**POLYTECHNIQUE MONTRÉAL**
affiliée à l'Université de Montréal

# Early-Stage Lung Cancer Detection Using Deep Leaning Algorithms

**YADOLLAH ZAMANIDOOST**
Département de génie informatique et génie logiciel

Thèse présentée en vue de l'obtention du diplôme de *Philosophiæ Doctor*
Génie informatique

Octobre 2025

**POLYTECHNIQUE MONTRÉAL**
affiliée à l'Université de Montréal

Cette thèse intitulée :

**Early-Stage Lung Cancer Detection Using Deep Leaning Algorithms**

présentée par **Yadollah ZAMANIDOOST**
en vue de l'obtention du diplôme de *Philosophiæ Doctor*
a été dûment acceptée par le jury d'examen constitué de :

**Farida CHERIET**, présidente
**Tarek OULD BACHIR**, membre et directeur de recherche
**Sylvain MARTEL**, membre et codirecteur de recherche
**Hervé LOMBAERT**, membre
**Carlos VAZQUES**, membre externe

## ACKNOWLEDGEMENTS

First and foremost, I would like to express my heartfelt thanks to my supervisor, Tarek Ould-Bachir, whose steadfast support, insightful feedback, and patient guidance were essential to the completion of this thesis. His expertise, encouragement, and belief in my work inspired me to grow as a researcher and to pursue bold ideas with confidence. I am especially grateful for the trust he placed in me and for the invaluable discussions we had throughout these years.

I am also deeply indebted to my co-supervisor, Sylvain Martel, whose generous financial support and mentorship made this research feasible. His commitment to fostering high-impact research and his unwavering encouragement were crucial during the most challenging phases of my Ph.D. His continuous support extended far beyond academic supervision—he was a pillar of strength, especially when I needed it most. I am truly honored to have had the opportunity to work under his guidance.

This research would not have been possible without the help of two exceptional interns, Nada Alami-Chentoufi and Matis Rivon. Their dedication and assistance in data collection, annotation, and preliminary analysis played a significant role in several stages of this work. I am thankful for their diligence, teamwork, and collaborative spirit, which helped move the project forward and enriched the overall research experience. I would also like to acknowledge all my colleagues and friends who were part of this journey in different ways—through technical discussions, emotional support, and countless hours spent together in the lab and beyond.

Finally, and most importantly, I owe the deepest gratitude to my wife, whose unwavering love, patience, and constant support were my anchor throughout the ups and downs of this demanding path. She stood by me during the most stressful moments, offered encouragement when I doubted myself, and celebrated every small victory along the way. I am forever grateful for her strength and companionship. To my family, especially my parents, thank you for your endless love and encouragement. Your belief in my potential and your sacrifices over the years laid the foundation for everything I have achieved. This accomplishment is as much yours as it is mine.

To all those who supported me on this journey, thank you from the bottom of my heart.

# RÉSUMÉ

Le cancer du poumon demeure la principale cause de mortalité liée au cancer dans le monde, et sa détection précoce est essentielle pour améliorer les taux de survie des patients. Les systèmes d'aide au diagnostic assisté par ordinateur (CAD), alimentés par l'apprentissage profond (DL), offrent des outils prometteurs pour aider les radiologues à identifier les nodules pulmonaires à un stade précoce. Cependant, des défis tels qu'une sensibilité limitée aux petits nodules, un taux élevé de faux positifs, un manque d'interprétabilité et des inefficacités computationnelles continuent de freiner leur adoption clinique à grande échelle.

Cette thèse présente une série de contributions basées sur l'apprentissage profond visant à relever ces défis et à améliorer l'efficacité de la détection précoce du cancer du poumon. La recherche est structurée selon quatre axes principaux : l'amélioration de la sensibilité, la réduction des faux positifs, l'efficacité computationnelle et l'interprétabilité clinique. Dans ce cadre, quatre nouvelles architectures de détection sont proposées et évaluées.

Premièrement, un mécanisme amélioré de propositions de régions est introduit en modifiant les couches d'extraction de caractéristiques de VGG16, ce qui améliore le rappel pour les petits nodules. Deuxièmement, un réseau de neurones convolutif multi-échelle optimisé (OMS-CNN) est développé à l'aide de stratégies métaheuristiques — Harmony Search et Beetle Antennae Search — pour une configuration et une initialisation efficaces des couches. Troisièmement, ce cadre est étendu avec des modules à double attention, un mécanisme DA-RoIPooling et des ensembles de Transformers Swin 3D afin de réduire les faux positifs tout en maintenant la sensibilité. Enfin, un modèle hybride interprétable est proposé, intégrant des connaissances anatomiques issues d'un U-Net préentraîné dans le flux CNN, améliorant à la fois la précision diagnostique et la transparence.

Tous les modèles sont rigoureusement évalués à l'aide de jeux de données de référence tels que LUNA16 et PN9. Les résultats montrent des améliorations constantes du score CPM, de la sensibilité à des taux de faux positifs cliniquement pertinents et de la généralisation entre ensembles de données. Des études d'ablation confirment la contribution de chaque amélioration architecturale. Notamment, le modèle final atteint un score CPM de 0,9112 sur LUNA16 et démontre une forte capacité de généralisation sur PN9, soulignant son potentiel clinique.

Cette thèse se conclut par l'identification de limitations clés, telles que la diversité des données et les changements de domaine, et propose des pistes futures, notamment l'intégration de techniques avancées d'intelligence artificielle explicable. Collectivement, les contributions de ce travail posent les bases de systèmes CAD précis, interprétables et déployables, favorisant leur intégration dans les flux cliniques courants et contribuant à l'amélioration des résultats pour les patients.

# ABSTRACT

Lung cancer remains the leading cause of cancer-related mortality worldwide, with early detection being critical to improving patient survival rates. Computer-Aided Diagnosis (CAD) systems, powered by deep learning (DL), offer promising tools to assist radiologists in identifying pulmonary nodules at early stages. However, challenges such as limited sensitivity to small nodules, high false-positive rates, lack of interpretability, and computational inefficiencies continue to hinder widespread clinical adoption.

This thesis presents a series of deep learning-based contributions designed to address these challenges and improve the effectiveness of early-stage lung cancer detection. The research is organized along four primary axes: sensitivity improvement, false-positive reduction, computational efficiency, and clinical interpretability. To this end, four novel detection frameworks are proposed and evaluated.

First, an enhanced region proposal mechanism is introduced by modifying VGG16's feature extraction layers, improving recall for small nodules. Second, an Optimized Multi-Scale Convolutional Neural Network (OMS-CNN) is developed using metaheuristic strategies—Harmony Search and Beetle Antennae Search—for efficient layer configuration and initialization. Third, the framework is extended with dual-attention modules, DA-RoIPooling, and 3D Swin Transformer ensembles to reduce false positives while preserving sensitivity. Finally, an interpretable hybrid model is proposed by integrating anatomical priors from a pretrained U-Net into the CNN stream, enhancing both diagnostic accuracy and transparency.

All models are rigorously evaluated using benchmark datasets such as LUNA16 and PN9. Results show consistent improvements in CPM score, sensitivity at clinically relevant false positive rates, and cross-dataset generalization. Ablation studies confirm the contribution of each architectural enhancement. Notably, the final model achieves a CPM of 0.9112 on LUNA16 and demonstrates strong generalization on PN9, highlighting its clinical potential.

This thesis concludes by identifying key limitations, such as data diversity and domain shift, and proposes future directions including the integration of advanced explainable AI techniques. Collectively, the contributions of this work lay a foundation for accurate, interpretable, and deployable CAD systems that support integration into routine clinical workflows and ultimately improve patient outcomes.

## TABLE OF CONTENTS

## LIST OF TABLES

## LIST OF FIGURES

# LIST OF SYMBOLS AND ACRONYMS

| | |
|---|---|
| CT | Computed Tomography |
| CXR | Chest Radiograp |
| MRI | Magnetic Resonance Imagin |
| PET | Positron Emission Tomography |
| CAD | Computer-Aided Detection |
| CNN | Convolutional Neural Networks |
| PSF-HS | Parameter-Setting-Free Harmony Search |
| BAS | Beetle Antenna Search |
| DCNN | Deep Convolutional Neural Network |
| Faster R-CNN | Faster Region-based Convolutional Neural Network |
| OMS-CNN | Optimizes Multi-Scale CNN |
| DA OMS-CNN | Dual Attention OMS-CNN |
| WHO | World Health Organization |
| RPN | Region Proposal Network |
| IoU | Intersection Over Union |
| RoI | Region of Intrest |
| TCIA | The Cancer Imaging Archive |
| LIDC | Lung Image Database Consortium |
| LUNA | Lung Nodule Analysis |
| NLST | National Lung Screening Trial |
| ROC | Receiver Operating Characteristic |
| FROC | Free-Response Operating Characteristic |
| CPM | Competition Performance Metric |
| HU | Hounsfield Unit |
| CNDET | Candidate Nodule Detection |
| FPRED | False Positive Reduction |
| WP | Work Package |
| XAI | Explainable AI |
| Grad-CAM | Gradient-weighted Class Activation Mapping |
| LIME | Local Interpretable Model-Agnostic Explanations |
| SHAP | SHapley Additive ExPlanations |
| AAG | Anatomical Attention Gate |
| SwinT | Shift Window Transformer |

# CHAPTER 1 INTRODUCTION

This dissertation is submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Computer Engineering at Polytechnique Montréal. The research was conducted between May 2021 and June 2025 at Polytechnique Montréal (Montréal, Québec, Canada). The work focuses on the early-stage detection of lung cancer using deep learning methodologies, with a particular emphasis on feature extraction, multi-scale analysis, false-positive reduction, and model interpretability in CT scan interpretation. This thesis comprises four peer-reviewed research papers that explore various strategies to enhance the accuracy, sensitivity, and clinical transparency of lung nodule detection systems. The following chapters provide background, methodological details, experimental results, and a discussion of the clinical and research contributions of each study.

## 1.1 Context and Motivation

Lung cancer remains one of the most lethal types of cancer, accounting for approximately 27% of all cancer-related deaths worldwide [1]. In a comprehensive 2015 study, the number of cancer-related fatalities reached 589,430, with lung cancer identified as the leading cause [1]. Despite advancements in medical imaging and oncology, the prognosis for lung cancer remains poor, particularly when diagnosed at advanced stages. However, early-stage detection can significantly improve survival rates, increasing the five-year survival rate from approximately 14% to 49% [2].

The early identification of pulmonary nodules—small, often cancerous growths in lung tissue—is essential to improving outcomes in patients with lung cancer. Computed tomography (CT) imaging has become the standard screening tool due to its high resolution and ability to capture subtle structural changes in the lung. Among various diagnostic techniques such as chest radiograph (CXR), magnetic resonance imaging (MRI), positron emission tomography (PET), sputum cytology, and breath analysis, CT is preferred for its balance of speed, accuracy, and radiation safety [3]. Nevertheless, interpreting CT scans remains a challenging task, particularly in detecting nodules as small as 3 mm in diameter, which are often present in early-stage cancer. Manual analysis is not only time-consuming but also susceptible to inter-observer variability and human error [3].

To address these limitations, computer-aided detection (CAD) systems have been developed to support radiologists in identifying and classifying pulmonary nodules. These systems typically consist of two main stages: (1) nodule candidate detection, aimed at maximizing sensitivity, and (2) false positive reduction, focused on enhancing specificity and overall accuracy. Although recent advances in deep learning, especially convolutional neural networks (CNNs), have significantly improved the performance of CAD systems in medical imaging, the detection of small nodules remains a major challenge due to their subtle appearance and similarity to surrounding tissues.

With the recent rise of deep learning and improvements in computational power, object detection techniques have emerged as powerful tools for medical image analysis. These methods have opened new avenues for automatically identifying anatomical structures and abnormalities within complex imaging data. In particular, the Faster Region-based Convolutional Neural Network (Faster R-CNN) framework [4] has gained considerable attention for its effectiveness in object detection tasks. Faster R-CNN is a two-stage detector that first uses a Region Proposal Network (RPN) to generate candidate object regions, followed by a convolutional network to classify and refine these regions. This architecture enables precise localization of small and low-contrast objects, such as pulmonary nodules, making it highly suitable for early-stage cancer detection in CT scans. Motivated by these advancements, this research investigates how deep object detection models, such as Faster R-CNN, can be adapted and optimized to enhance the performance of CAD systems, with a particular focus on detecting small pulmonary nodules in CT images.

Beyond detection accuracy, a growing challenge in medical AI systems is the need for transparency and explainability. Clinicians require not only reliable predictions but also a clear understanding of how those predictions are made [5]. To meet this need, this thesis introduces a hybrid interpretable architecture called SwinT-CNN, which combines the global modeling power of Swin Transformers with the local precision of CNNs, guided by anatomical priors from segmentation. This interpretable design allows the model to focus on clinically meaningful regions. It offers visual explanations of its decisions, making it a more trustworthy tool for early-stage lung cancer detection.

## 1.2 Problem Statement

Early detection of lung cancer remains a major clinical challenge despite significant advances in imaging technologies and artificial intelligence. Small pulmonary nodules, which often indicate the early stages of lung cancer, are particularly difficult to detect and characterize accurately. Radiologists frequently struggle to distinguish these nodules from surrounding anatomical structures in CT scans due to their small size, low contrast, and variability in shape and appearance. This diagnostic complexity leads to missed detections or false positives, both of which can have profound implications for patient outcomes. Therefore, developing robust and automated systems that can detect small nodules with high sensitivity and precision is a critical and ongoing research priority.

Deep learning-based object detection models have shown considerable promise in addressing this problem. In particular, the Faster R-CNN framework has emerged as one of the most effective solutions for detecting small objects within complex backgrounds [4]. However, its performance in the medical domain—especially in lung CT scans—faces several persistent limitations. The standard feature extraction process may not adequately capture the subtle and fine-grained details of small nodules, resulting in reduced sensitivity. Moreover, the classification stage of the network often yields a high number of false positives, which adversely affects the model's precision and limits its clinical applicability.

Furthermore, most existing CAD models operate as black-box classifiers, which limits their acceptance in real-world clinical workflows. The inability to visually explain AI decisions reduces radiologists' trust and complicates integration into diagnostic routines. The newly proposed SwinT-CNN addresses this challenge by incorporating segmentation-derived anatomical priors and visual saliency mechanisms to enhance model transparency and interpretability.

This thesis aims to address these challenges by proposing a series of architectural and algorithmic enhancements to the Faster R-CNN framework, tailored explicitly for early-stage lung nodule detection. First, a multi-scale feature extraction strategy is optimized using metaheuristic algorithms, such as advanced PSF-HS [6] and BAS [7], to enhance the network's ability to localize nodules of varying sizes. Second, a combination of 3D deep convolutional networks is integrated into the false positive reduction stage, utilizing volumetric spatial context to improve discrimination between true nodules and non-nodular structures. Third, a dual-attention mechanism is introduced at both the feature extraction and classification stages to refine spatial and channel-level focus, enhancing sensitivity and reducing misclassifications. Finally, the adoption of 3D Swin Transformers in the false positive reduction phase enables richer volumetric representation and further minimizes false detections.

Despite the success of conventional Faster R-CNN in general object detection tasks, its adaptation to the medical imaging domain—where detecting minute, ambiguous targets is crucial—remains limited. This research addresses this gap by systematically improving each stage of the Faster R-CNN pipeline to enhance both the sensitivity for small nodule detection and the precision through effective false positive suppression. The outcome is a more reliable and clinically relevant framework for aiding radiologists in early lung cancer screening.

## 1.3 Research Objectives

In response to the challenges associated with early-stage lung cancer detection (particularly the identification of small pulmonary nodules) and the limitations of conventional Faster R-CNN frameworks, this thesis establishes three primary research objectives, each divided into specific sub-objectives:

1. **Enhancing Sensitivity in Early Detection of Small Lung Nodules**

   - **A1:** Design and develop an optimized multi-scale convolutional neural network (OMS-CNN) to improve the feature extraction capability of the Faster R-CNN framework, especially for small nodules.

   - **A2:** Integrate dual-attention mechanisms into the CNN feature extractor to capture both spatial and channel-level dependencies for more accurate identification of critical regions.

   - **A3:** Employ DA-RoIPooling in the classification stage to refine region-wise feature representation and highlight diagnostically relevant characteristics.

2. **Reducing False Positives and Improving Precision**

   - **B1:** Develop a hybrid false positive reduction module by incorporating multiple 3D convolutional neural networks (3D DCNNs) to leverage volumetric spatial context.

