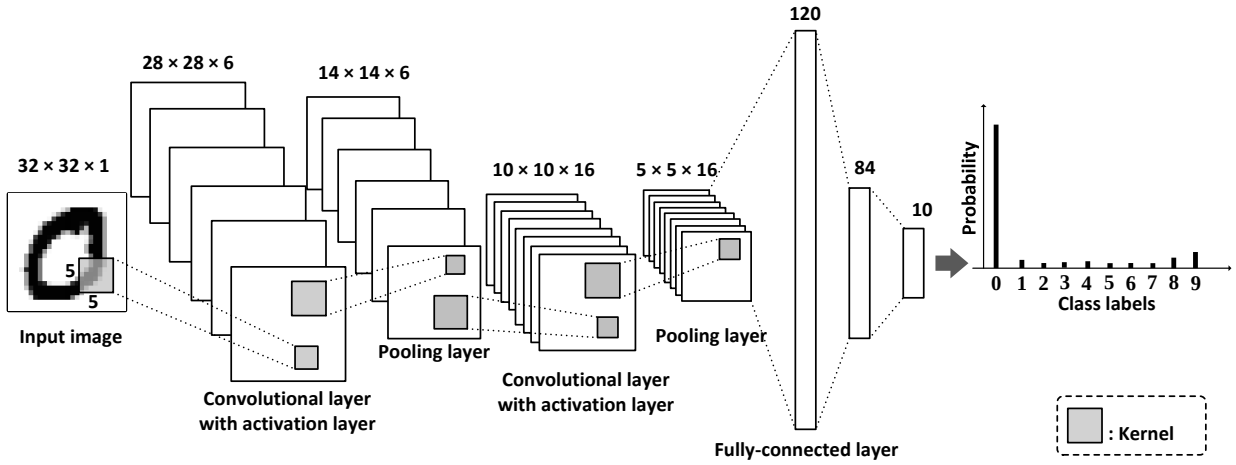


Figure 2.2: An example of LeNet CNN [45] architecture for digits image classification.

depth of feature maps is equal to the total number of applied kernels. Each kernel has a three-dimensional shape ($m \times m \times n$). The side length (m) of each kernel should be smaller than the side length of the input data. Besides, the depth (n) of each kernel should be smaller or equal to the depth of the input data. In the LeNet, a total number of 6 kernels with a size of $5 \times 5 \times 1$ are used to generate output feature maps of the first Conv layer by taking a grey-scale digits image as input, where the size of the input image is $32 \times 32 \times 1$. Moreover, kernels are the foundation blocks to build local connections by sharing their trainable parameters, including weights and bias. Specifically, each kernel convolves multiple fixed-scale local regions of the input data by reusing its weights and bias. As shown in equation (2.1), a feature map (f) is calculated by performing a dot product between the kernel's weights (w) and the input data (x), followed by adding the kernel's bais (b). In addition, when an input data has a shape of $w \times w \times d$, convolving a kernel with a size of $m \times m \times d$ generates a feature map having a side length of $(w - m + 1)$. For example, in the first Conv layer of the LeNet, a total number of 6 kernels with a size of $5 \times 5 \times 1$ convolve the input image with a size of $32 \times 32 \times 1$ and generates a group of feature maps with a size of $28 \times 28 \times 6$.

$$f = W \cdot x + b \tag{2.1}$$

After a Conv layer, an activation layer is commonly followed to apply the non-linearity activation function to the output of the Conv layer. The activation function makes CNN

possible to be trained and enables the feasibility of predicting probabilities as output.

A pooling layer is generally inserted between two consecutive Conv layers. The purpose of employing the pooling layer is to reduce the size of Conv layer output by down-sampling its corresponded feature maps. Hence, the overall CNN parameters are reduced, which improves training convergence speed and prevents overfitting issues. A group of kernels are utilized for pooling operation. By sliding kernels over each feature map, features within the region covered by kernels are summarized. For example, the first Conv layer output of LeNet is down-sampled via pooling operation. The pooling operation uses 6 kernels with a size of 2×2 to generate new feature maps with a size of $14 \times 14 \times 6$.

FC layers are normally placed at the end of CNN. The FC layers have an MLP-based structure that builds full connections for each other. The first FC layer converts the feature maps into one-dimensional vectors, which describe the high-level abstraction of features produced by the preceding Conv layers. In addition, the last FC layer generates the class probability scores. The number of neurons used in the last FC layer should match the number of class labels. For example, to predict digits from 0 to 9, the last FC layer of LeNet consists of 10 neurons.

2.1.1 Convolutional layer

In this section, the mechanism of the Conv layer is described in detail. As the core component of the CNN, the Conv layer employs a group of convolutional kernels to create feature maps by convolving the input data. Each convolutional kernel contains a set of trainable weights and one trainable bias. For instance, a convolutional kernel with a size of 2×2 consists of four weights and one bias. One novel aspect of the convolutional layer is that kernel weights and biases can be trained to describe the significance of the input images similar to how humans interpret images. In practice, to train a convolutional layer, its weights and biases are initialized with random values. During the training phase, these weights are gradually altered and learned to extract substantial features from input data. Instead of random initialization, there are several weight initialization techniques available to reduce the training time, such as He initializer [46] or Xavier initializer [47].

For a Conv layer, the output feature maps are generated by sliding convolutional kernels

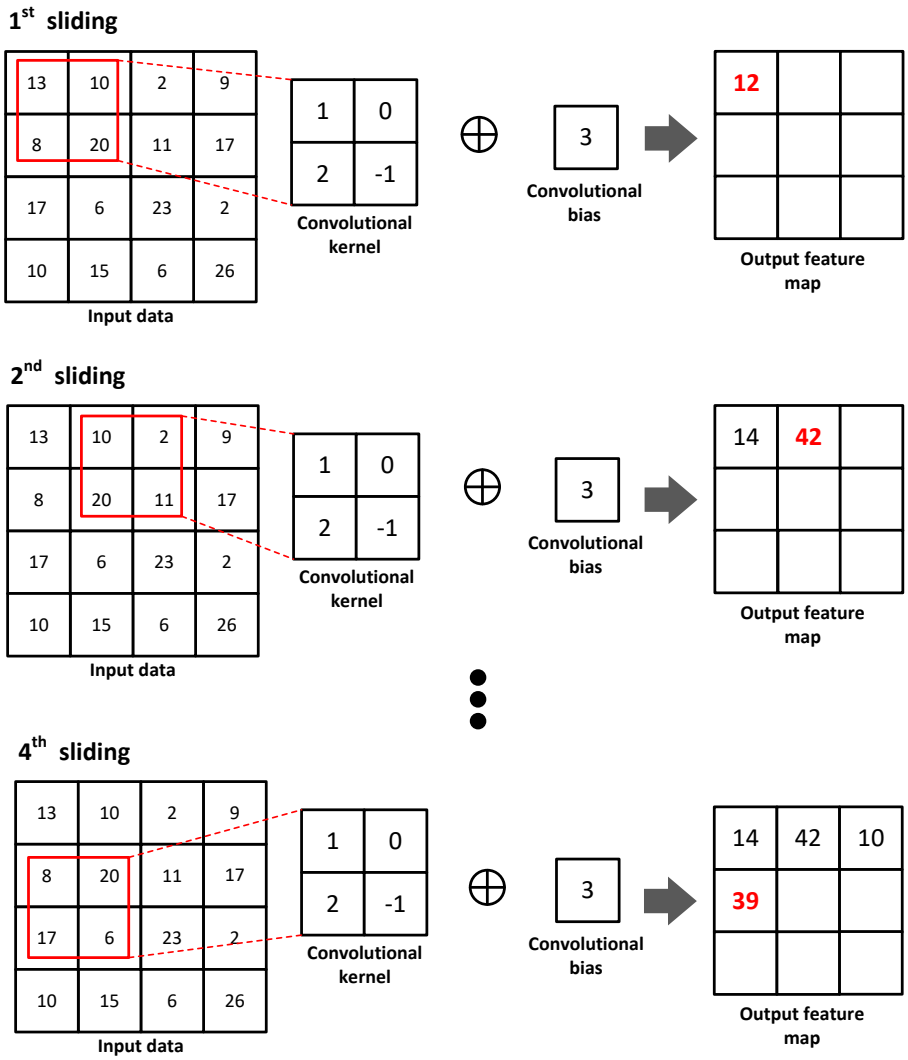


Figure 2.3: An example of feature map generation using a 2×2 convolution kernel to slide a input data with a size of 4×4.

over input data. During each time of sliding operation, a feature value is determined. The sliding operation is repeated until the feature map is generated. An example of feature map generation is illustrated in Figure 2.3. In this example, the input data with a size of 4×4 is convolved with a convolution kernel with a size of 2×2. The convolutional kernel slides over the input data by following a specific direction, which is left-to-right and then top-to-bottom. The first feature value is determined by performing a dot product between the convolutional kernel and a region of input data selected by the first sliding operation and then summing the convolutional bias up. By repeating the sliding operations, the output feature map is

generated, which has a size of 3×3.

In addition, the step size of each sliding operation is controlled by the stride parameter. For the previous example, the stride is set to 1. When stride is equal to 2, each sliding operation moves two grids. Thus, the resulted output feature map is reduced to 2×2. Furthermore, padding is commonly applied to CNN to conserve the size of the feature maps. Without padding, a reduction in feature maps may cause losses of important features. In practice, zero-padding is widely used by adding zeros at each border of input data.

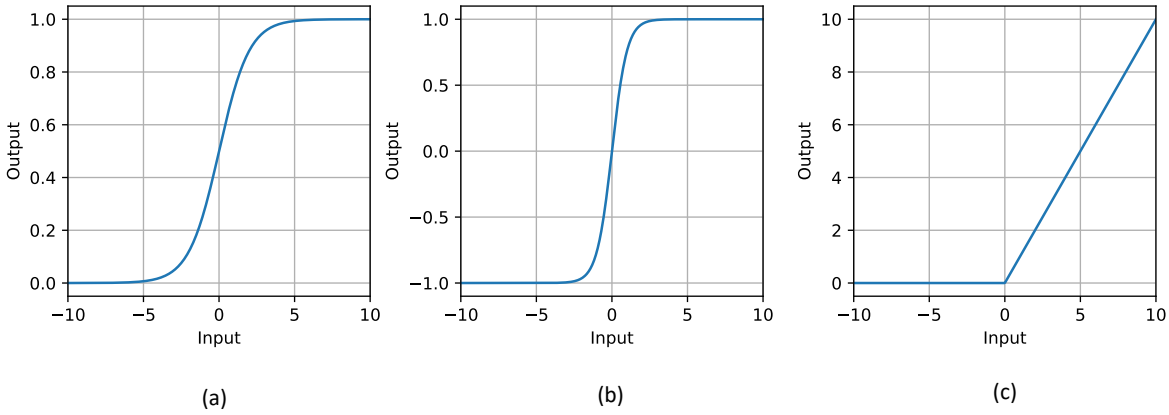


Figure 2.4: Common non-linearity functions. (a) Sigmoid function; (b) Tanh function; (c) Rectified linear unit (ReLU) function.

2.1.2 Activation layer

The insight of the activation layer is to map the input to the output by using a non-linearity function. The use of the non-linearity function is to allow the model to create non-linear decision boundaries. Without employing activation layers, the CNN devolves into a linear model, which is difficult to learn complex features from input and prone to overfitting. The common non-linearity functions are:

- Sigmoid: It is also known as logistic function. The output of the sigmoid function is in the range from zero to one. The sigmoid function is an S-shaped curve as illustrated in Figure 2.4(a). The mathematical form of the sigmoid function is shown in equation (2.2). Since the output of the sigmoid function is limited between zero and one, it

can be used for binary classification. Nevertheless, the sigmoid function has a vanishing gradient problem that may cause weights to stop updating during the training phase. This issue could happen when the input values are large positive numbers or small negative numbers.

$$f(x) = \frac{1}{1 + e^{-x}} \tag{2.2}$$

- Tanh: It is also called hyperbolic tangent function. The mathematical representation of the tanh function is illustrated in equation (2.3). The output of the Tanh function is in the range from -1 to 1. The tanh can be used to classify binary categories. Because the output shape of the tanh function (Figure 2.4(b)) is similar to the sigmoid function (Figure 2.4(a)), the vanishing gradient problem remains.

$$f(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \tag{2.3}$$

- ReLU: Rectified linear unit (ReLU) is the most commonly used non-linearity function in CNN. The ReLU function suppresses negative input values to 0 and keeps positive input values, as demonstrated in Figure 2.4(c). The mathematical form of the ReLU function is shown in equation (2.4). Compared to the sigmoid and tanh functions, ReLU solves the vanishing gradient problem. Besides, ReLU is less computationally intensitive since there is no need for exponent calculation. Nevertheless, ReLU may be confronted with a dying ReLU problem. When the input of the ReLU is less than zero, weights prevent updating. To overcome this problem, weights can be initialized via Xavier initializer [47] while the learning rate should avoid using a large number.

$$f(x) = \begin{cases} 0 & x \leq 0 \\ x & x > 0 \end{cases} \tag{2.4}$$

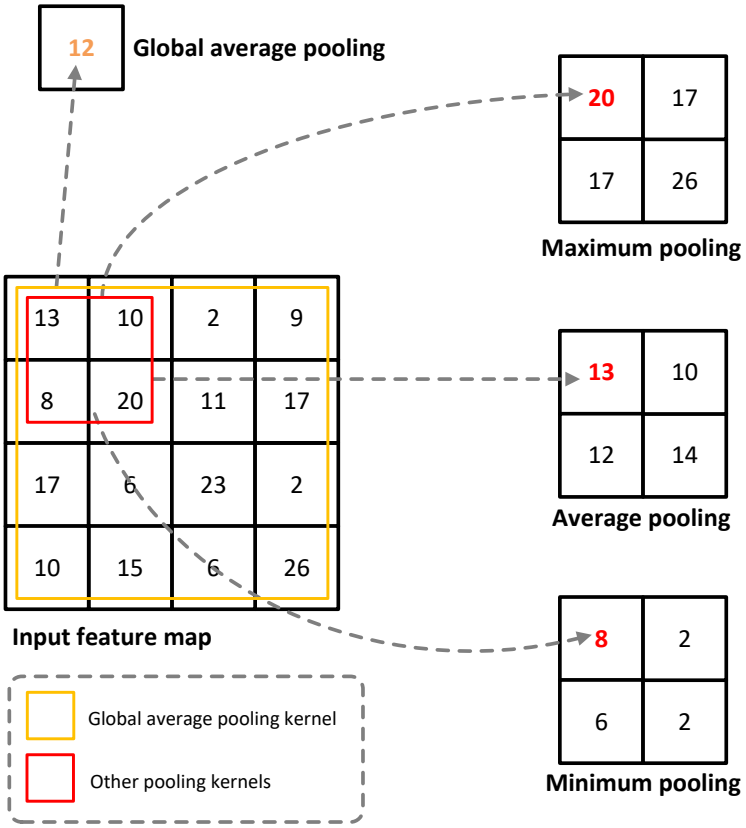


Figure 2.5: Illustration of common pooling operations including maximum pooling, minimum pooling, average pooling and global average pooling.

2.1.3 Pooling layer

The pooling layer is commonly followed after each Conv layer. By taking the output feature maps of the Conv layer as input, the pooling layer performs sub-sampling over these feature maps. One novel aspect of the pooling layer is to preserve useful spatial information which improves the generalizability of the CNN, especially for spatial invariance. In addition, the pooling layer reduces the total number of trainable parameters. Therefore, the CNN computation complexity is reduced, which speeds the training process up and prevents overfitting during training. Similar to the Conv layer, the pooling layer adopts a set of kernels. However these kernels do not contain any trainable parameters. By sliding these kernels over input, the output feature maps of the pooling layer are determined. During each time of the sliding operation, the pooling function is performed by down-sampling the input region covered by

these kernels. Particularly, there are three common types of pooling functions, including maximum pooling, minimum pooling and average pooling. An example of performing different types of pooling functions is shown in Figure 2.5. In this example, the input feature map with a size of 4×4 is pooled by a 2×2 pooling kernel with a stride of 2. For maximum pooling, the input region covered by the kernel is down-sampled by keeping the maximum feature value. In a similar manner to the maximum pooling, the minimum pooling keeps the minimum feature value while the average pooling averages the whole feature values. In addition, global average pooling (GAP) is commonly used before FC layers. Instead of averaging local regions of feature maps, the GAP adopts a kernel with the same size as the input feature maps to summarize each feature map into one value. An example of GAP is demonstrated in Figure 2.5. The input feature map with a size of 4×4 is reduced to a single feature value by employing the GAP.

2.1.4 Fully-connected layer

The ending layers of a CNN are FC layers. The structures of the FC layers are similar to ANN as shown in Figure. 2.1. Each FC layer consists of a group of neurons that builds full connections to the previous FC layer. For an FC layer, each neuron is connected to all neurons of the previous FC layer. A trainable weight is placed in each connection path that allows the neuron to learn useful features from the previous FC layer. The output of each neuron can be expressed as:

$$f = \sum^{i} x_i w_i + b \tag{2.5}$$

Where x_i is the input signal transmitted from the neuron i of the previous FC layer. w_i refers to the weight of the connection path used to connect the output neuron and the neuron i of the previous FC layer. b is a trainable bias. In addition, an activation function is commonly added to create non-linearity for the output.

In a CNN, the first FC layer is connected to the last Conv layer or pooling layer depending on the network configuration. The output feature maps of the last Conv layer or pooling layer are first converted into one-dimensional vectors. Each neuron in the first FC layer of

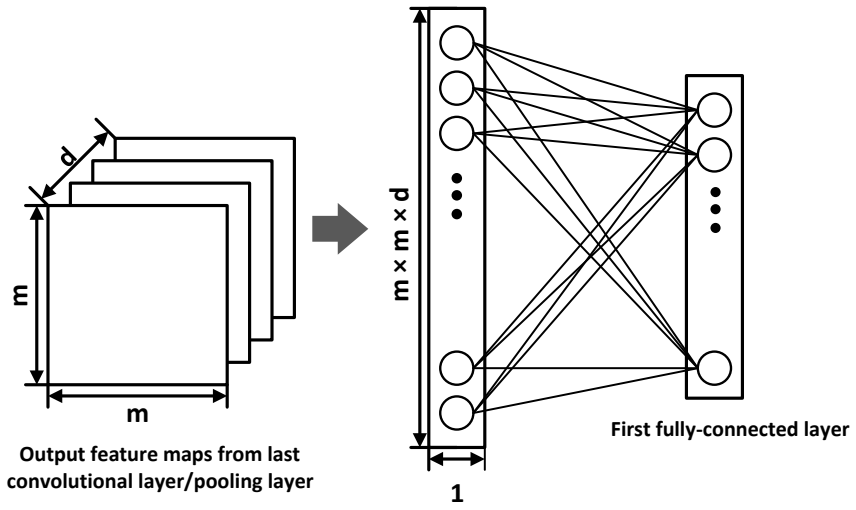


Figure 2.6: Illustration of mapping features maps to neurons of a fully-connected layer.

the CNN is connected to all vectors. The process of connecting feature maps to the first FC layer of a CNN is shown in Figure 2.6. In this example, the output feature maps have a size of $m \times m \times d$. The output features maps are flattened into a sequence of one-dimensional vectors. Each vector refers to one particular value of the feature maps. An additional FC layer having a size of $m \times m \times d$ is used to map these vectors. Then, each neuron of the first FC layer is connected to all neurons of the additional FC layer.

The last FC layer is used to output predicting outcomes of the CNN. Therefore, the number of neurons in the last FC layer of a CNN should match the total number of class labels. However, for binary classification, there are two approaches: (1) A single neuron with a sigmoid function can be used for binary classification. (2) Two neurons with a softmax function can be used to classify the binary class. Compared to the sigmoid function, the softmax function can also be employed for multiclass classification. The mathematical representation of the softmax function is shown in equation (2.6). When predicting a total number of i classes, i neurons are utilized to form the last FC layer of a CNN. The outputs of these neurons are $(x_1, x_2, x_3, ..., x_i)$, where x_i is the ith neuron's output. By applying the softmax function, the probability of the ith class is equal to the exponent of the ith neuron's output divided by the summation of the exponent of each neuron's output.

$$P(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}} \tag{2.6}$$

2.2 Training Convolutional Neural Network

The major advantage of CNN is that there is no need to design a feature extractor explicitly since CNN learns features automatically from training data. Therefore, understanding the training mechanism is the key to obtaining an effective CNN model. Without a training process, the CNN acts as a random guessing model. Intuitively, training a CNN is a process to adjust trainable parameters of the Conv layers and FC layers by referring to the training data. The goal of the training process is to tune the CNN to fit the training data. In the following sections, the essential steps of the training process are described, including estimating training error via loss function, tunning trainable parameters via optimizer and backpropagation algorithm.

2.2.1 Loss function

During the training process, the classification performance of the CNN on the training data is evaluated by a loss function. Practically, the loss function is placed at the end of the CNN and used to estimate the error between the actual outputs and the predicting outcomes. Then, the error is minimized by employing an optimizer to adjust the trainable parameters of the CNN.

During each iteration of the training process, the loss (error) is calculated by a loss function. Two input parameters are required to calculate the loss. One input parameter is the output of the CNN, which refers to the probability of each class. The second input parameter is the actual output by referring the ground truth of the input. The common loss functions are:

- Euclidean loss function: It is also known as mean square error. The Euclidean loss function is commonly used for solving regression problems. However, it is not recommended in CNNs due to slow convergence speed at the beginning of the training process. The

mathematical representation of the Euclidean loss function is shown in equation (2.7), where Y_i is the desired output for ith class. N is the total number of neurons in the last FC layer, which refers to the number of classes that need to be classified. P_i represents the output of the ith neuron.

$$L(P,Y) = \frac{1}{2N}\sum_{i=1}^{N}(Y_i - P_i)^2 \tag{2.7}$$

- Hinge loss function: It is widely used in support vector machines (SVMs) to solve binary classification problem because the Hinge loss is effective to create a maximum margin between the decision boundary and input samples. Therefore, it allows the optimizer to ensure that each input sample is classified correctly. The mathematical representation of the Hinge loss is illustrated in equation (2.8).

$$L(P,Y) = \sum_{i=1}^{N} max(0, 1 - Y_i P_i) \tag{2.8}$$

- Cross-Entropy loss function: For the multiclass classification problem, the mathematical expression of the cross-entropy is shown in equation (2.9). In addition, for the binary classification problem, the mathematical expression of the cross-entropy can be expressed as equation (2.10), where Y is equal to 1 when the input data is a positive label. When input data is a negative label, Y is equal to 0. P refers to the probability of the input data as a positive label. The cross-entropy is commonly used in CNN, and it solves the convergence problem raised by the Euclidean loss.

$$L(P,Y) = -\frac{1}{N}\sum_{i=1}^{N} Y_i \cdot log(P_i) \tag{2.9}$$

$$L(P,Y) = Y \cdot log(P) + (1 - Y) \cdot log(1 - P) \tag{2.10}$$

2.2.2 Optimizer

The use of a loss function is to evaluate the effectiveness of the CNN that fits the input data. When the trainable parameters of a CNN are assigned with random values, the CNN

most likely has a poor classification performance to predict the correct classes of input data. Therefore, it is crucial to define a mechanism to optimize the loss function for achieving a promising classification performance. The gradient-based optimizers are commonly utilized in CNN. During each epoch of the training process, the trainable parameters are updated via a gradient-based optimizer. The gradient-based optimizer seeks local optima that model achieves the lowest loss on training data.

The gradient descent algorithm is the earliest gradient-based optimizer, which is widely used for linear regression algorithms. The purpose of the gradient descent algorithm is to adjust trainable parameters to reach a minimal loss during each epoch of the training process. To update the trainable parameters in every epoch, the gradient descent algorithm firstly computes the gradient of the loss function by performing a first-order derivative with respect to the trainable parameters. Then, these trainable parameters are adjusted by following a backward direction of the gradient, where is from the output nodes of the network to the input nodes of the network. The backward updating process is also known as backpropagation, which is described in Section 2.2.3. Intuitively, the back propagation allows a neuron to pass its gradient to all neurons of the preceding layer. Thus, all trainable parameters are updated. The mathematical expression of the gradient descent algorithm is shown in equation (2.11).

$$\begin{aligned} W_t &= W_{t-1} - \Delta W_t \\ \Delta W_t &= \mu \cdot \frac{\partial L_{t-1}}{\partial W_{t-1}} \end{aligned} \tag{2.11}$$

By employing the gradient descent algorithm, the weight (W_{t-1}) at the $(t-1)$th epoch is updated to a new weight (W_t) after the tth epoch. The weight updating is controlled by the term ΔW_t as shown in equation (2.11), where L refers to the loss function output at the $(t-1)$th epoch. μ denotes the learning rate, which is a hyper-parameter used to adjust the step size of the weight updating. In practice, selecting a proper learning rate is challenging. A large learning rate prevents convergence while a small learning rate takes a significant longer time to reach the global minima.

The gradient descent algorithm is designed to update all trainable parameters at each epoch. Therefore, these parameters stop updating until the gradients of the entire training

data is computed. However, it costs a significant amount of computation power when the training data has a large size. Thus, the model takes a long time to train while it is easy to converge the trained model to local minima. To overcome these issues, the Stochastic Gradient Descent (SGD) [48] has been introduced. Instead of altering all trainable parameters at each epoch, the SGD updates these parameters on each training sample. Hence, the computation power is reduced due to less memory requirement compared to the conventional gradient descent algorithm. In addition, the model is easier to converge since parameters are updated more often. Nevertheless, due to frequent updates, each step of the update is noisy to reach the global minima. Therefore, the training convergence becomes unstable.

To improve the reliability of the SGD, a batch is commonly used. The batch is a small portion of training data. For example, a total number of 10 batch is used to hold a training dataset having 100 images. Each batch consists of 10 images without overlapping other batches. The batch-based SGD updates the trainable parameters after the gradients of one batch is calculated. Therefore, compared to the conventional gradient descent algorithm, the batch-based SGD is computational effective. In contrast to SGD, the training convergence is more stable by using the batch-based SGD.

The major disadvantage of the aforementioned optimizers is that identifying an optimal learning rate is difficult. It may require a series of additional steps to obtain one. By considering this factor, adaptive optimizers have been developed by automatically adjusting the learning rate based on the changes in gradients. The common adaptive optimizers are:

- Adagrad [49]: Instead of using a constant learning rate as the step size to control the weight updating, the Adagrad performs a big step update for the parameters associated with the neurons that have small gradients. When a neuron produces a large gradient, the associated parameters are updated by a small step. The mathematical representation of the Adagrad algorithm is shown in equation (2.12), where g_{t-1} refers to the gradients at $(t-1)$th iteration or epoch. ϵ is a constant ($\epsilon = 1e-8$) to prevent division by zero. μ is the initial learning rate, which is commonly set to 1e-3. At (t)th iteration or epoch, the learning rate is adjusted according to the term G_{t-1}, which is the sum of the squares of all past gradients.

$$W_t = W_{t-1} - \Delta W_t$$

$$\Delta W_t = \frac{\mu}{\sqrt{G_t + \epsilon}} \cdot g_{t-1} \tag{2.12}$$

$$G_t = \sum_{i=1}^{t-1} (g_i)^2$$

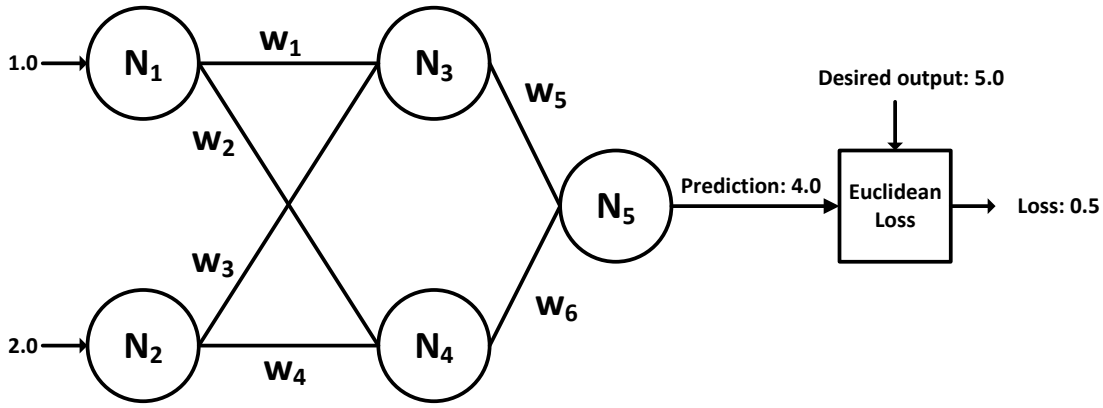
- Adadelta [50]: It is a variant of the Adagrad. Without using all past gradients to alter the learning rate, the Adadelta accumulates past gradients with a fixed scale. In addition, instead of taking squares of all past gradients, the exponential moving average is utilized. In the Adadelta, the term G_t defined in equation (2.12) is replaced as shown in equation (2.13), where $E[g]_{t-1}$ is denoted as the exponential moving average for past gradients at $t-1$th iteration or epoch. γ is a constant, which is commonly set to 0.9.

$$G_t = \gamma E[g]_{t-1} + (1 - \gamma) g_t \tag{2.13}$$

2.2.3 Backpropagation

In a CNN, the gradient-based optimizer provides a direction to alter the trainable parameters for achieving minimum loss of loss function. To propagate the gradient of a neuron to all neurons of the preceding layer, backpropagation is utilized. The back propagation allows the gradient-based optimizer to compute gradient recursively from the output of the CNN to the input of the CNN. Therefore, the trainable parameters of each layer can be updated while the loss of the loss function is minimized.