   - **B2:** Utilize a set of 3D Swin Transformers to analyze CT volumes from multiple perspectives, reducing the likelihood of misclassification.

   - **B3:** Introduce attention-guided classification layers to suppress irrelevant background features within candidate regions.

3. **Improving Computational Efficiency and Clinical Applicability**

   - **C1:** Apply metaheuristic optimization techniques—specifically parameter setting-free harmony search (PSF-HS) and beetle antenna search (BAS)—to determine the optimal configuration of composite layers.

   - **C2:** Optimize the initialization of network parameters to accelerate model training and enhance stability.

- **C3:** Evaluate the proposed framework across multiple public datasets (LUNA16 and PN9) to ensure generalizability and readiness for real-world clinical deployment.

4. **Improving Interpretability and Clinical Transparency**

- **D1:** Integrate Swin Transformer modules with CNNs through a dual-path architecture to jointly capture global semantic context and fine-grained local details.

- **D2:** Incorporate anatomical attention gates (AAG) that inject segmentation-derived voxel-level priors from a pretrained 3D U-Net into the CNN branch to enhance transparency and decision explainability.

## 1.4 Novelty and Impact

This thesis is based on four peer-reviewed research articles that collectively propose novel approaches to enhance the early-stage detection of lung cancer using deep learning, with a particular focus on identifying small pulmonary nodules and improving the interpretability of model predictions in CT scans. The contributions of this work encompass architectural enhancements to Faster R-CNN, optimization techniques for enhanced learning efficiency, and advanced attention mechanisms for improved feature localization. Each article directly addresses key research objectives outlined in Section 1.3.

### 1.4.1 Efficient Region Proposal Extraction

Presented in Chapter 4 and published in an IEEE conference, this study explores the limitations of conventional feature extraction in detecting small nodules and proposes a modified VGG16-based feature map generation technique. By combining the final three convolutional layers and integrating a region proposal network (RPN), the approach enhances sensitivity in identifying small nodules. The method demonstrates higher recall rates at various Intersection-over-Union (IoU) thresholds while maintaining robustness against reduced proposal counts. This work addresses Objective A1 and contributes to improving the sensitivity of early detection.

### 1.4.2 Optimized Multi-Scale CNN (OMS-CNN)

Presented in Chapter 5 and published in an IEEE journal, this article introduces an optimized multi-scale convolutional neural network (OMS-CNN) as part of the Faster R-CNN framework. Metaheuristic algorithms, such as parameter-setting-free harmony search (PSF-HS) and beetle antenna search (BAS), are employed to configure and initialize composite convolutional layers. These techniques enhance detection accuracy and computational efficiency by enabling the model to localize candidate nodules with greater accuracy. This contribution addresses objectives A1, B1, and C1.

### 1.4.3 Dual Attention OMS-CNN

Presented in Chapter 6 and published in an MDPI journal, this article extends the previous work by incorporating dual-attention mechanisms into the OMS-CNN backbone and introducing a novel DA-RoIPooling method in the classification stage. Additionally, it integrates multiple 3D Swin Transformers in the false-positive reduction stage to enhance volumetric feature representation. The combined approach improves both sensitivity and precision while significantly reducing false positives. This article addresses objectives A2, A3, B2, and B3.

### 1.4.4 Interpretable Hybrid SwinT-CNN Model

Presented in Chapter 7 and published in the IEEE Transactions on Biomedical Engineering, this study introduces an interpretable dual-path deep learning architecture that combines 3D CNNs with 3D Swin Transformers. The model integrates anatomical attention gates (AAG), which inject voxel-level priors from a pretrained U-Net segmentation model into the CNN stream. The design enhances diagnostic precision and transparency by enabling the model to focus on clinically significant areas. The interpretability is validated through 3D Grad-CAM visualizations, sensitivity, and entropy metrics. This article addresses objectives D1 and D2, contributing to the improvement of model trustworthiness in clinical decision support.

### 1.4.5 Overall Impact

The collective contributions of this thesis advance the state of research in automated lung cancer detection by addressing critical bottlenecks in four key areas: sensitivity, false-positive reduction, computational efficiency, and clinical interpretability. Across the four proposed models, this work presents end-to-end deep learning frameworks that not only enhance detection accuracy but also improve trustworthiness through transparent decision-making processes. The models were thoroughly evaluated on benchmark datasets such as LUNA16 and PN9, demonstrating consistent improvements over existing approaches in both diagnostic performance and generalizability. The outcomes of this thesis provide a robust foundation for the future development of AI-based clinical decision support systems, particularly those intended for early-stage lung cancer screening and diagnosis.

## 1.5 Ethical Statement on the Use of AI Tools

Throughout the preparation of this thesis, ChatGPT, a generative AI tool, was used occasionally as an auxiliary tool to enhance the clarity and readability of the English text. The tool helped improve grammar, rephrase complex sentences, and check language consistency. However, all scientific content, ideas, experimental designs, analyses, and interpretations were fully developed by the author.

The use of ChatGPT was strictly limited to linguistic refinement and formatting assistance, and every generated or modified text was carefully reviewed and validated by the author.

# CHAPTER 2 LITERATURE REVIEW

## 2.1 Introduction

Cancer remains one of the leading causes of death worldwide and, according to the World Health Organization (WHO) [8], is the second most common cause of death globally. Among all types of cancer, lung cancer imposes the greatest burden on healthcare systems due to its high incidence and mortality rates [8].

### 2.1.1 Epidemiology of Lung Cancer

As reported in the 2021 registry [9], lung cancer accounted for over 34,000 cases in males and more than 10,000 cases in females, marking it as the most frequently diagnosed cancer across both genders (Figure 2.1). The figure illustrates both the absolute number of lung cancer cases (in thousands) and their relative proportion among all cancer types for males, females, and the combined population. Notably, the disease constitutes a significantly higher proportion of all cancers in males (10.8%) compared to females (3.7%), reflecting both greater incidence and possibly higher exposure to risk factors such as smoking. These disparities highlight the importance of gender-specific strategies for early detection and prevention.

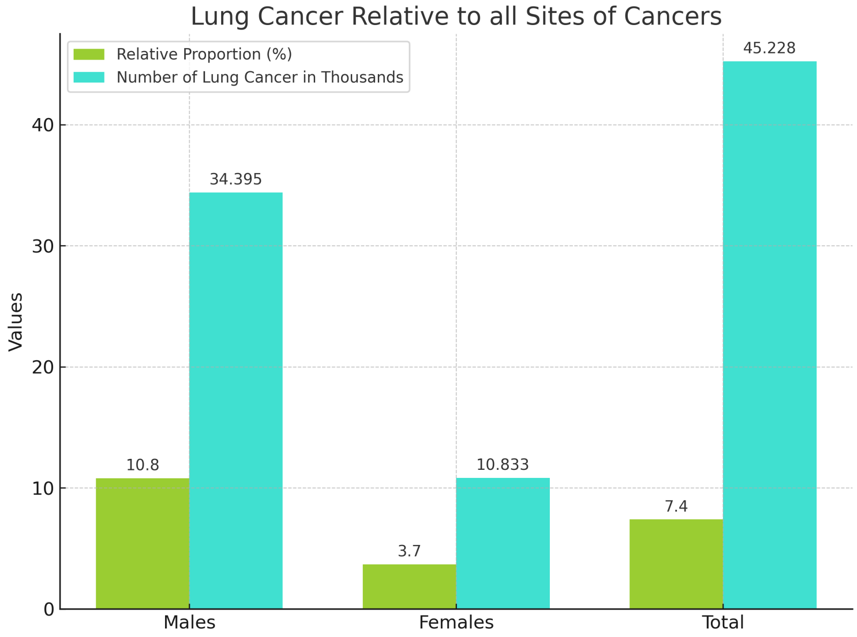


Figure 2.1 Relative proportion and number of lung cancer cases in males and females [9].

As illustrated in Figure 2.2, lung cancer not only ranks among the most frequently diagnosed cancers but also shows a disproportionately high mortality rate (18.4%) compared to its incidence rate (11.6%). This contrast is more pronounced than in other cancers like breast (11.6% incidence vs. 6.6% mortality) or prostate cancer (7.1% vs. 3.8%). The figure clearly highlights that, despite similar or lower incidence rates, lung cancer leads to a higher proportion of deaths, emphasizing its severity and the critical need for early detection and effective treatment strategies.

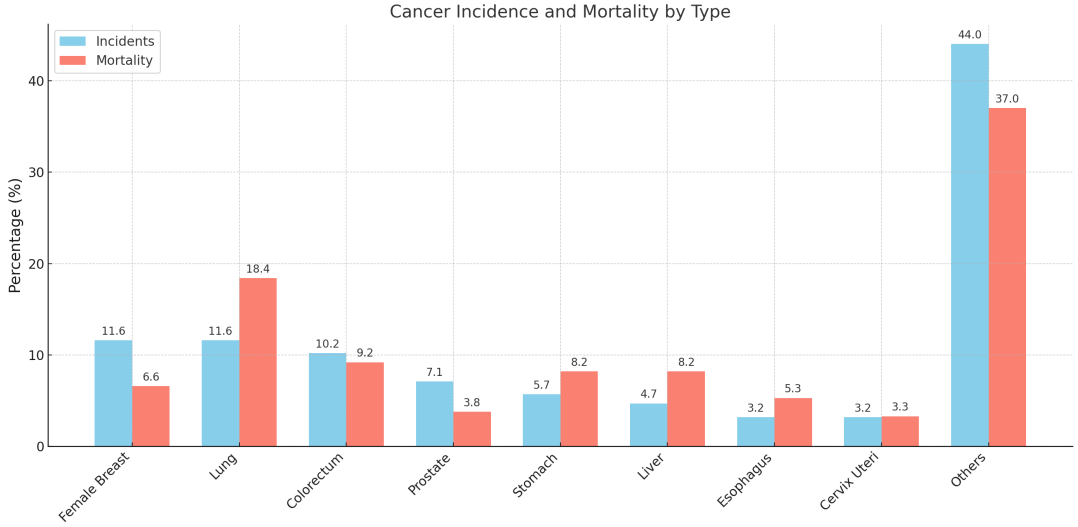


Figure 2.2 Incidence and mortality rates of various cancer types (%) [9].

The significant gap between incidence and mortality is largely attributed to the asymptomatic and insidious nature of lung cancer, which is often diagnosed only in advanced stages. Consequently, primary prevention, screening, and early diagnosis (particularly through the detection and classification of pulmonary nodules) play a vital role in mitigating the physical, emotional, and financial burden on patients while improving treatment outcomes [10].

### 2.1.2 Characteristics and Clinical Relevance of Pulmonary Nodules

Pulmonary nodules are defined as localized abnormalities within the lungs that typically measure from 3 to 30 millimeters in diameter. They may occur as a single lesion or appear in clusters dispersed throughout the pulmonary parenchyma. These nodules demonstrate considerable heterogeneity. Their dimensions can be smaller or larger than 8 millimeters, their geometric form may be rounded, polygonal, or irregular, and their margins can present as smooth, lobulated, or spiculated. In addition, their position varies, with some located centrally, others near the pleural surface, and some adjacent to vascular structures. Variations

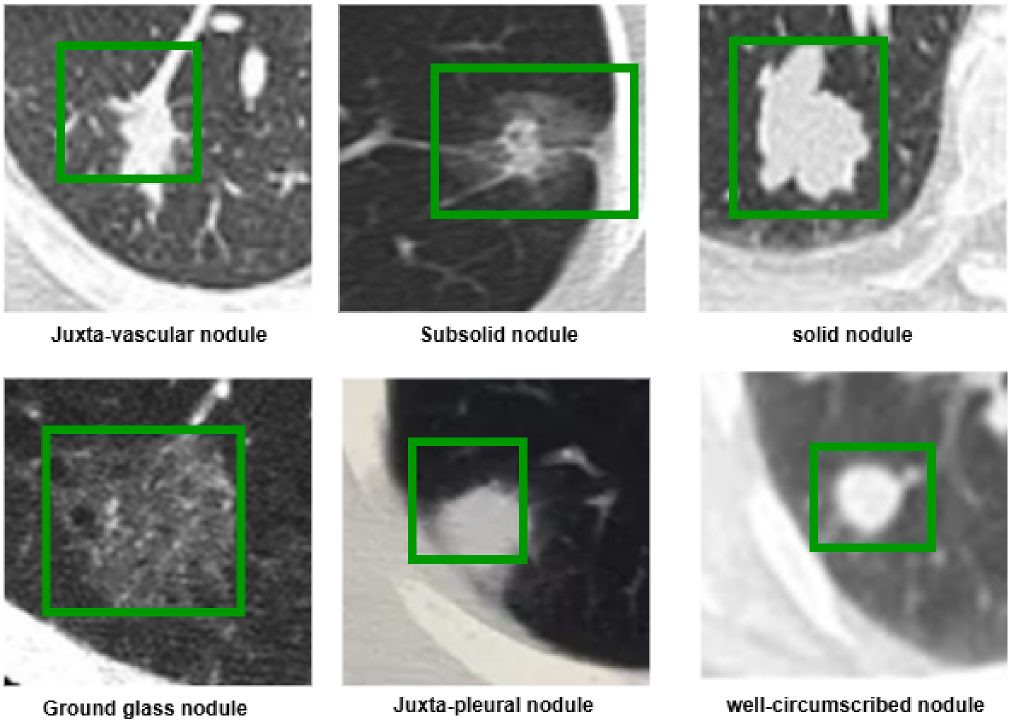


Figure 2.3 Different type of lung nodules [11].

are also observed in their density, which may appear solid, partially solid, or as a ground-glass opacity. Illustrative examples of these different categories are provided in Figure 2.3.

Although many nodules are non-malignant and remain undetected clinically, radiological patterns such as large diameter, subsolid composition, and irregular or lobulated contours are frequently linked to malignant transformation. These characteristics complicate early recognition, especially in asymptomatic cases [12].

Clinical studies consistently demonstrate the importance of timely intervention. When malignant nodules are discovered and surgically treated at an early stage, patients show a five-year survival rate as high as 65 to 80 percent. In contrast, survival drops dramatically to approximately 10 to 15 percent when diagnosis occurs at more advanced stages [13]. For this reason, the prompt detection of malignant pulmonary nodules remains one of the most critical and, at the same time, one of the most challenging aspects of lung cancer management.

### 2.1.3 Medical Imaging for Lung Cancer Screening

Medical imaging plays a pivotal role in the early detection and diagnosis of lung cancer, particularly for individuals at high risk, such as long-term smokers or those with a history of smoking. Among various imaging modalities, low-dose computed tomography (LDCT) has emerged as the most effective screening technique due to its ability to capture high-resolution cross-sectional images with minimal radiation exposure [14]. LDCT enables the identification of small pulmonary nodules that may not be visible on standard chest X-rays, thus improving the chances of early intervention and reducing mortality rates. In addition to LDCT, positron emission tomography (PET) scans provide functional imaging that helps assess metabolic activity of suspicious lesions. PET is often used in combination with CT to enhance diagnostic accuracy and assist in staging the disease. This multimodal approach allows clinicians to make more informed decisions regarding treatment strategies, especially when evaluating the malignancy of nodules detected in early screenings [15].

The primary goal of lung cancer screening is to accurately detect malignant cases while minimizing the risks associated with overdiagnosis, unnecessary treatments, and the psychological burden caused by false-positive findings [16]. One of the most influential factors in achieving this balance is the size of the detected lung nodule, which significantly affects the rate of false positives [17].

### 2.1.4 Computer-Aided Diagnosis Systems

With the widespread adoption of low-dose computed tomography (LDCT) for lung cancer screening, radiologists are increasingly overwhelmed by the substantial volume of CT scans generated during routine screenings. Manual interpretation of such large-scale imaging data is not only labor-intensive but also prone to fatigue-related errors. To address this growing challenge, Computer-Aided Diagnosis (CAD) systems have emerged as a critical tool in automating the detection and interpretation of pulmonary abnormalities [18]. CAD systems are typically categorized into two main components: CADe (Computer-Aided Detection), which focuses on identifying regions of interest such as potential nodules, and CADx (Computer-Aided Diagnosis), which aids in determining the nature and malignancy of the detected lesions. A standard CAD workflow for lung cancer includes three core stages: preprocessing, nodule detection, and classification [19]. During preprocessing, the system performs lung segmentation, noise reduction, and normalization to enhance the quality of the input data. The detection phase then identifies potential nodule candidates—prioritizing high sensitivity, even at the expense of increased false positives. To mitigate this, a dedicated false-positive reduction module is typically applied before proceeding to the final classification stage, where the likelihood of malignancy is estimated. By streamlining this workflow, CAD systems play a pivotal role in supporting early and accurate diagnosis while alleviating the workload of medical professionals.

### 2.1.5 The Role of Artificial Intelligence in Lung Cancer Screening

Recent advancements in artificial intelligence (AI) have introduced powerful tools capable of identifying subtle patterns and anomalies in medical images, patterns that may not be easily perceived by radiologists. These AI-based systems have shown great potential in enhancing early detection of serious health conditions, including lung cancer, by assisting clinicians in evaluating suspicious nodules more effectively. As a result, healthcare professionals are increasingly incorporating AI-driven algorithms into diagnostic workflows to improve accuracy and efficiency in identifying lung abnormalities at early stages [20].

In recent years, deep learning has emerged as a transformative force in the development of advanced CAD systems for pulmonary nodule analysis. Among various techniques, Convolutional Neural Networks (CNNs) have demonstrated remarkable success in numerous computer vision benchmarks, such as ImageNet and MS COCO challenges. Owing to their strong adaptability and feature learning capabilities, a range of CNN-based architectures—including U-Net, Faster R-CNN, Mask R-CNN [4, 21, 22], and RetinaNet—have been extensively adopted for nodule detection and classification tasks. These models significantly enhance the sensitivity and precision of CAD systems, contributing to more reliable medical image analysis.

Object detection within the realm of medical imaging focuses on identifying clinically relevant structures, such as tumors, lesions, or pulmonary nodules, from complex visual data. By leveraging the powerful feature extraction capabilities of deep learning models, particularly CNNs, researchers have enabled automated systems to detect subtle patterns and anomalies that may not be readily visible to the human eye [23]. Furthermore, techniques from computer vision—such as region segmentation, image registration, and 3D reconstruction—enhance the interpretability of imaging data [24]. Three-dimensional modeling, in particular, provides detailed visualizations of anatomical structures, which are essential for surgical planning and clinical decision-making. The integration of object detection and deep learning methodologies into CAD systems not only improves diagnostic accuracy but also reduces radiologist workload and minimizes the potential for human error [25].

To provide a comprehensive understanding of the landscape of lung nodule detection and diagnosis using deep learning techniques, the subsequent sections of this chapter are organized as follows. First, the datasets used in this study are described in detail, including their characteristics and relevance to early-stage lung cancer analysis. This is followed by a discussion of the evaluation metrics commonly used to assess detection performance. The structure of the proposed Computer-Aided Diagnosis (CAD) system is then presented, consisting of three main stages: data preprocessing, lung nodule detection, and false positive reduction.

Each stage is explained in terms of its technical process and clinical importance. Finally, we present a discussion of key limitations and challenges, followed by a summary conclusion of the chapter.

## 2.2 Dataset Description

The development and training of reliable pulmonary nodule detection and classification models require access to large-scale and high-quality annotated CT datasets. Publicly available datasets are therefore critical to advancing research in this domain. These datasets typically contain a combination of medical imaging data, clinical annotations, and patient information related to lung cancer cases. In this section, we present an overview of the most widely used and impactful public datasets that have been instrumental in the progress of computer-aided lung cancer diagnosis systems.

- **The Cancer Imaging Archive (TCIA):** TCIA is a widely used public repository that offers a broad range of medical imaging datasets, including data relevant to lung cancer research. Among its most valuable contributions is the Lung Image Database Consortium and Image Database Resource Initiative (LIDC-IDRI), which provides annotated CT scans of the lungs with detailed markings of nodules. This dataset serves as a foundational benchmark for developing and evaluating computer-aided diagnosis systems in lung cancer. The archive is freely accessible through the official TCIA platform[1] [26].

- **Lung Image Database Consortium (LIDC):** The LIDC dataset is a specialized subset of TCIA dedicated to lung cancer detection and research. It includes 399 thoracic CT scans with detailed annotations of pulmonary nodules provided by multiple radiologists. These annotations enable the training and evaluation of automated lung nodule detection algorithms. LIDC remains one of the most frequently used datasets for benchmarking CAD systems in the early diagnosis of lung cancer[2] [27].

- **Lung Nodule Analysis (LUNA) Challenge:** The LUNA dataset was developed as part of a public challenge aimed at advancing the performance of automated lung nodule detection systems. It consists of 1,018 CT scans with annotated pulmonary nodules from 1,010 patients. LUNA has been widely used in benchmark studies and

[1] https://wiki.cancerimagingarchive.net/display/NBIA/Downloading+TCIA+Images

[2] https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=1966254

machine learning competitions focused on lung cancer diagnosis, offering a standardized platform for performance comparison[3] [28].

- **National Lung Screening Trial (NLST) Dataset:** The NLST dataset originates from a large-scale clinical trial conducted to assess the effectiveness of low-dose CT in reducing lung cancer mortality. It provides a valuable resource for developing and evaluating computer-aided diagnostic tools, containing CT scan data from 1,058 patients diagnosed with lung cancer and 9,310 individuals with non-cancerous nodules[4] [29].
- **PN9 Dataset:** PN9 is a recently released and large-scale pulmonary nodule dataset specifically curated for nodule detection tasks. It includes 8,798 thoracic CT scans and 40,439 annotated nodules across nine common nodule types, making it one of the most comprehensive and diverse public resources available conferences. This wide coverage supports the development of robust algorithms capable of detecting a variety of nodule morphologies, especially in challenging clinical situations [30].

## 2.3 Evaluation Metrics

The commonly used evaluation metrics of pulmonary nodule detection is listed below:

- **Sensitivity and Precision:** In pulmonary nodule detection, sensitivity and precision are two fundamental metrics used to evaluate the effectiveness of a diagnostic system. Sensitivity, also known as recall, measures the model's ability to correctly identify actual positive cases (i.e., true nodules). It is defined as:

  $$\text{Recall(Sensitivity)} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2.1}$$

  A higher sensitivity indicates that the system can detect a greater proportion of real nodules, including small and subtle ones, which is critical for early-stage lung cancer detection.

  On the other hand, precision assesses the proportion of true positives among all the predicted positives. It is defined as:

  $$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2.2}$$

[3] https://academictorrents.com/details/58b053204337ca75f7c2e699082baeb57aa08578

[4] https://cdas.cancer.gov/nlst/

High precision means the system makes fewer false positive predictions, which helps reduce unnecessary follow-up procedures, false alarms, and clinical workload. Balancing sensitivity and precision is essential because increasing one may often reduce the other. Therefore, improving both simultaneously remains a key objective in the development of robust CAD systems.

- **ROC and FROC Curves:** The Receiver Operating Characteristic (ROC) curve and the Free-Response Operating Characteristic (FROC) curve are widely used tools to visualize and evaluate the performance of computer-aided detection (CADe) systems.

  The ROC curve plots sensitivity (true positive rate) against the false positive rate (FPR) at various threshold settings. It helps to understand the trade-off between detecting true nodules and generating false alarms. A model with better performance will have a ROC curve that bends closer to the top-left corner of the graph.

  In contrast, the FROC curve replaces the FPR with the average number of false positives per scan on the X-axis, while keeping sensitivity on the Y-axis. This approach is particularly useful in medical image analysis tasks, where false positives are counted per image or scan rather than as a global rate.

  FROC analysis is more informative in scenarios like pulmonary nodule detection, where the number of nodules per scan can vary. Consequently, it has become the standard evaluation method in many challenges and benchmarks involving lesion or abnormality detection.

- **Competition Performance Metric (CPM):** The Competition Performance Metric (CPM) is a widely adopted evaluation criterion in pulmonary nodule detection challenges. It reflects the average sensitivity of a computer-aided detection (CADe) system at seven predefined false positive rates: 1/8, 1/4, 1/2, 1, 2, 4, and 8 false positives per scan.

  The CPM score is derived from the FROC curve, summarizing the model's ability to maintain high sensitivity while limiting false positives across different operating thresholds. Formally, the CPM is calculated as:

$$\mathrm{CPM} = \frac{1}{N} \sum_{i \in \mathrm{I}} \mathrm{Recall}_{\mathrm{fpr}=i} \tag{2.3}$$

  With $\mathrm{I} = \{0.125, 0.25, 0.5, 1, 2, 4, 8\}$ and where the value of $N$ is set at seven, the variable $fpr$ represents the average number of false positives per scan, while $Recall_{fpr=i}$ signifies the recall rate associated with $fpr = i$.

A higher CPM score indicates better overall system performance in balancing sensitivity and false positives.

## 2.4 Structure of CAD System

CAD systems have become a vital component in supporting the early diagnosis of lung cancer [18]. By utilizing high-resolution thin-slice CT images, CAD systems function as a supplementary tool to assist radiologists in identifying pulmonary nodules more efficiently and consistently. Over the years, various CAD frameworks have been developed, each differing in structure and algorithmic complexity [4]. However, a typical CAD pipeline generally comprises three essential stages:

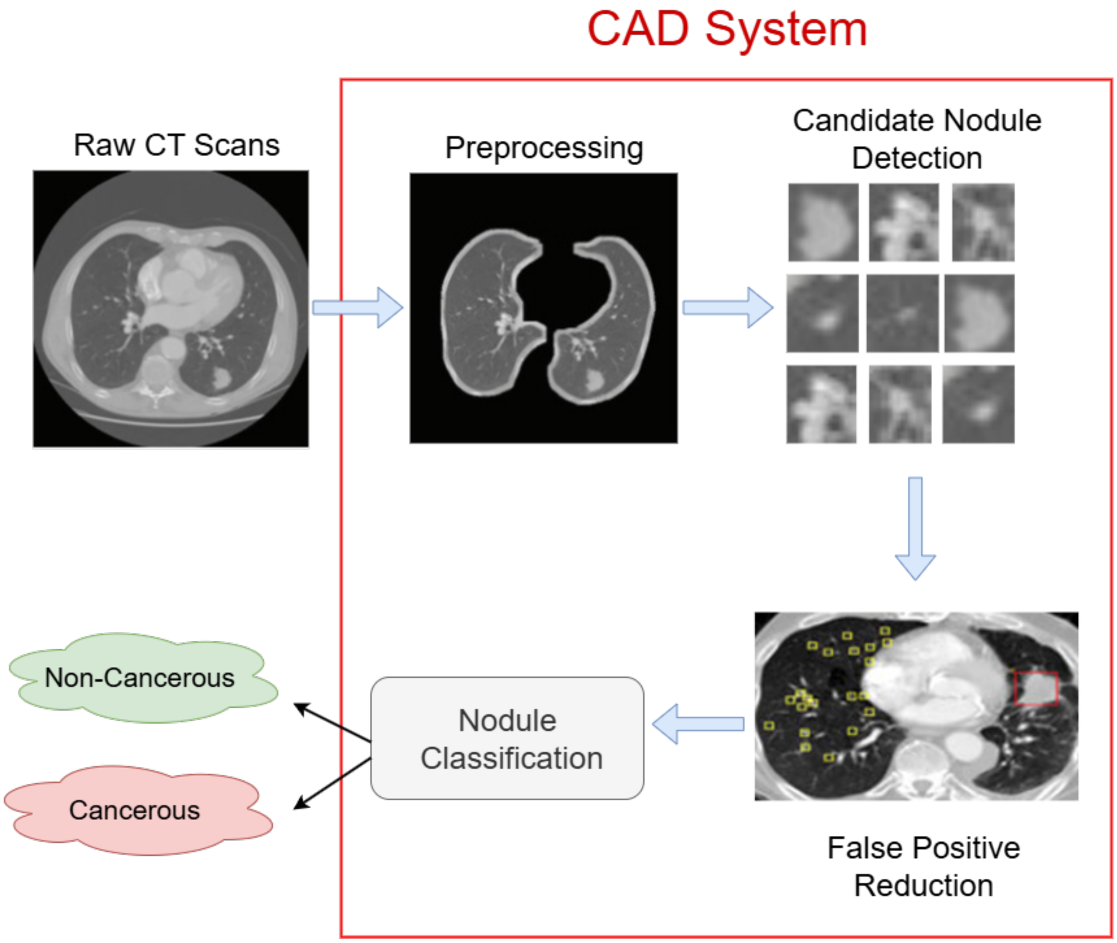


Figure 2.4 The complet workflow of a CAD system

1. **Data Preprocessing:** This stage involves enhancing image quality, segmenting relevant anatomical regions such as lungs, and standardizing the input data to reduce

variability and noise.

2. **Nodule Detection:** It typically includes two sub-phases: (a) *candidate nodule detection*, where potential regions of interest are identified, often with high sensitivity but at the cost of false positives, and (b) *false positive reduction*, which aims to refine the detections and reduce redundancy.

3. **Nodule Classification:** In this phase, detected nodules are analyzed to determine their likelihood of malignancy using a variety of machine learning or deep learning techniques.

An overview of the CAD system workflow is illustrated in Figure. 2.4. The performance of these systems can vary significantly depending on several factors, including CT image acquisition parameters, the diversity in nodule characteristics, and the underlying algorithms used in each component [18]. Recent advancements primarily focus on improving sensitivity and specificity by enhancing the false positive reduction and classification modules. In the subsequent subsections, we provide a comprehensive discussion of each component along with widely adopted algorithms that have demonstrated reliable results on public benchmark datasets.

### 2.4.1 Data Preprocessing

The preprocessing stage plays a pivotal role in the analysis of lung CT scans, as the raw images often contain substantial irrelevant structures that can hinder the performance and reliability of CAD systems. Since pulmonary nodules typically appear within the main lung volume—considered the region of interest (ROI)—it is crucial to eliminate unnecessary components such as surrounding tissues and imaging artifacts. Moreover, preprocessing aims to enhance or preserve clinically meaningful information to improve detection outcomes. One of the fundamental tasks at this stage is lung segmentation, which isolates the lung fields to narrow the search space for subsequent analysis. Studies have shown that incorporating proper lung segmentation methods in the preprocessing pipeline can significantly reduce the number of missed nodules, with improvements ranging from 5% to 17% [31]. Most segmentation approaches rely on the contrast differences in Hounsfield Unit (HU) values between the lung parenchyma and adjacent tissues. Based on the technique used, segmentation methods are commonly categorized into two groups: rule-based approaches and data-driven (machine learning-based) approaches [32].

Among traditional segmentation techniques, rule-based approaches are widely adopted due to their simplicity and effectiveness in specific scenarios. These methods typically rely on

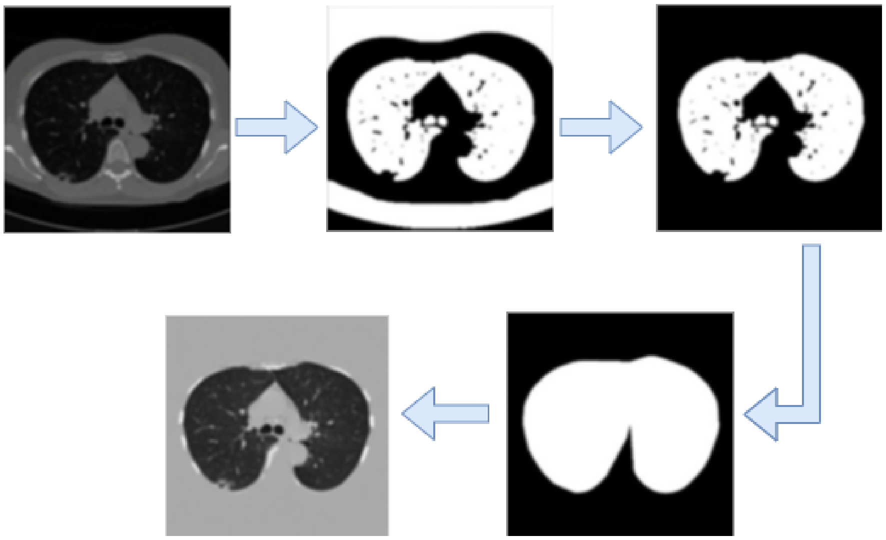


Figure 2.5 Outline of the preprocessing methodology [33].

handcrafted heuristics that exploit predefined thresholds in Hounsfield Unit (HU) values to distinguish lung regions from surrounding tissues. Common steps include thresholding, morphological operations (e.g., erosion and dilation), and connected component analysis to remove irrelevant structures such as bones, airways, or chest wall artifacts [21,33,34]. For example, a rule-based pipeline proposed by Liao et al [33]. utilizes a sequence of traditional image processing techniques to effectively segment the lung region. The process begins with the application of a Gaussian filter to reduce noise, followed by thresholding operations based on both intensity and spatial distance to isolate the lung area and exclude non-relevant anatomical structures. Subsequent morphological operations, such as convex hull computation and dilation, are used to refine the binary lung masks and ensure coverage of the entire lung volume. This multi-step approach illustrates how rule-based techniques can be systematically organized to yield accurate region-of-interest extraction, especially in standard CT scan settings. Figure 2.5 shows a complete preprocessing procedure using rule-based approaches. Rule-based techniques are particularly efficient when processing standard thoracic CT scans with relatively consistent acquisition parameters. However, their performance can degrade in the presence of low image contrast, irregular nodule morphology, or inter-patient variability. Despite these limitations, they are still utilized as a baseline or preprocessing step in more advanced systems due to their low computational cost and ease of implementation.