The CNN is a feedforward neural network, which propagates data from the input of the network to the output of the end. The feedforward behaviour allows the CNN to take an image as input and outputs the predicted class associated with the input image. Prior to performing backpropagation, the network requires to infer a result from input data via feedforward. As shown in Figure 2.7, the mechanism of feedforward is explained by using a 3-layered ANN. The ANN takes two signals as input and computes the output of each neuron in a feedforward manner. For example, the output of the neuron N_3 depends on the outputs of the neuron N_1 and the neuron N_2. To apply backpropagation, the loss of the final



Weights	Value
W_1	1.0
W_2	0.5
W_3	2.0
W_4	0.5
W_5	0.5
W_6	1.0

Neuron outputs	Value
N_1	1.0
N_2	2.0
N_3	5.0
N_4	1.5
N_5	4.0

$N_3 = N_1 W_1 + N_2 W_3$
$N_4 = N_1 W_2 + N_2 W_4$
$N_5 = N_3 W_5 + N_4 W_6$

Figure 2.7: An example of illustrating feedforward inference on a 2-layered artificial neural network.

output (N_5) of the ANN is calculated by employing a Euclidean loss function as described in Section 2.2.1. Then, the weights, w_5 and w_4, are updated by computing its gradient with respect to the loss function. The mathematical expression of the gradient of the weight w_5 is shown in equation (2.14). A chain rule is applied to decompose the gradient of the weight w_5 into two components. The first component is the gradient of the loss function with respect to the final output of the ANN, which is denoted as $\frac{\partial L}{\partial N_5}$. Another component is the gradient of the neuron N_5 with respect to the weight w_5, which is annotated as $\frac{\partial N_5}{\partial W_5}$. After the gradient of the weight w_5 is calculated, a gradient-based optimizer is utilized to update the weight w_5. Accordingly, the weight w_4 is adjusted by computing its gradient with respect to the lost function.

$$g_{w_5} = \frac{\partial L}{\partial W_5} = \frac{\partial L}{\partial N_5} \cdot \frac{\partial N_5}{\partial W_5}$$
$$\frac{\partial N_5}{\partial W_5} = \frac{\partial (N_3 W_5 + N_4 W_6)}{\partial W_5} = N_3 \tag{2.14}$$

Thereafter, the weights connecting between the first layer and the second layer are up-

dated by computing its gradients with respect to the lost function. For example, the gradient of weight w_1 with respect to the loss function can be obtained by:

$$g_{w_1} = \frac{\partial L}{\partial W_1} = \frac{\partial L}{\partial N_5} \cdot \frac{\partial N_5}{\partial N_3} \cdot \frac{\partial N_3}{\partial W_1}$$

$$\frac{\partial N_5}{\partial N_3} = \frac{\partial (N_3 W_5 + N_4 W_6)}{\partial N_3} = W_5 \tag{2.15}$$

$$\frac{\partial N_3}{\partial W_1} = \frac{\partial (N_1 W_1 + N_2 W_3)}{\partial N_1} = W_1$$

2.3 Training Convolutional Neural Network with a limited data size

Training CNN is a process to learn the representations of the input images. To achieve a promising classification performance, CNN requires large-scale training data. In most cases, the available data is sufficient to train the CNNs. For example, ImagNet [51], a database including 1.2 million training images, is widely used to train CNNs for solving natural image classification problems. Nevertheless, it is possible to have a database with a limited size, which may saturate the performance of the CNN. Especially, medical imaging databases usually have a limited data size due to restrictions on privacy and difficulties to acquire the data from daily life. To alleviate this problem, data augmentation and transfer learning are commonly suggested. Data augmentation is a process to apply image transformation to the training data, such as rotation or flipping. Thus, the size of the training dataset is increased while the augmented data produces extra spatial variance that improves the generalization of the trained model.

Another fruitful direction is to apply transfer learning. The concept of transfer learning was introduced by a psychologist, C.H. Judd. Transfer learning is a process that uses a person's experience previously learned from one particular area to perform a task in another area. For example, a person who knows how to ride a bicycle can learn to ride a motorbike faster than others since there is shared knowledge between the bicycle and motorbike. Similarly, a CNN, previously trained on a large-scale dataset, can generalize learned knowledge

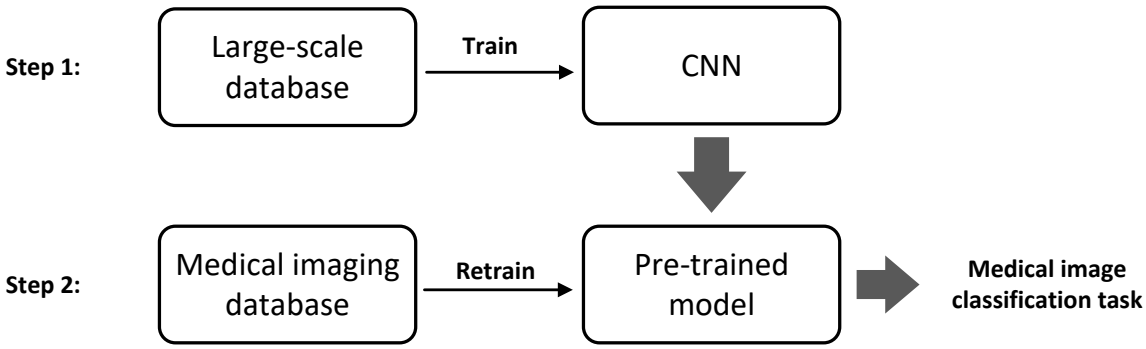


Figure 2.8: An example of applying transfer learning on medical imaging database.

to perform a similar task on other datasets. An example of employing transfer learning on medical imaging databases is illustrated in Figure 2.8. A deep CNN is first trained on a large-scale database such as ImageNet. Without training the deep CNN from scratch, the weights of the deep CNN are assigned by using the weights previously tuned on the large-scale database, which is also called pre-trained model. Next, the pre-trained model is retrianed on the medical imaging database to perform classification tasks on medical imaging.

A deep CNN is commonly used as the pre-trained model for transfer learning. A deep CNN typically consists of numerous Conv layers to provide effective feature extraction. However, training a deep CNN from scratch is a time-consuming task that may take several weeks to obtain a well-trained model. Besides, the trained model may lack generalization ability when the training dataset is small. To overcome these problems, the pre-trained model is effective to retrain fine generalization ability while accelerating the convergence speed. To retrain the pre-trained model on a small database, two common methods are suggested. First, all weights are tuned simultaneously. In practice, a small learning rate is utilized to make small adjustments to the weights. However, it may take a long time to converge. Second, the weights of the FC layers are tuned while the rest of the weights remain unchanged. In most cases, the original FC layers of the pre-trained model are replaced by a series of new FC layers. Therefore, the weights associated with these new FC layers need to be adjusted. On the other hand, the remaining weights are associated with the backbone of the pre-trained model. The backbone contains a set of Conv layers to extract features from the input image. Hence, freezing the weights of the backbone is an effective way to retain the generalization

ability, previously learned on a large-scale database.

The common pre-trained models utilized for solving medical imaging tasks are:

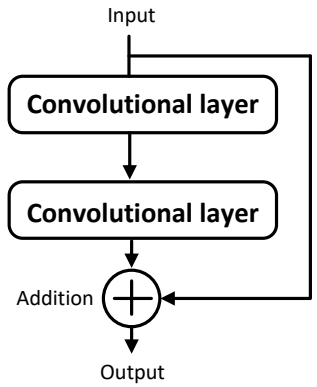


Figure 2.9: Architecture of a residual block.

- ResNet [9]: Compare to conventional CNN architecture, ResNet introduces residual blocks to overcome the vanishing gradient problem when the network is ultra-deep. The architecture of the residual block is shown in Figure 2.9. The residual block creates a shortcut that allows the preceding feature maps to combine with the current feature maps when the network employs a series of consecutive Conv layers. The output of the residual block bypassing a single Conv layer can be identified by summing the output feature map of the Conv layer and the input of the Conv layer. The input of the Conv layer refers to the output feature maps of the previous layer applied activation function. Depending on the number of Conv layers employed in the ResNet, there are several variants. ResNet50 is the most commonly used ResNet architecture including 49 Conv layers and 1 FC layer.

- GoogleNet [52]: As shown in Figure 2.10(a), the GoogleNet is a 22-layered deep CNN, designed for natural image classification. GoogleNet is also known as the Inception-based CNN. There are five variants of the GoogleNet, including Inception-v2, Inception-v3, Inception-ResNet and Inception-v4 [8, 53]. The insight of the Inception-based CNN is its use of the inception module. The architecture of the inception module used in the GoogleNet is shown in Figure 2.10(b). Instead of applying a fixed-scale kernel to

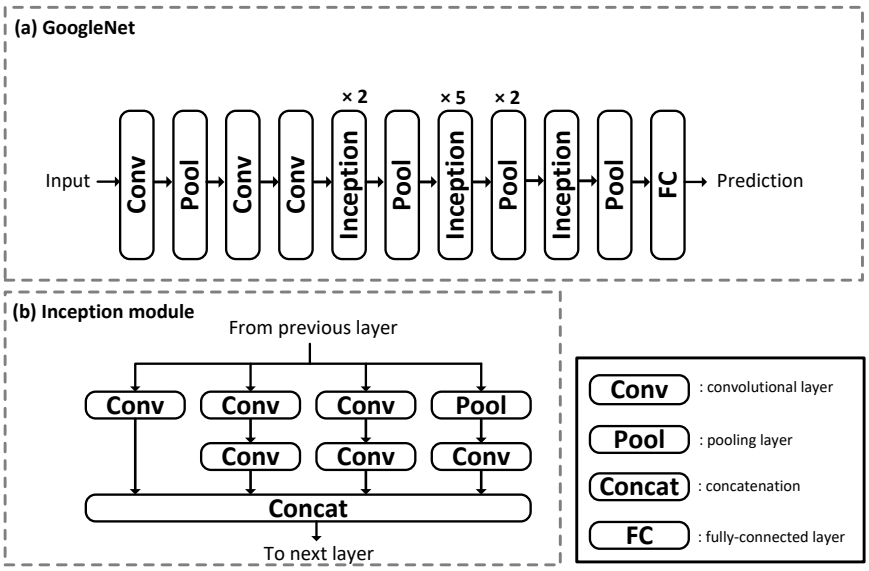


Figure 2.10: (a) Architecture of GoogleNet [52]; (b) Architecture of Inception block.

perform feature extraction in each Conv layer, the inception module adopts multiple kernels with different sizes to extract multi-scaled features.

- DenseNet [54]: To realize an ultra-deep CNN, the DenseNet employs several dense blocks to avoid gradient diminishing rapidly. Similar to the ResNet, the dense block creates multiple shortcuts to pass feature maps extracted from the preceding layers to the posterior layers. The architecture of a dense block is demonstrated in Figure 2.11. Instead of summing feature maps up, the dense block concatenates output feature maps generated from the current layer and the preceding layers.

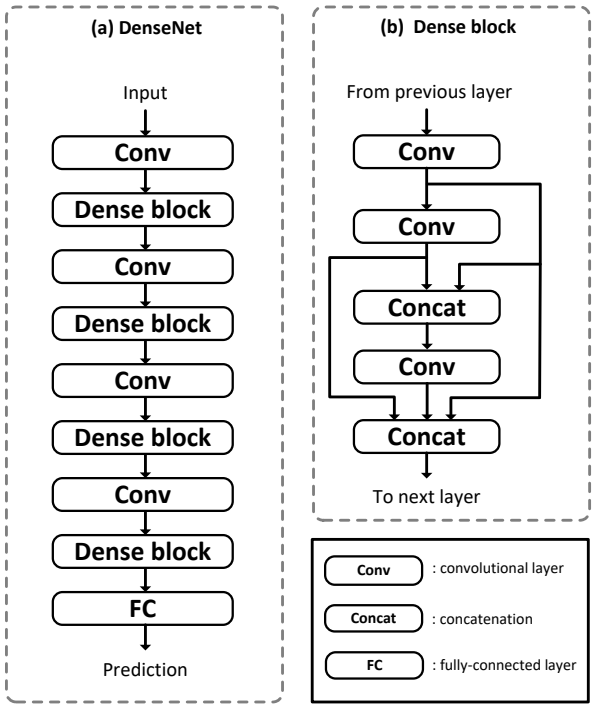


Figure 2.11: (a) Architecture of DenseNet [54]; (b) Architecture of a dense block.

Chapter 3

Convolutional Neural Network Based Computer-aided Diagnosis System for Pulmonary Nodule Classification on Computed Tomography[1]

This chapter presents the design of a novel convolutional neural network (CNN) architecture for pulmonary nodule classification on computed tomography (CT). By employing the proposed multi-scale convolutional layer, more effective features are extracted from pulmonary nodules thereby improving the classification perofmrnace of the proposed CNN. To retain features extracted in early layers, the multi-path feature extraction scheme is proposed to mix features extracted from different layers. By using the proposed multi-path feature extraction scheme, the proposed CNN shows an improved classification performance.

Section 3.1 explains the motivations to employ CNN for pulmonary classification on CT imaging. Section 3.2 describes the pre-processing steps to transform the CT imaging from the LUNGx Challenge database [55] to nodule patches, and the characteristics of the LUNGx Challenge database used to train and evaluate the proposed CNN are also provided in this section. Section 3.3 presents the architecture of the proposed CNN and the training and evaluation schemes. Section 3.4 provides the results of the proposed CNN and a classification performance comparison with other state-of-the-art works, including CNN-based approaches

[1] The major portion of this chapter is originally published as "Novel Convolutional Neural Network Architecture for Improved Pulmonary Nodule Classification on Computed Tomography" in Multidimensional Systems and Signal Processing.

Yi Wang (YW), Hao Zhang (HZ) and Seok-Bum Ko (SK) made the conception and design of the study. YW developed and optimized the network architecture, wrote the code of the network, and performed result analysis. Kum Ju Chae (KJC) and Gong Yong Jin (GYJ) annotated data. KJC performed the observation performance test. HZ and Younhee Choi (YC) provided suggestions to improve the network architecture. YW drafted the manuscript. SK provided suggestions on improving the manuscript structure.

and machine learning-based approaches.

3.1 Introduction

According to the statistical data published by the American Cancer Society [56], lung cancer is the leading cause of cancer death. Early detection and diagnosis is an important way to increase the survival rate of cancer patients. In the field of lung cancer diagnosis, CT, due to its high sensitivity to detect pulmonary nodules, is widely used by radiologists. It shows a higher detection rate compared to chest x-ray (radiograph), and thus it is more effective at reducing lung cancer mortality [57].

In practice, although the CT scans have high sensitivity in nodule detection, it is still not easy for a radiologist to determine whether the nodule is benign or malignant. [58] reported mortality from lung cancer was reduced 20% by screening low-dose CT scans while a total of 96.4% of positive screenings showed a false positive. A false positive diagnosis leads to unnecessary follow up medical examinations, mostly with further CT examination that increases radiation exposure for patients.

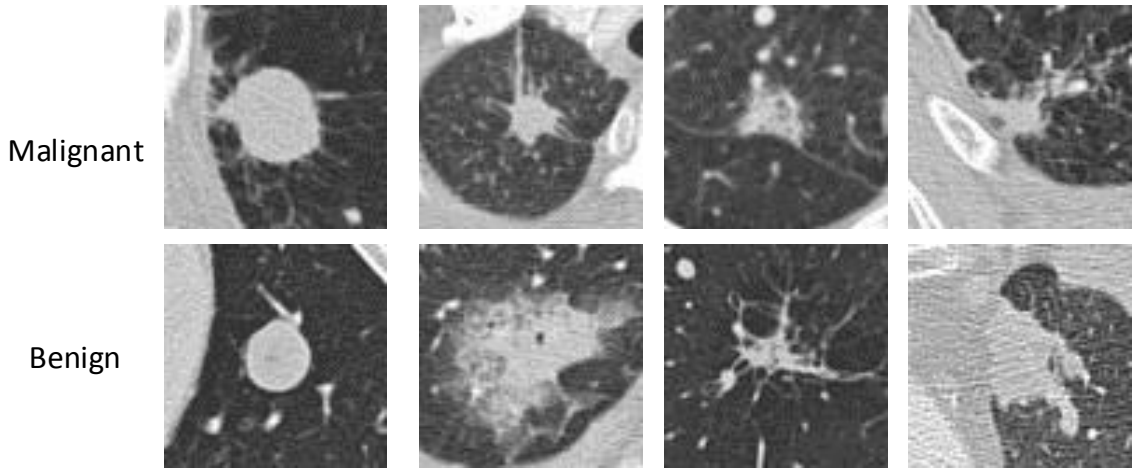


Figure 3.1: Examples of pulmonary nodules. Nodules are located in the center of image boxes (80mm × 80 mm). These examples illustrate that pulmonary nodule classification is a challenge due to the variety in sizes, shapes, and similar visual representation between malignant and benign nodules.

Computer-Aided Diagnosis (CADx) systems have been developed to assist thoracic radiologists in distinguishing between malignant and benign nodules. In general, such CADx systems translate the nodule features into benign nodule or malignant nodule decisions via a classifier. Because the CADx system improves the diagnostic accuracy of the radiologists, it becomes a good choice for preliminary diagnosis. Early developed CADx systems [17–19, 59, 60] use low-level features of medical imaging, which represent the global structures in the images such as contour, shape and size. These methods prove the usefulness of low-level features to classify pulmonary nodules because pulmonary nodules have various sizes and shapes as shown in Fig. 3.1. However, only relying on low-level features for nodule classification is less robust because different types of the nodules have similar visual representations, such as the two nodules of the first column in the Fig. 3.1, which cannot be distinguished with only global structures. More recent studies adopt unsupervised learning schemes such as principal component analysis (PCA) with support vector machines (SVM) [20] or Bag-of-frequencies [61]. These unsupervised feature learning schemes enable high-level image feature extraction and can achieve better classification performance than the feature extractors based on low-level features. However, unsupervised learning schemes require manual feature extractions, so it is complex to find the optimal feature combinations [62].

In recent years, deep learning has achieved great success in image classification, objective detection, image segmentation, and natural language processing. In many of these fields, deep neural networks (DNNs) can achieve near human performance [63]. CNN, the most popular DNN architecture, adopts supervised learning schemes [7, 52, 64] and has the ability to extract high-level features from raw input images without any manual intervention and thus is expected to be helpful in improving the performance of pulmonary nodule classification. An early study [22] adopts a shallow CNN with one convolutional layer to classify pulmonary nodules and shows a better accuracy compared to the conventional feature extraction methods. To express the full potential of the CNN, more recent studies employ deeper CNNs [23, 65, 66]. Instead of implementing tailored CNN architecture, transfer learning scheme also shows promising results for the nodules classification [24, 25].

[1] Global structures refer to low-level features which are utilized to describe global properties of the input such as edges or corners.

In the conventional CNN, global structures[1] are generated from early layers. When reaching the deeper layers, local structures[2] are gradually generated. A conventional CNN usually has only one path and can only use the local structures to perform classification. However, in pulmonary nodule classification, combining global and local structures, through a skip connection [9], is expected to improve the classification performance. In addition, the conventional CNN adopts a single-scale filter to extract features. However, due to variation in nodule sizes, analyzing nodule images using multi-scale filters [52] could generate more effective nodule features.

To fulfill the need of an accurate and robust pulmonary nodule classification, in this chapter, a novel convolutional neural network architecture is proposed. The proposed CNN adopts a multi-path feature extraction scheme to preserve both local and global structures[1]. In addition, in order to cover more effective nodule features, the proposed CNN uses multi-scale convolutional layers by using multiple filters with various sizes. Compared to the previous unsupervised learning approach [20], the proposed CNN achieves 14% improvement in an area under the receiver operating characteristic curves (AUC). The proposed CNN can also achieve up to 13% higher accuracy and 11% higher AUC than the previous models based on CNN [22–25, 67].

3.2 Data description and preparation

The LUNGx Challenge database [55] provides 70 patients' CT scans, and the entire database consists of 83 pulmonary nodules including 42 benign and 41 malignant nodules. Based on the measurements of the two thoracic radiologists (Co-authors of this paper, K.J.C and G.Y.J with 4 and 14 years of experiences, respectively), the sizes of malignant nodules are between 5.7 mm and 45.0 mm. Within the database, the sizes of benign nodules vary from 4.6 mm to 34.6 mm.

Each CT scan from the LUNGx Challenge database [55] is obtained under 120kV or 140kV tube peak potential energy with tube current in the range from 240 to 500 mA and

[2] In contrast to global structure, local structures [52] refer to high-level features that are more sparse and abstract patterns and are used to generalize the semantic representation of the input.

tube current-exposure time product of 200-325 mA. The CT scans are reconstructed as the digital imaging and communication in medicine (DICOM) format containing various 2D slice images, and each slice image has size of 512×512 pixels and the slice thickness is 1 mm.

To generate nodule patches which are the 2D representations of the nodules at each slice image, three steps are followed: 1) By referring the ground truth labels provided by the LUNGx Challenge database, the two thoracic radiologists (K.J.C and G.Y.J) confirmed the types of the nodules (benign or malignant), and the centers of the nodules were identified. 2) Depending on the size of each nodule, the two thoracic radiologists obtained the minimum bounding box that covers the entire nodule. The centers of the bounding boxes are aligned to the centers of the nodules, and the background information is preserved in the bounding boxes. The sample of the bounding boxes are showed in Fig. 3.2 as red rectangles. 3) By cropping the slice images with respected to the bounding boxes, the nodule patches are extracted. Since one nodule patch is extracted from one slice, multiple effective nodule patches can also be generated from different slices of the same nodule scan. As illustrated in Fig. 3.2, nodule patches captured from the same nodules have different visual representations. In total, 460 nodule patches are extracted from 42 benign nodules, and 671 nodule patches are extracted from 41 malignant nodules. Since the proposed CNN is trained and evaluated through 5-fold cross validation (as described in Section 3.3.2), 10% of 41 malignant nodules and 10% of 42 benign nodules are reserved in each fold.

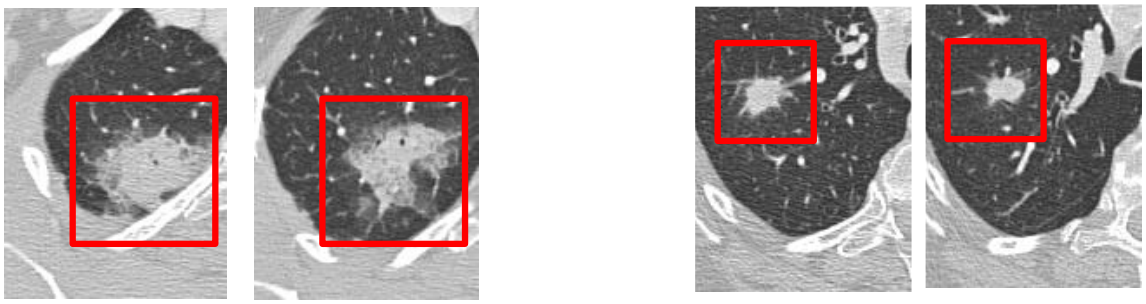


Figure 3.2: Extracted nodule patches via bounding boxes (red rectangles). (a) Two nodule patches captured from two slices of the same benign nodule. (b) Two nodule patches captured from two slices of the same malignant nodule.

3.2.1 Data augmentation

To achieve high classification performance, CNNs need a large dataset to tune their huge amount of learnable parameters. For example, the ImageNet database [51], used to trained AlexNet [7], contains 1.2 million training images. As discussed in Section 3.2, only 1,131 nodule patches are extracted from the LUNGx Challenge database. When the amount of data is limited, the trained network can neither model the training data well nor generalize to predict well for new data. In order to increase the amount of training images, a data augmentation method is usually applied [23, 67]. To gain unique spatial variance for each augmented nodule patch, in this chapter, the data augmentation involves different image processing algorithms by random selections of imaging rotation, flipping and translation. The rotation is done by rotating the nodule patches with a random angle from 0 degrees to 359 degrees. The nodule patches are randomly flipped vertically or horizontally. The translation randomly shifts the nodules towards the edges of the bounding boxes. To perform translation, the original bounding box is enlarged by 20% as shown in Fig. 3.3(b). Then, the translated nodule patch is extracted by shifting the enlarged bounding box with a shifting amount less than 20% of the enlarged bounding box's size.

3.2.2 Contrast normalization

Contrast normalization is applied by subtracting the global mean from the each pixel in all images. To obtain the global mean, intensities of all pixels are averaged by first summing the intensities of all pixels in all images, followed by dividing by the total number of the images. Normalized images are invariant to changes in illumination, which commonly occurs when CT scans are generated using different equipment. On the other hand, since the CNN is optimized by a gradient descent algorithm, normalized images have less variance and thus a large gradient is less likely to be produced [68]. Because a large gradient may lead the CNN towards a local minimal point, the performance of the CNN can be affected.

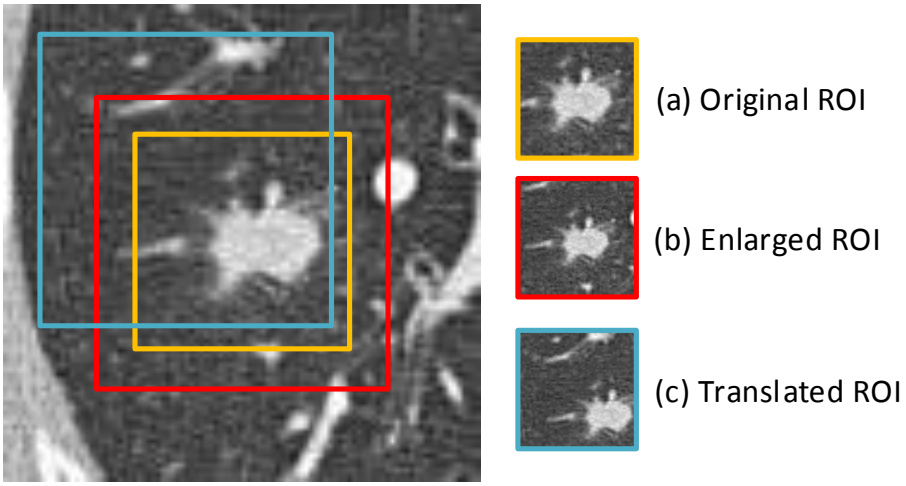


Figure 3.3: Illustration of translation operation. (a) Extracted nodule patch by the original bounding box defined by the two experienced radiologists. (b) Extracted nodule patch by the enlarged bounding box with 1.2 times sizes of the original one. (c) Translated nodule patch by shifting the enlarged bounding box within 20% of its size.

3.3 Proposed CNN

A typical CNN uses a series of computational layers, such as convolutional layers, pooling layers, and fully-connected layers, to extract features from raw input images and then use them to classify input images into different categories. Because of effective high-level feature extraction, many previous studies, including [23, 65, 66], demonstrate the usefulness of the conventional CNN in diagnosing lung cancer from CT images. In this chapter, a novel CNN architecture is proposed for nodule classification in lung CT scans. The proposed CNN applies two proposed techniques to improve classification performance: 1) Multi-path feature extraction scheme (Section 3.3.1) creates an additional path to combine features generated from early layers and later layers to preserve the losses of global structures (e.g., object contour or shape) in later layers. 2) The conventional convolutional layer is replaced by the proposed multi-scale convolutional layer (Section 3.3.1) to provide multi-scale features. Instead of using a conventional (single-scale) convolutional layer in which the feature extraction is performed using fixed-scale filters, the multi-scale convolutional layer adopts multiple

filters with various sizes to cover more local structures.

3.3.1 Architecture details

The architecture of the proposed CNN is shown in Fig. 3.4(a). The input size of the proposed CNN is 96×96. The proposed CNN has two feed-forward paths. The main path performs multi-scale feature extraction, and the fundamental structure of the main path is based on the hierarchical neural network that consists of consecutive 4 single-scale convolutional layers followed by 3 fully-connected layers. On the other hand, the second path is used for multi-path feature extraction by creating a shortcut from the first single-scale convolutional layer to the first fully-connected layer in the main path. Regarding the main path, the multi-scale feature extraction is done after two single-scale convolutional layers. The first single-scale convolutional layer performs a 7×7 convolution operation to extract 12 output feature maps. Similarly, the second single-scale convolutional layer uses the same filter size to generate 24 output feature maps. Then, this is followed by two multi-scale convolutional layers.

The two paths are eventually merged and connected to the first fully-connected layer. After three fully-connected layers, the final output features are used to classify pulmonary nodules into two classes which are benign and malignant. The first fully-connected layer has 256 hidden neurons. The second fully-connected layer consists of 128 hidden neurons while the last fully-connected layer has 2 hidden neurons in order to match the number of output classes. After the last fully-connected layer, the predicted outcomes are normalized via SoftMax [69]. The denser fully-connected layer often leads to overfitting due to full connectivity structure. Here in the proposed CNN architecture, in order to avoid overfitting, a dropout [70] scheme is applied to the first and second fully-connected layers with a dropout ratio of 0.5.

The rectified linear unit (ReLU) is used as the activation function of each convolutional layer. The ReLU suppresses negative values to 0 and keeps positive values. The ReLU shows better gradient changes compared to the sigmoid function and the tanh function [7]. Besides, it is easy to be implemented which can help improve the speed performance of training and inference of the network.

A max-pooling layer is inserted after each convolutional layer. The max-pooling layers are

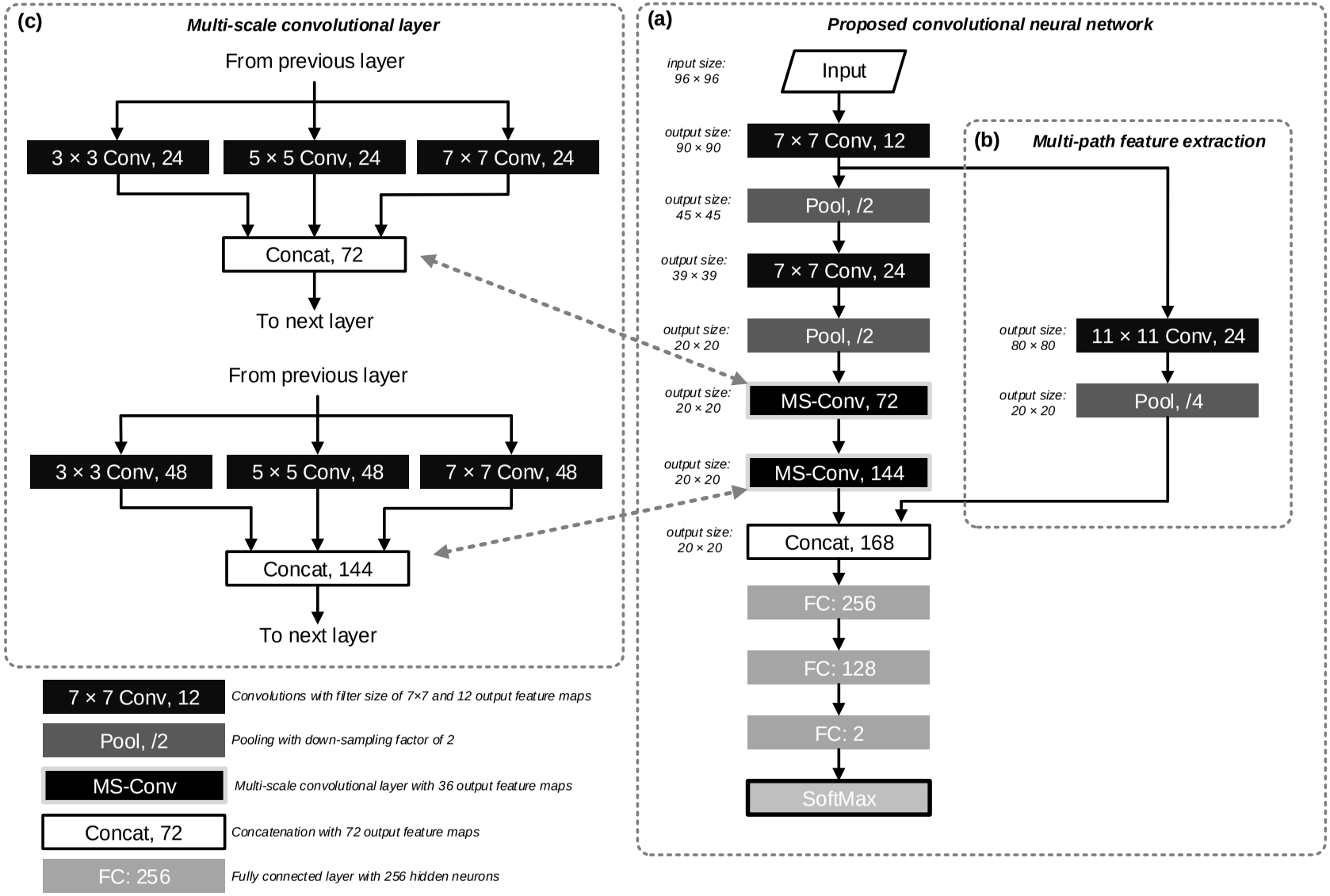


Figure 3.4: (a) Architecture of the proposed CNN. (b) Illustration of the multi-path feature extraction. (c) Configuration detail of the multi-scale convolutional layer. Conv, single-scale convolutional layer followed by rectified linear unit; pooling, maximum pooling layer; MS-Conv, multi-scale convolutional layer; Concat, depth-wise concatenation; FC, fully-connected layer.

used to reduce the spatial resolution of the following layers of the network in order to maintain more relevant local structures. Moreover, the pooling operation increases the receptive field size so that the network can learn more complex local structures from the input.

Multi-path feature extraction

In this chapter, a multi-path feature extraction scheme is realized by concatenating feature maps of different levels of layers. The multi-path feature extraction allows the CNN to combine more robust features with respect to the fine global structures and sparse local structures.

The conventional CNN has a straightforward path to pass extracted features into the fully-connected layers. The earlier layers of the conventional CNN extract features with respect to the fine global structure. After the CNN propagates to the deeper layer, the extracted features become more sparse and localized, while the global structures are diminished. Although the conventional CNN shows is well-suited to extract features that describe local structures, the performance could be limited in the analysis of the pulmonary nodules. Since the visual representations of the pulmonary nodules have various shapes and sizes, local structures have less power to represent such low-level features compared to global structures.

The proposed multi-path feature extraction remedies losses of global structures via a shortcut. As shown in Fig. 3.4(b), the shortcut branches the output of the first convolutional layer and creates a skip connection to the first fully-connected layer. To merge the features from two branches, a maximum pooling layer is inserted in the shortcut. By passing the features of the early layer into the first fully-connected layer, the final combined features can describe both global and local structures and thus are expected to provide more nodule features, such as global contour of the nodule and local unique textures.

However, one problem for the shortcut is that passing the features of the early layer may mislead the CNN to analyze unrelated global structures, such as normal tissues surrounded to the nodule. Also, the CNN with a shortcut is prone to interference from global noise due to low resolution of the CT scans. To resolve the above mentioned two issues, in this chapter, an additional convolutional layer with filter size of 11×11 is inserted into the shortcut to refine the features obtained from early layers for better representations of global structures.

Multi-scale convolutional layer

A naive methodology to improve CNN performance is to increase its depth via additional convolutional layers. However, this method would cause overfitting because the dataset is small. Also, choosing an appropriate filter size for convolution is difficult. The right filter size significantly improves accuracy for pulmonary nodule classification [71]. In this chapter, a multi-scale convolutional layer is proposed to better extract local sparse structures with respect to different receptive fields. Instead of using a single-scale filter, the filter bank in the multi-scale convolutional layer is configured with three different filter sizes.

The configuration detail of the multi-scale convolutional layer is shown in Fig. 3.4(c). The multi-scale convolutional layer consists of three branches. The first branch performs a 3×3 convolution operation. The second branch is responsible to extract relatively lager local structures by employing a 5×5 convolutional layer. Similarly, the last branch has one convolutional layer except the size of filter is further enlarged to 7×7 in order to cover more local structures. The feature maps generated via three branches are merged via depth-wise concatenation. In order to perform an effective concatenation, the input of 3×3 convolutional layer is padded with a 1×1 border of zeros while zero paddings of 2 or 3 are used for the 5×5 or 7×7 convolutional layer respectively.

The insight of multi-scale convolutional layer is to extract features via different filter sizes. When the local structure cannot be described by a single-scale filter, by applying a relatively large filter size, such structure can be covered due to increased receptive field. In this chapter, filter sizes of 3×3, 5×5 and 7×7 are used to extract more local structures simultaneously. Empirically, standard filter sizes for effective nodule feature extraction are: 3×3, 5×5, 7×7 or 9×9 [22,23,71]. However, it is still hard to conclude the optimal filter combination among all previous studies. The proposed CNN extends a single-scale convolutional layer to a multi-scale convolutional layer which allows the network to generate more local structures with varying sizes.

3.3.2 Model training and evaluation

In order to select a trained model with robust generalization, the proposed CNN is trained and evaluated via 5-fold cross validation as described in [67]. Each fold consists of 10% of 41 malignant nodules and 10% of 42 benign nodules. During each round of the cross validation, three folds are reserved for training, while leaving one fold for validation and one fold for testing. The augmentation is done within the training data set. The accuracies, sensitivities, specificities, and AUCs across each of the 5 cross validation folds are averaged to obtain overall performance results.

Since the augmented nodule patches contain more malignant samples than benign ones, class imbalance might have a detrimental effect on training convergence and the generalization of the trained model [72]. One of the most common ways to deal with imbalance is to over-sample the minority class, and it shows performance improvement in deep learning [73,74]. To balance the proposed training and validation datasets, over-sampling is applied by duplicating benign nodule patches with random selections.

At the beginning of the training process, the weights of all convolutional layers and fully-connected layers are initialized via Gaussian initialization. For all convolutional layers, initial weights have standard deviation of 0.001 and a constant bias of 0, while the weights in the fully-connected layers have standard deviation of 0.005 and constant bias of 0. During the training phase, the model is trained up to 50 epochs, and is validated and tested after each epoch. The loss function is optimized by Adagrad [49] with a batch size of 128.

3.4 Results and analysis

3.4.1 Experimental environment

The proposed CNN is implemented using Caffe [75] deep learning framework. Other convolutional neural network experiments are also implemented using Caffe environment. Network training and testing is performed on an Ubuntu 16.04 Linux system with Intel Xeon E5-2630 @ 2.60 GHz, 32 GB RAM and Nvidia K40 GPU.

3.4.2 Analysis of proposed CNN

In this section, relevant experiments are performed to tune the hyper-parameters and justify the performance of the multi-path feature extraction scheme and the multi-scale convolutional layer.

Tuning of hyper-parameters

A series of experiments is conducted to justify the choices of input scale, number of hidden neurons in the fully-connected layers, and training parameters in the proposed CNN. The results are summarized in Table 3.1. Increasing the input scale from 96×96 to 128×128 reduces the classification accuracy, roughly by 1.5%. Reducing the input scale to 64×64 drops accuracy by 4%. Furthermore, when number of hidden neurons in the fully-connected layers is doubled or halved, the resulting classification accuracies are lower than that proposed. Without applying dropout scheme, the proposed CNN shows approximately 1.5% reduced accuracy. Moreover, increasing the batch size from 96 to 128 shows an accuracy improvement of approximately 0.5%. Initializing the convolutional layer weights using a Gaussian distribution with zero mean and a standard deviation of 0.01 shows improved accuracy compared to using standard deviations of 0.05 or 0.005.

Analysis of multi-path feature extraction

In this section, the performance of the multi-path feature extraction is evaluated by comparing the classification performances between the proposed CNN with and without the shortcut. The shortcut improves the accuracy from 86.77% to 90.38%, the AUC from 0.914 to 0.948, and contributes to a 2% sensitivity and 6% specificity increase. These results are shown in Table 3.2. This improvement over the proposed CNN without a shortcut confirms that the multi-path feature extraction is well-suited to extracting low-level features with respect to the global structures, while keeping the high-level features to represent sparse local structures. Intuitively, the shortcut passes more global structures which have potential usefulness to describe the shapes or sizes of the nodules.

The effectiveness of the proposed multi-path feature extraction is investigated by creating

Table 3.1: Classification accuracies of the proposed CNN with different hyper-parameters.

Input scale	FC layer factor	Dropout ratio	Batch size	Weight initialization	Accuracy (%)
64	1	0.5	128	0.01	86.60
128	1	0.5	128	0.01	88.77
96	1/2	0.5	128	0.01	89.17
96	2	0.5	128	0.01	87.61
96	1	0	128	0.01	88.50
96	1	0.5	96	0.01	89.93
96	1	0.5	128	0.005	89.28
96	1	0.5	128	0.05	88.50
96	**1**	**0.5**	**128**	**0.01**	**90.38**

Underline represents the changed hyper-parameters.

Bold represents the best configuration of hyper-parameters.

Table 3.2: Classification accuracies of the proposed CNN with and without shortcut.

Network	Accuracy (%)	Sensitivity	Specificity	AUC
Proposed CNN without shortcut	86.77	0.869	0.863	0.914
Proposed CNN with shortcut	**90.38**	**0.887**	**0.924**	**0.948**

multiple branches in parallel. Intuitively, all branches share the same input features and perform feature extraction via independent convolution operation, followed by a maximum pooling. As shown in Table 3.3, the classification performance is reduced by adding branches into the proposed multi-path feature extraction. Specifically, three branches' multi-path feature extraction has lower classification performance than the one employing two branches. Additionally, employing multi-scaled convolution filters (e.g., {11×11, 13×13}) has better classification performance than the one applying a fixed-scale convolution filter (e.g., {11×11, 11×11}). However, it still cannot reach the performance of the proposed multi-path feature extraction. The results confirmed two aspects that creating multiple branches has no benefit for the proposed multi-path feature extraction. First, to pass more effective global structures from the early layer, a large convolution filter is used to refine the global noise. However, since the output features of each branch are concatenated as the final output features, using multiple branches may enhance the global noise. Secondly, employing multiple branches increase the complexity of the model due to large filter size, which makes the network more difficult to be fine-tuned.

Effectiveness of multi-scale convolutional layer

To verify the effectiveness of the proposed filter configurations in the multi-scale convolutional layer, the proposed multi-scale convolutional layer is replaced by single-scale convolutional layer with filter size of 3×3, 5×5 or 7×7. In the best scenario, replacing multi-scale convolutional layer with 5×5 convolution operation achieves an accuracy of 88.06% and an AUC of 0.931 as shown in Table 3.4, but not even being able to reach the accuracy and AUC in case of using two branches' multi-scale convolutional layer with filter sizes of {3×3, 5×5} or

Table 3.3: Classification performances of the proposed multi-path feature extraction and multiple branches' multi-path feature extraction with different filter configuration.

Filter configuration	Accuracy (%)	Sensitivity	Specificity	AUC
{11×11}	**90.38**	**0.887**	**0.924**	**0.948**
{11×11, 11×11}	88.40	0.870	0.892	0.920
{11×11, 13×13}	<u>88.79</u>	<u>0.877</u>	<u>0.901</u>	<u>0.938</u>
{11×11, 15×15}	87.63	0.861	0.895	0.922
{11×11, 17×17}	86.34	0.847	0.884	0.916
{11×11, 19×19}	86.21	0.849	0.878	0.911
{11×11, 21×21}	85.18	0.833	0.875	0.910
{11×11, 11×11, 11×11}	87.11	0.866	0.878	0.914
{11×11, 13×13, 15×15}*	87.79	0.870	0.887	0.929
{11×11, 13×13, 17×17}	85.95	0.838	0.887	0.921

{11×11, 13×13, 15×15}*: three branches' multi-path feature extraction with filter sizes of 11×11, 13×13 and 15×15.

Bold represents the top-1 values among all filter configurations.

<u>Underline</u> represents the top-2 values among all filter configurations.

{3×3, 7×7}. The results conclude that proposed multi-scale convolution improves classification performance over single-scale convolution since multi-scale convolution varies the sizes of the filters to cover more possible local structures, while the single-scale convolution uses a fixed-scale filter regardless of the effective sizes of the local structures.

Furthermore, to identify the optimal multi-scale feature extraction scheme, a series of experiments is conducted by varying the size of the filter at each branch of the multi-scale convolutional layer. In particular, the experimental choices of filter sizes are 3×3, 5×5 and 7×7, which the single-scale convolutional layer adopting those filter sizes showed effective feature extraction in the previous studies of the pulmonary nodule classification [23, 71, 76]. Table 3.4 summarizes the performances of the proposed CNN with respect to different filter configurations in the branches of the multi-scale convolutional layer. Based on the

Table 3.4: Classification performances of the multi-scale convolutional layers with different filter configurations, or substituted by single-scale convolutional layers.

Convolutional layer	Filter configuration			Accuracy (%)	Sensitivity	Specificity	AUC
	3×3	5×5	7×7				
Single-scale	✓			87.89	0.881	0.876	0.926
		✓		88.06	0.876	0.886	0.931
			✓	87.63	0.876	0.876	0.907
Multi-scale	✓	✓		89.09	0.870	0.917	0.932
	✓		✓	88.23	0.867	**0.942**	0.938
		✓	✓	87.54	0.866	0.888	0.935
	✓	✓	✓	**90.38**	**0.887**	0.924	**0.948**

Underline represents the best values among all configurations of single-scale convolutional layer.
Bold represents the best values among all configurations of multi-scale convolutional layer.

experimental result, the multi-scale configuration used in the proposed CNN, containing three branches with filter sizes of 3×3, 5×5 and 7×7, yields superior classification performance compared to different two branches' configurations. The proposed configuration achieves an accuracy of 90.38%, a sensitivity of 0.887, a specificity of 0.924, and an AUC of 0.948.

In addition, one previous study [77] adopts the concept of multi-scale feature learning from Inception modules [8, 53]. In order to show that the proposed multi-scale convolutional layer provides more efficient nodule feature extraction than the Inception modules used in [77], the multi-scale convolutional layers of the proposed CNN are replaced by the Inception-V2 or Inception-ResNet modules. As shown in Table 3.5, using either of the two Inception modules to replace the proposed multi-scale convolutional layer shows lower classification performance than the proposed one having three branches. The results indicate the effectiveness of the proposed multi-scale feature extraction as follows: 1) One branch of both Inception modules rely on 1×1 convolution operation, but the size of the 1×1 filter is too small to analyze the local structure entirely. Intuitively, the CNN cannot learn any correlated patterns from 1×1 convolution operation. Therefore, this convolution size is not considered in the proposed multi-scale method. 2) Although two branches in both Inception

modules perform 3×3 and 5×5 convolution operations in order to recover the loss of local structures from 1×1 convolution operation, based on the experimental results, Inception-V2 module shows no major performance difference by comparing with the two branches' multi-scale convolutional layer with filter sizes of {3×3, 5×5}. This comparison indicates that an additional branch applying maximum pooling has negligible benefit for multi-scale feature learning for pulmonary nodule classification. 3) The residual block in the Inception-ResNet module demonstrates the usefulness of solving the vanishing gradient problem when the CNN is very deep [9]. However, based on the experimental results, the residual block has no benefit to the proposed CNN because the number of layers is small.

Table 3.5: Classification performances of the multi-scale convolutional layers, or substituted by Inception-v2 or Inception-ResNet modules.

Method	Type of multi-scale convolutional layer	Accuracy (%)	Sensitivity	Specificity	AUC
Kang et al. [77]	Inception-V2	89.18	0.867	0.922	0.937
	Inception-ResNet	87.80	0.863	0.897	0.918
Proposed	{3×3, 5×5}[1]	89.09	0.870	0.917	0.932
	{3×3, 5×5, 7×7}	**90.38**	**0.887**	**0.924**	**0.948**

{3×3, 5×5}[1]: two branches' multi-scale convolutional layer with filter sizes of 3×3 and 5×5.

3.4.3 Comparisons with previous works

In this section, the diagnostic performance of the proposed CNN is evaluated against sate-of-the-art works applying unsupervised or supervised feature learning approaches.

Comparison with unsupervised feature learning approach

To show the superiority of the proposed method over unsupervised learning scheme, the proposed method is compared with Nishio's study [20], a nodule-based unsupervised feature learning approach. To extract robust unsupervised features, Nishio's study applies a

Table 3.6: Comparison of the proposed CNN with unsupervised feature learning approach.

	Method	Cross validation	ROC analysis	Sensitivity	Specificity	AUC
Nishio and Nagashima [20]	Hisogram + SVM	10-fold	nodule-based	0.867	0.488	0.640
	LBP-TOP + SVM			0.900	0.558	0.688
	RLBP + SVM			0.800	0.674	0.725
	Multi-stages + SVM			0.867	0.744	0.837
Proposed	CNN	5-fold	nodule-based	0.875	0.889	0.958
			patch-based	0.887	0.924	0.948

Histogram, histogram of CT density; LBP-TOP, local binary pattern on the three orthogonal planes; RLBP, LBP with 3D random sampling; Multi-stages, combination of PCA, convolution and pooling operations; SVM, support vector machine.

combination of PCA and convolution operation. Evaluating using the LUNGx Challenge database, Nishio's study demonstrates superior performance among other hand-crafted features, such as histogram of CT density, LBP on the three orthogonal planes (axial, coronal and sagittal CT images), and LBP with random sampling. Since the proposed method is a patch-based supervised feature learning approach, direct comparison between the proposed method and Nishio's method is not available. To perform nodule-based analysis in this chapter, the patch-based classification results are summarized into pre-nodule through committee-fusion as describe in [78]. Intuitively, the whole nodule image can be reconstructed by stacking the corresponding nodule patches along sagittal axis. Hence, distinguishing the nodule in-between malignant and benign can be divided into several classification tasks of the corresponded nodule patches. By averaging the corresponded nodule patches' classification results, the prediction outcome of the nodule is obtained. Based on the nodule-based analysis, the proposed method (Sensitivity = 0.875, Specificity = 0.889, AUC = 0.958) outperforms Nishio's method as shown in Table 3.6. The proposed method surpasses the unsupervised feature extraction as described in Nishio's work in two key ways: 1) Principal components are based on the covariance matrix. To provide sufficient high-level features, the covariance matrix cannot be evaluated in an accurate manner due to the complexities in high dimensional spaces. 2) PCA method does not have a strong ability to analyze variance in object location [79]. When the nodules appear in different locations of the input images during testing phase, the PCA method may generate inaccurate predictions. However, the CNN-based methods have a strong ability to process translational variance for objects. Therefore, the proposed method has better feature extraction ability.

Comparison with other CNN approaches

In this section, the classification performance of the proposed CNN is compared with the CNNs using single-scale convolutional layers [22, 23], pre-trained AlexNet and GoogLeNet with transfer learning [24, 25], and multi-cropped CNN [67]. Previous works' classification performance statistics are generated by training and evaluating their CNNs using the LUNGx Challenge database with the same data partitioning and evaluation scheme as used in the proposed architecture (as described in Section 3.3.2). The results are summarized in Ta-

Table 3.7: Comparison of the proposed CNN with other CNN architectures on LUNGx Challenge database.

	Methods	Accuracy (%)	Sensitivity	Specificity	AUC
Li et al. [22]	CNN with 1 single-scale convolutional layer	80.15	0.789	0.818	0.854
Zhao et al. [23]	CNN with 2 single-scale convolutional layers	84.97	0.843	0.858	0.902
Tajbakhsh et al. [24]	Transfer learning on pre-trained AlexNet	80.58	0.821	0.787	0.855
Shin et al. [25]	Transfer learning on pre-trained GoogLeNet	86.68	**0.906**	0.798	0.933
Shen et al. [67]	Multi-cropped CNN	86.77	0.846	<u>0.895</u>	<u>0.940</u>
Proposed Method	Multi-scale + multi-path	**90.38**	<u>0.887</u>	**0.924**	**0.948**

Bold represents the top-1 values among all CNNs.

<u>Underline</u> represents the top-2 values among all CNNs.

ble 3.7. Specifically, the proposed CNN achieves the highest accuracy (90.38%), specificity (0.924), and AUC (0.948) compared to other CNNs. The resulting sensitivity of the proposed CNN is 0.887, only 0.019 lower than the model achieving the highest sensitivity, a pretrained GoogLeNet [25] with transfer learning.

Compared against the CNNs applying single-scale convolutional layers [22, 23], the proposed CNN shows improved classification performance with a minimum 5% improvement in both of accuracy and AUC. The results confirm the improvement over single-scale CNNs in two aspects: 1) The multi-path feature extraction applied to the proposed CNN enhances the feature representation power. The proposed CNN does not only extract sparse high-level features with respect to local regions, but it also retains useful fine low-level features to describe the global structure such as the shape or size of the nodule. 2) Instead of using a single-scale filter to extract the local structure, the multi-scale convolutional layer efficiently captures local structures with different sizes. The performance of the proposed CNN is also compared with pre-trained AlexNet [24] and GoogLeNet [25] with transfer learning. AlexNet is an 8 layered single-scale CNN, does not perform as well as either the proposed CNN or the pre-trained GoogLeNet, both of which apply multi-scale feature learning. However, the pre-trained GoogLeNet still cannot reach the performance of the proposed one. By taking advantage of multi-path feature extraction scheme, the proposed CNN preserves the global structures of the nodule, while the extracted features of the GoogLeNet are limited in local regions.

Furthermore, the performance of the proposed CNN is compared with the multi-crop CNN [67], cropping feature maps towards multi-scale feature learning. Despite concerns that the multi-crop CNN is evaluated under nodule-based data, using multi-crop CNN to extract patch-based data should also be valid since the concept is to crop the feature maps with various sizes to enhance the central nodule feature regardless of the input shape. Based on the results as shown in Table 3.7, the proposed CNN shows better accuracy and AUC than the multi-crop CNN. Instead of cropping feature maps, the proposed multi-scale convolutional layer uses multiple filters with various sizes to provide more local structures; hence, the central nodule features are also enhanced. On the other hand, the proposed multi-path feature extraction passes the useful global structures to further enhance the nodule features

corresponding to size or shape of the nodule.

The receiver operating characteristic curves (ROCs) for different CNN architectures are shown in Fig. 3.5. To test the statistical significance of the AUC differences between the proposed CNN and other CNN approaches, a statistical analysis, the significance level under DeLong test [80], is performed. The results of DeLong test concludes that the proposed CNN showed statistical significance ($\rho < .05$) compared to other CNN approaches except pre-trained GoogLeNet [25] ($\rho = 0.0509$) and multi-crop CNN [67] ($\rho = 0.2749$).

The training time, inference time, and testing error of the proposed CNN are compared against other CNNs. It should be noted that all experiments were conducted on an NVIDIA K40 GPU. The proposed CNN takes 0.966 seconds to finish predicting one nodule patch. Each training epoch for the proposed CNN takes 3.750 milliseconds. As shown in Fig. 3.6, the proposed CNN has slower training and inference time compared to [22], [23] or [67]. Since the proposed CNN consists of more learnable parameters than these three CNNs, it is expected that the proposed CNN takes longer to train or perform inference. Nevertheless, compared to these three CNNs, the inference time of the proposed CNN is less than half milliseconds slower, which is negligible for real clinical usage. Most importantly, of all the compared CNNs, the proposed CNN has the lowest testing error.

3.4.4 Usefulness of the proposed CNN for clinical decision

Table 3.8: Comparison of the proposed CNN with thoracic radiologist.

	Review rounds	Sensitivity	Specificity	AUC
Thoracic radiologist	1st round	0.625	0.667	0.715
	2nd round	0.750	0.778	0.882
Proposed CNN		0.875	0.889	0.958

To demonstrate the potential usefulness of the proposed CNN in making clinical decisions, an additional experiment is conducted to compare the performance between the proposed CNN and a thoracic radiologist (co-author of this paper, K.J.C with 4 years of experiences in thoracic imaging). First, the nodule-based predictions of the proposed CNN are collected

by averaging the corresponded nodule patches' predictions of each nodule in the testing dataset used in this chapter. Then, the thoracic radiologist reviews the same nodules without knowing the patients' information and the predicting outcomes of the CNN, followed by a second-round review given the CNN's predictions. To make a valid comparison between the proposed CNN and the thoracic radiologist, the thoracic radiologist is asked to provide the malignancy rating of the nodule as 10-scale (from 0% to 100%). Table 3.8 shows sensitivities, specificities and AUCs of the proposed CNN and the thoracic radiologist. Fig. 3.7 shows the examples of classification results from the proposed CNN and the thoracic radiologist. The results indicate that the performance of the thoracic radiologist is clearly improved by refereeing the predictions of the CNN. Specifically, there are only 2 out of 17 cases that the proposed CNN successfully convinces the radiologist to change the wrong decisions in the first-round review. In most cases, both provide the same diagnostic decision, but the CNN tends to provide more confident results such as high probable malignant ($> 90\%$) or benign ($< 10\%$). Despite concerns that this experiment is conducted with a small number of nodules (17) and not being confirmed via external validation, the proposed CNN is still a good clinical tool and sober second thought for the radiologist and their decisions. There are circumstances that follow-up medical examinations and treatments may be required prior to the radiologists' interpretations. During the radiologists' interpretations, the proposed CNN can be used in decision making thus the follow-up time is shorted.

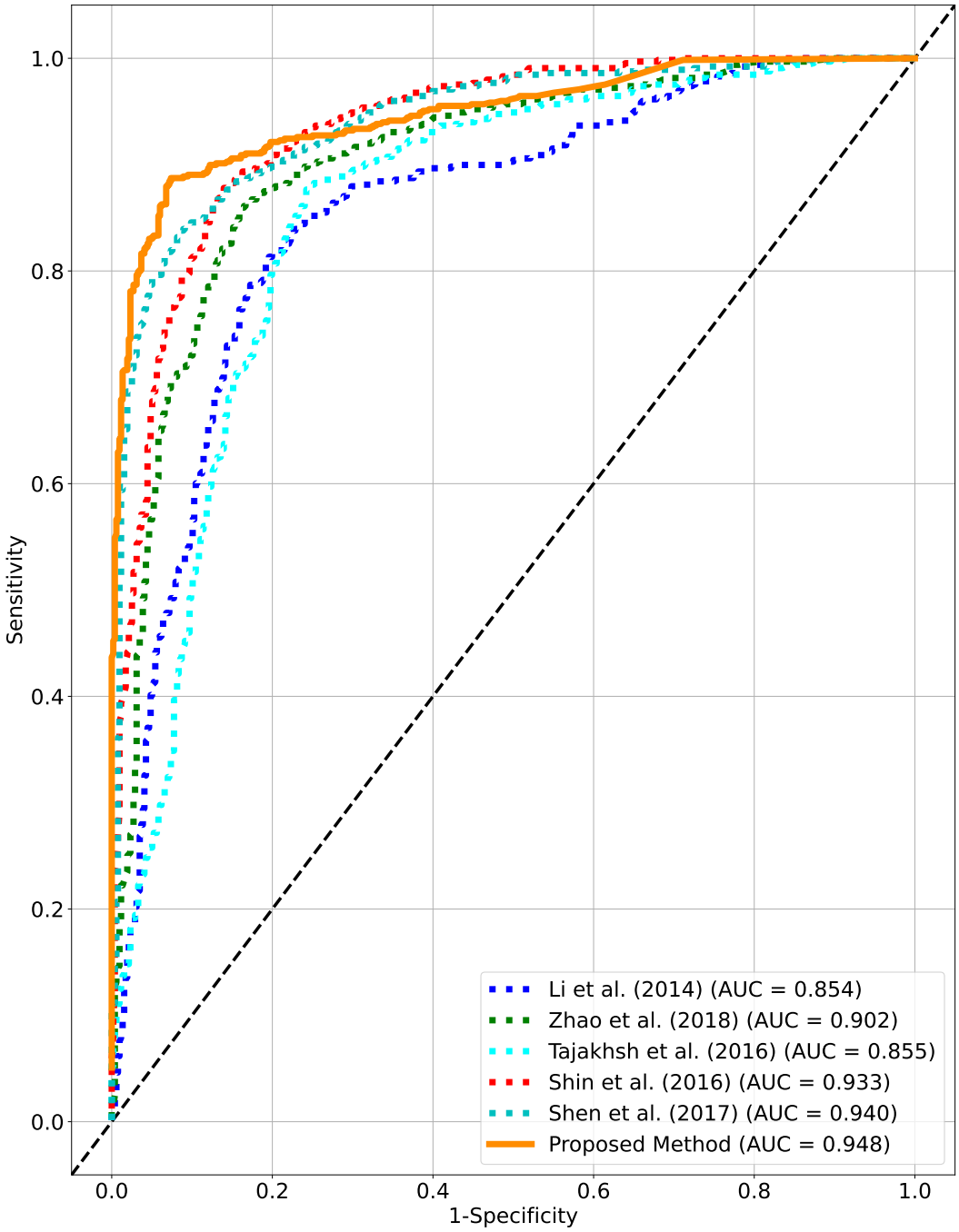


Figure 3.5: ROC curves for different CNN architectures.

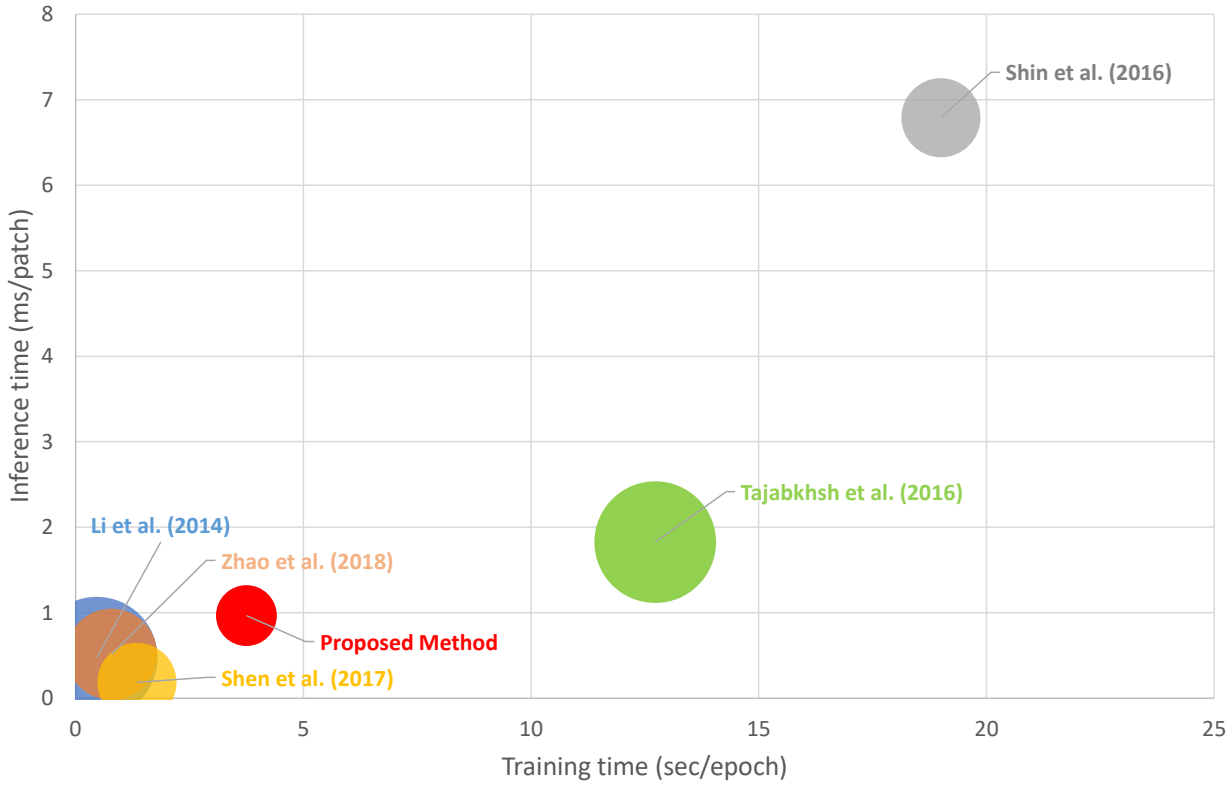


Figure 3.6: Inference time vs. training time and testing error for different CNNs.