In contrast to rule-based methods, data-driven approaches rely on statistical models or deep learning frameworks trained on large annotated datasets to perform lung segmentation more robustly and adaptively. These techniques have gained significant popularity due to their ability to generalize across diverse CT scan qualities, scanner types, and pathological variations. A widely adopted method is the U-Net architecture, which employs an encoder-decoder structure to learn spatial and semantic features for accurate pixel-level segmentation of lung regions. For instance, Çiçek et al. [35] extended the U-Net to 3D data (3D U-Net), making it well-suited for volumetric CT scan segmentation tasks. Furthermore, Zhou et al. [36] introduced UNet++, an enhanced version with nested skip pathways, improving performance on noisy and low-contrast scans. Additionally, ResUNet [37] and Attention U-Net [38] variants have been proposed to incorporate residual learning and attention mechanisms, further refining the segmentation accuracy by focusing on relevant lung areas and suppressing irrelevant structures.

While data-driven segmentation methods offer strong performance by learning patterns from labeled datasets, rule-based techniques can often achieve comparable accuracy through careful manual tuning of their parameters. Despite their accuracy, data-based models typically require considerable computational resources and time for training, making them less efficient for real-time or large-scale applications. In contrast, rule-based methods are generally faster, more lightweight, and easier to implement, making them a practical choice for many researchers working with lung CT images in CAD systems.

### 2.4.2 Nodule Detection

Pulmonary nodule detection is typically divided into two essential phases: candidate nodule identification and false positive reduction. Figure 2.6 shows the overall framework of the nodule detection system, including candidate nodule detection (CNDET) and false positive reduction (FPRED) stages. Due to the heterogeneous nature of nodules—varying in size, texture, shape, and anatomical position—numerous detection strategies have been developed over the years. These approaches can be broadly grouped into two main categories: conventional techniques and deep learning-based methods. Traditional approaches usually rely on handcrafted features and classic machine learning classifiers to locate nodules and discard false positives by optimizing the match between predefined feature sets and suspicious image regions. In contrast, deep neural network (DNN)-based methods, particularly convolutional neural networks (CNNs), learn feature representations automatically from the data through end-to-end training pipelines. These models act as powerful black-box systems capable of capturing complex patterns beyond human-defined rules. In the following sections, we provide an overview of representative algorithms that have been developed for both candidate detection and subsequent false positive elimination. The effective algorithms proposed in 2020-2025 are selected and summarized in Table 2.1.

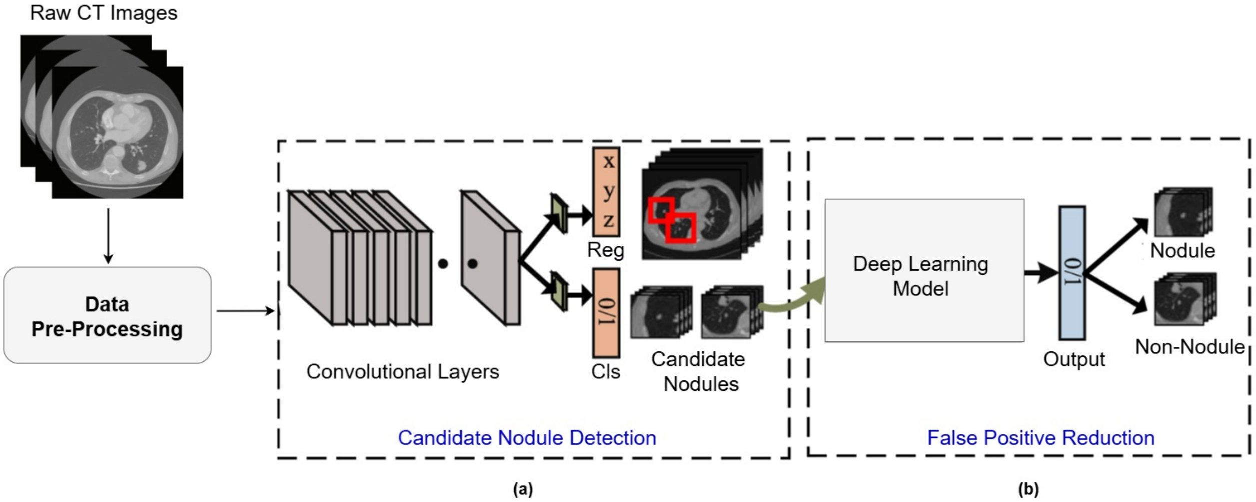


Figure 2.6 Deep learning-based nodule detection framework: (a) CNDET stage and (b) FPRED stage.

### Candidate Nodule Detection

At the initial stage of pulmonary nodule detection, the primary objective is to maximize sensitivity, ensuring that as many potential nodules as possible are captured, even at the expense of including false positives. This step, known as candidate nodule detection (CNDET), is designed to identify all regions that may represent nodular lesions. Instead of focusing on classification accuracy at this point, the system aims to generate a comprehensive set of possible candidates by scanning the entire lung area for abnormal patterns. Increasing the likelihood of detecting malignant nodules early can significantly improve patient outcomes, making this stage crucial in any CAD pipeline.

Traditional image processing techniques have long played a central role in the initial detection of pulmonary nodules. These methods typically rely on manually engineered features and pixel-level characteristics, such as intensity, shape, and texture. Approaches like region growing, morphological filtering, distance mapping, and thresholding have been commonly applied to isolate areas of interest within lung CT scans [34,39–42]. For instance, early systems often used segmentation heuristics to extract the lung region and identify attached or embedded nodules. While these classical techniques are computationally efficient and interpretable, their performance is often limited when dealing with highly variable nodule appearances, such as subtle margins or irregular densities. As a result, additional refinements—like geometric feature analysis or rule-based enhancements—are frequently necessary to improve candidate selection and reduce missed detections [43].

With the widespread adoption of deep learning, an increasing number of detection algorithms have been developed based on DNN frameworks. In particular, CNN-based models are widely utilized in candidate nodule detection tasks due to their capability to extract both low-level spatial details and high-level semantic features, thereby significantly enhancing detection sensitivity. Common network structures applied in pulmonary nodule detection include standard CNNs, U-Net and its variants, Feature Pyramid Networks (FPN), Region Proposal Networks (RPN), Residual Networks (ResNet), and hybrid architectures such as Retina-Net and Faster R-CNN extensions [4, 44–47]. Many of these approaches build upon these foundational models, introducing tailored modifications to improve performance for nodules of varying sizes and shapes. Some studies have proposed hybrid networks that combine multiple architectures in a cascade or parallel arrangement to leverage their complementary strengths. For example, MS-3DCNN [48] integrated a multi-scale 3D UNet++ architecture with RPNs and residual blocks to boost sensitivity, while MSM-CNN [49] combined Faster R-CNN with multiscale feature extraction for robust small nodule detection. Similarly, OMS-CNN [50] enhanced Faster R-CNN by integrating an optimized multi-scale CNN feature extraction model using metaheuristic algorithms, TiCNet [51] introduced a transformer module within a 3D CNN framework alongside multi-scale skip pathways to capture both local and global dependencies, and DA OMS-CNN [52] incorporated dual-attention mechanisms into a multi-scale CNN within an improved Faster R-CNN architecture combined with 3D Swin Transformers, achieving high sensitivity while effectively reducing false positives. These advanced architectures demonstrate that incorporating multi-scale learning, residual connections, attention, and transformer modules into candidate detection pipelines can significantly improve overall detection performance in lung cancer screening applications.

In the candidate nodule detection stage, the objective is to achieve high sensitivity by identifying all potential nodules, even at the cost of including many false positives. As shown in the top row of Figure 2.7, multiple candidates are detected in each CT slice. The red boxes represent the detected candidates, while the green boxes show the ground truth nodules. This stage ensures that no true nodules are missed before proceeding to further analysis.

### False Positive Reduction

Even after the candidate nodule detection stage, a considerable number of false positives (FPs) often remain, which can hinder the efficiency of lung nodule diagnosis. High rates of FPs may lead to unnecessary follow-up procedures, overdiagnosis, and increased healthcare costs. Thus, minimizing false positives is crucial to enhance the accuracy and clinical applicability of detection systems. False Positive Reduction (FPRED) involves distinguishing true

Table 2.1 Summary of recent CAD models for candidate nodule detection (CNDET) and false positive reduction (FPRED).

| **CAD Models** | **Year** | **Method** | **Dataset** | **Best Performance** |
|---|---|---|---|---|
| Zue et al. [53] | 2020 | **FPRED**: Multi-branch 3D CNN | LUNA16 | Sensitivity: 87.71%, CPM: 0.830 |
| AECS-CNN [54] | 2021 | **FPRED**: Attention-Embedded Complementary-Stream CNN | LUNA16 | Sensitivity: 0.92%, CPM: 0.762 |
| I3DR-Net [55] | 2022 | **CNDET**: One-stage detection using I3D + FPN | Public and Private CT datasets | CPM: 0.812 |
| MSM-CNN [49] | 2022 | **CNDET**: Faster R-CNN with multiscale features;<br>**FPRED**: 3D CNN with multiscale fusion | LUNA16 | Sensitivity: 98.6%, CPM: 0.829 |
| MS-3DCNN [48] | 2023 | **CNDET**: Multi-scale 3D UNet++ with RPN and residual connections;<br>**FPRED**: Multi-input fusion classification | LUNA16 | Sensitivity: 87.3%, CPM: 0.871 |
| MK-3DCNN [56] | 2024 | **CNDET**: Multi-kernel 3D CNN with residual encoder-decoder | LUNA16 | CPM: 0.859 |
| TiCNet [51] | 2024 | **CNDET**: Transformer + 3D CNN hybrid with attention and multi-scale skip pathways;<br>**FPRED**: Two-head detector | LUNA16, PN9 | CPM: 0.884 |
| OMS-CNN [50] | 2024 | **CNDET**: Optimized Multi-Scale CNN;<br>**FPRED**: Multiple 3D DCNNs | LUNA16, PN9 | Sensitivity: 94.89%, CPM: 0.892 |
| AttentNet [57] | 2025 | **CNDET**: 3D RPN with attention;<br>**FPRED**: Fully convolutional attention + joint spatial analysis | LUNA16 | CPM: 0.871 |
| DA OMS-CNN [52] | 2025 | **CNDET**: Improved Faster R-CNN with Dual-Attention OMS-CNN;<br>**FPRED**: Multiple 3D Swin Transformers (3D SwinT) | LUNA16, PN9 | Sensitivity: 96.93%, CPM: 0.911 |

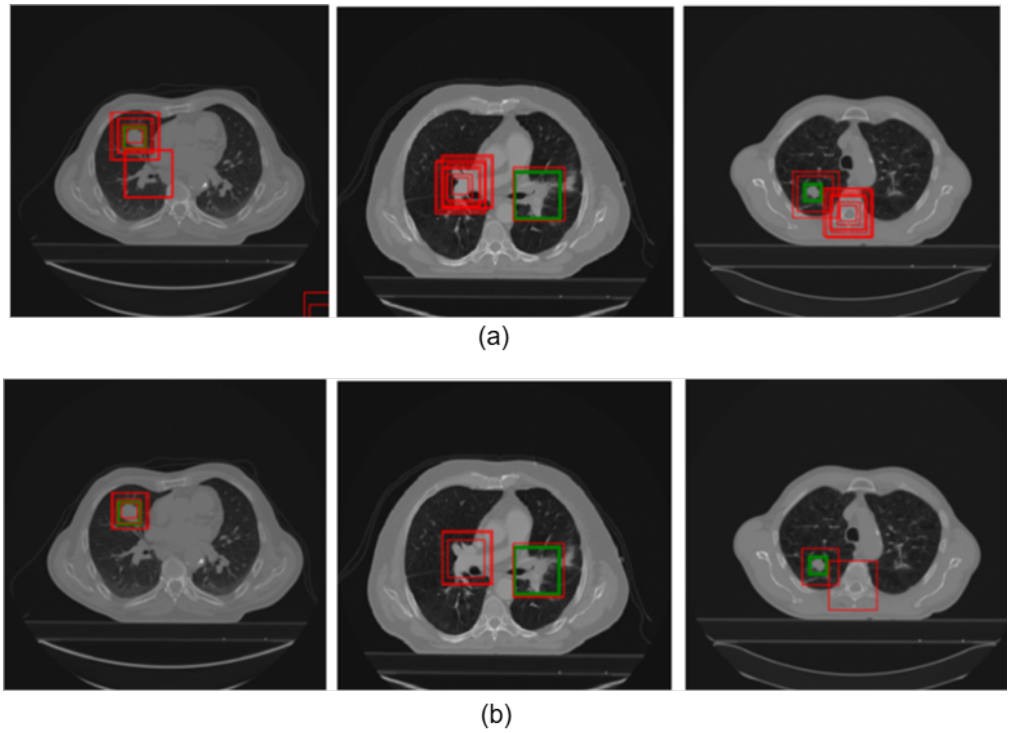


Figure 2.7 Nodule detection results: (a) CNDET outputs; (b) FPRED outputs.

nodules from non-nodules within the detected candidates, essentially functioning as a binary classification task. Numerous studies have been dedicated to developing effective methods for this critical stage.

In the FPRED stage, various handcrafted features, including intensity-based, morphological, and texture descriptors, are extracted from candidate nodule regions. These features are then used to train traditional machine learning classifiers to distinguish true nodules from non-nodule structures. Commonly employed classifiers in traditional approaches include Support Vector Machines (SVM), k-Nearest Neighbors (k-NN), linear discriminant analysis, and different boosting algorithms [39,58–60]. For instance, Naqi et al. [34] proposed combining geometric texture features with Histogram of Oriented Gradient features reduced by Principal Component Analysis (HOG-PCA) to construct a hybrid feature vector, which was subsequently classified using k-NN, Naive Bayes, SVM, and AdaBoost for effective false positive reduction.

In recent years, numerous deep neural network (DNN)-based approaches, particularly those leveraging convolutional neural networks (CNNs), have been introduced to improve classification accuracy in false positive reduction. Depending on their architectural designs, these methods can generally be grouped into two categories: advanced pre-trained CNN models [61–63]and multi-stream heterogeneous CNN architectures [53, 64]. For instance, an attention-embedded complementary-stream CNN (AECS-CNN) [54] employed multi-scale 3D CT inputs with attention-guided feature extraction to capture rich contextual information and enhance discriminative feature learning. This method achieved a sensitivity of 0.92 with 4 false positives per scan on the LUNA16 dataset. Similarly, an optimized multi-scale CNN (OMS-CNN) [50], multiple 3D deep CNNs were combined to effectively reduce false positives. This method, evaluated on the LUNA16 and PN9 datasets, achieved a CPM score of 0.892, highlighting its capacity to extract representative nodule features of varying sizes and improve both sensitivity and specificity for clinical use. Furthermore, dual-attention optimized multi-scale CNNs (DA OMS-CNN) [52] combined with 3D shift window transformers (3D SwinT) have been proposed to improve both feature extraction and spatial modeling capabilities, resulting in a CPM score of 0.911 on benchmark datasets. These advanced architectures demonstrate the effectiveness of integrating attention mechanisms, multi-scale learning, and transformer modules in enhancing FPRED performance for lung nodule detection systems.

In the FPRED stage, a dedicated classifier is applied to eliminate non-nodules from the detected candidates. As illustrated in the bottom row of Figure 2.7, most of the false positives are successfully removed while retaining the true nodules. This significant reduction in false alarms is critical to improve the overall specificity of the nodule detection system and to avoid unnecessary follow-up examinations by radiologists.

### 2.4.3 Classification

Nodule classification constitutes the final stage in CAD systems. While many CAD systems focus on predicting the malignancy of detected nodules to determine whether they are cancerous, some are designed to categorize nodules based on their types [65]. Malignant nodules generally exhibit larger sizes (typically with diameters greater than 8 mm) and irregular surface morphologies such as spiculation or lobulation. Therefore, precise measurements of nodule size and detailed analysis of their appearance remain crucial for estimating malignancy probability.