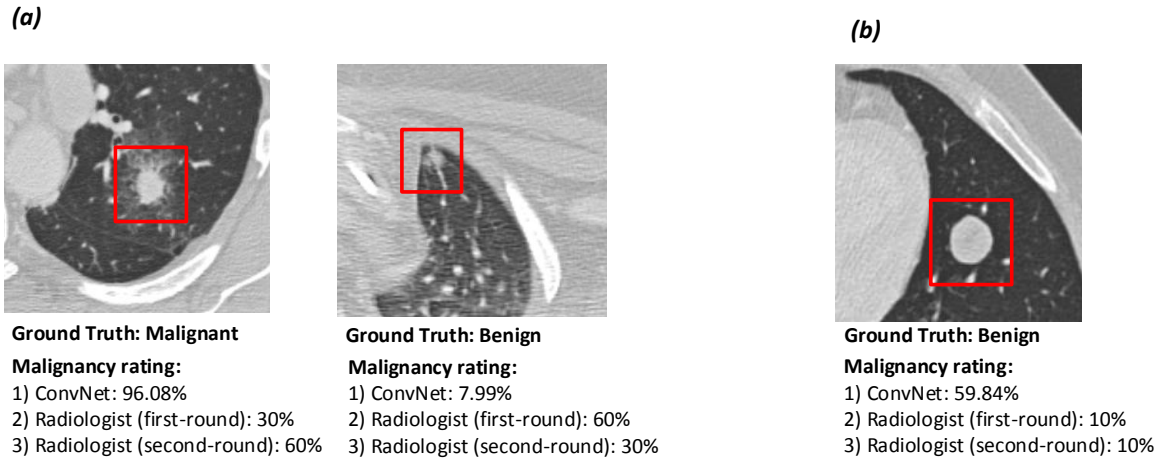


Figure 3.7: Examples of classification results from the proposed CNN and thoracic radiologist. a) Radiologist makes correct prediction after refereeing the CNN. b) Wrong prediction made by CNN. c) Wrong prediction made by radiologist. d) Wrong prediction made by both CNN and radiologist.

Chapter 4

Convolutional Neural Network Based Computer-aided Diagnosis System for Breast Lesion Classification on Automated Breast Ultrasound[1]

This chapter presents a convolutional neural network (CNN) based computer-aided diagnosis (CADx) system tailored for automated breast ultrasound (ABUS) imaging. Since ABUS imaging can be visualized on transverse and coronal views, two multiview CNNs are proposed to extract features from different views simultaneously.

Section 4.1 describes the motivations for designing a CNN-based CADx system for breast lesion classification on ABUS imaging. The characteristics of the ABUS imaging database used for this study and the architectures of the proposed multiview CNNs are presented in Section 4.2. Section 4.3 presents the results of the proposed multiview CNNs. Section 4.4 presents the detailed analysis of the proposed multiview CNNs.

4.1 Introduction

Breast cancer is the second leading cause of female cancer death [81]. To reduce the mortality rate of breast cancer, early detection and treatment are important. Mammography and

[1] The major portion of this chapter is originally published as "Breast Cancer Classification in Automated Breast Ultrasound Using Multiview Convolutional Neural Network with Transfer Learning" in Ultrasound in Medicine & Biology.

Yi Wang (YW) and Seok-Bum Ko (SK) made the conception and design of the study. YW developed and optimized the network architecture, wrote the code of the network, and performed result analysis. Eun Jung Choi (EJC) and Gong Yong Jin (GYJ) prepared the data. EJC designed and conducted the observation performance test. HZ and Younhee Choi (YC) provided suggestions to improve the network architecture. YW drafted the manuscript. SK provided suggestions on improving the manuscript structure.

ultrasound (US) are common screening modalities to diagnose breast cancer. Recent studies showed improved diagnostic performance in the dense breast by interpreting handheld US imaging in addition to mammography [82–84].

Nevertheless, interpreting the handheld US is operator dependent since the visualization of one handheld US image is limited in one certain orientation [83,84]. To minimize the operator dependence, the automated breast ultrasound (ABUS) has been introduced to provide the capability of breast visualization in a 3-D volume where it is helpful to locate breast lesion prior to diagnose breast cancer. One example of ABUS is shown in Figure 4.1. The ABUS imaging reconstructs the breast in transverse view, which is similar to the handheld US. Besides, the ABUS imaging offers a coronal view of the breast. In practice, although ABUS demonstrates better reproducibility and less time-consuming than the hand-held US [26], screening ABUS still takes a significantly longer time than mammography [85].

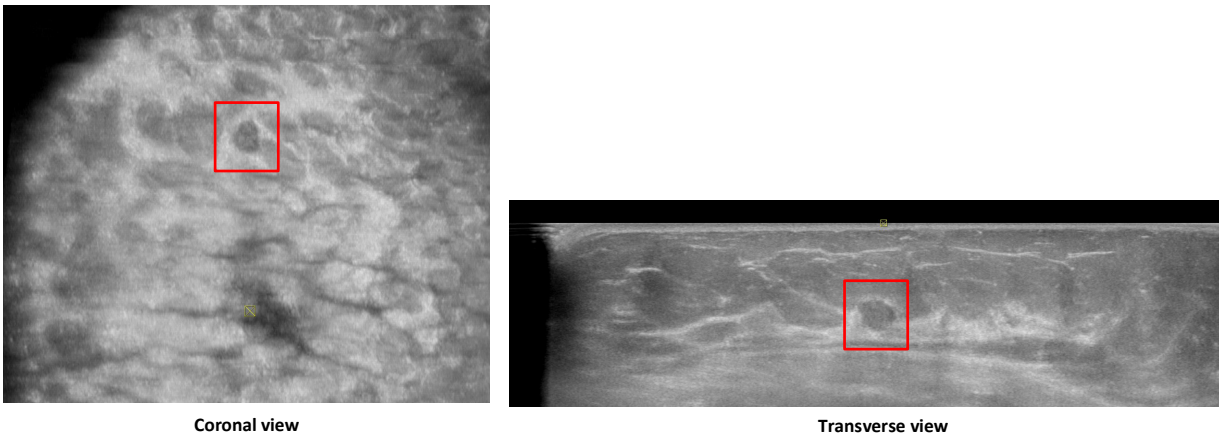


Figure 4.1: Examples of automated breast ultrasound images acquired during screening in a 50-y-old woman. A benign lesion located in both coronal and transverse views is enclosed by red rectangular boxes.

The usefulness of a computer-aided diagnosis (CADx) system as a second reviewer to support radiologists in decision making has also been reported. Furthermore, the screening time, especially for interobserver variation, is reduced [86–88]. To capture unique characteristics of lesion patches, machine learning-based feature extractors are commonly applied in CADx systems. The common feature extractors include histogram of oriented gradients (HOG) [89], local binary pattern (LBP) [90], and principal component analysis

(PCA) [91]. As an alternative, morphological operations are also used to extract breast lesion features [28, 92]. After feature extraction, a classifier is followed, such as support vector machine (SVM) [28, 54], k-nearest neighbor (KNN) [93], logistic regression [27] and linear discriminant analysis (LDA) [94, 95]. However, both machine learning-based and morphologic feature extraction schemes consist of several manual processing steps. For this reason, it is complicated to select optimal feature combinations, which require numerous trials and errors before a convincing result is obtained at each intermediate step [96, 97].

The convolutional neural network (CNN), one of the deep learning approaches, has achieved promising results in natural image classification tasks [98]. Instead of using the aforementioned feature extractors, the CNN extracts imaging features directly from input data without designing explicit feature extractors or tuning intermediate results at each intermediate step. In recent years, CNN has been effectively applied to the field of breast cancer classification [96, 99, 100]. In particular, transfer learning is a good solution for a CNN-based CADx system if the size of the database is limited. With transfer learning, the pre-trained models, explicitly trained by a large-scale natural image database such as ImageNet [98], are tuned to be applied for solving classification problems in the medical imaging domain. In [96], a pre-trained GoogLeNet [52] was adopted for discrimination between malignant and benign lesions. A pre-trained VGG [64] was used for lesion feature extraction [99], followed by the Fisher discriminant analysis to classify breast lesion. In the work of [100], the Inception-v3 [8] was used to classify breast lesions in US imaging.

Recently, multi-view CNN has been introduced and found to improved breast lesion classification in mammography [101]. In this method, mammography visualizes the same breast lesion in different mammographic views such as cranial-caudal and medio-lateral oblique views. To provide more useful features, multiple CNNs were used to extract features from different views independently. In terms of US imaging, the multiview strategy has also been applied to CNN by Han et al. [96]. For each breast lesion, multiple lesion patches are cropped by multiple scales. Through combination of lesion patches with multiple scales, classification performance was improved compared with that using a lesion patch with a single scale. Nevertheless, there are some limitations to the work of Han et al. [96]. First, the sizes of the lesions are required before generating the multiview lesion patches, which involves manual

intervention. Second, extraction of features from multiple scaled lesion patches may be less generic compared with that in [101], in which the overall extracted features are redundant.

In this article, a multiview CNN is proposed to classify breast lesions between malignant and benign. Without any manual pre-processing step, the proposed CNN extracts features from the lesion patch directly. To provide an efficient feature extraction scheme, more lesion features are extracted by the proposed CNN from different views of the automated breast ultrasound (ABUS) imaging. On the basis of our results, the proposed multi-view CNN outperformed conventional machine learning approaches and single-view CNNs. Furthermore, an observer performance test was conducted. The test results revealed that the proposed CNN could be used as a second reviewer to improve human reviewers' diagnostic performance in breast cancer classification. The main contributions of the proposed CNN are as follows:

- The proposed CNN is the first CNN model that has been successfully applied to breast cancer classification in ABUS imaging, to the best of our knowledge
- Multiview strategies have been utilized in the proposed CNN to provide robust and effective feature extraction.
- Comprehensive evaluations are performed to ensure the effectiveness of the proposed CNN model.
- An observer performance test is performed to justify the usefulness of the proposed CNN model from the clinical perspective.

4.2 Methods

4.2.1 Clinical data set

The data set used in this study was collected between March 2012 and March 2018 at Jeonbuk National University Hospital (JNUH). An ACUSON S2000 (Siemens Medical Solutions, Mountain View, CA, USA) automated breast volume scanner (ABVS) in combination with

a 15-cm-wide linear array transducer was used to acquire ABUS images by a single technologist who had more than 3 y of experience in operating ABVS. Depending on breast size, acquisition frequencies vary from 9 11 MHz. Each ABUS scan produces 15.4 cm × 16.8 cm × maximum 6 cm volume data with a slice thickness of 1 mm. The volume data obtained were reconstructed to 2-D slice images through multiplanar reconstruction. For our data set, slice images of each breast lesion were collected from both coronal and transverse views. For our retrospective study, the informed consent for data usage was approved by the institutional review board of JNUH.

Table 4.1: Distribution of the number of lesions in size.

Size (mm)	Malignant (n=135)	Benign (n=181)
1-5	4 (2.96%)	32 (17.68%)
5-10	37 (27.41%)	91 (50.28%)
10-20	94 (69.63%)	58 (30.04%)
Mean±SD	13.23±4.29	9.28±3.99

SD = standard deviation

A total of 316 breast lesions in 263 patients (ages: 28-76 years; mean age: 51.4±9.8 years) were included in our data set, which consists of 135 malignant and 181 benign lesions. Mean lesion size was 13.23 mm with a standard deviation of 4.29 mm. A detailed distribution of the number of lesions by size is provided in Table 4.1. All lesions were pathologically confirmed after surgery or biopsy. For each benign lesion, a 2-y follow-up examination was done to ensure the lesion was unchanged. The ground truth of each lesion was annotated by the physicians using the bounding box to cover the lesion. After annotation, the ground truth of each lesion was verified by a radiologist with 8 y of experience in US imaging. The size of the bounding box varied with the size of the lesion. The sample of bounding boxes is illustrated in Figure 4.2. Lesion patches were cropped along the bounding boxes, and the proposed CNN extracted the lesion features from the cropped lesion patches. To generate more effective lesion patches, multiple lesion patches were cropped from different slices of each lesion. For example, as illustrated in Figure 4.2, the same lesion at different slices has different visual representations. Thereafter, 743 malignant patches (359 and 384 patches from coronal and

transverse views, respectively) and 419 benign patches (233 and 186 patches from coronal and transverse views, respectively) were generated in this study.

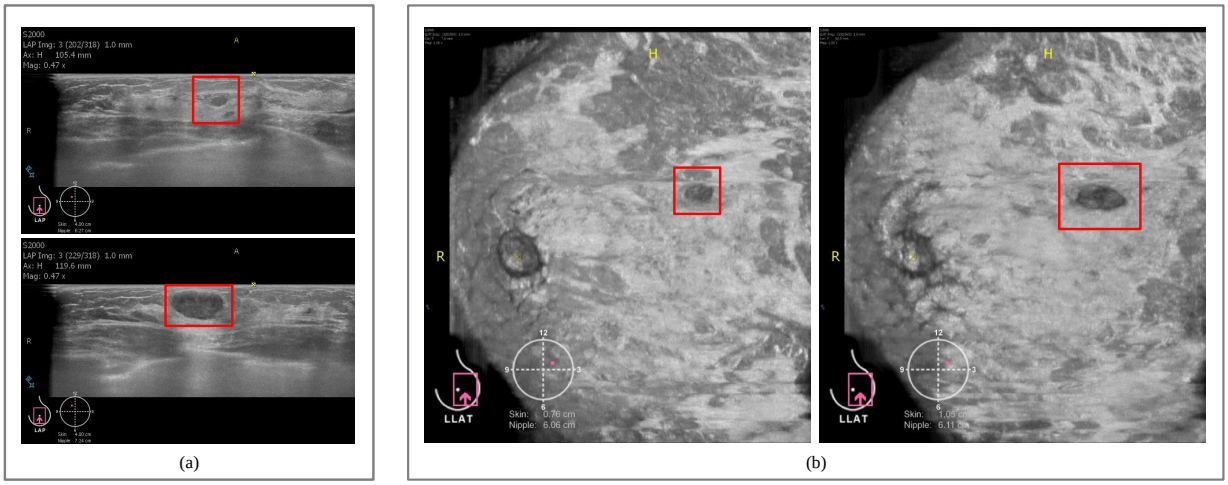


Figure 4.2: Lesion patches around bounding boxes in red. (a) Two lesion patches obtained from two slices of the same benign lesion in transverse view. (b) Two lesion patches obtained from two slices of the same benign lesion in coronal view.

4.2.2 CNN-based lesion feature extraction and classification

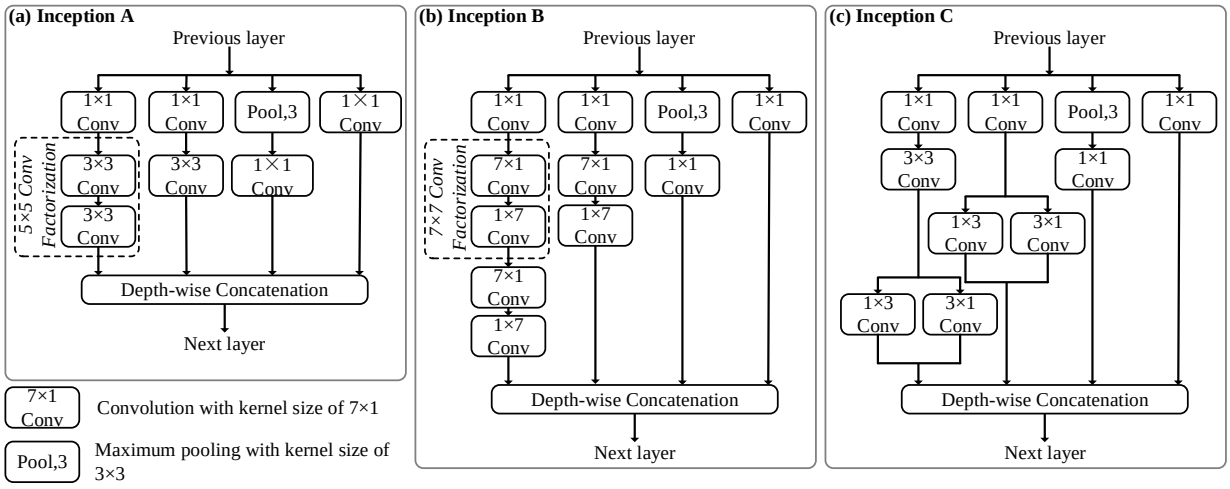


Figure 4.3: Architectures of (a) Inception module A, (b) Inception module B and (c) Inception module C.

Lesion features were extracted and classified by employing a modified Inception-v3 CNN [8], the third generation of GoogLeNet [52]. One novel aspect of Inception-v3 is the Inception

modules used to replace the conventional convolutional layer. The Inception module uses multiple conventional convolutional layers for feature extraction and concatenates extracted features as the output. The main difference between the Inception module and a conventional convolutional layer is that the Inception module allows extraction of features with different kernel sizes. Therefore, the extracted features are not limited to the fixed-scale local regions. Intuitively, local regions of varying sizes are covered. With respect to the usefulness of the Inception module in lesion feature extraction, various kernel sizes help the model to generalize lesions of different sizes effectively. In terms of the architecture of the Inception module, there are three types of Inception modules: Inception A, Inception B and Inception C. Inception A (Fig. 4.3a) is equivalent to the inception module used in GoogLeNet [52]; however, the 5×5 convolution is factorized to two 3×3 convolutions. For Inception B (Fig. 4.3b), the 7×7 convolution is factorized to two asymmetric convolutions with kernel sizes of 1×7 and 7×1.

With factorization, the total number of learnable parameters is reduced; on the other hand, it reduces the risk of overfitting. Inception C (Fig. 4.3c) adopts multiple kernels of different sizes to promote high-dimensional representations [8]. The kernel sizes of Inception C are 1×1, 1×3, 3×1 and 3×3.

To retain the powerful feature extraction of Inception-v3 from natural images to ABUS images, we have adopted transfer learning, which has been widely applied in medical imaging analysis [25, 102, 103]. With transfer learning, the CNN model pre-trained from a very large scale database, such as ImageNet [98],), can efficiently apply the pre-trained knowledge to the specific task. By re-training the pre-trained model with a small amount of data, the retrained model can achieve a promising result on the specific task domain. In our case, all fully-connected (FC) layers proposed in Inception-v3 were redesigned, leaving the convolutional structure as the backbone for lesion feature extraction. The architecture of the backbone is illustrated in Figure 4.4a. The input size of the backbone is 299×299×3. The first several layers of the backbone consist of six convolutional layers with kernel sizes of 3×3 and an average pooling layer with a kernel size of 3×3, followed by five Inception A, four Inception B, and two Inception C modules. The backbone outputs 2048 feature maps, and each feature map has a size of 8×8. A global average pooling layer is added on the top of the backbone's

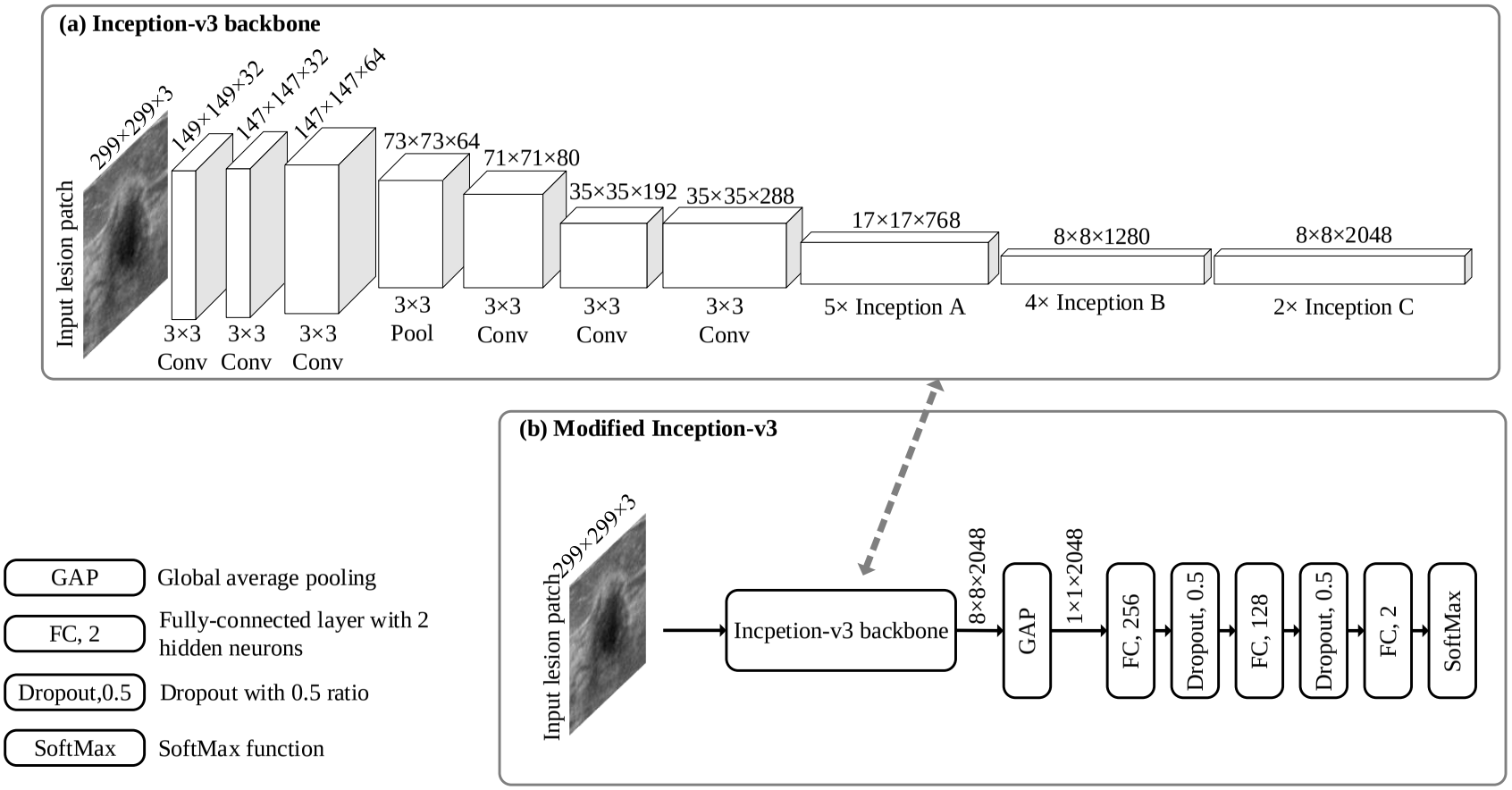


Figure 4.4: Architectures of (a) Inception-v3 backbone and (b) modified Inception-v3 convolutional neural network (CNN).

output feature maps, each of which is averaged to a single vector. The global average pooling transforms the features that can be more robust for interpretation as categories [104]. Then, three FC layers are added to bridge the convolutional features with a neural network classifier. The numbers of hidden neurons in the three FC layers are 256, 128 and 2. The dropout scheme is applied after the first and the second FC layers to relieve overfitting. Thereafter, the output of the last FC layer is normalized by the SoftMax function. The modified version of Inception-v3 is illustrated in Figure 4.4b. The modified version of Inception-v3 takes lesion patch as the input; then, the lesion features are extracted via the Inception-v3 backbone. Finally, it outputs the probabilities of malignant and benign.

4.2.3 Multiview CNNs

Two multiview CNNs, as illustrated in Figure 5, have been explored in this study. For the multiview CNN A (Fig. 4.5a), two lesion patches are cropped from the same lesion over transverse and coronal views independently. Then, the multiview lesion patch is generated by concatenating the cropped lesion patches into different image layers. For example, one breast lesion generates one lesion patch from the coronal view (CA) and two lesion patches from the transverse view (TA and TB). As a result, there are two multiview lesion patches with dual image layers combined, which are (CA-TA, CA-TB). By feeding the multiview lesion patch into the modified Inception-v3 model (Fig. 4.4b), the type of the lesion is classified. As the acquired ABUS images use gray-scale reconstruction, one multiview lesion patch requires two image layers to encode both transverse and coronal views. However, the modified Inception-v3 model requires three input channels because the pre-trained model is trained by RGB images. To fulfill the input requirement of the modified Inception-v3 model, the lesion patch obtained from the transverse view is used as the third image layer. For each combination of dual image layers, all lesion patches obtained from the transverse view are used. In the aforementioned example, TA and TB are attached to generate two dual-layered multiview patches, and thus four three-layered multiview patches can be generated (CA-TA-TA, CA-TA-TB, CA-TB-TA, CA-TB- TB). Therefore, more effective multiview lesion patches are generated and used for training purposes. In total, we obtained 3085 multiview patches (1767 malignant and 1318 benign) that can be used to train the network.

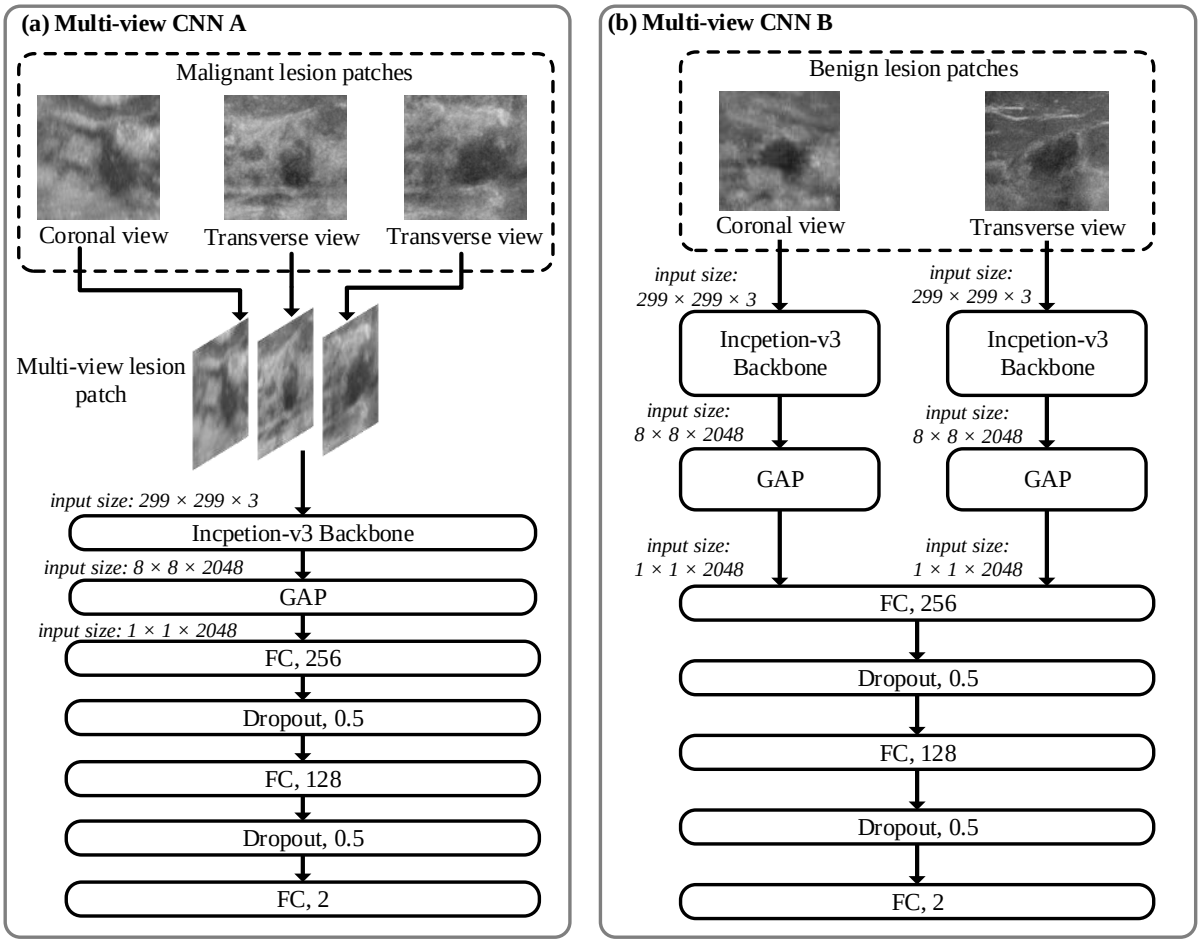


Figure 4.5: Architectures of proposed multiview convolutional neural networks (CNNs). FC = fully connected layer; GAP = global average pooling.

Instead of extracting lesion features from the multi-view lesion patch, the multiview CNN B (Fig. 4.5b) adopts two Inception-v3 backbones to enable multiview learning. Each Inception-v3 backbone corresponds to extract lesion features from a certain view. Then, the extracted features from the two Inception-v3 backbones are concatenated on top of the first FC layer. To match the input channels of the Inception-v3 backbone, the input lesion patch is transformed to three channels by duplicating the pixel value of the lesion patch. Thereafter, we obtained 1525 lesion samples (549 malignant and 468 benign) to train the multiview CNN B.

4.2.4 Network training and evaluation

A five-folder cross-validation was used to train, validate and test the multiview CNNs. Specifically, each folder contains 20% of the breast lesions in our ABUS data set, including 31 malignant lesions and 33 benign lesions. In each round of cross-validation, one folder is reserved as the testing data set. The remaining four folders are split into training and validation data sets; three of four folders belong to the training data set and the remaining folder is used as the validation data set. After each epoch in training, the trained model is evaluated by the validation data set. To select the final trained model, an early stopping strategy is applied based on the classification performance on the validation data set. Practically, the final trained model is selected when the area under the curve (AUC) value does not increase for five epochs. Then, the final trained model is evaluated by the testing data set. Thereafter, the classification performance for each round of cross-validation is obtained. In this study, the reported results were obtained by averaging the classification performance across five rounds of cross-validation, including sensitivities, specificities and AUC values.

To enhance the learning process within limited training samples, the training data set is augmented. For multiview CNN A, the multiview lesion patches used for training are rotated three times with rotation angles of 90°, 180° and 270°, followed by flipping operation horizontally and vertically. For multiview CNN B, the same augmentation method is applied to the lesion patches obtained from transverse and coronal views.

The weights of the FC layers in multiview CNNs are initialized by following the Xavier uniform initializer [47]. In addition, the weights of the Inception-v3 backbone are initialized by applying pre-trained weights optimized for ImageNet database. For transfer learning, we followed the approach of [105], in which the entire layers undergo fine-tuning during the training phase. Furthermore, the multi-view CNNs are trained with a batch size of 32. The losses are optimized by Adadelta with an adaptive learning rate [50]. The CNNs were implemented with Keras and were trained by using an Nvidia P6000 GPU on an Ubuntu 18.04 system.

4.3 Results

4.3.1 Classification performance of multiview CNNs

To select the effective multiview strategies as described under Multiview CNNs, the classification performance of the multiview CNNs with two different configurations are compared. Multiview CNN A achieved a sensitivity of 0.886, specificity of 0.876 and mean AUC value of 0.9468 with a standard deviation of 0.0164. Compared with multiview CNN A, multiview CNN B had roughly 2% reduced sensitivity (0.865) and specificity (0.848), and the mean AUC value dropped to 0.9346 with a standard deviation of 0.0095. The mean receiver operating characteristic curves for the multiview CNNs are illustrated in Figure 4.6.

Table 4.2: Comparison of multi-view CNNs with single-view CNNs.

Single-view		Multi-view		Sensitivity	Specificity	Area under curve*
Coronal	Transverse	A	B			
✓				0.831	0.800	0.8874±0.0054
	✓			0.832	0.838	0.9076±0.0233
		✓		**0.886**	**0.876**	**0.9468±0.0164**
			✓	0.865	0.848	0.9346±0.0095

CNN = convolutional neural network.

Significance of bold values indicate the best performance among all compared methods.

* Mean ± standard deviation.