A wide range of classification techniques have been employed in this stage. These include: (1) traditional machine learning classifiers such as support vector machines (SVM), k-nearest neighbors (k-NN), Bayesian classifiers, boosting algorithms, and optimal linear classifiers [66, 67]; (2) advanced off-the-shelf convolutional neural networks (CNNs) [68, 69]; (3) hybrid models integrating CNNs with machine learning classifiers [70]; (4) multi-stream heterogeneous CNN architectures [71]; and (5) CNNs trained using transfer learning strategies to leverage features from large-scale datasets [72]. Among these, multi-stream hybrid

CNNs combined with transfer learning have shown promising results due to their ability to extract discriminative features across multiple scales and views.

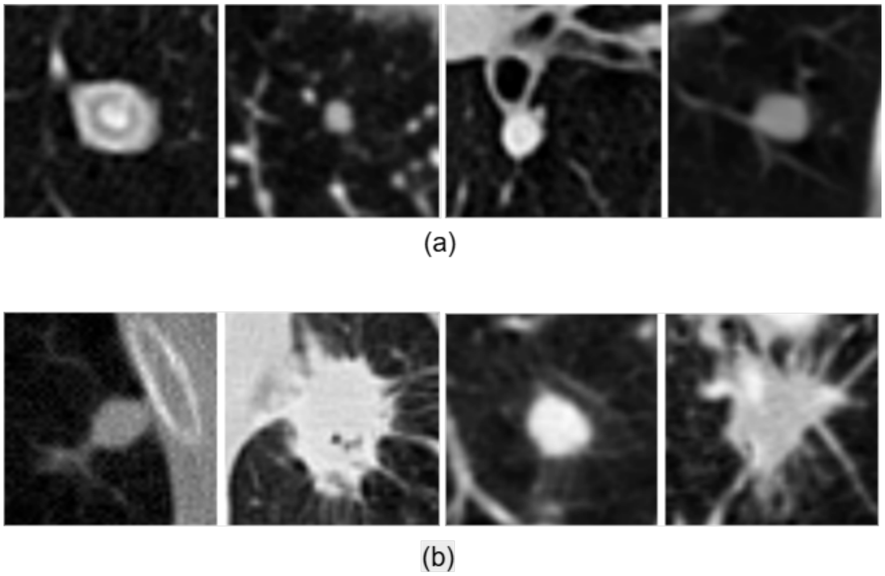


Figure 2.8 Examples of pulmonary nodules: (a) benign nodules with smooth margins; (b) malignant nodules with irregular or spiculated edges.

Figure 2.8 illustrates examples of benign and malignant pulmonary nodules. Subfigure (a) shows benign nodules, which generally appear with smooth and regular boundaries, whereas subfigure (b) presents malignant nodules characterized by larger sizes, spiculated or lobulated margins, and heterogeneous textures. Accurate differentiation between benign and malignant nodules is vital for early diagnosis and treatment planning, significantly impacting patient outcomes.

### 2.4.4 Explainability and Transparency in CAD Systems

While deep learning (DL) models have demonstrated impressive performance in detecting pulmonary nodules, their adoption in clinical workflows is limited due to their lack of interpretability. Traditional convolutional neural networks (CNNs) often function as "black-box" systems, offering little insight into how or why a particular decision is made. This opacity hinders clinical trust and raises concerns about accountability, especially in high-stakes applications such as lung cancer screening [73].

To address this, the field has increasingly focused on developing explainable AI (XAI) techniques tailored for medical imaging. These methods aim to provide visual or semantic explanations of model predictions, helping clinicians assess the reliability of automated decisions.

A common approach involves saliency-based techniques such as Gradient-weighted Class Activation Mapping (Grad-CAM) [74], which highlights regions in the input CT scan that most influenced the classification outcome. Other methods like Local Interpretable Model-Agnostic Explanations (LIME) [75] and SHapley Additive exPlanations (SHAP) [76] offer post-hoc interpretations, though their application in 3D medical imaging remains challenging.

Several DL frameworks have attempted to embed interpretability into their architecture. For instance, HSCNN [77] employed a hierarchical semantic structure to make intermediate predictions about nodule attributes (e.g., texture, shape) before producing a final malignancy score. Similarly, MTMR-Net [45] employed multi-task learning to jointly estimate attribute labels and malignancy probabilities, providing insights into the decision-making process. However, these methods often require additional supervision and auxiliary classifiers, which complicate the training pipeline and limit scalability.

More recent work has explored the use of attention mechanisms to focus the model's learning on informative regions. Fu et al. [78] proposed an attention-based multi-task CNN that weights cross-attribute features to emphasize diagnostically relevant information. Although attention maps offer a more integrated form of interpretability, they still lack direct anatomical grounding and are often difficult to validate clinically.

To address these limitations, the fourth study in this thesis (Chapter 7) proposes a hybrid SwinT-CNN model that integrates anatomical priors derived from a pretrained 3D U-Net segmentation model via an Anatomical Attention Gate (AAG). This design not only improves classification performance but also enhances interpretability by ensuring that the model's attention aligns with known anatomical structures. The proposed approach combines Grad-CAM saliency maps with quantitative interpretability metrics such as sensitivity and entropy to evaluate the quality of explanations. Experimental results demonstrate that this anatomically guided design improves both transparency and diagnostic trust, positioning the model for real-world deployment.

## 2.5 Challenges

Despite significant advancements in CAD and deep learning-based systems for pulmonary nodule detection and classification, several challenges remain that limit their widespread adoption and optimal performance. These challenges span multiple aspects, including the quality and availability of annotated data, the generalizability and robustness of model performance, the high computational demands of training and deployment, integration into clinical workflows, and various ethical and legal concerns [79–81]. Addressing these limitations is crucial to ensure that such systems can reliably assist radiologists in early lung cancer detection and ultimately improve patient outcomes [82].

### 2.5.1 Data Quality and Availability

One of the primary challenges in developing effective machine learning (ML) and deep learning (DL) models for lung nodule detection is the limited availability of annotated medical imaging datasets. High-quality annotations require extensive input from expert radiologists, which is both time-consuming and costly. Unlike natural image datasets, which can leverage large-scale crowdsourcing, medical image annotation requires specialized clinical expertise to ensure diagnostic accuracy and reliability. As a result, the lack of sufficiently large and well-annotated datasets often hinders the development and validation of robust models for clinical use.

Another significant issue is the class imbalance commonly observed in medical imaging datasets for lung cancer detection. Typically, there is a disproportionately lower number of positive cases (cancerous nodules) compared to negative cases (benign or non-nodules). This imbalance can lead to biased models that are highly sensitive to the majority class while failing to accurately detect malignant nodules, which are clinically the most critical. Addressing this challenge requires the use of data augmentation, synthetic data generation, or advanced loss functions designed to mitigate the effects of class imbalance during training.

Moreover, variations in imaging protocols, scanner types, and patient characteristics introduce considerable heterogeneity in medical image datasets. Differences in slice thickness, reconstruction algorithms, and scanning parameters across institutions can significantly impact the appearance and quality of images. Additionally, the diverse manifestation of lung cancer in different patients further complicates model generalization. Such variability makes it challenging to develop models that are robust across various clinical settings and imaging devices, highlighting the need for standardized imaging protocols and extensive cross-institutional datasets to improve generalizability.

### 2.5.2 Model Performance

While data imbalance during training affects the learning quality of models, another critical challenge during their deployment is achieving an optimal balance between false-positive and false-negative rates. High false-positive rates can lead to unnecessary diagnostic procedures, additional imaging, invasive biopsies, and increased patient anxiety, placing a burden on both patients and healthcare systems. Conversely, false negatives, where malignant nodules remain undetected, pose an even greater risk by delaying diagnosis and treatment, potentially resulting in poorer patient outcomes. Striking the right balance to maximize sensitivity without compromising specificity remains a complex and clinically significant challenge.

Another limitation in model performance arises from the complexity of identifying and extracting relevant features from medical images. While deep learning models, particularly convolutional neural networks (CNNs), are capable of learning hierarchical and discriminative features automatically, their effectiveness heavily depends on access to large and diverse datasets for training. In contrast, traditional machine learning methods require handcrafted features designed by experts, which might not capture subtle imaging characteristics critical for accurate classification. This trade-off highlights the challenge of developing models that can generalize well with limited annotated data while maintaining high diagnostic accuracy.

Furthermore, the interpretability of deep learning models poses a significant barrier to their clinical adoption. CNN-based models are often regarded as "black boxes" due to their complex internal representations and lack of transparent decision-making processes. For medical professionals to trust these models and for them to gain regulatory approval, it is essential to understand how such systems arrive at their predictions. Developing interpretable AI models or incorporating explainability frameworks to visualize feature importance and decision pathways is thus critical for fostering clinical trust and ensuring responsible deployment in healthcare settings.

### 2.5.3 Computational Resources

Training deep learning models for pulmonary nodule detection, especially those utilizing 3D medical images, requires substantial computational resources and memory capacity. Processing volumetric CT data involves handling large input sizes and complex network architectures, often necessitating high-performance GPUs or computing clusters to achieve feasible training times. This requirement can pose a significant barrier for many research institutions and healthcare facilities with limited access to such advanced computational infrastructure, thereby restricting the development and experimentation of state-of-the-art models.

Beyond training, deploying these models in clinical settings introduces additional computational challenges. CAD systems must analyze medical images and generate diagnostic results quickly, thereby integrating seamlessly into clinical workflows. However, optimizing deep learning models to run efficiently on available hospital hardware without compromising diagnostic accuracy remains a significant challenge. The need to balance computational efficiency with predictive performance is especially crucial in time-sensitive environments such as lung cancer screening and early detection programs.

### 2.5.4 Integration into Clinical Practice

Integrating computer-aided detection (CAD) systems into clinical practice faces notable challenges related to user training and acceptance. Healthcare professionals require adequate training to effectively use these new AI-based tools alongside their standard diagnostic procedures. However, introducing unfamiliar technologies can encounter resistance, particularly if clinicians perceive them as complex or disruptive to established workflows. Ensuring that CAD systems are user-friendly and accompanied by comprehensive training programs is therefore essential to facilitate their smooth adoption in real-world clinical environments.

### 2.5.5 Ethical and Legal Concerns

Ensuring the privacy and security of patient data is a fundamental ethical and legal requirement when developing and deploying AI-based medical systems. Medical imaging datasets often contain sensitive personal health information, and any breaches in data security can lead to serious consequences, including violations of patient confidentiality, legal penalties for institutions, and loss of public trust. Implementing strict data protection measures, adhering to regulatory standards such as HIPAA or GDPR, and employing secure data storage and transmission protocols are critical for safeguarding patient information in AI model development and deployment processes.

Another major ethical challenge is the potential for AI models to exhibit biases due to training on non-representative datasets. If models are developed using data that do not adequately cover diverse demographic groups, their performance may be inconsistent, leading to reduced diagnostic accuracy in underrepresented populations. This raises concerns about fairness and equity in healthcare delivery, as biased models can exacerbate existing disparities in health outcomes. Therefore, it is essential to ensure that training datasets are diverse and that fairness assessments are incorporated into the model development pipeline to mitigate biases and promote equitable AI applications in clinical practice.

## 2.6 Discussion

The literature reviewed in this study highlights the significant evolution of computer-aided diagnosis (CAD) systems for pulmonary nodule detection, particularly in the context of advancements in deep learning. Early-stage CAD systems primarily relied on traditional image processing and handcrafted features, often employing classical machine learning classifiers such as SVMs, k-NN, or decision trees. These methods were effective to a certain degree but were heavily dependent on expert-designed features and could not generalize well across diverse nodule appearances or imaging variations.

With the emergence of deep learning, particularly convolutional neural networks (CNNs), the field of pulmonary nodule analysis has undergone significant transformation. Deep learning models, such as U-Net and Faster R-CNN, along with their numerous extensions, now dominate both candidate detection and false positive reduction stages. These models provide end-to-end learning pipelines, enabling the automatic extraction of features from raw CT images without the need for manual engineering. Furthermore, recent innovations, such as attention mechanisms, multi-stream networks, and transformer-based modules (e.g., Swin Transformers), have significantly enhanced model sensitivity and specificity, especially when applied to challenging small or subsolid nodules.

An important observation from the reviewed methods is the apparent trend toward multi-scale and 3D analysis. Lung nodules vary substantially in size, location, and texture, and using 3D volumetric CT data provides a richer spatial context than 2D slices. Methods such as MS-3DCNN, AECS-CNN, and DA OMS-CNN have successfully integrated multi-scale feature extraction with 3D processing, thereby improving detection performance across various nodule types. Additionally, models such as OMS-CNN and TiCNet introduced metaheuristic optimization or hybrid CNN-transformer architectures, reflecting the growing emphasis on architectural flexibility and task-specific tuning to improve classification accuracy.

Another key theme is the increasing importance of false positive reduction (FPRED) in the overall performance of CAD systems. While early systems prioritized sensitivity, often generating large numbers of false positives, recent approaches strike a more effective balance. Advanced FPRED techniques now incorporate dedicated CNN classifiers or attention-guided feature fusion layers to retain true nodules while discarding irrelevant candidates selectively. The improvements observed in CPM scores across recent models demonstrate the effectiveness of such enhancements in reducing clinical burden.

Moreover, classification of nodules into benign and malignant categories remains a critical component of CAD workflows. Studies show that malignant nodules often exhibit spiculated, lobulated, or irregular margins, while benign nodules tend to be smooth and well-defined. Modern classification models leverage multi-view, multi-stream CNNs or transfer learning from large image datasets to learn subtle morphological and textural cues. These models are also increasingly integrated with attention and explainability mechanisms to improve clinical interpretability and decision support.

Overall, the reviewed literature suggests a strong trend toward developing more precise, interpretable, and clinically relevant AI models for lung cancer screening. However, while technical improvements in detection and classification are clear, the path to real-world clinical integration still requires addressing aspects such as interpretability, robustness across data sources, and regulatory validation. Future research may benefit from hybrid models that combine classical medical knowledge with data-driven learning, as well as from collaborative, multi-center dataset development to support generalizable and equitable model training.

## 2.7 Conclusion

This chapter presented a comprehensive review of the current landscape of CAD systems for pulmonary nodule detection and classification, with a particular focus on the integration of deep learning techniques. Beginning with an overview of the clinical significance of early lung cancer detection, we examined the unique challenges posed by pulmonary nodules in terms of their variability in size, shape, and appearance. The chapter then outlined the typical pipeline of CAD systems, including data preprocessing, candidate nodule detection (CNDET), false positive reduction (FPRED), and final classification. Recent advances in deep learning, particularly convolutional neural networks (CNNs) and their 3D and multi-scale variants, have significantly enhanced the performance of each stage in the CAD pipeline. Cutting-edge models, such as AECS-CNN, OMS-CNN, and DA OMS-CNN, demonstrate that integrating attention mechanisms, optimized feature fusion, and transformer-based modules can yield high sensitivity and specificity across diverse datasets. The utilization of large-scale public datasets, including LIDC-IDRI, LUNA16, and PN9, has further enabled the development and benchmarking of increasingly sophisticated algorithms. Evaluation metrics such as sensitivity, precision, ROC/FROC curves, and the CPM score were discussed as essential tools for quantifying model performance. The literature reveals a consistent effort to balance detection accuracy with computational efficiency and clinical interpretability, highlighting the importance of reducing false positives and ensuring explainability in real-world applications. In addition, this chapter identified several persistent challenges that must be addressed for successful clinical translation, including data quality and availability, model robustness, computational constraints, integration into clinical practice, and ethical and legal considerations. These issues underscore the need for continued research on interpretable, efficient, and fair AI models that can operate reliably across diverse populations and healthcare settings.

In conclusion, while deep learning has significantly advanced the capabilities of CAD systems for lung cancer detection, translating these innovations into routine clinical use will require addressing both technical and systemic challenges. The insights from this chapter provide a foundation for the development of next-generation CAD systems that are not only accurate and efficient but also trustworthy and clinically actionable.