The classification performance of the multiview CNNs were compared with that of the single-view CNNs. The single-view CNNs followed the architecture of the multiview CNN A; however, the inputs of the single-view CNNs take the lesion patches obtained from either a transverse view or a coronal view. To match the input size of the modified Inception-v3, the lesion patches were encoded by duplicating the gray-scale intensities. Table 4.2 summarized the classification performance of multiview and single-view CNNs. The single-view CNN using the coronal view achieved a sensitivity of 0.831, a specificity of 0.800 and a mean AUC

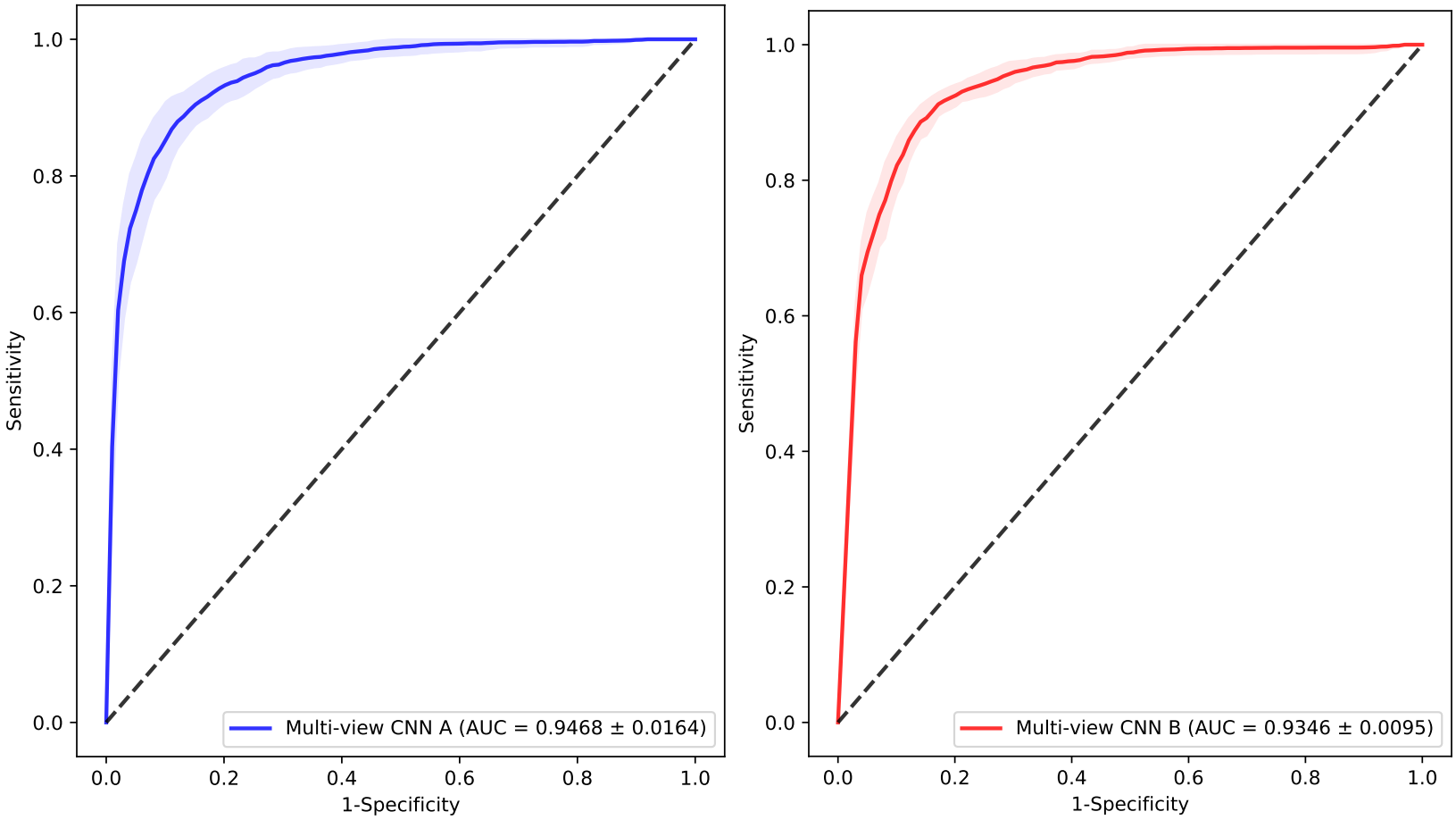


Figure 4.6: (a) Mean receiver operating characteristic (ROC) curve of multiview convolutional neural network (CNN A) (area under the ROC curve [AUC] = 0.9468 with standard deviation of 0.0164). (b) Mean ROC curve of multiview CNN B (AUC value = 0.9346 with standard deviation of 0.0095). The shadow of each ROC curve illustrates the variance of the ROC during five-folder cross-validation.

value of 0.8874 with a standard deviation of 0.0054. By employing the transverse view only, the single-view CNN's classification performance increased (sensitivity = 0.831, specificity = 0.800 and mean AUC = 0.8874±0.0233) compared with that using the coronal view, but it still could not reach the classification performance of the multiview CNNs.

Table 4.3: Classification performance of multi-view CNN A using different backbones.

Backbone	Pre-train weights	Sensitivity	Specificity	Area under curve*
ResNet	✓	0.809	0.830	0.9045±0.0153
DenseNet	✓	0.847	0.848	0.9221±0.0206
Inception-v4	✓	0.805	0.823	0.8851±0.0096
Inception-ResNet-v2	✓	0.866	0.873	0.9303±0.0156
Inception-v3	✗	0.822	0.859	0.9043±0.0173
Inception-v3	✓	**0.886**	**0.876**	**0.9468±0.0164**

CNN = convolutional neural network.

Significance of bold values indicate the best performance among all compared methods.

* Mean ± standard deviation.

Furthermore, the effectiveness of lesion feature extraction is justified by adopting different backbones. Specifically, the proposed backbone was replaced with the convolution structures of ResNet [9], DenseNet [54], Inception-v4 [53] and Inception-ResNet-v2 [53]. Table 4.3 summarizes the classification performance of multiview CNN A employing different backbones. Based on the results, the multiview CNN A with the Inception-v3 backbone outperformed those with ResNet, DenseNet, Inception-v4 or Inception-ResNet-v2 backbones. In addition, the effectiveness of the Inception-v3 backbone with the ImageNet pre-train weights was explored by training the multiview CNN A from scratch. To train the model from scratch, the weights of all convolutional layers were initialized by following the Xavier uniform initializer, and the biases were initialized to zeros. As outlined in Table 3, the multiview CNN A trained from scratch degraded classification performance compared with that using ImageNet pre-train weights.

4.3.2 Comparison with conventional machine learning feature extractors

In this section, the classification performance of multiview CNN A is compared with that of conventional machine learning feature extractors, including PCA and HOG, which both revealed the usefulness of lesion feature extractions in US imaging [89, 91]. To set up the experiment for PCA, we followed the pre-processing steps as described in [91]. First, the multi-view lesion patches were resized to 128×128×3; this was followed by histogram equalization for contrast enhancement. Furthermore, we added an additional pre-processing step, local contrast normalization, as described in [106]. Such normalization can improve the learning process of the classifier. After the pre-processing steps, PCA was applied to extract the lesion features. An exhaustive search was performed to find the optimal number of principal components. By following the same pre-processing steps used for PCA, the lesion features were also extracted by employing HOG. The optimal parameters of HOG were obtained through an exhaustive search on different numbers of orientation bins and pixels per block as suggested in [89], which are {8, 16} and {3×3, 5×5}, respectively. Thereafter, we used a SVM with a radial basis function kernel to classify extracted lesion features. To obtain optimal classification performance, the penalty parameter (C) was determined by performing an exhaustive search in (0.001, 0.01, 0.1, 1, 10, 100, 1000), and the possible parameters for the kernel coefficient were (0.0001, 0.001, 0.1, 1 ,10). The SVM was trained and evaluated by five-folder cross-validation with the same data partitioning used for multiview CNN A. The PCA and SVM were implemented with Scikit-learn [107], and the HOG was implemented with Scipy [108].

As outlined in Table 4.4, the PCA with SVM had inferior classification performance, a sensitivity of 0.719, a specificity of 0.764 and a mean AUC value of 0.8195 with a standard deviation of 0.0644. HOG outperformed PCA (sensitivity = 0.782, specificity = 0.771, mean AUC = 0.8537±0.0201), but still could not reach the classification performance of the multiview CNN. Figure 7 illustrates the mean ROCs of the multiview CNN, PCA and HOG approaches.

Table 4.4: Comparison of the multi-view CNN A with conventional machine learning approaches.

Feature extractor	Classifier	Sensitivity	Specificity	Area under curve*
PCA	SVM	0.719	0.764	0.8195±0.0644
HOG	SVM	0.782	0.771	0.8537±0.0201
Multi-view CNN A		**0.886**	**0.876**	**0.9468±0.0164**

CNN = convolutional neural network.

Significance of bold values indicate the best performance among all compared methods.

* Mean ± standard deviation.

4.3.3 Observer performance test

Table 4.5: Results of observer performance test.

	Accuracy		Sensitivity		Specificity	
	Round 1*	Round 2†	Round 1	Round 2	Round 1	Round 2
Special radiologist	0.864	0.883	0.881	0.933	0.851	0.845
Senior resident 1	0.797	0.823	0.867	0.926	0.746	0.746
Senior resident 2	0.842	0.900	0.874	0.926	0.818	0.873
First-year resident	0.816	0.848	0.881	0.941	0.768	0.780
Physician	0.604	0.876	0.874	0.911	0.403	0.851
Multi-view CNN A	0.880		0.886		0.876	

* Round 1: independent interpretation.

† Round 2: interpretation with aid of the multi-view CNN A.

To justify the usefulness of the proposed multiview CNNs, an observer performance test was conducted by comparing the diagnostic performance of human reviewers before and after referring to the predicting outcomes of the multiview CNN A. There were five human reviewers, including one breast special radiologist, two senior radiology residents, one first-year radiology resident and one physician. It should be noted that the reviewers who partici-

pated in the observed performance test did not annotate the ground truth of each lesion. Without having the patients' information, the reviewers independently reviewed the ABUS images used in this study. Each lesion was interpreted by reporting the malignancy rating on a 7-point scale (1 = definitely not malignant, 2 = almost definitely not malignant, 3 = probably not malignant, 4 = may be malignant, 5 = probably malignant, 6 = almost definitely malignant and 7 = definitely malignant). After the first round of interpretation, the reviewers were able to change their previous decisions by considering predicted outcomes of the multi-view CNN A. Table 4.5 summarizes the changes in diagnostic performance of the five human reviewers from round 1 to round 2 of the observer performance test. With the aid of the multiview CNN A, all human reviewers improved in diagnostic accuracy (mean accuracy improvement: 0.102; range: 0.083 to 0.170). Also, all human reviewers' diagnoses had increased sensitivity (mean sensitivity improvement: 0.052; range: 0.015 to 0.008). The physician, a non-radiologist, achieved significant improvement in diagnostic specificity (from 0.403 0.851). In contrast, the proposed multiview CNN was found to be less advantageous in improving the diagnostic specificity of the radiologists (mean specificity improvement: 0.015; range: 0.021 to 0.040). Practically, the special radiologist had a reduced diagnostic specificity (from 0.851 0.845). Regardless of the support of the multiview CNN, senior resident 1 exhibited no difference in performance in terms of specificity.

The AUC values of the five human reviewers and the multiview CNN A are illustrated in Figure 4.8. AUC values of four of the five human reviewers improved from round 1 to round 2; the exception was the special radiologist. Specifically, the AUC value of the non-radiologist (physician) was significantly increased by 38% (AUC improvement: 0.245). To evaluate the statistical significance of the changes in AUC changes for each reviewer, we performed the Delong test under 1000 bootstrap samples. For the special radiologist, the AUC value decreased by 0.024 from round 1 to round 2 was decreased; however, the reduced AUC value for the special radiologist was not statistically significant. In turn, the remaining four human reviewers exhibited significant improvement statistically ($p < 0.05$) after referring to the classification results of the multiview CNN A.

4.4 Discussion

In this study, the proposed method employing a modified Inception-v3 CNN with multiview strategies achieved promising classification performance for discrimination of breast lesion patches between malignant and benign. We explored two different multiview CNN architectures (multiview CNNs A and B). Both multiview CNNs outperformed the single-view CNNs using either transverse or coronal lesion patches. The improvement over the single-view CNNs may have occurred because the multi-view strategies enable feature extraction over both transverse and coronal views, which increases the amount of extracted lesion features. Therefore, the features learned in the multiview CNNs had better discriminative power than those learned from the single-view CNNs.

4.4.1 Analysis of multiview CNNs

It should be noted that the concept of the multiview CNN B was first used for mammography [101], and it also exhibited great discriminating performance in ABUS imaging based on our experimental results. However, the multiview CNN A would be preferred for two reasons. First, the multiview strategy applied in multiview CNN B requires two backbones; thus, the increased model complexity is not easily fine-tuned. Second, the inference time of multiview CNN A was faster than that of multi-view CNN B (34.34 ms/lesion vs. 73.79 ms/lesion), which indicated computational efficiency.

The Inception-v3 backbone, using Inception modules for convolution, exhibited improved lesion feature extraction compared with ResNet and DenseNet, which both employ conventional convolutional layers. Our experimental results confirmed that the Inception modules could extract features from lesions of different sizes; however, the conventional convolutional layer extracted fixed-scale features regardless of the size of the lesions. Compared with Inception-based CNNs, the proposed multiview CNN A employing the Inception-v3 backbone outperformed the proposed one using the Inception-v4 or Inception-ResNet- v2 backbone. The effectiveness of the Inception-v3 backbone justifies two aspects: (i) The main difference from Inception-v3 to Inception-v4 or Inception-ResNet-v2 is that Inception-v3 consists of fewer Inception modules. However, based on our experimental results, adding more

Inception modules has no benefit in improving classifica- tion performance in ABUS imaging. (ii) When CNN goes deeper, adding more Inception modules, the performance of the deeper CNN could be affected by the vanishing gradient. To avoid the vanishing gradient, Inception-ResNet-v2 applies a residual connection from the input to the output of each Inception module. In contrast, there is no residual connection in Inception-v4, which could explain why the vanishing gradient may cause further performance degradation.

4.4.2 Comparison with previous works

With five-folder cross-validation of 316 breast lesions, the proposed method employing multiview CNN A achieved a mean AUC value of 0.9468. In previous studies using conventional machine learning feature extraction schemes, Moon et al. [27] evaluated 147 breast lesions and reported the best AUC value of 0.9466, and Tan et al. [28] reported the best AUC value of 0.93 based on 201 breast lesions. However, we cannot make a direct comparison. Because we use a different private database, it is impossible to estimate the complexity of each database based on the number of lesions. Nevertheless, the previous studies could be less generic than the proposed method because the performance of the previous methods could be affected by the quality of the input data, in which the lesion has to be segmented precisely. In contrast, the proposed method takes a raw lesion patch as input.

As mentioned earlier, this is the first study using CNNs for breast cancer classification in ABUS imaging. For US imaging, some previous studies explored the usefulness of CNNs. One previous study [99] used pre-trained VGG and obtained an AUC value of 0.847 in evaluating 100 breast lesions. In a previous method used on 251 breast lesions [109], the AUC value was 0.857 by employing pre-trained Inception-v3.

Because the database is different, we still cannot make a direct comparison. However, considering the AUCs reported from the two previous studies, the proposed method using the multiview strategy could perform better without sacrificing efficiency of computation. Because the multiview strategy was applied to the input of the CNN, there was no additional cost of computation compared with previous studies. Furthermore, one previous study adopted a multiview strategy to classify 829 breast lesions and achieved a best AUC of 0.9601 [96]. Similar to the multiview CNN A, the multiview patches were generated by

combining multiple lesion patches with different scales. However, to generate these patches, radiologists must measure the distance between the boundary of the lesion and the boundary of the patch itself. When the amount of cases is large, this is a significant time-consuming task. In contrast, the proposed method directly crops the lesion patch from ABUS imaging; thus, it is expected to reduce the workflow of the radiologist. On the other hand, on comparison of AUCs, the best AUC value we obtained was 0.9704, which is similar to that in the previous study. Moreover, similar to the multiview CNN B, a previous study by Xiao et al. combined lesion features extracted from three different CNNs, including an Inception-v3, an Xception and a ResNet [110]. By evaluating 2058 breast lesions, the previous method achieved the best AUC of 0.93. Nevertheless, using multiple CNNs for feature extraction may produce redundant features as all CNNs used the same lesion patch as the input. In turn, the proposed method generates more features by performing feature extraction over lesion patches of different views.

4.4.3 Analysis of observer performance test

In our observer performance test, the results confirmed that human reviewer increases diagnostic accuracy after referring to the proposed CNN. In addition, our results suggest that the proposed CNN is more effective in improving the diagnostic performance of reviewer compared with a reviewer who has more experience in diagnostic ABUS imaging. In the real clinical situation, follow-up medical examination or treatment is required before interpretation. Therefore, the proposed CNN could be used as a second reviewer to shorten the follow-up interval. Furthermore, our results confirmed that the proposed method could help reviewers make the correct diagnostic decision when there is a difference between transverse and coronal views. For example, as illustrated in Figure 4.9a, four of the five reviewers decided to interpret the malignant lesion as benign because the lesion shape on the transverse view appeared benign, and reviewers can see the suspicious shape of the lesion on the coronal view only. Another example is illustrated in Figure 4.9b. A benign lesion was misclassified by four of the five reviewers because the lesion shape on the transverse view was a suspicious finding, and the lesion shape appeared benign on the coronal view. In fact, ABUS stores the volume data digitally and reconstructs multiplanar images, which allows radiologists to simultane-

ously evaluate breast lesions using reconstructed coronal planes as well as transverse planes. Previous studies have reported that retraction phenomenon, an especially useful feature of coronal planes, had high diagnostic accuracy for breast malignancy [111, 112]. However, the reviewers in this study were less familiar with coronal planes than transverse planes because ABUS has become popular only recently. In contrast, the proposed CNN evaluated breast lesions with transverse and coronal planes equally.

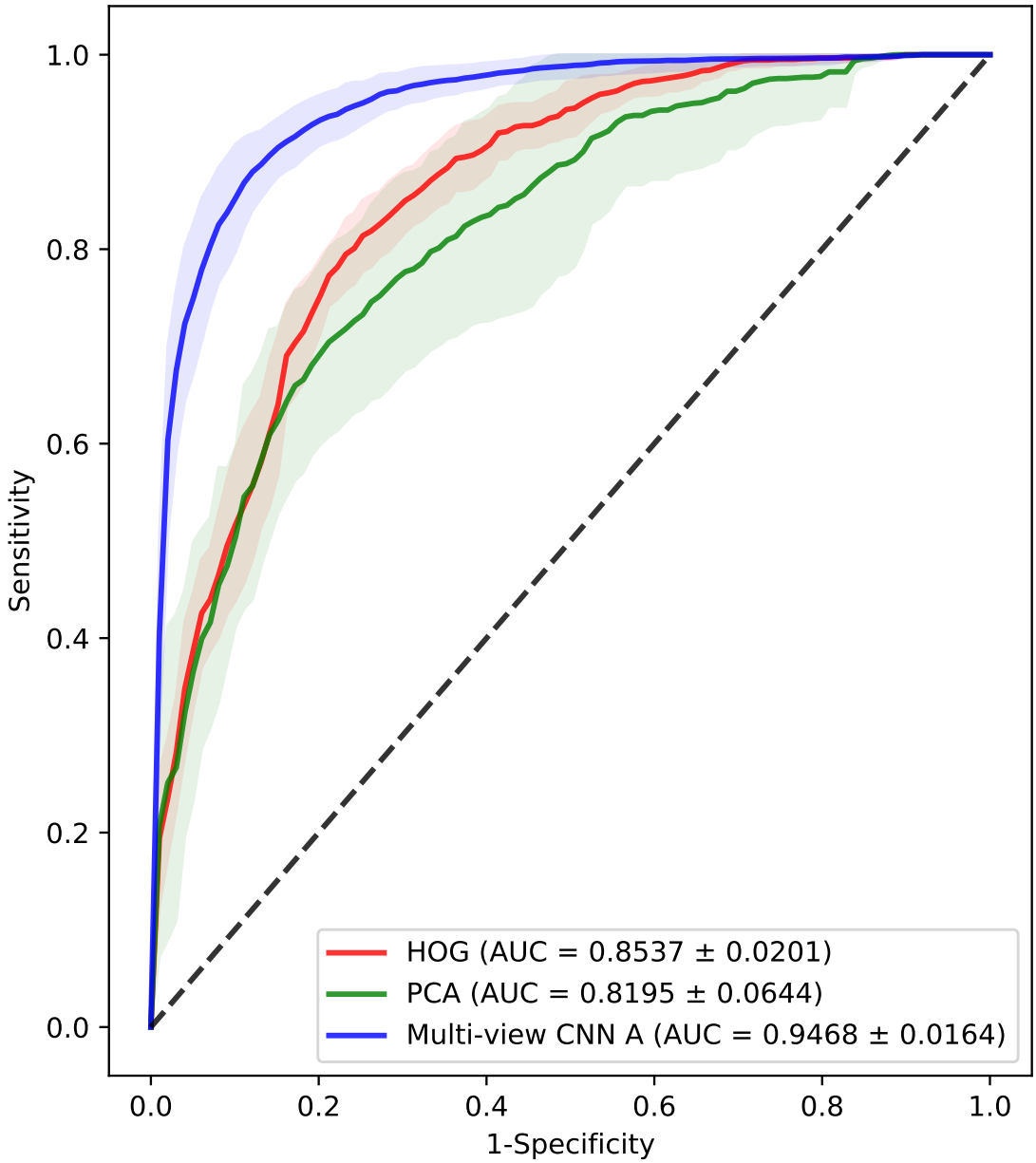


Figure 4.7: Mean receiver operating characteristic (ROC) curves of histogram of oriented gradients (HOG) with support vector machine (SVM), principal component analysis (PCA) with SVM and multiview convolutional neural network (CNN) A. The shadow of each ROC curve illustrates the variance of the ROC curve during five-folder cross-validation.

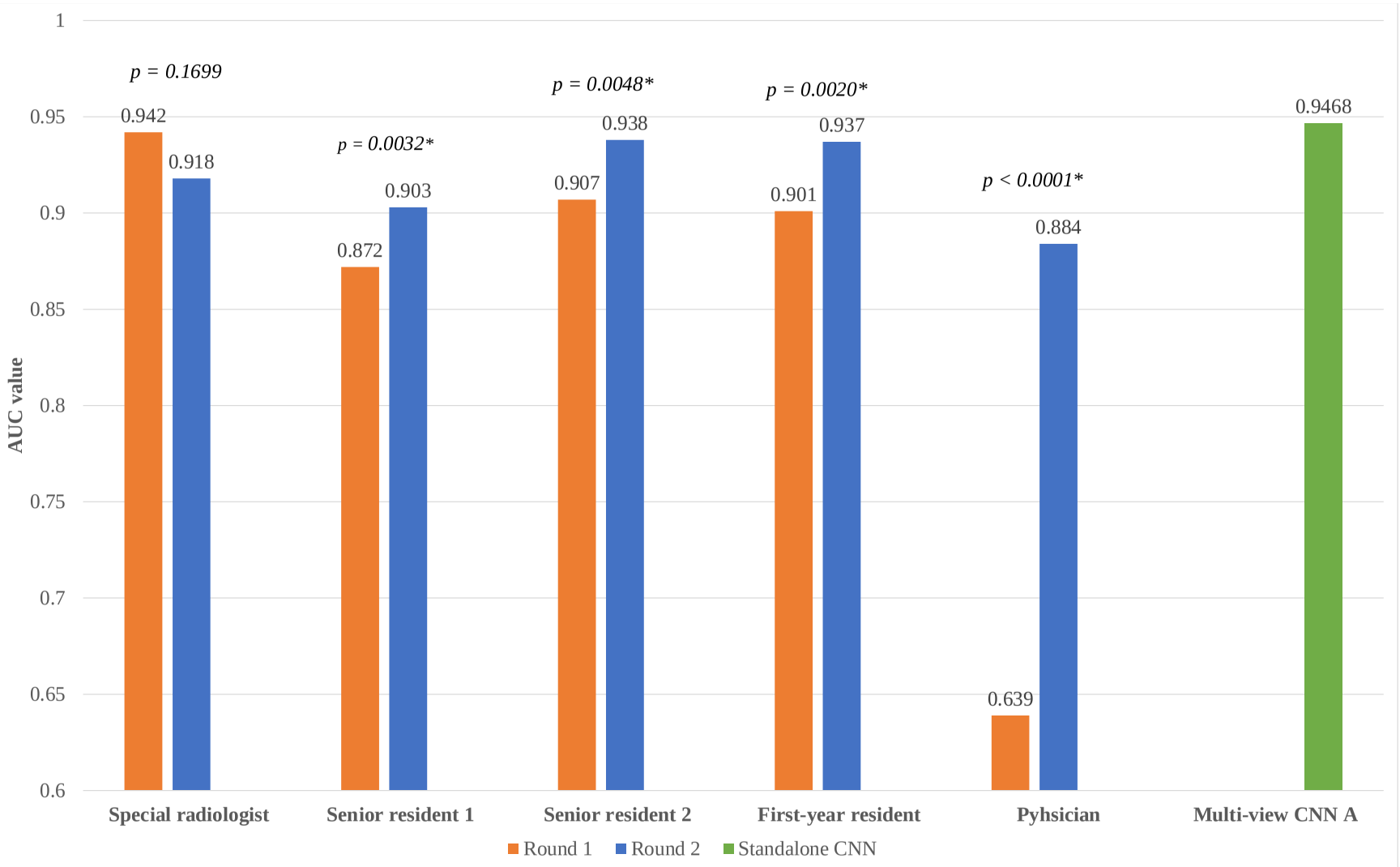
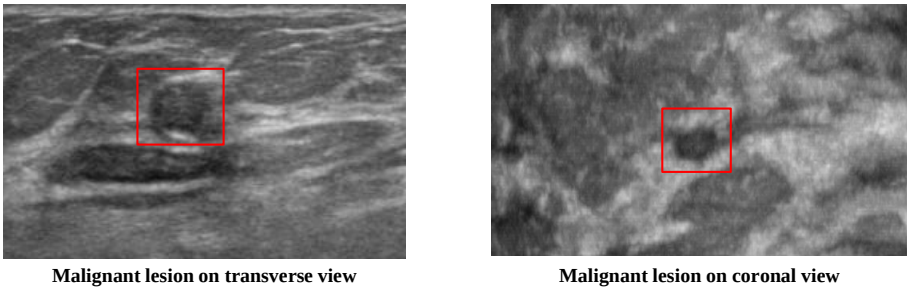
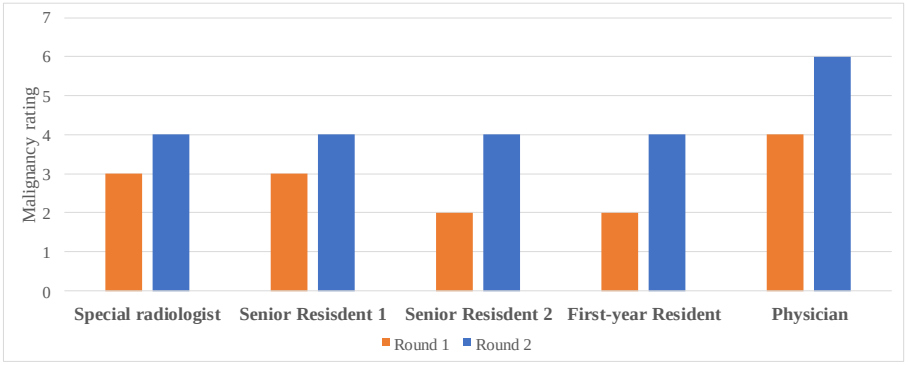


Figure 4.8: Changes in the five human reviewers' areas under the curve (AUCs) in the observer performance test. Round 1: independent interpretation. Round 2: interpretation with aid of the multi-view convolutional neural network (CNN) A. *Significant difference in AUCs between rounds 1 and 2.

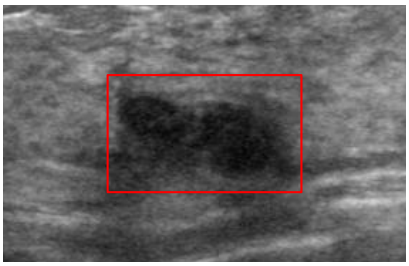


Malignant lesion on transverse view

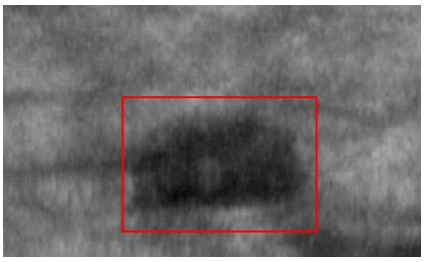
Malignant lesion on coronal view



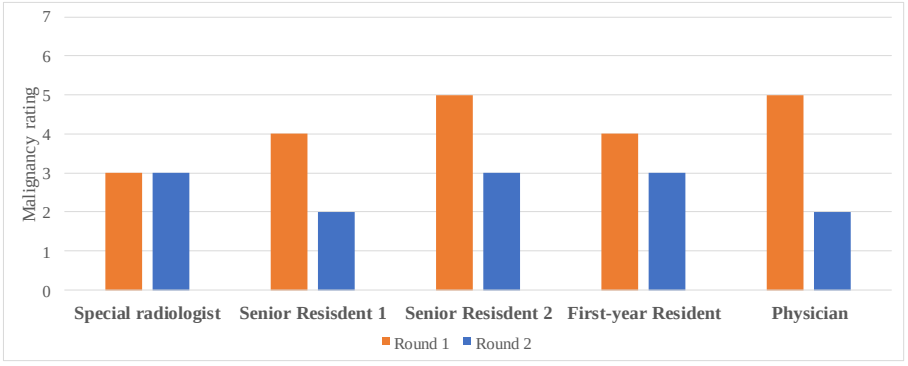
(a)



Benign lesion on transverse view



Benign lesion on coronal view



(b)

Figure 4.9: Samples of decision changed by the five human reviewers. Round 1, independent interpretation. Round 2, interpretation with aid of the multi-view CNN A. (a) A malignant lesion was classified as malignant by the multi-view CNN A with a malignancy rating of 7. (b) A benign lesion was classified as benign by the multi-view CNN A with a malignancy rating of 1.

Chapter 5

Fully Convolutional Neural Network Based Computer-aided Detection System for Anterior Mediastinal Nodular Lesion Segmentation from Chest Computed Tomography[1]

This chapter presents a fully convolutional neural network (FCN) based computer-aided detection system utilized to segment anterior mediastinal nodular lesions (AMLs) from chest computed tomography (CT) imaging. The general architecture of the proposed network follows UNet, the most widely used segmentation network for medical imaging. To improve the feature extraction power of the proposed network, different attention mechanism are utilized.

Section 5.1 describes the motivations for designing a CNN-based CADe system for AML segmentation from chest CT imaging. Section 5.2 presents the characteristics of the AML dataset used in this study and the architecture of the proposed network. The experimental results are provided in Section 5.3. Section 5.4 presents a detailed analysis of the proposed network.

[1] The major portion of this chapter is originally submitted as "Anterior Mediastinal Nodular Lesion Segmentation from Chest Computed Tomography Imaging Using UNet based Neural Network with Attention Mechanisms" in Journal of Academic Radiology.

Yi Wang (YW), Won Gi Jeong (WJ) and Seok-Bum Ko (SK) made the conception and design of the study. YW developed and optimized the network architecture, wrote the code of the network, and performed result analysis. WJ and Gong Yong Jin (GYJ) prepared the data. Hao Zhang (HZ) and Younhee Choi (YC) provided suggestions to improve the network architecture. YW drafted the manuscript. SK provided suggestions on improving the manuscript structure.

5.1 Introduction

Chest computed tomography (CT) scans are rapidly increasing, which screening purpose such as lung cancer screening account for one of the main reasons as a significant mortality reduction was reported to be achievable with lung cancer screening [113]. Therefore, the incidence of incidentally observed anterior mediastinal nodular lesions (AMLs) is also increasing though the prevalence is extremely rare [29]. In general, chest CT scans performed for screening purposes including lung cancer screening are performed as unenhanced CT [29]. AMLs are often overlooked by radiologists interpreting unenhanced chest CT. The reasons are as follows: 1) A significant number of AMLs show similar attenuation to adjacent structures. 2) Radiologists generally focus on pulmonary findings, especially in the case of lung cancer screening. 3) The majority of incidental AMLs are small in size. Approximately 80% of incidental AMLs were less than 2 cm according to previous report [29]. Nonetheless, detecting AMLs on CT is important as it is the first step to guide subsequent management [32]. At a time when radiologists' burnout is becoming a issue [114], it would be of great help if automatic detection for AMLs is possible.

Computer-aided detection (CADe) systems have been introduced to assist radiologists during the reading process. By marking suspicious regions inside computed tomography (CT) imaging, the CADe systems make screening more cost-effective [115, 116]. One fruitful direction to realize a CADe system is to follow segmentation procedures. By highlighting abnormalities of CT imaging at the pixel level, segmentation procedures are effective to delineate contours of lesions, which are the important determinants to evaluate the malignancy of the lesions [33]. Early techniques such as region growing [117], graph cut [118] and the activate contour-based method [119] are often used to perform lesion segmentation from chest CT imaging. However, those approaches rely on handcrafted features which require several manual processing steps to obtain the final results. For this reason, it is not feasible to realize an automatic segmentation procedure.

Fully convolutional neural networks (FCNs) [120], derived from deep learning (DL) methods, have lately become widespread in medical imaging segmentation. Compared to handcrafted feature-based approaches, FCNs automatically extract and learn features directly

from the input image without any manual intervention. More recently, UNet [10], a variant of FCNs, has been introduced. The UNet is a U-shaped convolutional network, which consists of symmetric encoders and decoders. Based on the UNet architecture, numerous success has been achieved in pulmonary nodule segmentation from chest CT imaging [34,35,121] and mediastinal lymph node segmentation from chest CT imaging [36, 37]. Nevertheless, there was only one existing study that has explored the feasibility of applying the DL-based method, particularly for AML segmentation from chest CT imaging. Specifically, Huang et al. [38] introduced a two-staged 3D ResUNet to detect and segment AMLs from chest CT imaging.

Attention mechanisms have emerged to improve the performance of a neural network by allowing it to focus on the most important parts of the input. Many efforts have been made to integrate attention mechanisms into the original UNet architecture. For example, AttentionUNet [122] has been introduced and showed improved segmentation performance over the original UNet in organ segmentation. By adding the attention mechanism into the original UNet decoders, the AttentionUNet is effective to identify small objects of interest or the ones which are difficult to distinguish from the background. Lately, inspired by the self-attention mechanisms introduced in Transformer [123], TransUNet [124] has been introduced and surpassed the segmentation performance of the original UNet in a wide range of medical applications. By adding transformer blocks into the original UNet encoders, the TransUNet allows the network to capture global information, which is effective to describe long-range dependencies of the input [125].

In this study, to assist radiologists with AML diagnosis, a modified UNet architecture is proposed to segment AMLs from chest CT imaging. Without any pre-processing or post-processing step, the proposed network takes a 2D slice image as input and predicts the masks of the AMLs. To provide a robust feature extraction, attention mechanisms are utilized in the proposed network. Based on our experimental results, the proposed network outperformed the state-of-the-art segmentation networks, including UNet, ResUNet [126], AttentionUNet, TransUNet and UNet++ [127]. The main contributions of this study are as follows:

- A novel UNet-based neural network architecture is proposed. A self-attention mechanism is established to enhance extracted features from the encoders of the proposed network. Besides, an image-grid attention mechanism, convolutional block attention

module (CBAM), is incorporated with the decoders of the proposed network to allow the proposed network to focus on reconstructing lesion semantics.

- A multi-path feature extraction scheme is utilized to preserve both discriminative features extracted from the self-attention mechanism and the convolution operations.
- This is very first study to investigate the feasibility of applying DL techniques to segment AMLs from 2D CT slice image. Comprehensive evaluations are performed to justify the effectiveness of the proposed network.

5.2 Materials and methods

5.2.1 Dataset

The dataset used in this study was retrospectively collected between February 2014 and August 2022 at Chonnam National University Hwasun Hospital (CNUHH). For this study, the informed consent of data usage has been approved from the institutional review board of CNUHH. All CT scans were obtained using the following multidetector CT scanners: LightSpeed 16 (GE Healthcare, Chicago, IL, USA), LightSpeed VCT (GE Healthcare, Milwaukee, Wis, USA), Somatom Definition Flash (Siemens Healthineers, Erlangen, Germany), Somatom Definition Edge (Siemens Healthineers, Forchheim, Germany) and Revolution (GE Healthcare, Waukesha, WI, USA). Detailed parameters were as follows: reconstruction thickness of the enhanced CT scan, 2.5 mm; rotation time, 0.5 to 0.8 s; peak kilovoltage, 120 kVp; and tube current, 60–220 mAs, with automatic exposure control. For the Somatom Definition Flash and Somatom Definition Edge scanners, the following parameters were used: reconstruction thickness, 2.5–3.0 mm; rotation time, 0.5 s; peak kilovoltage, 120 kVp; tube current, 60–220 mAs, with automatic exposure control.

We collected the data of thymic cysts and thymomas as they are the most common benign disease and malignancy among incidentally detected AMLs [128]. A total of 180 patient scans were selected, which comprised 90 patients with thymomas and 90 patients with thymic cysts. Thymomas were confirmed on histopathological specimens taken with thymectomy. Of the 90 patients with thymic cysts, 35 patients were diagnosed with magnetic resonance imaging

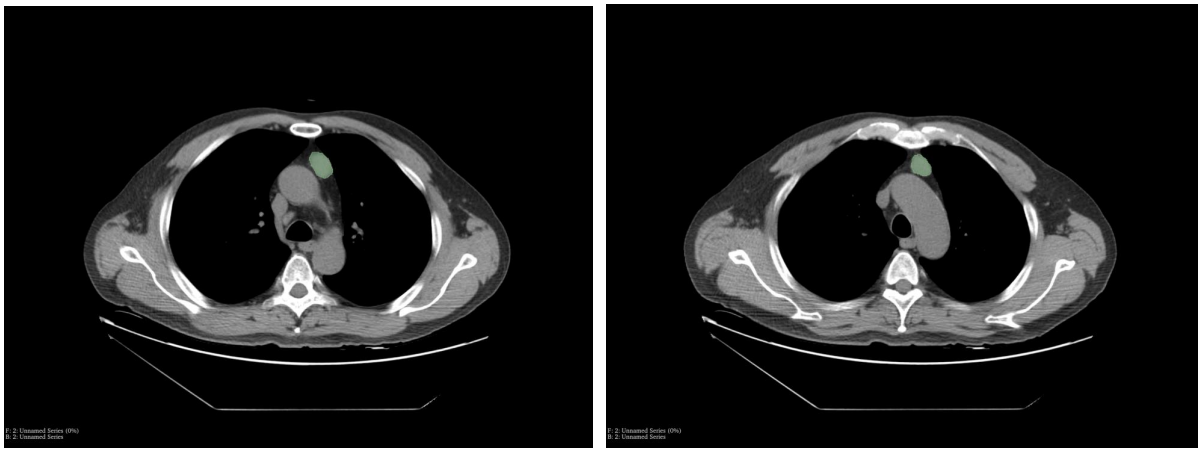


Figure 5.1: Two slices images containing same anterior mediastinal nodular lesion (thymoma type). Annotations are marked in green color.

(MRI) while 65 patients were diagnosed with surgery. Mediastinal MRI is a reliable modality in diagnosing benign mediastinal cysts of anterior mediastinum [129]. Detailed demographic information is provided in Table 5.1. Each scan generated 2D slice images with spatial resolutions of 512×512. For each scan, standard mediastinal window settings with a window width of 400 Houndsfield Unit (HU) and a window level of 30 HU were applied. All scans were manually annotated by an experienced physician using the software 3D Slicer [130]. To collect sufficient abnormal slice images that contain AMLs, multiple slice images having the same AML were included in our dataset. As shown in Figure. 5.1, a thymoma that appears in two slice images has different visual representations. Thereafter, a total of 1,017 abnormal slice images were collected in this study. Besides, the dataset used in this study is composed of 3,000 normal slice images, which were randomly selected from the 180 patient scans.

5.2.2 Network architecture

The foundation of the proposed network is based on UNet [10], which is the most used FCN-based architecture for medical image segmentation. The architecture of the proposed network is shown in Figure 5.2. The input size of the proposed network is $256 \times 256 \times 1$, which is equal to the halved size of a slice image. By taking a slice image as input, the proposed network generates the corresponding mask image that marks the locations of the AMLs at the pixel

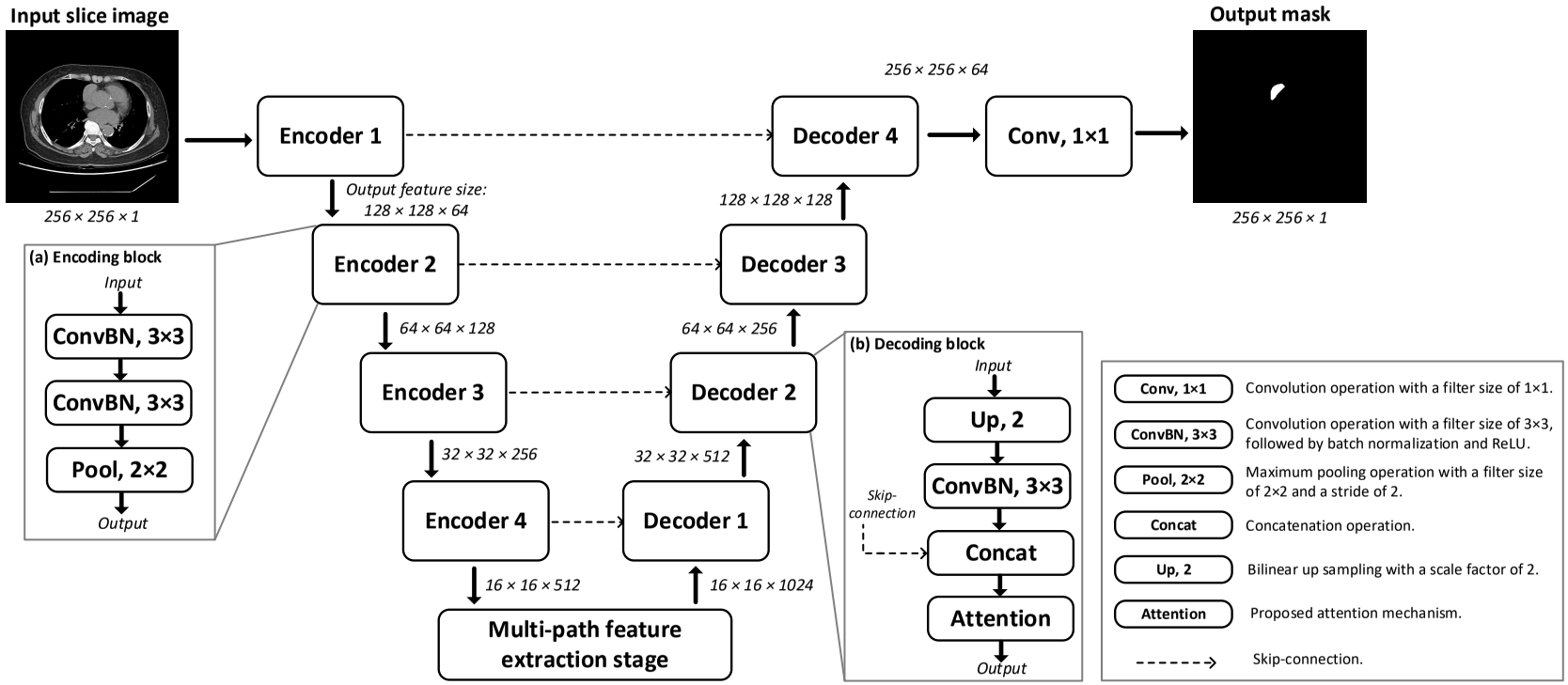


Figure 5.2: Architecture of proposed network tailored for anterior mediastinal nodular lesion segmentation from computed tomography imaging. (a) Architecture of encoding block. (b) Architecture of decoding block.

Table 5.1: Patient information and nodule characteristics of dataset used in this study.

	Thymic cyst (90)	Thymoma (90)	Total (180)
Age (mean±SD)(years)	60.6 ± 13.5	56.7 ± 12.1	58.5 ± 13.0
Gender (Male:Female)	52:38	48:42	100:80
Nodule size (mean±SD)(mm)	34.9 ± 19.6	51.5 ± 23.1	43.2 ± 22.9
Location*			
Maubrium	15 (16.7)	4 (4.4)	19 (10.6)
Sternomanubrial junction	20 (22.2)	5 (5.6)	25 (13.9)
Body	55 (61.1)	79 (87.8)	134 (74.4)
Xiphoid process	0 (0)	2 (2.2)	2 (1.1)

* location was divided according to the anatomical sternal level.

Note: values in parentheses are percentages.

level.

The proposed network is composed of a symmetric encoding path and decoding path. Each path consists of four sequential encoding and decoding blocks. At the encoding path, each encoding block follows convolutional neural network (CNN) architecture as shown in Figure 5.2(a). Specifically, each encoding block consists of two consecutive convolutional layers with a filter size of 3×3, followed by a batch normalization layer and a ReLU activation function. Additionally, a 2×2 maximum pooling layer with a stride of 2 is inserted to downsample the output feature maps. By cascading four encoding blocks, the input slice image is transformed to feature maps with a size of 16 × 16 × 512. The feature maps are then refined via the multi-path feature extraction stage, which aims to enhance features by employing the self-attention mechanism and dilated convolution. The details of the proposed multi-path feature extraction stage are presented in Section 5.2.3.

The refined feature maps are fed to the decoding blocks of the proposed network to generate the mask image output. The main purpose of the decoding path is to recover the spatial resolution of feature maps generated via the encoding path. In the original UNet,

transpose convolution is utilized to upsample feature maps. However, it is prone to generate artifacts due to uneven overlap and checkerboard problem [37]. To avoid artifacts, each decoding block of the proposed network is composited of a bilinear upsampling operation with a scale factor of 2, followed by a convolution layer with a filter size of 3×3, a batch normalization layer and a ReLU activation function, sequentially. The upsampled feature maps are concatenated with the feature maps generated from the encoding blocks via skip-connections. Skip-connections allow the decoding path to use early feature maps, which enhance detail retention [120]. Thereafter, the proposed attention mechanism is followed. Section 5.2.4 presents the proposed attention mechanism in detail. The architecture of a decoding block used in the proposed network is shown in Figure 5.2(b). At the end of the decoding path, a 1×1 convolution layer is followed to transform final feature maps to mask image with a spatial resolution of $256 \times 256 \times 1$.

5.2.3 Multi-path feature extraction stage

In the proposed network, a multi-path feature extraction stage is utilized to bridge the encoding path with the decoding path. An overview of the proposed multi-path feature extraction stage is illustrated in Figure 5.3. The proposed multi-path feature extraction stage consists of two paths, where one path is composed of a self-attention block that learns global semantic correlations. Another path consists of a dilated convolution block to model robust local information. Two paths are merged via a concatenation operation and connected to the first decoding block of the proposed network.

The general architecture of the proposed self-attention block follows the Transformer structure [123]. One novel aspect of the Transformer is effective to learn the long-range dependencies in a global space. For semantic segmentation, both local and global information is crucial to classify and label the pixels in an image precisely. Therefore, integrating the Transformer allows the proposed network to capture sufficient local and global information simultaneously. The architecture of the proposed self-attention block is shown in Figure 5.3(a). The input feature map is first encoded to learnable patch spatial information via a patch embedding. Intuitively, the patch embedding splits an input image into several fixed-size patches. However, common patch embedding treats the feature map as a sequence of non-

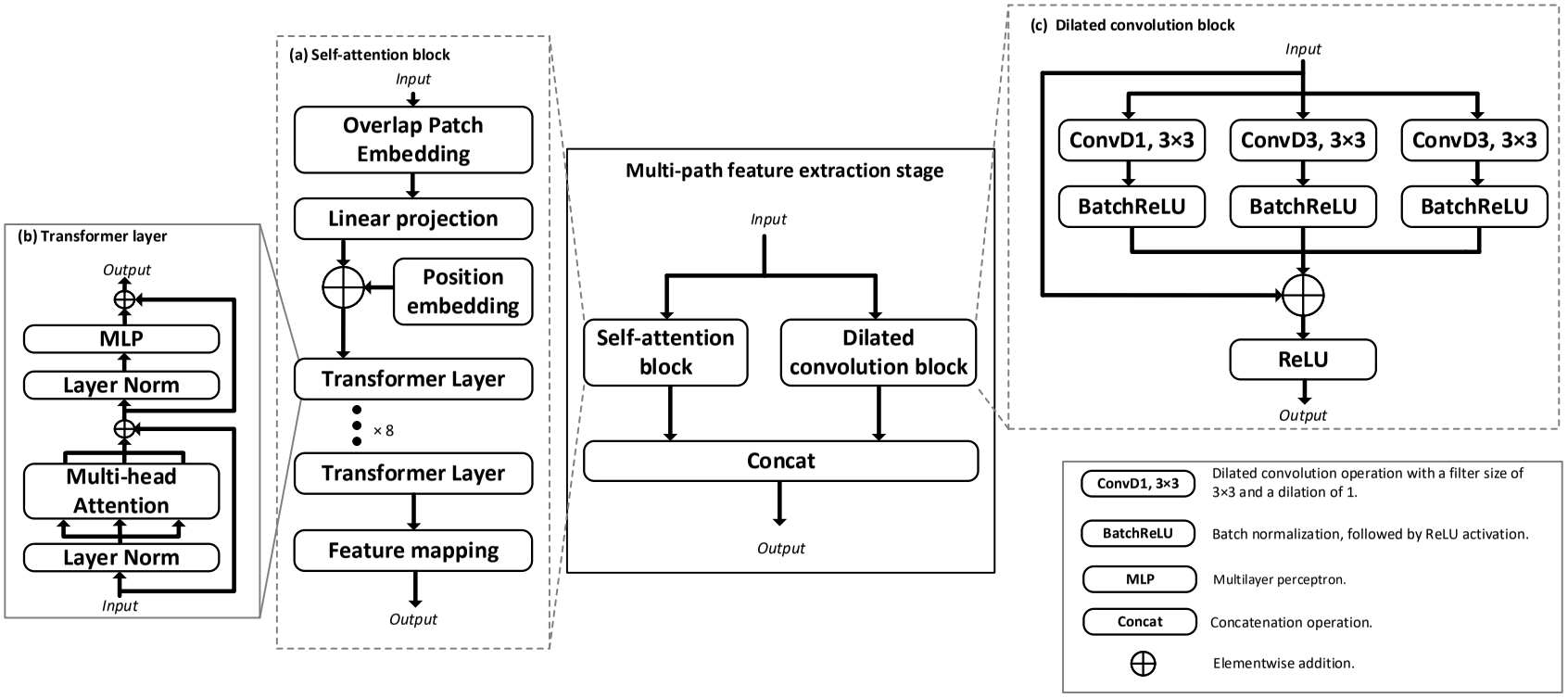


Figure 5.3: Architecture of proposed multi-path feature extraction stage. (a) Architecture of self-attention mechanism block. (b) Architecture of transformer layer. (c) Architecture of dilated convolution block.

overlapping feature patches [39]. Therefore, to maintain the local continuity around adjacent feature patches, we adopt an overlap patch embedding as introduced in [131]. Specifically, a 3×3 convolution operation is utilized to realize the overlap patch embedding. To preserve the original size of the input feature map, we set the stride to 1 and use zero padding. After patch embedding, a linear projection is utilized to transform the feature map into 1D sequences. For instance, given a feature map having a size of $W \times H \times D$, linear projection flattens the feature map into a total of d ($d = D$) sequences. Besides, the length of each sequence is N ($N = W \cdot H$). In the proposed network, the linear projection converts a $16 \times 16 \times 512$ feature map into 512 sequences. Each sequence has a length of 256. To encode the location information of each sequence, learnable position embedding is followed according to:

$$z_0 = P_{linear} \times f + E_{pos} \tag{5.1}$$

where $z_0 \in \mathbb{R}^{d \times N}$ refers to embedded sequences, $E_{pos} \in \mathbb{R}^{d \times N}$ denotes position embedding, P_{linear} is the linear projection, and f refers to the feature map. Thereafter, the embedded sequences are fed into L transformer layers. In the proposed network, we set L to 8. The main purpose of transformer layers is to connect every element of the feature map. Therefore, it is expected to access an expansive receptive field that concludes global information.

In the proposed network, each transformer layer follows the architecture of the Transformer encoder as described in [123]. As illustrated in Figure 5.3(b), each transformer layer consists of a Multi-Head Attention (MHA) block and a Multilayer perceptron (MLP) block with one hidden layer. Layer normalization (LN) [132] is applied before every block while the residual connection is utilized after every block. By default, the number of heads in the MHA block is set to 8. The number of hidden neurons in the MLP block is set to 2048. To fulfil the input dimension of the proposed decoding block, the features extracted from the transformer layers are mapped into 3D feature map. Specifically, the output sequence of the last Transformer layer is reshaped from $d \times N$ (512×256) to $W \times H \times D$ ($16 \times 16 \times 512$).

Dilated convolution is a type of convolution operation that uses expanded filters by inserting zeros between non-zero values of a filter [133]. As shown in Figure 5.4, the dilated convolution increases the receptive field without reducing the spatial resolution of the output. This can be particularly useful to capture fine-grained details. By taking the advantage of

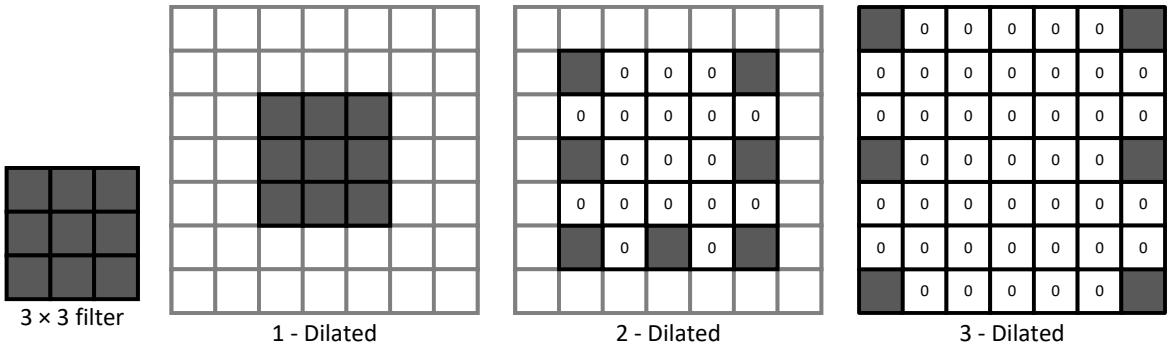


Figure 5.4: Illustration of dilated convolution operation with different dilation rates. From left to right, 3 × 3 convolution operations with a dilation rate of 1, 2 and 3, respectively.

the dilated convolution, in the proposed dilated convolution block, we adopt three parallel dilated convolutions to aggregate multi-scale contextual information, which is expected to improve the detection sensitivity for small AMLs. The architecture of the proposed dilated convolution block is shown in Figure 5.3(c). The three dilated convolutions follow dilated rates of 1, 3 and 5, respectively. A filter size of 3 × 3 is applied to all dilated convolutions. A batch normalization and ReLU activation are followed after each dilated convolution. Instead of stacking all dilated convolutions' features, we propose a residual connection to sum all features with input. One novel aspect of the residual connection is to reduce the complexity of the network, which is beneficial to avoid over-fitting.

5.2.4 Attention mechanism in decoding block

Image-grid attention mechanisms are widely utilized in CNNs, which allow the networks to focus on important information instead of learning irrelevant information, such as the background [122, 134, 135]. For semantic segmentation, the image-grid attention mechanism is beneficial to capture more useful information from the regions of interest in the feature map. In the original UNet, skip-connections concatenate features from the respective encoding path and decoding path. However, concatenated features may be redundant. To alleviate this problem, an image-grid attention mechanism is applied to the proposed decoding block. Specifically, we adopt the convolutional block attention module (CBAM), which

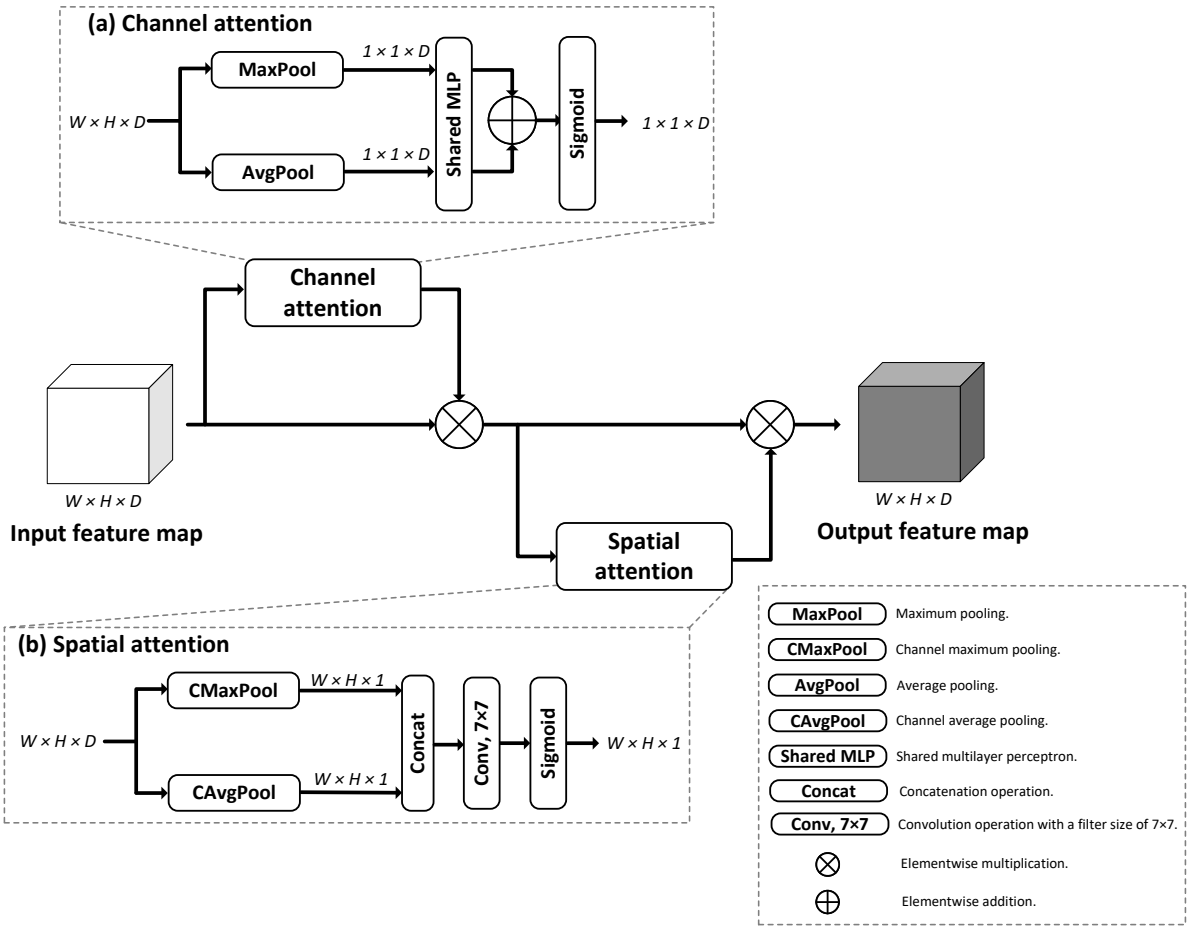


Figure 5.5: Illustration of a convolutional block attention module (CBAM). (a) Architecture of channel attention module. (b) Architecture of spatial attention module.

was originally introduced for CNN-based image classification [122]. As shown in Figure 5.5, the CBAM consists of two stages, which are the channel attention module and the spatial attention module. The channel attention module is utilized to strengthen the most important channel information by redistributing the channel features. The architecture of the channel attention module is illustrated in Figure 5.5(a). The spatial dimension of the input feature map is first squeezed by employing a maximum pooling operation and an average pooling operation in parallel. Squeezed features are fed into a shared MLP, in which the weights of the layers are shared across all of the input samples. Then, the outputs of the shared MLP are merged via an elementwise addition operation, followed by a sigmoid activation function. Finally, an elementwise multiplication operation is applied to combine the outputs of the channel attention module with the input feature map. After the channel attention module,

a spatial attention module is followed. The spatial attention module creates the inter-spatial relationship of features, which allows the network to focus on more important spatial locations than others. As shown in Figure 5.5(b), the input of the spatial attention module is first transformed by applying both maximum pooling and average pooling operations along the channel axis, followed by a concatenation operation. The fused features are convolved by applying a 7×7 convolution operation, followed by a sigmoid activation function. Thereafter, the outputs of the spatial attention module are merged with the input feature map via a concatenation operation.

5.2.5 Network training and evaluation

A 5-fold cross-validation is utilized to train and evaluate the proposed network. Specifically, each fold is composed of 36 patient cases. During each round of cross-validation, four folders are used to train and validate the proposed network while leaving one fold to evaluate the performance of the trained model. During the training phase, each model is trained for up to 100 epochs. The default batch size is 12. The loss function is a combination of soft dice loss [136] and binary cross-entropy loss. The loss is optimized by employing the Adam optimizer with an initial learning rate of 0.00001. Besides, the learning rate is reduced by 10 times when the validation loss stops increasing for continuous 5 epochs. All networks used in this study were implemented via Pytorch, and the networks were trained on an Nvidia P6000 GPU.

In addition, an early-stopping strategy is applied based on the validation performance of each epoch. Specifically, the training process is terminated when the dice coefficient stops increasing for consecutive 15 epochs. The final trained model is obtained by selecting the one achieving the best dice coefficient on the validation dataset. In this study, the reported results were obtained by averaging the testing performances across the five rounds of cross-validation, including the dice similarity coefficient (DSC), Intersection over Union (IoU), sensitivity and precision.

Augmentation is applied to improve the robustness of the learning process. Specifically, we randomly apply horizontal and vertical flipping, as well as angle rotation (a range from -60 to 60 degrees) to the training images.