# CHAPTER 3 RESEARCH APPROACH

## 3.1 Methodology

This thesis follows a structured research methodology grounded in iterative development, experimental evaluation, and clinical relevance. While each contribution targets specific technical challenges in lung nodule detection, all studies share a unified methodological backbone designed to ensure practical impact and reproducibility. The methodology consists of the following stages:

- **Problem Identification:** Identify key limitations in current detection systems, including limited sensitivity to small nodules, high false positive rates, inefficient computational architectures, and a lack of interpretability that hinders clinical trust and adoption.
- **Hypothesis Formulation and Solution Design:** Propose deep learning-based solutions such as architectural modifications (e.g., VGG16 or Faster R-CNN), metaheuristic optimization, and attention mechanisms.
- **Incremental Model Development:** Build minimal viable solutions and refine them iteratively through multi-stage experimentation.
- **Explainability and Visual Validation:** Evaluate model interpretability through Grad-CAM saliency maps, anatomical attention analysis, and quantitative metrics such as heatmap sensitivity and entropy.
- **Evaluation on Benchmark Datasets:** Test all models on public datasets such as LUNA16 and PN9 using metrics like sensitivity, CPM score, and false positives per scan.
- **Ablation and Comparative Analysis:** Perform controlled comparisons with baselines and ablated versions to isolate performance gains introduced by each module.

This unified methodology ensures each contribution builds upon the previous work in a coherent and scientifically rigorous manner, while also promoting clinical trust through explainability-focused validation.

## 3.2 Research Contributions and Work Packages

The core contributions of this thesis are structured into four research articles, each mapped to specific objectives related to early-stage lung cancer detection and clinical interpretability. These contributions are presented as individual work packages (WPs) aligned with three overarching research axes:

- **Sensitivity Improvement**
- **False Positive Reduction**
- **Efficiency and Applicability**
- **Interpretability and Clinical Transparency**

**WP1: Enhanced Region Proposal with VGG16 (Chapter 4)**
This work investigates the limitations of conventional feature extractors in detecting small nodules. By combining the final layers of VGG16 into a unified feature map and integrating it with a Region Proposal Network (RPN), the model improves the recall of small nodules.
*Addresses Objective A1.*

**WP2: OMS-CNN with Metaheuristic Optimization (Chapter 5)**
This study introduces OMS-CNN, an optimized multi-scale CNN embedded within Faster R-CNN. It employs harmony search (PSF-HS) and beetle antenna search (BAS) for layer optimization and kernel initialization, improving accuracy and efficiency.
*Addresses Objectives A1, B1, and C1.*

**WP3: Transformer-Based False Positive Reduction (Chapter 6)**
This article expands OMS-CNN using dual-attention modules, DA-RoIPooling for region-wise feature refinement, and 3D Swin Transformers for robust false positive reduction.
*Addresses Objectives A2, A3, B2, and B3.*

**WP4: Interpretable SwinT-CNN with Anatomical Priors (Chapter 7)**
This work proposes an interpretable hybrid model combining 3D CNNs and Swin Transformers in a dual-path architecture, guided by anatomical priors from a pretrained U-Net. The model improves diagnostic transparency by focusing on clinically meaningful regions, validated through Grad-CAM heatmaps and attention-based metrics such as sensitivity and entropy.
*Addresses Objectives D1 and D2.*

## 3.3 Document Structure

The remainder of this thesis is organized as follows:

- **Chapter 4** – Presents the first research article on VGG16-based region proposal enhancement.
- **Chapter 5** – Introduces OMS-CNN and its metaheuristic optimization framework.
- **Chapter 6** – Describes the extended model with dual attention, RoIPooling, and 3D Swin Transformers.
- **Chapter 7** – Presents an interpretable hybrid model (SwinT-CNN) that integrates anatomical priors and attention mechanisms to enhance diagnostic transparency.
- **Chapter 8** – Provides a general discussion of findings and cross-article insights.
- **Chapter 9** – Concludes the thesis and outlines future research directions.

# CHAPTER 4 ARTICLE 1: EFFICIENT REGION PROPOSAL EXTRACTION OF SMALL LUNG NODULES USING ENHANCED VGG16 NETWORK MODEL

**Preface:** This chapter presents a novel method for improving region proposal extraction in the context of early-stage lung nodule detection. The proposed approach leverages the VGG16 convolutional network, enhanced through a multi-layer feature map aggregation strategy, to better capture features of small-sized nodules. This work has been peer-reviewed and was published in the proceedings of the *2023 IEEE 36th International Symposium on Computer-Based Medical Systems (CBMS)*.

**Contributions:** This research originated as part of my doctoral work at Polytechnique Montréal and was developed in close collaboration with my co-authors. I contributed to the formulation of the research problem, designed the improved region proposal method, implemented the experimental pipeline using VGG16 and RPN, and conducted extensive evaluations on benchmark datasets. I also led the writing of the manuscript and coordinated the preparation of the final submission. My co-authors provided valuable input on experimental design, data interpretation, and manuscript revisions.

**Full Citation:** Yadollah Zamanidoost, Nada Alami-Chentoufi, Tarek Ould-Bachir, and Sylvain Martel, *"Efficient Region Proposal Extraction of Small Lung Nodules Using Enhanced VGG16 Network Model,"* in *Proceedings of the 2023 IEEE 36th International Symposium on Computer-Based Medical Systems (CBMS)*, L'Aquila, Italy, June 22–24, 2023. IEEE.

**DOI**: 10.1109/CBMS58004.2023.00266

## 4.1 Abstract

The efficiency of state-of-the-art convolutional networks trained to detect lung cancer nodules depends on their feature extraction model. Various feature extraction models have been proposed based on convolutional networks, such as VGG-Net, or ResNet. It has been demonstrated that such models effectively extract features from objects in an image. However, their efficacy is limited when the objects of interest are very small, such as lung nodules. One of the widely used feature extraction models for detecting small objects is the VGG16 network. The model, which has a small kernel of $3 \times 3$ and optimal layers, can extract the features of small

objects with reasonable accuracy. In this article, feature maps are created by combining the last three layers of the VGG16 network to extract features of various sizes of nodules. This study utilizes a Region Proposal Network (RPN) to compare the accuracy of the feature map created in the proposed method and the original VGG16. An RPN is a fully-convolutional network that simultaneously predicts object bounds and objectness scores at each position. RPNs are trained end-to-end to generate high-quality region proposals, which Faster R-CNN uses for detection. In this article, we select 300, $1,000$ and $2,000$ regions chosen by the RPN network for each method; then, we calculate the recall for different Intersection over Union (IoU) ratios with ground-truth boxes. The results show that the feature map of the proposed method works more optimally than the feature map of different layers of VGG16 for extracting various sizes of nodules. Also, by reducing the number of selected region proposals, the recall of the proposed method has fewer changes than other methods.

## 4.2 Introduction

Lung cancer is one of the most severe cancers. It has devastating effects on human life [83] and was declared one of Europe's most significant causes of death in 2019 [84]. Radiotherapy and chemotherapy are suitable and effective methods to treat the disease. However, the 5-year survival rate for people with lung cancer is only 16% [85]. Early detection of lung cancer is an effective and essential way to increase the chance of survival [86]. Today, computer-aided detection (CAD) systems are vital in helping radiologists diagnose cancerous tumours. Also, this technique helps to improve the accuracy of detecting lung nodules, reducing the number of missed nodules and misdiagnoses [87]. Recent advances in object detection are driven by success in region proposal methods [88] and region-based convolutional neural networks (R-CNNs) [89]. One of the difficulties of CAD systems is detecting small nodules. We can use different feature extraction methods by convolutional neural networks to address this problem. Three optimal and practical models of convolutional neural networks are VGG16 [90], ResNet-50 [91], and MobileNet [92], which are used in different articles to detect lung nodules.

Our method utilizes VGG16 with 5-group convolution as the main feature extraction network. As mentioned before, the detection of small nodules in the CT image of the lung scan is a very challenging task. The minimum size of the nodules is 3mm, and the maximum is about 30mm. After a series of convolutions and pooling in VGG16, the size of the feature map of the last layer is reduced, which leads to limited performance in the ROIs detection of nodules. It has been observed that the utilization of small feature maps does not provide sufficient resolution to represent the features of small nodules accurately. We can create a feature map that demonstrates the feature resolution of various sizes of nodules in the proposed method by combining the last three layers of VGG16. The proposed feature map enters a region proposal network (RPN) and obtains a set of rectangular-shaped nodule proposals, each of which has a score in the output. The proposed network of the region consists of a fully convolutional network. This paper presents the best set of CT scan image ROIs for input into the Faster R-CNN network using improved VGG16 to extract deeper features of lung nodules and region proposal network (RPN). The experimental results show that the proposed method is more accurate and stable compared to other common VGG16 methods.

The remainder of this paper is structured as follows. Section 4.3 explains the proposed method. Section 4.4 presents the implementation of the proposed method and its results. Section 4.5 offers a conclusion.

## 4.3 Method

Our research proposes a method for extracting efficient region proposals of lung nodules, utilizing a three-step approach. (1) automatic lung segmentation; (2) feature extraction; (3) Region proposal network (RPN). The paper presents a framework based on 3D Convolutional Neural Networks (3D CNN) (Fig. 4.1). The methodology employed is described in detail below.

### 4.3.1 Automatic Lung Segmentation

Automatic lung segmentation removes irrelevant regions, such as clothes, machine objects, tissues, spines, or ribs, from chest CT scan images. The stages of this segmentation can be seen in Fig. 4.2. First, we adjust the CT scan images from -1000 HU to 400 HU and then normalize the CT scan in the range between 0 and 1. We use the mean value of CT scan images as a threshold to divide the chest into outside and inside regions (Fig. 4.2(b)). Then we remove the isolated pixels attached to the white label's border. We also use a 4-connected neighborhood operator to remove unrelated tissues and noises. As shown in Fig. 4.2(c), the unimportant outside region of the chest is removed using a mask created from the chest. More boundary informations are preserved by morphological operations, such as opening and closing, to prevent the loss of lesions attached to the wall that may contain nodules. In addition, small noise areas such as graininess, vessels and tissues in the processed image are removed using the "Binary Fill Holes" [93] morphology operation. After performing the steps mentioned, the mask can be seen in Fig. 4.2(d), and the lung parenchyma is well segmented (Fig. 4.2(e)).

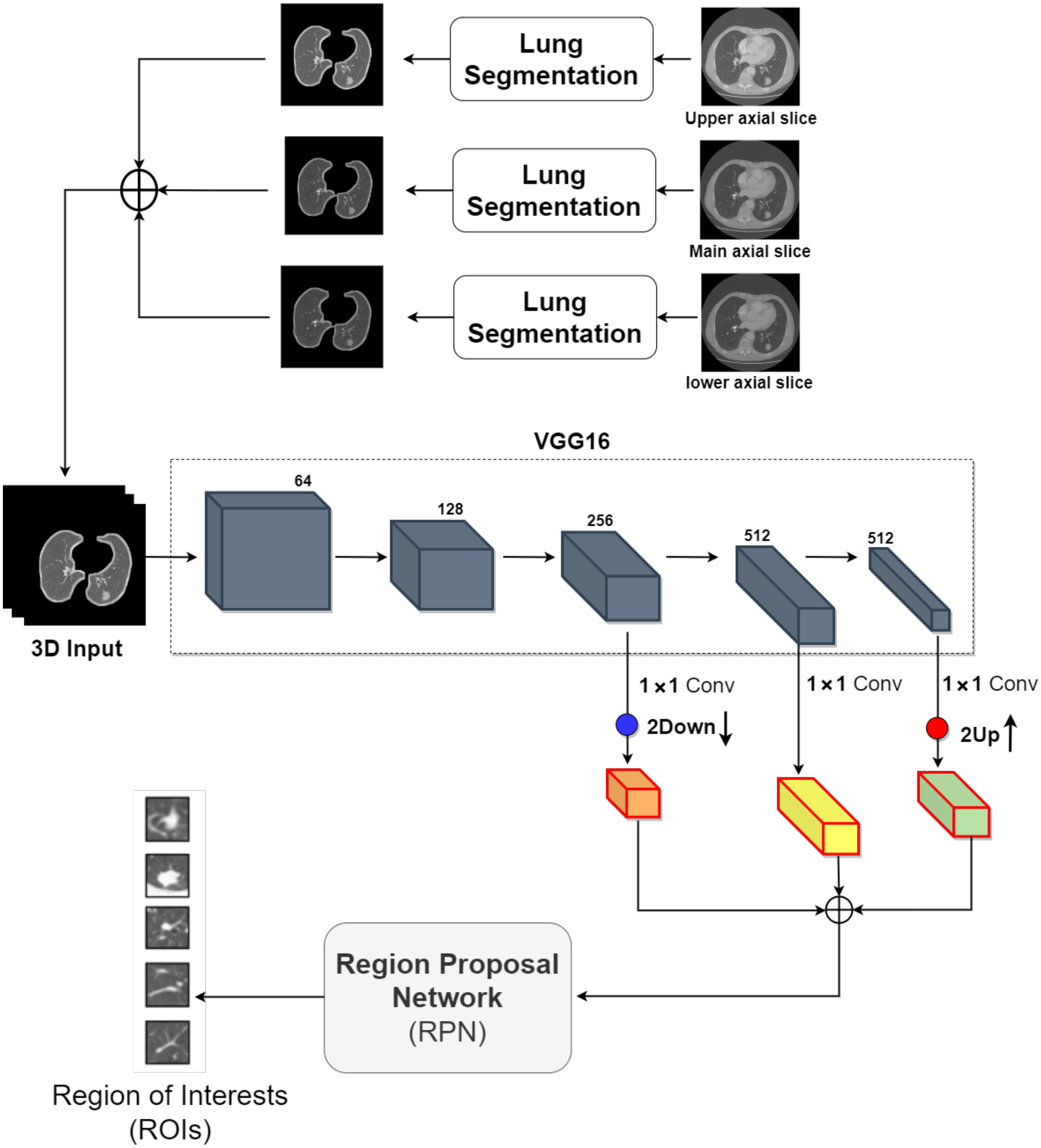


Figure 4.1 A proposed overview to detect candidate nodules using an improved VGG16 model

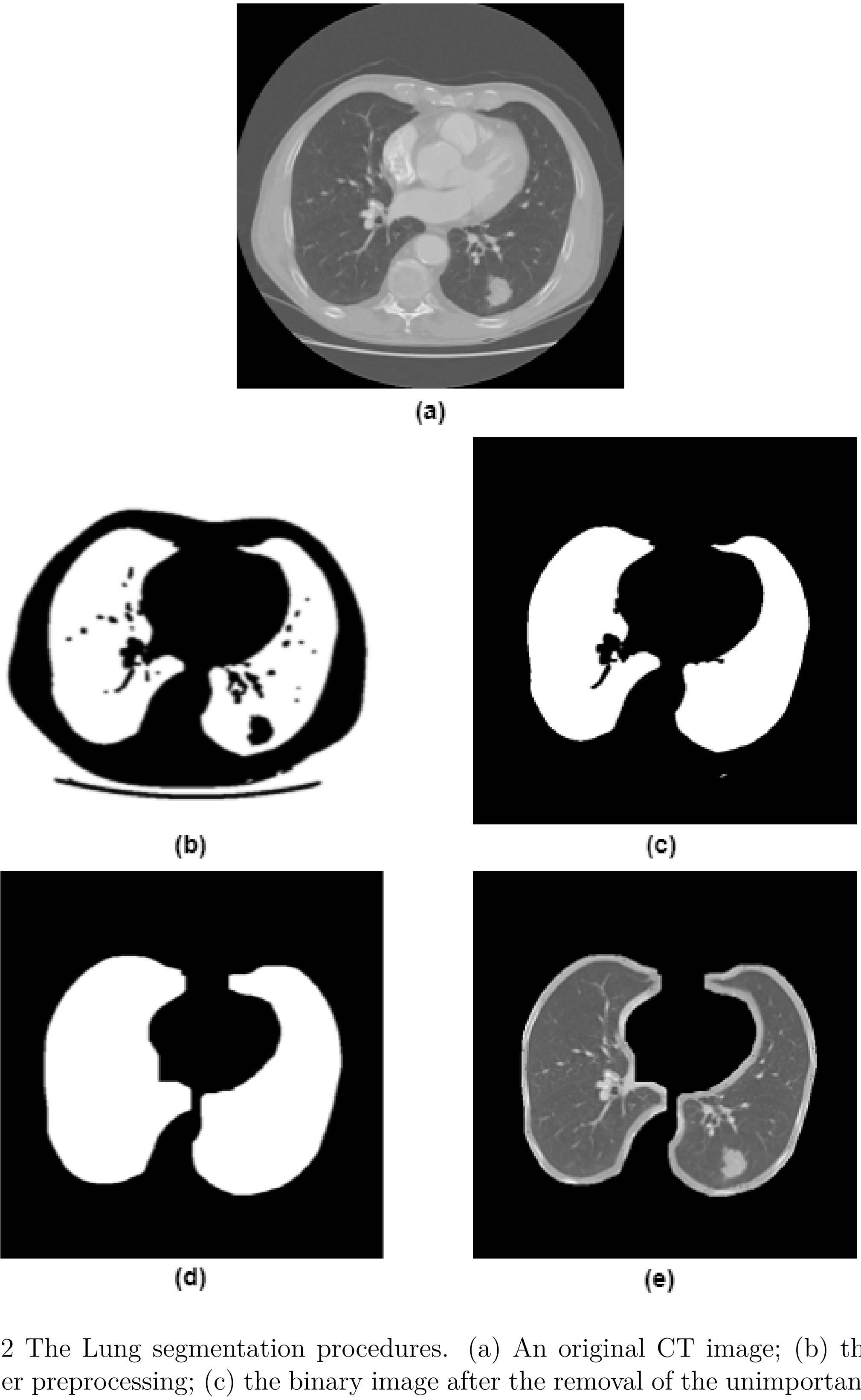


Figure 4.2 The Lung segmentation procedures. (a) An original CT image; (b) the binary image after preprocessing; (c) the binary image after the removal of the unimportant outside region; (d) the mask of lung parenchyma; (e) the image of segmented lung parenchyma.