5.3 Experiments and results

5.3.1 Effectiveness of proposed multi-path feature extraction stage

To identify the effectiveness of the proposed multi-path feature extraction stage as described in Section 5.2.3, the performances of the proposed network employing different bridge connections to connect the encoding path with the decoding path were compared. First, the proposed multi-path feature extraction stage was replaced with the original UNet configuration as described in [10]. Specifically, two consecutive 3×3 convolutional layers were utilized. Second, we compared the proposed one with another multi-path based bridge connection as introduced in [137]. Similar to the proposed one, it is composed of two paths, where one path adopts a self-attention mechanism using two consecutive transformer layers. A conventional patch embedding is utilized to transform the input feature map into sequences. On another path, a group of four dilated convolutions are stacked in parallel. The dilated rates are 1, 3, 6 and 9. Features generated from the four dilated convolutions are merged via a concatenate operation, followed by an additional 1×1 convolution operation.

Table 5.2: Segmentation performance for different bridge connections utilized in the proposed network.

Bridge Method	Multi-path	DSC	IoU	Sensitivity	Precision
Original UNet	✗	90.93	87.87	92.48	94.74
Thomar et al. [137]	✓	91.92	88.82	93.53	94.59
Proposed	✓	**93.23**	**90.29**	**93.98**	**95.68**

Table 5.2 summarizes the AML segmentation performances of the proposed network employing different bridge connections. Based on our experimental results, with the proposed multi-path feature extraction stage, the proposed network achieved a DSC of 93.23, an IoU of 90.29, a sensitivity of 93.98 and a precision of 95.68. By replacing the proposed multi-path feature extraction stage with the original UNet's bridge connection, the network demonstrated a performance downgrade (DSC = 90.93; IoU = 87.87; Sensitivity = 92.48; Precision = 94.74). In contrast, the multi-path based bridge connection adopted in [137] showed an

improved segmentation performance (DSC = 91.92; IoU = 88.82; Sensitivity = 93.53; Precision = 94.59) over the one with the original UNet's bridge connection. However, it still cannot reach the performance of the proposed multi-path feature extraction.

Table 5.3: Segmentation performance for different number of transformer layers (L), patch embedding strategy and dilated convolution merging strategy.

	Number of transformer layers (L)	Patch embedding	Dilated convolution merging	DSC	IoU	Sensitivity	Precision
	2	Overlap	Residual	92.91	89.92	93.78	95.42
	8	Overlap	Concatenation	91.81	88.74	92.83	95.28
	8	Non-overlap	Residual	92.43	89.40	93.72	95.08
Proposed	8	Overlap	Residual	**93.23**	**90.29**	**93.98**	**95.68**
	12	Overlap	Residual	92.21	89.05	93.26	95.21

In addition, a series of experiments was conducted to justify the effectiveness of the optimal number of transformer layers, overlap patch embedding and residual connection utilized in the proposed dilated convolution block. The results are summarized in Table 5.3. With 2 consecutive transformer layers, the network achieved better performance than the one utilizing 12 transformer layers. However, compared to our suggested configuration (L = 8), using 2 transformer layers showed around 0.3% and 0.4% lower on DSC and IoU, respectively. On the other hand, without employing overlap patch embedding, the proposed network showed performance downgrade with approximately 0.8% reduction on both DSC and IoU. Moreover, by replacing the proposed residual connection scheme with a direct concatenation, the proposed network yield lower DSC (91.81 vs 93.23) and IoU (88.74 vs 90.29).

5.3.2 Effectiveness of proposed attention mechanism

The effectiveness of the proposed attention mechanism was justified by removing it from the proposed network or replacing it with attention gates as described in [122]. Similar to the proposed attention mechanism, the attention gate is based on image-grid attention. To realize the attention gate, the decoding block of the proposed network is modified. As shown in Figure 5.6, the attention gate takes two inputs from the upsampling operation and the

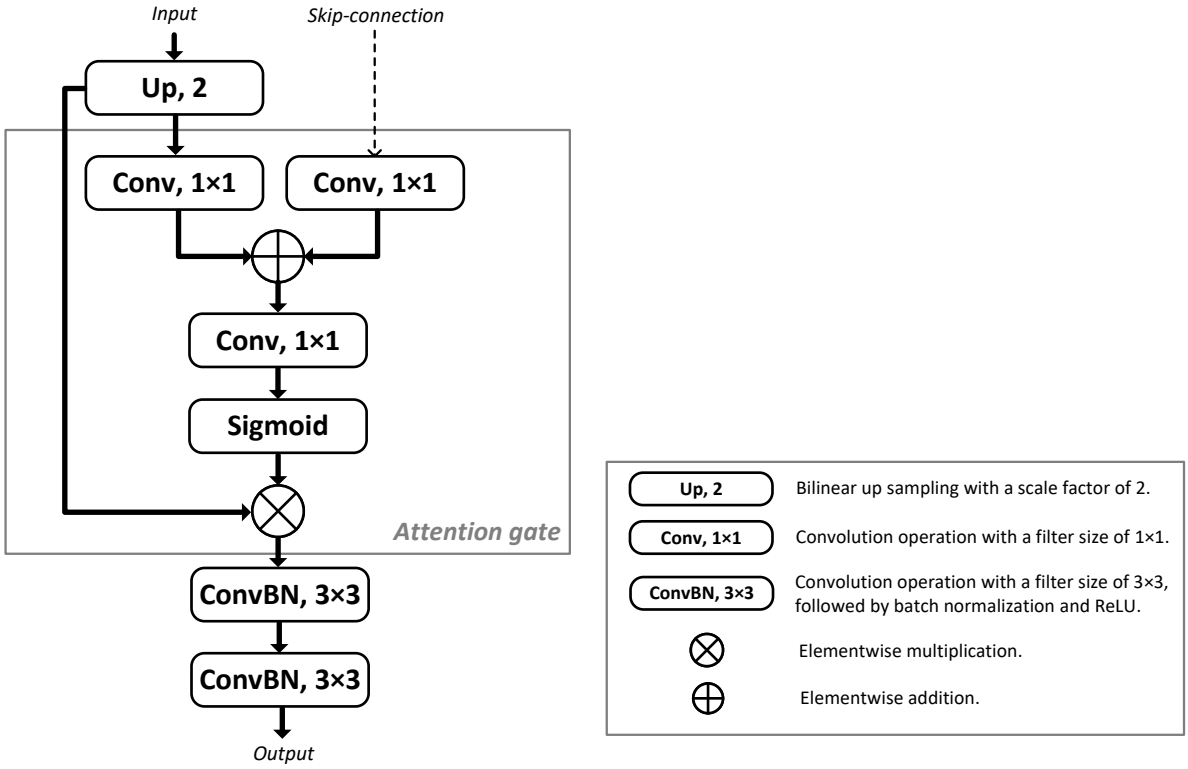


Figure 5.6: An overview of a decoding block utilizing attention gate.

corresponding skip connection. Two inputs are merged via 1×1 convolution operations, followed by an elementwise addition. After merging, the attention coefficients are obtained by employing a sigmoid function. Thereafter, the upsampled feature map is combined with the attention coefficients via an elementwise multiplication. As suggested in [122], two consecutive 3×3 convolution operations are followed to refine the feature map after passing the attention gate.

Table 5.4: Segmentation performance for the proposed network with or without employing attention mechanisms into decoding blocks.

Type of Attention	Attention	DSC	IoU	Sensitivity	Precision
N/A	✗	91.59	88.45	92.54	93.35
Attention gate	✓	91.96	88.94	92.89	93.50
Proposed	✓	**93.23**	**90.29**	**93.98**	**95.68**

As shown in Table 5.4, without adopting any attention mechanism, the proposed network

showed inferior AML segmentation performance with a DSC of 91.59, an IoU of 99.45, a sensitivity of 92.54 and a precision of 93.35. By integrating the attention gates, the proposed network outperformed the proposed one without applying any attention mechanism. With attention gates, the proposed network achieved a DSC of 91.96, an IoU of 88.94, an sensitivity of 92.89 and a precision of 93.50. Yet, the proposed attention mechanism achieved better segmentation performance than the proposed one using attention gates.

5.3.3 Comparison to mainstream segmentation networks

In this section, the AML segmentation performance of the proposed network is compared with state-of-the-art DL-based networks, including UNet, ResUNet, AttentionUNet, TransUNet and UNet++. All networks were trained and evaluated on our AMLs dataset by following the same data partitioning and evaluation scheme as described in Section 5.2.5. It is worth noting that we added batch normalization into UNet since early works have demonstrated the usefulness of improving the UNet performance by integrating batch normalization [138]. For the rest of the networks, batch normalization was utilized by default. Besides, all networks used a combination of soft dice loss and binary cross-entropy loss.

Table 5.5: Segmentation performance for different methods.

	Architecture	DSC	IoU	Sensitivity	Precision
Ronneberger et al. [10]	UNet	90.14	86.75	90.19	**95.87**
Zhang et al. [139]	ResUNet	89.13	85.84	89.98	95.22
Oktay et al. [122]	AttentionUNet	91.04	87.97	92.60	94.33
Chen et al. [124]	TransUNet	90.74	87.39	92.70	93.67
Zhou et al. [127]	UNet++	91.25	88.21	92.80	94.88
Proposed		**93.23**	**90.29**	**93.98**	95.68

Table 5.5 summarizes the segmentation performances of different DL-based networks on our AMLs dataset. Compared to other networks, the ResUNet showed inferior performance for all metrics (DSC = 89.13; IoU = 85.84; Sensitivity = 89.98; Precision = 95.22). The UNet achieved a DSC of 90.14, an IoU of 86.75, a sensitivity of 90.19 and a precision of

95.87. Compared to the rest of the networks, the UNet showed the highest precision. However, compared to the proposed network, the UNet led a roughly 2% lower on DSC, and a 3.5% lower on both IoU and sensitivity. By integrating the attention mechanism into the UNet architecture, the AttentionUNet showed improved DSC (91.04 vs 90.14), IoU (87.97 vs 86.75) and sensitivity (92.60 vs 90.19). However, it still cannot reach the performance of the proposed network. In addition, we compared the proposed network with TransUNet, a transformer-based U-shape network adopting the self-attention mechanism. The TransUNet yielded the lowest precision, which was 2% lower than the proposed network. The DSC, IoU and sensitivity of the TransUNet are 90.74, 87.39 and 92.70, respectively. The UNet++ achieved the highest DSC (91.25), IoU (88.21) and sensitivity (92.80) among other networks except the proposed one.

The qualitative comparison of different networks on our AMLs dataset is presented as shown in Figure 5.7. When the AMLs have small visual sizes in slice images, both UNet and ResUNet are more likely to fail to detect those lesions. One example is shown in the first row of Figure 5.7. In addition, for small lesions that overlap with surrounding tissues (Second row of Figure 5.7), the proposed network and the TransUNet demonstrated better performance than other networks. The results explain that the self-attention mechanism provides a greater ability to encode global contexts, which allows the network to differentiate the lesion from the surrounding tissues. In the third row of Figure 5.7, both AttentionUNet and the proposed network resulted in few false positives compared to others, which elucidates that the image-grid attention mechanisms (e.g., CBAM and attention gate.) are effective to allow the network to focus on desired local regions and suppress noisy predictions. As illustrated in the last three rows of Figure 5.7, UNet, ResUNet and AttentionUNet are more likely to over-segment or under-segment the AMLs compared to others when the sizes of the lesions are large. Besides, compared to other networks, the proposed network showed the nearest visual representation of the ground truths.

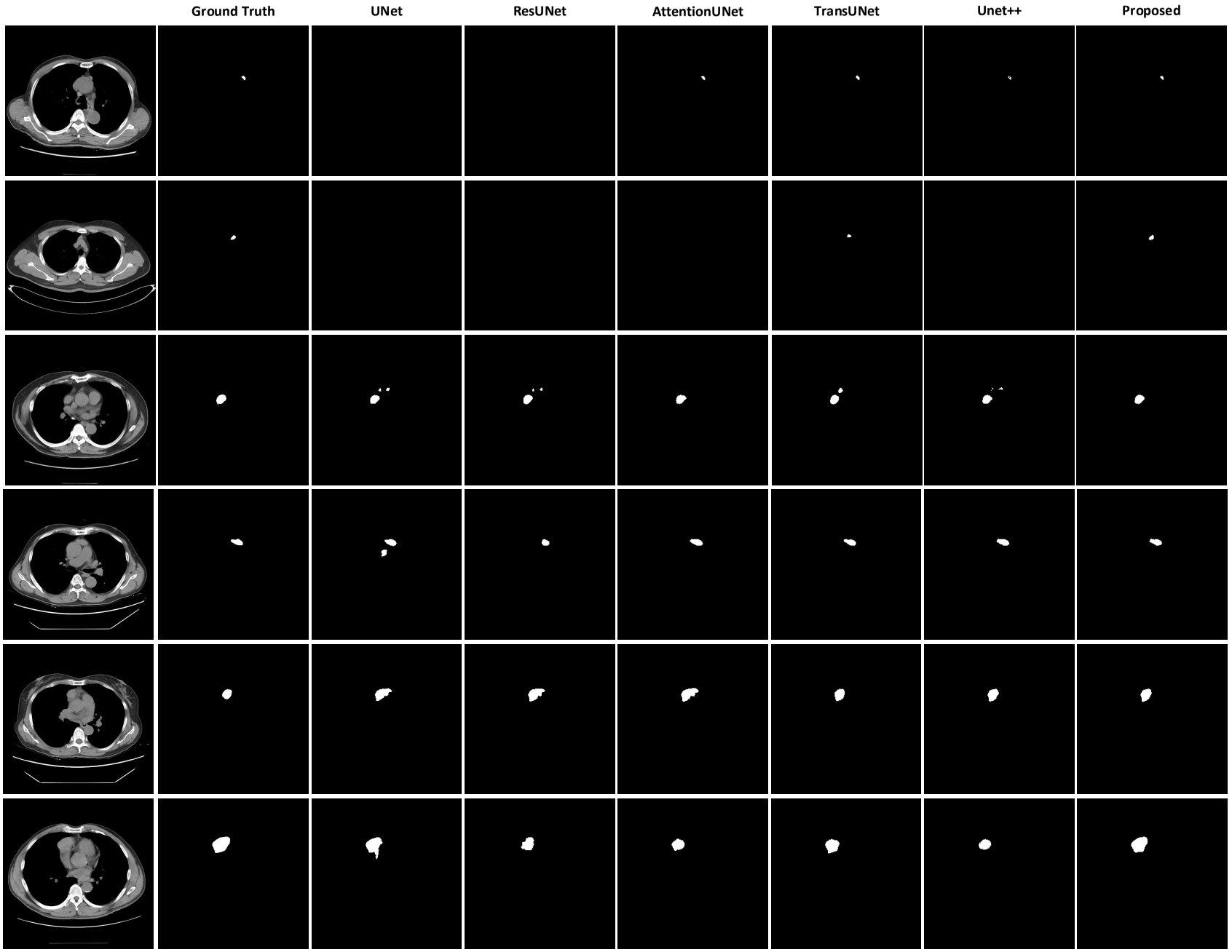


Figure 5.7: Qualitative comparison of different networks by visualization. From left to right: 1) Input slice image, 2) Ground truth, 3) UNet, 4) ResUNet, 5) AttentionUNet, 6) TransUNet, 7) UNet++, 8) Proposed.

5.4 Discussion

Clinically, radiologists rely on chest CT imaging to detect and diagnose AMLs. However, it is challenging to distinguish AMLs accurately by visual inspection. The reasons are as follows: 1) Due to the low contrast of CT imaging, AMLs may not stand out clearly from the surrounding tissue. Hence, it makes them difficult to see. 2) The gray distribution of AMLs may be uneven. This could cause some areas to appear lighter or darker than others. 3) The gray level of a AML may be similar to that of the surrounding tissue. Although additional magnetic resonance examination is recommended to justify the radiologists' findings [32], the entire process is still significantly time-consuming. Therefore, an automated procedure to detect AMLs from chest CT imaging is desired. In this study, a deep learning-based CADe system is proposed to assist radiologists in AML detection from chest CT imaging. Specifically, we adopt the self-attention and the image-grid attention mechanisms into the original UNet architecture to realize automated AML segmentation. Without any requirement for pre-processing step or manual intervention, the proposed network takes a 2D slice image as input and predicts the locations of the AMLs at the pixel level. The segmentation procedure does not only identify the locations of the lesions, but it also outlines areas of detected lesions. As the contours of the AMLs are the important diagnostic index to differentiate thymic cysts (benign) from thymomas (malignant), radiologists could be benefited by using the predicting outcomes of the proposed network without any further manual annotations. Besides, to establish a fully automated diagnosis procedure, a computer-aided diagnosis (CADx) system is commonly followed after a CADe system. Many previous works have demonstrated the effectiveness of utilizing CADx systems to improve the diagnostic performance of radiologists for AMLs [140–142]. However, lesions must be segmented from the images prior to deploying those systems. Manual segmentation is a time-consuming task and requires an experienced radiologist to verify the annotations. Therefore, the proposed network can be integrated into those CADx systems to provide automated lesion segmentation.

While numerous attempts have demonstrated the effectiveness of applying FCN-based networks for pulmonary nodule segmentation or mediastinal lymph node segmentation, there is currently one existing study that focused on AML segmentation by utilizing an FCN-based

network. [38] adopt a two-staged 3D ResUNet to realize AML segmentation. By evaluating 114 AML, the 3D ResUNet achieved a DSC of 87.73. It should be noted that a direct comparison is not possible since we used different private datasets. For this reason, it is impossible to determine the relative difficulty of each dataset based on the number of AMLs. Nevertheless, the previous study could be less generic than the proposed one because the performance of the 3D ResUNet was highly correlated with the quality of the input data. Specifically, a comprehensive pre-processing step is required to generate a lung lobe mask and remove irrelevant anatomical structures inside the CT imaging. Without the pre-processing step, the 3D ResUNet tended to under-segment AMLs. In contrast, the proposed network takes raw CT slice images as input.

Moreover, the feasibility of employing an FCN-based network to segment AMLs from chest CT imaging is justified. Specifically, we evaluated the performances of various state-of-the-art networks on our AMLs dataset, including UNet, ResUNet, AttentionUNet, TransUNet and UNet++. Compared to those former networks, the proposed network achieved a roughly 2% higher DSC and IoU. The improvement over those former networks could be explained by several reasons as follows: 1) Our results conclude that the self-attention mechanism utilized in-between the encoding and decoding paths improves the network performance, particularly for AML segmentation. The self-attention mechanism preserves the long-range dependencies in a global space, which improves the generalization of the network by capturing larger and more complex receptive fields [143]. 2) Our results confirmed that incorporating an image-grid attention mechanism into the decoding block of the network is beneficial to improve the performance of segmenting AMLs from chest CT imaging. The image-grid attention mechanisms provide an efficient way to allow the network to extract more discriminative features within the regions of the AMLs. 3) Original UNet uses a series of convolutional layers to pass the high-level features from the encoding path to the decoding path. To provide more robust high-level features, we suggest using the proposed multi-path feature extraction stage to bridge a connection that links the encoding path with the decoding path. Compared to the bridge connection of the original UNet, the proposed multi-path feature extraction stage replaces the conventional convolutional layers with a group of dilated convolutions, which allows the network to capture more fine-grained details from enlarged

local regions. Besides, our results confirmed that the proposed multi-path feature extraction enhances the high-level features by combing the rich multi-scale local features produced from the dilated convolutions with global features generated via the proposed self-attention mechanism. It is worth noting that the concept of the multi-path feature extraction stage was first introduced in [137]. Nevertheless, the proposed multi-path feature extraction stage would be preferred due to two reasons: 1) We adopt an overlap patch embedding with additional position embedding to provide a better description of the sequences' relations in the vector space. In addition, the proposed overlap patch embedding allows the exploitation of better representations of continuous information between feature patches. 2) To merge the feature maps of dilated convolutions, we propose a residual connection method instead of stacking the feature maps via a concatenation operation. The proposed residual connection is beneficial to reduce the complexity of the network while reducing the redundant features produced by the concatenation operation.

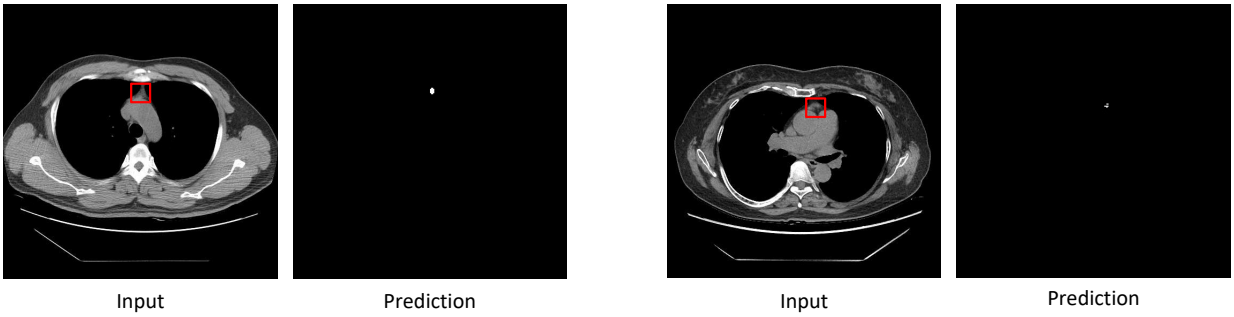


Figure 5.8: Two examples show false positives are generated within the shadow regions.

Although the proposed network yielded superior performance in the experiments, it still has some limitations. As shown in Figure 5.8, normal tissues overlayed in the shadow regions may be misclassified as AMLs by the proposed network. Such false positives only appear when the shadow regions are apart from the artery or organs. In our future work, a study of how to conduct extra analysis in those areas is desired. Specially, as suggested in [144], investigating the feasibility of integrating inter-slice spatial information into the proposed network is a fruitful direction. Moreover, since the incidence of AML is extremely low, the size of our AML dataset is relatively smaller compared to other common chest diseases, such

as pulmonary nodule [145]. This may impact the generalization ability of the model. To remedy this problem, we adopt comprehensive image augmentation techniques to generate spatial variances for each training image. In the future, further prospectively designed studies with larger data scales can be conducted to confirm the effectiveness of the proposed network.

Chapter 6

Summary and Future Works

Up until now, cancer remains the first leading cause of death worldwide. Hence, early detection and diagnosis are important to reduce the mortality rate of cancer death. Medical imaging is a commonly used screening modality for various cancer diagnoses. However, screening and interpreting medical imaging is still a challenge even for experienced radiologists due to time-consuming. Computer-aided detection and diagnosis (CAD) systems have been developed that adopt computerized procedures to analyze medical imaging and provide supplemental diagnostic opinions to assist radiologists in medical imaging interpretation. Nevertheless, there are currently not many CAD systems utilized for real clinical situations due to the high false-positive rate to detect lesions and inconsistent performance to categorize lesions with a wide range of variance. To overcome the limitations of the existing CAD systems, the feasibility of adopting deep learning techniques into CAD systems is explored in this thesis. In the following sections, Section 6.1 summarizes the presented works in this thesis. Section 6.2 presents the plan of future works.

6.1 Summary

A CAD system generally consists of two sub-systems, including a computer-aided detection (CADe) system and a computer-aided diagnosis (CADx) system. CADe system is utilized for lesion detection. Next, the detected lesions are classified into benign or malignant by employing a CADx system. In this thesis, two CNN-based CADx systems are proposed to support radiologists for lung cancer and breast cancer diagnoses. In addition, a CNN-based CADe system is proposed to help radiologists for diagnosis of anterior mediastinal nodular lesions from chest CT imaging.

A CNN-based CADx system designed for pulmonary nodule classification on computed tomography (CT) imaging is presented in Chapter 3 of this thesis. A multi-path feature extraction scheme is proposed that allows the features extracted at the benign of the CNN to mix with the features extracted at the end of the CNN. With the proposed multi-path feature extraction scheme, the final extracted features of the CNN are effective to describe the characteristics of different types of nodules, including fine global structures and sparse local structures. The proposed CNN adopting the multi-path feature extraction scheme shows a 4% improvement in classification accuracy compared to the proposed one without employing the multi-path feature extraction scheme. In addition, multi-scale convolutional layers are proposed to perform feature extraction at different scales. Compared to the proposed CNN employing conventional convolutional layers, the proposed one adopting multi-scale convolutional layers shows 2% increased classification accuracy. The effectiveness of the proposed CNN is justified by comparing the classification performance of the proposed CNN with existing machine learning (ML) approaches and state-of-the-art CNN methods. Specifically, the proposed CNN shows a 14% area under the curve (AUC) improvement over the previous ML approaches while the proposed CNN achieves up to 13% higher classification accuracy and 11% higher AUC compared to the existing CNN methods. Furthermore, the clinical benefits of the proposed CNN is confirmed. By referring to the predicting outcomes of the proposed CNN, the thoracic radiologist shows improved diagnostic performance with a 23% improvement in AUC. The results conclude that the proposed CNN is well-suited to support radiologists for lung cancer diagnosis by providing accurate predictions to differentiate different types of pulmonary nodules.

A CNN-based CADx system tailored for breast lesion classification on automated breast ultrasound (ABUS) imaging is proposed in Chapter 4 of this thesis. This was the first deep learning-based approach utilized as a breast cancer diagnosis tool on ABUS imaging. The proposed CNN adopts the Inception-v3 [8] backbone to provide effective lesion feature extraction. Compared to other state-of-the-art backbones, the proposed backbone shows up to 8% improvement in sensitivity and 4% increased specificity. To enhance feature extraction of the proposed CNN, two multiview learning strategies are designed that allow the proposed CNN to extract features from different views simultaneously. By adopting the proposed

multiview learning strategies, the proposed CNN improves its sensitivity and specificity by 5% and 4%, respectively. Compared to previous methods which require to segment breast lesions from the input images manually or define the feature extractor explicitly, the proposed CNN is designed to take raw images as input and predict the types of breast lesions without any manual intervention. In contrast to ML approaches, the proposed CNN achieves 10% higher sensitivity and specificity. Furthermore, the clinical usefulness of the proposed CNN is explored by comparing the diagnostic performances of the human reviewers with and without the aid of the proposed CNN. The results demonstrate that the non-breast specialists show 38% diagnostic performance improvement (AUC improvement: 0.245) by referring to the predicting outcomes of the proposed CNN. The Therefore, the proposed CNN has advantages as a second reviewer to provide diagnostic suggestions for non-breast specialists.

In Chapter 5 of this thesis, a fully convolutional neural network (FCN) based CADe system is proposed to segment anterior mediastinal nodular lesions (AMLs) from chest CT imaging. The proposed network is composed of a modified UNet to take a 2D slice image as input and predict the contours of AMLs at the pixel level. For the proposed network, a multi-path feature extraction stage is utilized to combine the strengths of the Transformer for modeling long-range dependencies and dilated convolutions for capturing multi-scale context. By integrating the proposed multi-path feature extraction stage, the proposed network shows a roughly 2.3% improvement on both dice similarity coefficient (DSC) and intersection-over-union (IoU) compared to the one without adopting the proposed multi-path feature extraction stage. Besides, an image-grid attention mechanism, the convolutional block attention module (CBAM), is adopted to provide robust feature learning for discriminating the regions of interest from dense scenes. By employing the CBAM, the proposed network yields improved segmentation performance for AMLs, achieving approximately 1.5% higher DSC and 1.8% better IoU than the proposed one without utilizing the image-grid attention mechanism. Furthermore, the effectiveness of the proposed network is justified by comparing the performance of the proposed network with state-of-the-art CNN-based segmentation networks, including UNet, ResUNet, AttentionUNet, TransUNet and UNet++. Compared to those former networks, the proposed network achieved up to 2% of improvements on both DSC and IoU. Based on the experimental results, the proposed network can potentially as-

sist radiologists in AML detection while providing the contours of the detected lesions for malignancy measurement.

6.2 Future work

In the future, the presented works of this thesis can be extended or improved from several aspects as follows:

First, both proposed CADx systems presented in Chapter 3 and Chapter 4 of this thesis can be incorporated with candidate detectors to realize end-to-end CAD systems. Similarly, the proposed CADe system described in Chapter 5 of this thesis can integrate with a CADx system to establish a fully automated detection and diagnosis system. Compared to a standalone CADe system or CADx system, a completed CAD system is more generic and suitable for real clinical usage. Similar to most common CADx systems, the proposed CADx systems as described in Chapter 3 and Chapter 4 of this thesis are designed to use patch images as input. Each patch image is cropped by referring to the radiologists' annotations, which is a time-consuming task and operator depended. Besides, the common CADe systems including the proposed CADe system as described in Chapter 5 of this thesis predict areas of abnormalities without suggesting the types (e.g., benign or malignant) of detected regions. Therefore, to make the final diagnostic decision, radiologists have to interpret those areas suggested by CADe systems. To reduce the workflow of radiologists, a completed CAD system is desired. Existing solutions to realize an end-to-end CAD system can be considered [146, 147].

Second, the feasibility of applying 3D-based feature extraction can be explored. For the CNNs presented in thesis, the input format is 2D-based data. To realize 3D-based feature extraction, CNN can be designed to use volumetric data as input. Compared to 2D-based data, the volumetric data may contain more features since the data includes multiplanar information. To realize feature extraction over volumetric data, the 3D convolution operation can be developed. Nevertheless, there are some limitations of employing 3D-based feature extraction. The reasons are as follows: (1) Computation complexity is considerably higher compared to 2D-based feature extraction. Thus, it could result in a slow inference time, which is not desired for actual clinical usage. (2) Samples available to train the network could have

a smaller size compared to 2D-based feature extraction. For instance, one 3D sample can create multiple 2D slices or patches. As the size of the medical imaging database is commonly small, it may be difficult to train a 3D-based network to have effective generalization ability.

Bibliography

[1] D. R. Brenner, A. Poirier, R. R. Woods, L. F. Ellison, J. M. Billette, A. A. Demers, S. X. Zhang, C. Yao, C. Finley, N. Fitzgerald, N. Saint-Jacques, L. Shack, D. Turner, and E. Holmes, "Projected estimates of cancer in Canada in 2022," *CMAJ*, vol. 194, pp. E601–E607, may 2022.

[2] J. C. van Zelst, T. Tan, R. M. Mann, and N. Karssemeijer, "Validation of radiologists' findings by computer-aided detection (CAD) software in breast cancer detection with automated 3D breast ultrasound: a concept study in implementation of artificial intelligence software," *Acta Radiologica*, vol. 61, pp. 312–320, mar 2020.

[3] K. J. Chae, G. Y. Jin, S. B. Ko, Y. Wang, H. Zhang, E. J. Choi, and H. Choi, "Deep Learning for the Classification of Small (2 cm) Pulmonary Nodules on CT Imaging: A Preliminary Study," *Academic Radiology*, vol. 27, pp. e55–e63, apr 2020.

[4] B. Van Ginneken, B. M. Ter Haar Romeny, and M. A. Viergever, "Computer-aided diagnosis in chest radiography: A survey," *IEEE Transactions on Medical Imaging*, vol. 20, pp. 1228–1241, dec 2001.

[5] W. H. Wolberg, W. N. Street, and O. L. Mangasarian, "Machine learning techniques to diagnose breast cancer from image-processed nuclear features of fine needle aspirates," *Cancer Letters*, vol. 77, pp. 163–171, mar 1994.

[6] W. Z. Liu, A. P. White, M. T. Hallissey, and J. W. Fielding, "Machine learning techniques in early screening for gastric and oesophageal cancer," *Artificial Intelligence in Medicine*, vol. 8, pp. 327–341, aug 1996.

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Communications of the ACM*, vol. 60, pp. 84–90, Curran Associates Inc., 2017.

[8] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016-Decem, pp. 2818–2826, 2016.

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

[10] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," *CoRR*, vol. abs/1505.0, 2015.

[11] E. Shelhamer, J. Long, and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, pp. 640–651, nov 2017.

[12] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016-Decem, pp. 779–788, jun 2016.

[13] 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, dec 2016.

[14] M. Z. Alom, T. M. Taha, C. Yakopcic, S. Westberg, P. Sidike, M. S. Nasrin, M. Hasan, B. C. Van Essen, A. A. Awwal, and V. K. Asari, "A state-of-the-art survey on deep learning theory and architectures," *Electronics (Switzerland)*, vol. 8, p. 292, mar 2019.

[15] R. L. Siegel, K. D. Miller, H. E. Fuchs, and A. Jemal, "Cancer statistics, 2022," *CA: A Cancer Journal for Clinicians*, vol. 72, pp. 7–33, jan 2022.

[16] American Lung Association, "Lung Cancer Fact Sheet."

[17] M. L. Giger, K. T. Bae, and H. Macmahonr, "Computerized detection of pulmonary nodules in computed: Tomography images," *Investigative Radiology*, vol. 29, no. 4, pp. 459–465, 1994.

[18] S. G. Armato, M. L. Giger, C. J. Moran, J. T. Blackburn, K. Doi, and H. MacMahon, "Computerized detection of pulmonary nodules on CT scans," *Radiographics*, vol. 19, no. 5, pp. 1303–1311, 1999.

[19] F. Zhang, Y. Song, W. Cai, M. Z. Lee, Y. Zhou, H. Huang, S. Shan, M. J. Fulham, and D. D. Feng, "Lung nodule classification with multilevel patch-based context analysis," *IEEE Transactions on Biomedical Engineering*, vol. 61, pp. 1155–1166, mar 2014.

[20] M. Nishio and C. Nagashima, "Computer-aided Diagnosis for Lung Cancer: Usefulness of Nodule Heterogeneity," *Academic Radiology*, vol. 24, pp. 328–336, mar 2017.

[21] Z. Fan, H. Sun, C. Ren, X. Han, and Z. Zhao, "Texture recognition of pulmonary nodules based on volume local direction ternary pattern," *Bioengineered*, vol. 11, pp. 904–920, jan 2020.

[22] Q. Li, W. Cai, X. Wang, Y. Zhou, D. D. Feng, and M. Chen, "Medical image classification with convolutional neural network," *2014 13th International Conference on Control Automation Robotics and Vision, ICARCV 2014*, pp. 844–848, 2014.

[23] X. Zhao, L. Liu, S. Qi, Y. Teng, J. Li, and W. Qian, "Agile convolutional neural network for pulmonary nodule classification using CT images," *International Journal of Computer Assisted Radiology and Surgery*, vol. 13, pp. 585–595, mar 2018.

[24] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang, "Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning?," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1299–1312, 2016.

[25] H. C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers, "Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning," *IEEE Transactions on Medical Imaging*, vol. 35, pp. 1285–1298, may 2016.

[26] H. J. Shin, H. H. Kim, and J. H. Cha, "Current status of automated breast ultrasonography," *Ultrasonography*, vol. 34, no. 3, pp. 165–172, 2015.

[27] W. K. Moon, Y. W. Shen, C. S. Huang, L. R. Chiang, and R. F. Chang, "Computer-Aided Diagnosis for the Classification of Breast Masses in Automated Whole Breast Ultrasound Images," *Ultrasound in Medicine and Biology*, vol. 37, no. 4, pp. 539–548, 2011.

[28] T. Tan, B. Platel, H. Huisman, C. I. Sánchez, R. Mus, and N. Karssemeijer, "Computer-aided lesion diagnosis in automated 3-D breast ultrasound using coronal spiculation," *IEEE Transactions on Medical Imaging*, vol. 31, no. 5, pp. 1034–1042, 2012.

[29] S. H. Yoon, S. H. Choi, C. H. Kang, and J. M. Goo, "Incidental Anterior Mediastinal Nodular Lesions on Chest CT in Asymptomatic Subjects," *Journal of Thoracic Oncology*, vol. 13, no. 3, pp. 359–366, 2018.

[30] A. C. Adler and S. Modlin, "Anterior Mediastinal Masses," *Case Studies in Pediatric Anesthesia*, pp. 86–91, sep 2019.

[31] C. I. Henschke, I. J. Lee, N. Wu, A. Farooqi, A. Khan, D. Yankelevitz, and N. K. Altorki, "CT screening for lung cancer: Prevalence and incidence of mediastinal masses," *Radiology*, vol. 239, pp. 586–590, may 2006.

[32] R. F. Munden, B. W. Carter, C. Chiles, H. MacMahon, W. C. Black, J. P. Ko, H. P. McAdams, S. E. Rossi, A. N. Leung, P. M. Boiselle, M. S. Kent, K. Brown, D. S. Dyer, T. E. Hartman, E. M. Goodman, D. P. Naidich, E. A. Kazerooni, L. L. Berland, and P. V. Pandharipande, "Managing Incidental Findings on Thoracic CT: Mediastinal and Cardiovascular Findings. A White Paper of the ACR Incidental Findings Committee," *Journal of the American College of Radiology*, vol. 15, pp. 1087–1096, aug 2018.

[33] W. Jung, S. Cho, S. Yum, Y. K. Lee, K. Kim, and S. Jheon, "Differentiating thymoma from thymic cyst in anterior mediastinal abnormalities smaller than 3 cm," *Journal of Thoracic Disease*, vol. 12, pp. 1357–1365, apr 2020.

[34] G. Aresta, C. Jacobs, T. Araújo, A. Cunha, I. Ramos, B. van Ginneken, and A. Campilho, "iW-Net: an automatic and minimalistic interactive lung nodule segmentation deep network," *Scientific Reports*, vol. 9, pp. 1–9, aug 2019.

[35] M. Usman, B. D. Lee, S. S. Byon, S. H. Kim, B. il Lee, and Y. G. Shin, "Volumetric lung nodule segmentation using adaptive ROI with multi-view residual learning," *Scientific Reports*, vol. 10, pp. 1–15, jul 2020.

[36] H. Oda, K. K. Bhatia, H. R. Roth, M. Oda, T. Kitasaka, S. Iwano, H. Homma, H. Takabatake, M. Mori, H. Natori, J. A. Schnabel, and K. Mori, "Dense volumetric detection and segmentation of mediastinal lymph nodes in chest CT images," *Proc. SPIE*, vol. 10575, p. 1, feb 2018.

[37] A. A. Nayan, B. Kijsirikul, and Y. Iwahori, "Mediastinal Lymph Node Detection and Segmentation Using Deep Learning," *IEEE Access*, vol. 10, pp. 89289–89307, 2022.

[38] S. Huang, X. Han, J. Fan, J. Chen, L. Du, W. Gao, B. Liu, Y. Chen, X. Liu, Y. Wang, D. Ai, G. Ma, and J. Yang, "Anterior Mediastinal Lesion Segmentation Based on Two-Stage 3D ResUNet With Attention Gates and Lung Segmentation," *Frontiers in Oncology*, vol. 10, p. 3290, feb 2021.

[39] L. Chen, S. Li, Q. Bai, J. Yang, S. Jiang, and Y. Miao, "Review of image classification algorithms based on convolutional neural networks," *Remote Sensing*, vol. 13, p. 4712, nov 2021.

[40] Z. Q. Zhao, P. Zheng, S. T. Xu, and X. Wu, "Object Detection with Deep Learning: A Review," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, pp. 3212–3232, jul 2019.

[41] S. Minaee, Y. Y. Boykov, F. Porikli, A. J. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image Segmentation Using Deep Learning: A Survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, jan 2021.

[42] X. Sun, D. Yang, X. Li, T. Zhang, Y. Meng, H. Qiu, G. Wang, E. Hovy, and J. Li, "Interpreting Deep Learning Models in Natural Language Processing: A Review," oct 2021.

[43] E. Ahishakiye, M. B. Van Gijzen, J. Tumwiine, R. Wario, and J. Obungoloch, "A survey on deep learning in medical image reconstruction," *Intelligent Medicine*, vol. 1, pp. 118–127, sep 2021.

[44] Q. Zhang, M. Zhang, T. Chen, Z. Sun, Y. Ma, and B. Yu, "Recent advances in convolutional neural network acceleration," *Neurocomputing*, vol. 323, pp. 37–51, dec 2019.

[45] 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. 2278–2323, 1998.

[46] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2015 Inter, no. 1, pp. 1026–1034, 2015.

[47] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," *Journal of Machine Learning Research*, vol. 9, pp. 249–256, 2010.

[48] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proceedings of COMPSTAT 2010 - 19th International Conference on Computational Statistics, Keynote, Invited and Contributed Papers*, pp. 177–186, Springer Science and Business Media Deutschland GmbH, 2010.

[49] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *COLT 2010 - The 23rd Conference on Learning Theory*, vol. 12, pp. 257–269, jul 2010.

[50] M. D. Zeiler, "ADADELTA: An Adaptive Learning Rate Method," *CoRR*, vol. abs/1212.5, 2012.

[51] Jia Deng, Wei Dong, R. Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, jun 2009.

[52] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 07-12-June, pp. 1–9, 2015.

[53] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-ResNet and the impact of residual connections on learning," *31st AAAI Conference on Artificial Intelligence, AAAI 2017*, pp. 4278–4284, 2017.

[54] M. W. Huang, C. W. Chen, W. C. Lin, S. W. Ke, and C. F. Tsai, "SVM and SVM ensembles in breast cancer prediction," *PLoS ONE*, vol. 12, no. 1, pp. 1–14, 2017.

[55] S. G. Armato, K. Drukker, F. Li, L. Hadjiiski, G. D. Tourassi, R. M. Engelmann, M. L. Giger, G. Redmond, K. Farahani, J. S. Kirby, and L. P. Clarke, "LUNGx Challenge for computerized lung nodule classification," *Journal of Medical Imaging*, vol. 3, no. 4, p. 044506, 2016.

[56] A. C. Society, "Cancer Facts & Figures 2017. Atlanta, Ga." Available online: https://www.cancer.org/research/cancer-facts-statistics/all-cancer-facts-figures/cancer-facts-figures-2017.html, 2006.

[57] E. J. van Beek, "Lung cancer screening: Computed tomography or chest radiographs?," *World Journal of Radiology*, vol. 7, no. 8, p. 189, 2015.

[58] J. Abraham, "Reduced lung cancer mortality with low-dose computed tomographic screening," *Community Oncology*, vol. 8, no. 10, pp. 441–442, 2011.

[59] A. Kamiya, S. Murayama, H. Kamiya, T. Yamashiro, Y. Oshiro, and N. Tanaka, "Erratum to: Kurtosis and skewness assessments of solid lung nodule density histograms: Differentiating malignant from benign nodules on CT (Japanese Journal of Radiology

(2014) 32, 1 (14-21) DOI 10.1007/s11604-013-0264-y)," *Japanese Journal of Radiology*, vol. 32, p. 251, jan 2014.

[60] H. Zhu, H. Cheng, and Y. Fan, "Random local binary pattern based label learning for multi-atlas segmentation," *Medical Imaging 2015: Image Processing*, vol. 9413, p. 94131B, mar 2015.

[61] F. Ciompi, C. Jacobs, E. T. Scholten, M. M. Wille, P. A. De Jong, M. Prokop, and B. Van Ginneken, "Bag-of-frequencies: A descriptor of pulmonary nodules in computed tomography images," *IEEE Transactions on Medical Imaging*, vol. 34, pp. 962–973, apr 2015.

[62] M. Anthimopoulos, S. Christodoulidis, L. Ebner, A. Christe, and S. Mougiakakou, "Lung Pattern Classification for Interstitial Lung Diseases Using a Deep Convolutional Neural Network," *IEEE Transactions on Medical Imaging*, vol. 35, pp. 1207–1216, mar 2016.

[63] T. Graepel, "AlphaGo - Mastering the game of go with deep neural networks and tree search," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9852 LNAI, p. XXI, jan 2016.

[64] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, pp. 1–14, 2015.

[65] Q. Z. Song, L. Zhao, X. K. Luo, and X. C. Dou, "Using Deep Learning for Classification of Lung Nodules on Computed Tomography Images," *Journal of Healthcare Engineering*, vol. 2017, 2017.

[66] P. Monkam, S. Qi, M. Xu, F. Han, X. Zhao, and W. Qian, "CNN models discriminating between pulmonary micro-nodules and non-nodules from CT images," *BioMedical Engineering Online*, vol. 17, pp. 1–16, mar 2018.

[67] W. Shen, M. Zhou, F. Yang, D. Yu, D. Dong, C. Yang, Y. Zang, and J. Tian, "Multi-crop Convolutional Neural Networks for lung nodule malignancy suspiciousness classification," *Pattern Recognition*, vol. 61, pp. 663–673, 2017.

[68] K. Jarrett, K. Kavukcuoglu, M. Ranzato, and Y. LeCun, "What is the best multi-stage architecture for object recognition?," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2146–2153, sep 2009.

[69] N. Abramson, D. Braverman, and G. Sebestyen, *Pattern recognition and machine learning*, vol. 9. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 1963.

[70] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 56, pp. 1929–1958, 2014.

[71] J. Chen and Y. Shen, "The effect of kernel size of CNNs for lung nodule classification," in *2017 9th International Conference on Advanced Infocomm Technology, ICAIT 2017*, pp. 340–344, nov 2018.

[72] M. A. Mazurowski, P. A. Habas, J. M. Zurada, J. Y. Lo, J. A. Baker, and G. D. Tourassi, "Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance," *Neural Networks*, vol. 21, no. 2-3, pp. 427–436, 2008.

[73] G. Levi and T. Hassncer, "Age and gender classification using convolutional neural networks," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, vol. 2015-Octob, pp. 34–42, jun 2015.

[74] A. Janowczyk and A. Madabhushi, "Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases," in *Journal of Pathology Informatics*, vol. 7, p. 29, 2016.

[75] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," *MM 2014 - Proceedings of the 2014 ACM Conference on Multimedia*, pp. 675–678, 2014.

[76] W. Li, P. Cao, D. Zhao, and J. Wang, "Pulmonary Nodule Classification with Deep Convolutional Neural Networks on Computed Tomography Images," *Computational and Mathematical Methods in Medicine*, vol. 2016, p. 6215085, dec 2016.

[77] G. Kang, K. Liu, B. Hou, and N. Zhang, "3D multi-view convolutional neural networks for lung nodule classification," *PLoS ONE*, vol. 12, p. e0188290, mar 2017.

[78] B. Van Ginneken, A. A. Setio, C. Jacobs, and F. Ciompi, "Off-the-shelf convolutional neural network features for pulmonary nodule detection in computed tomography scans," in *Proceedings - International Symposium on Biomedical Imaging*, vol. 2015-July, pp. 286–289, 2015.

[79] C. Li, Y. Diao, H. Ma, and Y. Li, "A Statistical PCA Method for face recognition," in *Proceedings - 2008 2nd International Symposium on Intelligent Information Technology Application, IITA 2008*, vol. 3, pp. 376–380, dec 2008.

[80] R. W. Johnson, *An Introduction to the Bootstrap*, vol. 23 of *Monographs on statistics and applied probability (Series) ; 57*. New York: Chapman & Hall, 2001.

[81] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2019," *CA: A Cancer Journal for Clinicians*, vol. 69, pp. 7–34, jan 2019.

[82] T. M. Kolb, J. Lichy, and J. H. Newhouse, "Comparison of the performance of screening mammography, physical examination, and breast US and evaluation of factors that influence them: An analysis of 27,825 patient evaluations," *Radiology*, vol. 225, no. 1, pp. 165–175, 2002.

[83] R. F. Brem, M. J. Lenihan, J. Lieberman, and J. Torrente, "Screening breast ultrasound: Past, present, and future," *American Journal of Roentgenology*, vol. 204, pp. 234–240, mar 2015.

[84] D. Thigpen, A. Kappler, and R. Brem, "The role of ultrasound in screening dense breasts - A review of the literature and practical solutions for implementation," *Diagnostics*, vol. 8, mar 2018.

[85] R. Rella, P. Belli, M. Giuliani, E. Bufi, G. Carlino, P. Rinaldi, and R. Manfredi, "Automated Breast Ultrasonography (ABUS) in the Screening and Diagnostic Setting: Indications and Practical Use," *Academic Radiology*, vol. 25, pp. 1457–1470, mar 2018.

[86] Y. Wang, S. Jiang, H. Wang, Y. H. Guo, B. Liu, Y. Hou, H. Cheng, and J. Tian, "CAD algorithms for solid breast masses discrimination: Evaluation of the accuracy and interobserver variability," *Ultrasound in Medicine and Biology*, vol. 36, pp. 1273–1281, aug 2010.

[87] K. D. Marcomini, E. F. Fleury, H. Schiabel, and R. M. Nishikawa, "Proposal of semi-automatic classification of breast lesions for strain sonoelastography using a dedicated CAD system," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (A. Tingberg, K. Lång, and P. Timberg, eds.), vol. 9699, pp. 454–460, Springer International Publishing, 2016.

[88] J. C. van Zelst, T. Tan, B. Platel, M. de Jong, A. Steenbakkers, M. Mourits, A. Grivegnee, C. Borelli, N. Karssemeijer, and R. M. Mann, "Improved cancer detection in automated breast ultrasound by radiologists using Computer Aided Detection," *European Journal of Radiology*, vol. 89, pp. 54–59, 2017.

[89] D. C. Moura and M. A. Guevara López, "An evaluation of image descriptors combined with clinical data for breast cancer diagnosis," *International Journal of Computer Assisted Radiology and Surgery*, vol. 8, no. 4, pp. 561–574, 2013.

[90] D. K. Iakovidis, E. G. Keramidas, and D. Maroulis, "Fuzzy local binary patterns for ultrasound texture characterization," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 5112 LNCS, pp. 750–759, 2008.

[91] Y. L. Huang, S. J. Kuo, C. S. Chang, Y. K. Liu, W. K. Moon, and D. R. Chen, "Image retrieval with principal component analysis for breast cancer diagnosis on various ultrasonic systems," *Ultrasound in Obstetrics and Gynecology*, vol. 26, no. 5, pp. 558–566, 2005.

[92] J. Z. Cheng, Y. H. Chou, C. S. Huang, Y. C. Chang, C. M. Tiu, K. W. Chen, and C. M. Chen, "Computer-aided US diagnosis of breast lesions by using cell-based contour grouping," *Radiology*, vol. 255, no. 3, pp. 746–754, 2010.

[93] S. AhmedMedjahed, T. Ait Saadi, and A. Benyettou, "Breast Cancer Diagnosis by using k-Nearest Neighbor with Different Distances and Classification Rules," *International Journal of Computer Applications*, vol. 62, no. 1, pp. 1–5, 2013.

[94] H. Rajaguru and S. Kumar Prabhakar, "Bayesian linear discriminant analysis for breast cancer classification," *Proceedings of the 2nd International Conference on Communication and Electronics Systems, ICCES 2017*, vol. 2018-Janua, pp. 266–269, 2018.

[95] A. Sadeghi-Naini, H. Suraweera, W. T. Tran, F. Hadizad, G. Bruni, R. F. Rastegar, B. Curpen, and G. J. Czarnota, "Breast-Lesion Characterization using Textural Features of Quantitative Ultrasound Parametric Maps," *Scientific Reports*, vol. 7, no. 1, p. 13638, 2017.

[96] S. Han, H. K. Kang, J. Y. Jeong, M. H. Park, W. Kim, W. C. Bang, and Y. K. Seong, "A deep learning framework for supporting the classification of breast lesions in ultrasound images," *Physics in Medicine and Biology*, vol. 62, no. 19, pp. 7714–7728, 2017.

[97] T. C. Chiang, Y. S. Huang, R. T. Chen, C. S. Huang, and R. F. Chang, "Tumor detection in automated breast ultrasound using 3-D CNN and prioritized candidate aggregation," *IEEE Transactions on Medical Imaging*, vol. 38, no. 1, pp. 240–249, 2019.

[98] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[99] M. Byra, "Discriminant analysis of neural style representations for breast lesion classification in ultrasound," *Biocybernetics and Biomedical Engineering*, vol. 38, no. 3, pp. 684–690, 2018.

[100] M. Byra, T. Sznajder, D. Korzinek, H. Piotrzkowska-Wroblewska, K. Dobruch-Sobczak, A. Nowicki, and K. Marasek, "Impact of Ultrasound Image Reconstruction Method on Breast Lesion Classification with Deep Learning," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11867 LNCS, pp. 41–52, 2019.

[101] K. J. Geras, S. Wolfson, Y. Shen, N. Wu, S. G. Kim, E. Kim, L. Heacock, U. Parikh, L. Moy, and K. Cho, "High-Resolution Breast Cancer Screening with Multi-View Deep Convolutional Neural Networks," *CoRR*, vol. abs/1703.0, 2017.

[102] M. Byra, G. Styczynski, C. Szmigielski, P. Kalinowski, Ł. Michałowski, R. Paluszkiewicz, B. Ziarkiewicz-Wróblewska, K. Zieniewicz, P. Sobieraj, and A. Nowicki, "Transfer learning with deep convolutional neural network for liver steatosis assessment in ultrasound images," *International Journal of Computer Assisted Radiology and Surgery*, vol. 13, no. 12, pp. 1895–1903, 2018.

[103] J. Xie, R. Liu, J. Luttrell, and C. Zhang, "Deep learning based analysis of histopathological images of breast cancer," *Frontiers in Genetics*, vol. 10, no. FEB, p. 80, 2019.

[104] M. Lin, Q. Chen, and S. Yan, "Network in network," *2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings*, pp. 1–10, 2014.

[105] H. Azizpour, A. S. Razavian, J. Sullivan, A. Maki, and S. Carlsson, "From generic to specific deep representations for visual recognition," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, vol. 2015-Octob, pp. 36–45, 2015.

[106] J. Arevalo, F. A. González, R. Ramos-Pollán, J. L. Oliveira, and M. A. Guevara Lopez, "Representation learning for mammography mass lesion classification with convolu-

tional neural networks," *Computer Methods and Programs in Biomedicine*, vol. 127, pp. 248–257, 2016.

[107] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[108] S. and technology Group, "Python & SciPy Modules SciPy Open Source Scientific Tools for Python," 2007.

[109] M. Byra, T. Sznajder, D. Korzinek, H. Piotrzkowska-Wroblewska, K. Dobruch-Sobczak, A. Nowicki, and K. Marasek, "Impact of Ultrasound Image Reconstruction Method on Breast Lesion Classification with Deep Learning," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11867 LNCS, pp. 41–52, apr 2019.

[110] T. Xiao, L. Liu, K. Li, W. Qin, S. Yu, and Z. Li, "Comparison of Transferred Deep Neural Networks in Ultrasonic Breast Masses Discrimination," *BioMed Research International*, vol. 2018, 2018.

[111] L. Chen, Y. Chen, X. H. Diao, L. Fang, Y. Pang, A. Q. Cheng, W. P. Li, and Y. Wang, "Comparative Study of Automated Breast 3-D Ultrasound and Handheld B-Mode Ultrasound for Differentiation of Benign and Malignant Breast Masses," *Ultrasound in Medicine and Biology*, vol. 39, pp. 1735–1742, oct 2013.

[112] F. Y. Zheng, L. X. Yan, B. J. Huang, H. S. Xia, X. Wang, Q. Lu, C. X. Li, and W. P. Wang, "Comparison of retraction phenomenon and BI-RADS-US descriptors in differentiating benign and malignant breast masses using an automated breast volume scanner," *European Journal of Radiology*, vol. 84, pp. 2123–2129, nov 2015.

[113] H. J. de Koning, C. M. van der Aalst, P. A. de Jong, E. T. Scholten, K. Nackaerts, M. A. Heuvelmans, J.-W. J. Lammers, C. Weenink, U. Yousaf-Khan, N. Horeweg, S. van 't Westeinde, M. Prokop, W. P. Mali, F. A. Mohamed Hoesein, P. M. van Ooijen,

J. G. Aerts, M. A. den Bakker, E. Thunnissen, J. Verschakelen, R. Vliegenthart, J. E. Walter, K. ten Haaf, H. J. Groen, and M. Oudkerk, "Reduced Lung-Cancer Mortality with Volume CT Screening in a Randomized Trial," *New England Journal of Medicine*, vol. 382, pp. 503–513, feb 2020.

[114] C. R. Bailey, A. M. Bailey, A. S. McKenney, and C. R. Weiss, "Understanding and Appreciating Burnout in Radiologists," *Radiographics*, vol. 42, pp. E137–E139, jul 2022.

[115] K. Murphy, B. van Ginneken, A. M. Schilham, B. J. de Hoop, H. A. Gietema, and M. Prokop, "A large-scale evaluation of automatic pulmonary nodule detection in chest CT using local image features and k-nearest-neighbour classification," *Medical Image Analysis*, vol. 13, pp. 757–770, oct 2009.

[116] H. R. Roth, L. Lu, J. Liu, J. Yao, A. Seff, K. Cherry, L. Kim, and R. M. Summers, "Improving Computer-Aided Detection Using Convolutional Neural Networks and Random View Aggregation," *IEEE Transactions on Medical Imaging*, vol. 35, pp. 1170–1181, may 2016.

[117] T. Sandor, D. Metcalf, and Y. J. Kim, "Segmentation of brain CT images using the concept of region growing," *International Journal of Bio-Medical Computing*, vol. 29, pp. 133–147, nov 1991.

[118] X. Ye, G. Beddoe, and G. Slabaugh, "Automatic graph cut segmentation of lesions in CT using mean shift superpixels," *International Journal of Biomedical Imaging*, vol. 2010, 2010.

[119] R. Hemalatha, T. Thamizhvani, A. J. A. Dhivya, J. E. Joseph, B. Babu, and R. Chandrasekaran, "Active Contour Based Segmentation Techniques for Medical Image Analysis," *Medical and Biological Image Analysis*, jul 2018.

[120] J. Long, E. Shelhamer, and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," *CoRR*, vol. abs/1411.4, 2014.

[121] K. Suzuki, Y. Otsuka, Y. Nomura, K. K. Kumamaru, R. Kuwatsuru, and S. Aoki, "Development and Validation of a Modified Three-Dimensional U-Net Deep-Learning

Model for Automated Detection of Lung Nodules on Chest CT Images From the Lung Image Database Consortium and Japanese Datasets," *Academic Radiology*, vol. 29, pp. S11–S17, feb 2022.

[122] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, B. Glocker, and D. Rueckert, "Attention U-Net: Learning Where to Look for the Pancreas," *CoRR*, vol. abs/1804.0, 2018.

[123] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 2017-Decem, pp. 5999–6009, jun 2017.

[124] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation," feb 2021.

[125] D. J. Zhang, K. Li, Y. Wang, Y. Chen, S. Chandra, Y. Qiao, L. Liu, and M. Z. Shou, "MorphMLP: An Efficient MLP-Like Backbone for Spatial-Temporal Representation Learning," pp. 230–248, nov 2022.

[126] Z. Zhang, Q. Liu, and Y. Wang, "Road Extraction by Deep Residual U-Net," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 5, pp. 749–753, 2018.

[127] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11045 LNCS, pp. 3–11, jul 2018.

[128] J. Choe, S. M. Lee, Y. Ahn, C. H. Kim, J. B. Seo, and H. Y. Lee, "Characteristics and outcomes of anterior mediastinal cystic lesions diagnosed on chest MRI: implications for management of cystic lesions," *Insights into Imaging*, vol. 13, pp. 1–12, dec 2022.

[129] J. B. Ackman, W. Chintanapakdee, D. P. Mendoza, M. C. Price, M. Lanuti, and J. A. O. Shepard, "Longitudinal CT and MRI characteristics of unilocular thymic cysts," *Radiology*, vol. 301, no. 2, pp. 443–454, 2021.

[130] R. Kikinis, S. D. Pieper, and K. G. Vosburgh, "3D Slicer: A Platform for Subject-Specific Image Analysis, Visualization, and Clinical Support," *Intraoperative Imaging and Image-Guided Therapy*, pp. 277–289, 2014.

[131] W. Wang, E. Xie, X. Li, D. P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "PVT v2: Improved baselines with Pyramid Vision Transformer," *Computational Visual Media*, vol. 8, pp. 415–424, jun 2022.

[132] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer Normalization," jul 2016.

[133] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *4th International Conference on Learning Representations, ICLR 2016 - Conference Track Proceedings*, 2016.

[134] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11211 LNCS, pp. 3–19, 2018.

[135] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-Excitation Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, pp. 2011–2023, sep 2020.

[136] F. Milletari, N. Navab, and S. A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," *Proceedings - 2016 4th International Conference on 3D Vision, 3DV 2016*, pp. 565–571, jun 2016.

[137] N. K. Tomar, A. Shergill, B. Rieders, U. Bagci, and D. Jha, "TransResU-Net: Transformer based ResU-Net for Real-Time Colonoscopy Polyp Segmentation," jun 2022.

[138] G. Tong, Y. Li, H. Chen, Q. Zhang, and H. Jiang, "Improved U-NET network for pulmonary nodules segmentation," *Optik*, vol. 174, pp. 460–469, dec 2018.

[139] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "MixUp: Beyond empirical risk minimization," *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*, oct 2018.

[140] H. Sui, L. Liu, X. Li, P. Zuo, J. Cui, and Z. Mo, "CT-based radiomics features analysis for predicting the risk of anterior mediastinal lesions," *Journal of Thoracic Disease*, vol. 11, no. 5, pp. 1809–1818, 2019.

[141] L. Yang, W. Cai, X. Yang, H. Zhu, Z. Liu, X. Wu, Y. Lei, J. Zou, B. Zeng, X. Tian, R. Zhang, H. Luo, and Y. Zhu, "Development of a deep learning model for classifying thymoma as Masaoka-Koga stage I or II via preoperative CT images," *Annals of Translational Medicine*, vol. 8, no. 6, pp. 287–287, 2020.

[142] Z. Liu, Y. Zhu, Y. Yuan, L. Yang, K. Wang, M. Wang, X. Yang, X. Wu, X. Tian, R. Zhang, B. Shen, H. Luo, H. Feng, S. Feng, and Z. Ke, "3D DenseNet Deep Learning Based Preoperative Computed Tomography for Detecting Myasthenia Gravis in Patients With Thymoma," *Frontiers in Oncology*, vol. 11, p. 631964, may 2021.

[143] D. Linsley, J. Kim, V. Veerabadran, C. Windolf, and T. Serre, "Learning long-range spatial dependencies with horizontal gated recurrent units," *Advances in Neural Information Processing Systems*, vol. 2018-Decem, pp. 152–164, may 2018.

[144] 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, may 2016.

[145] 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, pp. 915–931, feb 2011.

[146] R. Sun and Y. Pang, "Efficient lung cancer image classification and segmentation algorithm based on improved swin transformer," 2022.

[147] S. Ramesh, S. Sasikala, S. Gomathi, V. Geetha, and V. Anbumani, "Segmentation and classification of breast cancer using novel deep learning architecture," *Neural Computing and Applications*, vol. 34, pp. 16533–16545, oct 2022.