### 4.3.2 Feature Extraction Structure

In the proposed method, the basic structure of the feature extractor is VGG16. This structure consists of 5 convolution groups. The features extracted in the last layer are suitable for detecting large lung nodules. On the other hand, these features can not be used to detect small nodules because small feature maps cannot clearly represent the features of small nodules. The upper layers have a semantic feature map and get more detailed features from nodules. The lower layers have more resolution but cannot extract finer details from the nodules. To use the detailed features with higher resolution, we combine the upper and lower layers of VGG16, as shown in Fig. 4.3. Using the proposed structure, we can select features from feature maps of 3 different layers, which include features with high accuracy and resolution.

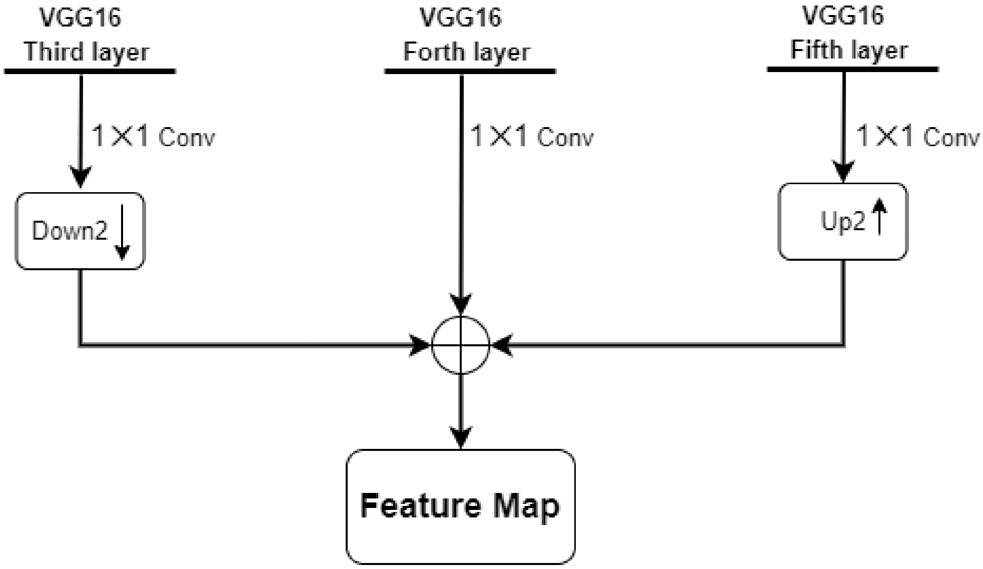


Figure 4.3 The proposed composition structure of the upper and lower layers of VGG16

Also, compared to the feature pyramid network (FPN) method [94], the proposed structure has a lower computational cost because two upsamples are used in FPN (Fig. 4.4). In contrast, only one upsample is used in the proposed method to combine three layers. In addition, the feature map of the FPN method is merged with the adjacent higher-resolution feature maps by element-wise addition. In contrast, the feature map of our method is combined with the high-resolution and high-semantic information of the three last layers.

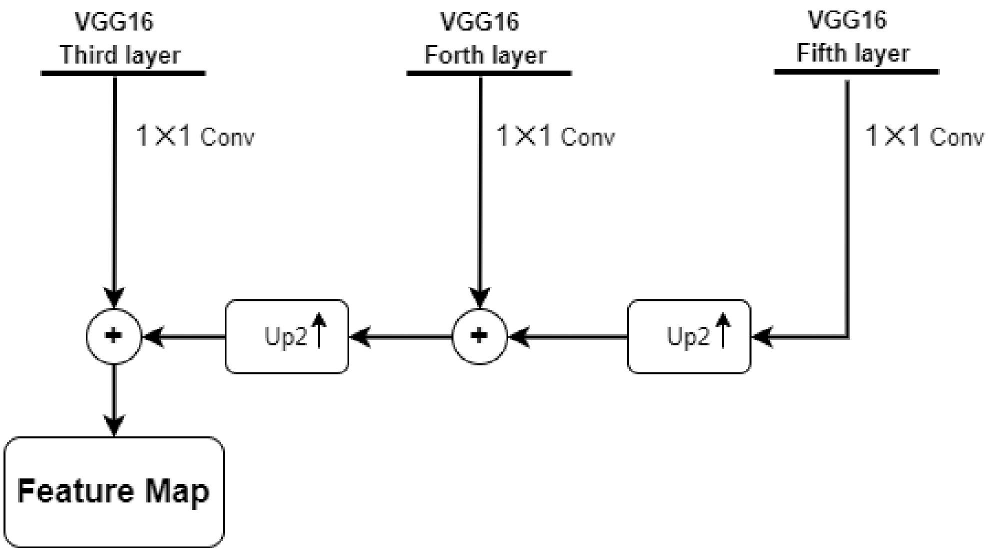


Figure 4.4 The Feature Pyramid Network (FPN) structure

### 4.3.3 Region Proposal Network (RPN)

An RPN aims to suggest potential nodule regions (called region-of-interest — ROI). As shown in Fig. 4.5, RPN receives a feature map as an input and outputs a set of rectangular object maps, each of which has an object score. In the RPN structure, a sliding window takes the feature map as input, obtains its convolution values using a 3 × 3 spatial window, and then maps each sliding window to a 512-dimensional feature. These features finally enter into two fully connected layers a box-regression layer (`reg`) and a box-classification layer (`cls`).

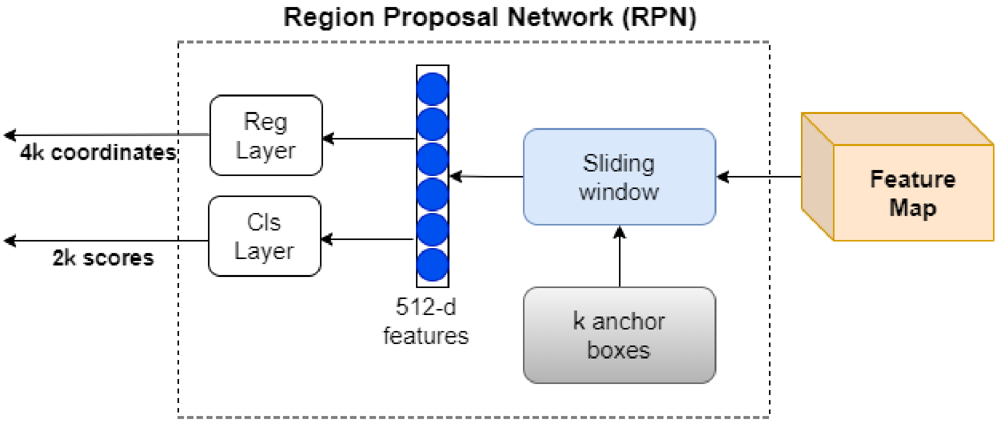


Figure 4.5 The Region Proposal Network (RPN) structure

Simultaneously, $K$ region proposals are predicted at each sliding window location. So the `reg` layer has $4K$ output that predicts the spatial coordinates of each box. The `cls` layer has $2K$ outputs that estimate each proposed region's nodule/non-nodule probability. Each box is called an anchor in the centre of the sliding window. To train RPNs, a binary class label is assigned to each anchor. The anchor with the highest intersection over-union (IoU) overlaps with the ground truth box, or the anchor with an IoU overlap higher than 0.6 is assigned a positive label. On the other hand, if the overlap of the IoU with the ground truth box is less than 0.3, the anchor is assigned a negative label. Anchors that are neither positive nor negative do not contribute to the training goal.

The *reg* layer has $4K$ outputs. The $x$ and $y$ are the centres of the box, and the $w$ and $h$ are its width and height. For a region proposal $(P)$ and a ground truth $(G)$, these four parameters compute as follows:

$$\begin{cases} t_x = \frac{(G_x - P_x)}{P_w} \\ t_y = \frac{(G_y - P_y)}{P_h} \\ t_w = \log \frac{G_w}{P_w} \\ t_h = \log \frac{G_h}{P_h} \end{cases} \tag{4.1}$$

where $P^i = (P^i_x, P^i_y, P^i_w, P^i_h)$ specifies the pixel coordinates of the center of proposal $P^i$ and $G = (G_x, G_y, G_w, G_h)$ specifies the ground-truth bounding box.

## 4.4 Experiment and results

### 4.4.1 Dataset

CT is one of the most sensitive imaging techniques in diagnosing lung nodules with fastness, cost-effectiveness and availability features. Different CAD systems are trained using various databases. One popular and useful database is Lung Image Database Consortium and Image Database Resource Initiative (LIDC–IDRI) [95]. It is publicly available via the Cancer Imaging Archive (TCIA) website. The primary purpose of the database is to develop CAD methods to automatically detect lung nodules or even for classification and quantitative assessment. Since 2011, the dataset has contained 1018 patients' diagnoses. Nodules whose diameter is equal to or above 3mm, nodules whose diameter is below 3mm, and non-nodules whose diameter is equal or above 3mm. The diameters of the nodules determine their sizes. Because multiple radiologists may have annotated a nodule, the average of the diameters is retained. However, the number of pixels of the pulmonary nodules is $4 - 56$ based on the diameters of the pulmonary nodules 3mm-30mm [96].

### 4.4.2 Implementation Details and Results

In this paper, a 3D image is used to input the VGG16 network. Since using the original 3D volume of the CT scan as input to the nodule detection network has a very high computational cost, we use axial slices as input instead. Therefore, for each axial slice in CT images, we extract its two adjacent slices and then change it to an 800 × 800 × 3 image.

In the automatic lung segmentation section, the threshold value is 0.99. To get the accuracy of this operation, we divide the results into four groups. The first group is the whole extracted lungs (Fig. 4.6(a)). The second group is the whole extracted lungs with their surrounding noises (Fig. 4.6(b)). The extracted lungs are incomplete in the third group, but there is a nodule in the extracted lung part (Fig. 4.6(c)). The fourth group of incompletely extracted lungs is that there is no nodule in the extracted lung part (Fig. 4.2(d)). Table. 4.1 shows the percentage of each group in total. To carry out this project, it is crucial to extract the lung

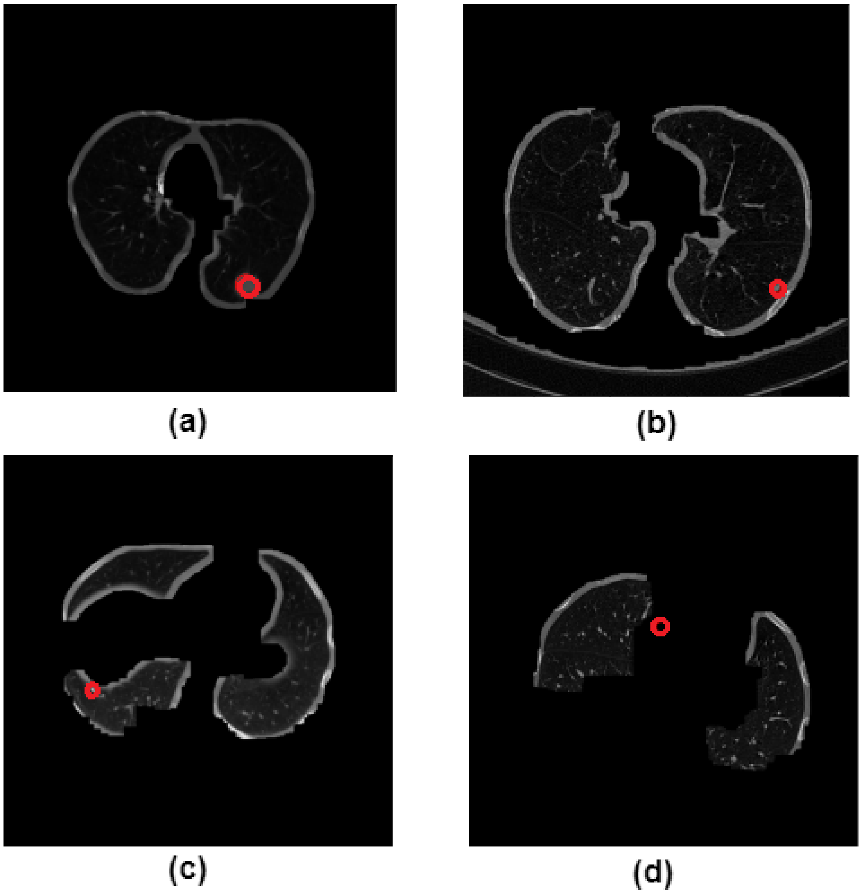


Figure 4.6 Automatic lung segmentation result groups; (a) The first group, (b) The second group, (c) The third group, (d) The fourth group

if there is a nodule, so the ILDC-IDRI and Luna16 datasets [97]' accuracy are approximately 94% and 97%, respectively. Compared to the automatic lung extraction methods that use deep learning, such as U-Net [98], besides having a much lower computational cost, this method also has acceptable accuracy.

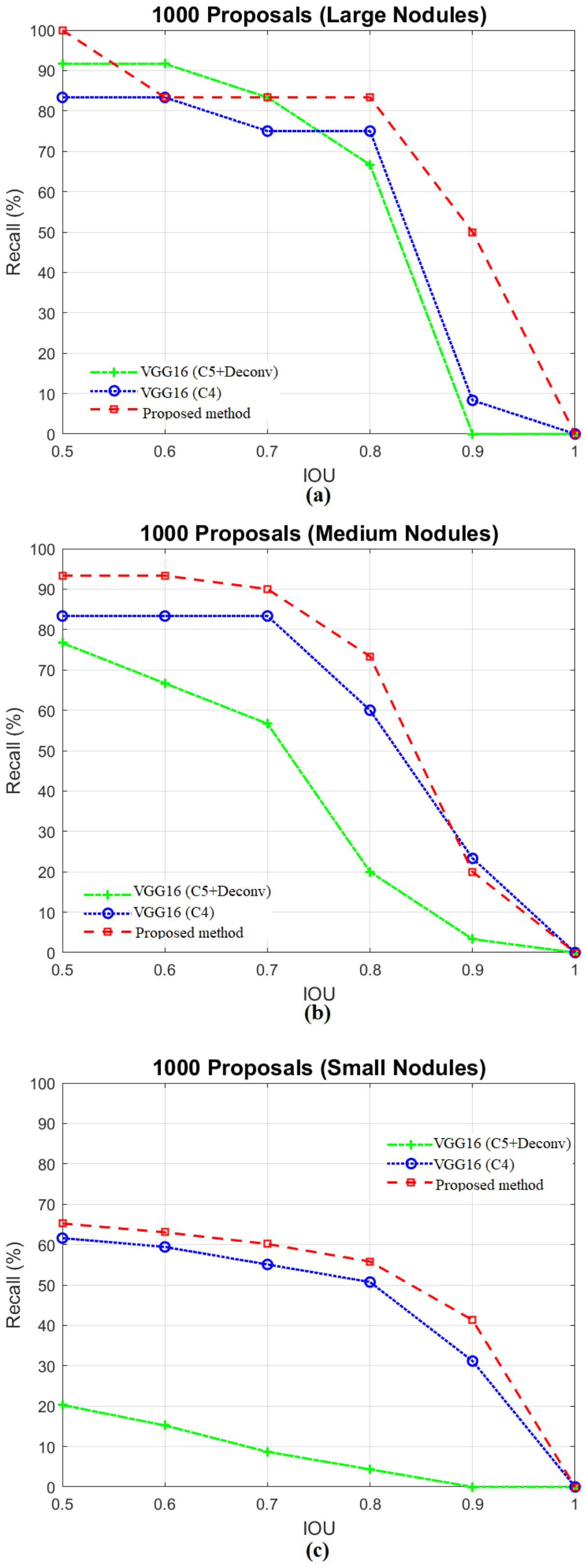
Table 4.1 Automatic lung segmentation accuracy (%)

| Dataset | Group1 | Group2 | Group3 | Group4 |
|---|---|---|---|---|
| LIDC-IDRI | 55.22 | 33.20 | 5.31 | 6.18 |
| LUNA16 | 72.64 | 20.30 | 4.12 | 2.93 |

After the automatic lung extraction stage, the images are re-scaled for better resolution of small nodules. This section converts the $512 \times 512$ CT scan images into $800 \times 800$ images with the cubic-interpolation method. The input for feature extraction utilizes 3D lung-extracted CT scan images. Due to the tiny size of lung nodules compared to normal objects in natural images, the original RPN network that uses VGG16-Net for feature extraction cannot extract the features of lung nodules with high accuracy, causing limited performance in detecting ROIs of nodules.

To solve this problem, in the proposed method, the structure of VGG16-Net layers has been improved to extract objects smaller than normal scale, such as lung nodules. In this method, by concatenating the upsampling of the last layer, the fourth layer, and the downsampling of the third layer of VGG16 and tuning the number of kernels, we can obtain the best performance to detect the ROI of the lung nodules. Combining layers recovers more fine-grained features compared to the original feature map. Therefore, the proposed model has better detection results than the original RPN method [99].

This article uses the RPN network to compute the recall of ROIs of the lung nodule and compare it with other methods. It also uses six anchors of different sizes, including $4 \times 4$, $6 \times 6$, $10 \times 10$, $16 \times 16$, $22 \times 22$ and $32 \times 32$ for each sliding window, as in [100]. The LIDC-IDRI dataset was utilized as the training data source. This study uses 800 cases. The CT images containing 2100 lung nodules are utilized in this study. We consider 10% of images as validation data and use the remaining images as training data. We employ validation data to modify the parameters of the training model. In addition, We train the network end-to-end in the RPN stage by the stochastic gradient descent (SGD) algorithm. We also randomly initialize all new layers using Gaussian distribution with mean 0 and variance 0.01. Also, we initialize the weight values of all VGG16 layers by pre-trained a model for ImagNet classification [101]. The network model trains in the 2 V100 GPUs environment, and the memory is 192G. Table. 7.2 shows the parameter of the training model.


1000 Proposals (Large Nodules)

1000 Proposals (Medium Nodules)

1000 Proposals (Small Nodules)

Figure 4.7 Recall *vs*. IoU overlap ration on; (a) Large nodules, (b) Medium nodules, (c) Small nodules

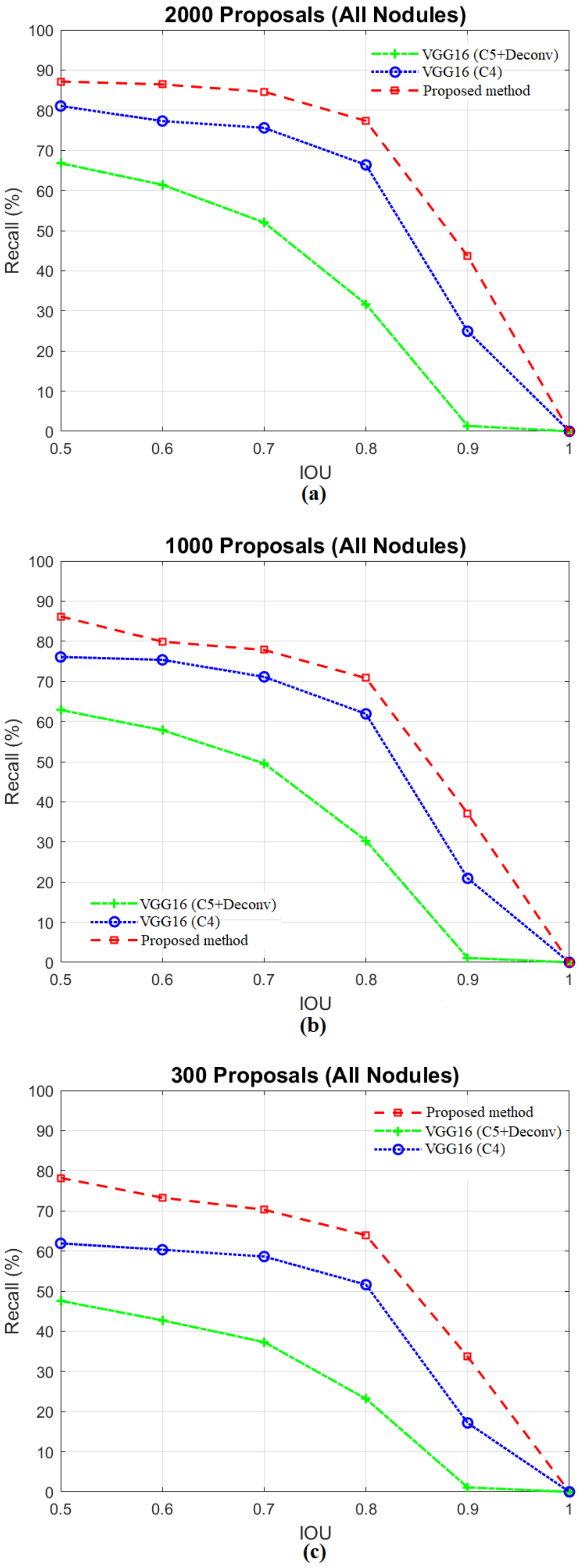


Figure 4.8 Recall *vs*. IoU overlap ration on all size of lung nodules; (a) 2K proposals, (b) 1K proposals, (c) 300 proposals

Table 4.2 Parameters for training RPN model

| | |
|---|---|
| Initialization of network | Guassian distribution |
| Batch size | 10 |
| Learning rate | 0.0001 |
| Momentum | 0.9 |
| Weight decay | 0.00001 |

Fig. 4.7 shows the recall of 1000 proposals on small, medium, and large lung nodules at different IoUs ratios with ground-truth boxes in the proposed method, Conv4 outputs (C4), and deconvolution of Conv5 outputs(C5+Deconv) [100] of the VGG16 model. The results have been obtained for three modes of CT scan images with small nodules ($< 10mm$), medium nodules ($10 - 20mm$)and large nodules($> 20mm$). As shown in Fig. 4.7, the recall of VGG16(C5+Deconv) [100] mode is acceptable only for CT images with large nodules. However, the recall of small nodules is negligible. In this case, the feature map of C4 has more representative features compared to the feature map of the last layer of VGG16. Therefore, this method improves the recall for medium and small nodules. In the proposed method, the combination of semantic and representation features taken from the last three layers of VGG16 is used to generate a feature map, which contains fine-grained features of different sizes. As shown in Fig. 4.7, the accuracy of the proposed method is improved compared to the mentioned methods for all sizes of the lung nodules.

Table. 7.3 shows the average recall (AR) on all sizes of lung nodules when the overlap of IoU with ground-truth boxes is more than 0.7. We report results for 100, 300, 1K, 2K, and 5K ($AR^{100}, AR^{300}, ..., AR^{5K}$). It shows our method has strong semantic and fine-resolution feature maps.

Table 4.3 Average recall (AR(%)) for various numbers of region proposals in IoU=0.7

| | $AR^{100}$ | $AR^{300}$ | $AR^{1K}$ | $AR^{2K}$ | $AR^{5K}$ |
|---|---|---|---|---|---|
| Proposed method | **61.8** | **70.3** | **77.8** | **84.6** | **86.1** |
| C4 | 43.5 | 58.6 | 71.1 | 75.6 | 85.1 |
| C5+Deconv | 30.7 | 37.3 | 49.6 | 52.1 | 61.7 |

In Fig. 4.8, we show the results of using 300, 1k, and 2k proposals for all sizes of nodules. The plots show that, by changing the number of proposal regions, our method has higher recall than other methods. Second, By reducing the number of proposal regions from 1000 to 300, the average recall difference of our method, VGG16(C4), and VGG16(C5+Deconv) is 5.4%, 9.3% and 8.4% respectively. So, our method has a more stable behaviour by reducing the number of proposal regions.

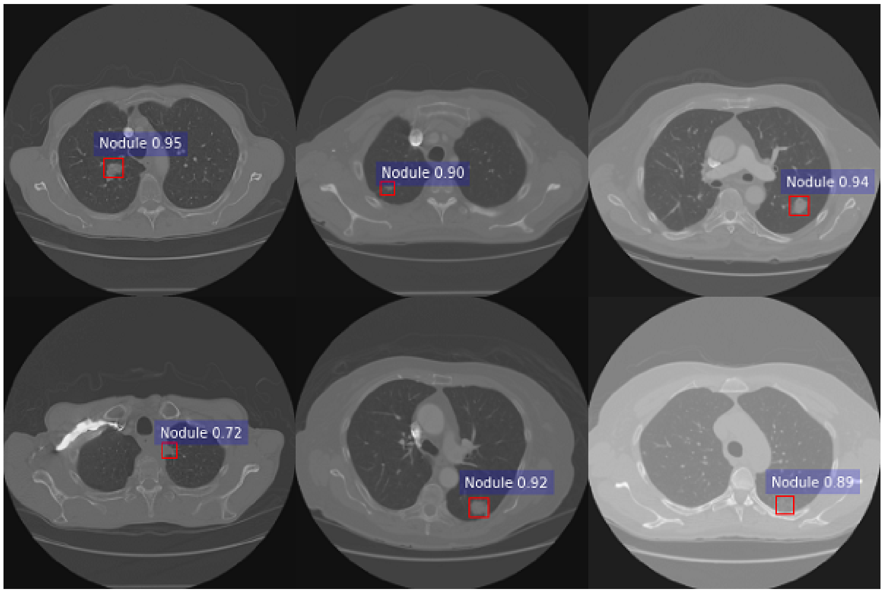


Figure 4.9 Some examples of Intersection-over-Union (IoU) ratios of detected pulmonary nodules with ground-truth bounding boxes

Fig. 4.9 shows some lung nodule candidates detected by our method. The depicted examples are some complex samples, which include solitary nodules, vascularized nodules and juxta-pleural nodules. As shown, the proposed method can detect those nodules with high IoU ratios.

## 4.5 Conclusion

In this study, we presented an improved feature extractor based on the VGG16 convolutional network, trained by a region proposal network (RPN). To improve the performance of the system, we first automatically extracted the lung from CT scans and then entered the 3D segmented lung images into the deep convolutional neural network (DCNN). The experimental results on the LIDC-IDRI dataset show that the feature map extracted by the proposed method performs better than that extracted from the original VGG16 layers. Also, by changing the number of region proposals, our method is more accurate and stable in behaviour than other methods. Our method considers fixed filter sizes of the new layers in VGG16. In future work, we will focus on the filter sizes as a hyperparameter and then change them during the training to get the best result to enhance the system's performance.

# CHAPTER 5 ARTICLE 2: OMS-CNN: OPTIMIZED MULTI-SCALE CNN FOR LUNG NODULE DETECTION BASED ON FASTER R-CNN

**Preface:** This chapter presents an improved Faster R-CNN architecture for early-stage lung cancer detection, incorporating a novel optimized multi-scale convolutional neural network (OMS-CNN). This approach utilizes metaheuristic algorithms to optimize the feature map generation process based on the VGG16 backbone, enhancing the detection of small pulmonary nodules in CT scans. This work has been peer-reviewed and was published in the *IEEE Journal of Biomedical and Health Informatics*, Volume 29, Issue 3, on November 27, 2024.

**Contributions:** This work was conducted during my PhD research at Polytechnique Montréal. I developed the OMS-CNN framework and integrated it within the Faster R-CNN detection pipeline. My responsibilities included designing the multi-scale feature aggregation strategy, applying the PSF-Harmony Search and Beetle Antenna Search algorithms for parameter optimization, implementing the experimental setup, and conducting evaluations on the LUNA16 and PN9 datasets. I also led the manuscript preparation and revision process. My co-authors provided valuable input on algorithm design and clinical validation.

**Full Citation:** Yadollah Zamanidoost, Tarek Ould-Bachir, and Sylvain Martel, *"OMS-CNN: Optimized Multi-Scale CNN for Lung Nodule Detection Based on Faster R-CNN," IEEE Journal of Biomedical and Health Informatics*, Vol. 29, No. 3, pp. 2148–2160, November 27, 2024.

**DOI**: 10.1109/JBHI.2024.3507360

## 5.1 Abstract

The global increase in lung cancer cases, often marked by pulmonary nodules, underscores the critical importance of timely detection to mitigate cancer progression and reduce morbidity and mortality. The Faster R-CNN approach is a two-stage, high-precision nodule detection method designed for detecting small nodules, particularly in computed tomography (CT) images. This paper presents an improved Faster R-CNN by introducing an optimized multi-scale convolutional neural network (OMS-CNN) technique for feature map generation. This approach aims to achieve an optimal feature map through metaheuristic optimization by combining the last three layers of the VGG16 architecture. The advanced parameter-setting-

free harmony search (PSF-HS) algorithm is utilized to implement this method, automatically adjusting the number of channels in the composite layers as a hyperparameter. The beetle antenna search (BAS) optimization algorithm is utilized to effectively initialize the kernel filter weights and biases in the composite layers, thereby enhancing training speed and detection accuracy. In the false-positive reduction stage, a combination of multiple 3D deep convolutional neural networks (3D DCNN) is designed to reduce false-positive nodules. The proposed model was evaluated using the LUNA16 and PN9 datasets. The results demonstrate that the OMS-CNN technique effectively extracted representative features of nodules at various sizes, achieving a sensitivity of 94.89% and a CPM score of 0.892. The comprehensive experiments illustrate that the proposed method can enhance detection sensitivity and manage the number of false positive nodules, thereby offering clinical utility and serving as a valuable point of reference.

## 5.2 Introduction

Lung cancer stands as one of the deadliest known diseases worldwide, comprising nearly two-thirds of all existing cancers [102]. The mortality rate of this type of cancer among all recognized tumours is 18% [103]. Studies indicate that early detection of lung cancer can improve treatment outcomes and increase patients' survival rates [104–106]. Among the available imaging modalities, computed tomography (CT) imaging is important in lung cancer detection and diagnosis [107,108]. With increased access to CT equipment, physicians review a substantial volume of CT images daily. However, due to the prolonged time required for physicians to examine each CT scan, errors in cancer detection may occur due to fatigue or external factors, posing significant risks to patients [109]. Therefore, to reduce individual errors, computer-aided detection (CAD) systems have been developed to assist physicians in rapidly and accurately identifying tumours.

Given the diversity of lung nodules in CT images, accurate diagnosis presents a significant challenge. Consequently, various methods have been proposed in recent years for detecting lung nodules [33,55,110–115]. One type of lung nodule that is particularly difficult to detect by proposed systems is the identification of very small lung nodules with a diameter of fewer than 6 millimetres [116]. Detecting these types of nodules will aid in early lung cancer diagnosis. Lung nodule detection systems typically identify nodules in two stages. The first stage involves extracting candidate nodules to increase the system's sensitivity. Traditional methods for extracting remaining nodules used threshold-based or region-based algorithms, which perform poorly in extracting nodules with lower contrast than the surrounding tissue. The second stage involves removing false-positive candidate nodules to increase the system's precision. However, lung nodule detection approaches need more efficiency, such as lengthy processes, lack of end-to-end detectability, and challenges with larger datasets [117].

In previous years, object detection in medical images has not seen significant advancements due to hardware limitations in machine performance. However, with improved processor speeds and the introduction of deep learning techniques, various object detection methods in images have been widely presented in recent years. For example, Faster R-CNN [4, 117], and Cascade R-CNN [118] are two-stage object detection techniques that accurately detect objects. Additionally, YOLO [119, 120] and SSD [121] are one-stage methods that rapidly detect objects. Attention mechanisms, such as the deep CNN with dual attention mechanism [122], enhance lung nodule detection by incorporating channel and spatial attention to refine feature representation. This allows the model to focus on the most significant details within lung nodule images, improving its ability to distinguish subtle nodules from surrounding tissues. In detecting lung nodules, the integration of region proposal networks in Faster R-CNN is utilized, leading to increased accuracy in detecting lung nodules, especially small ones.

The original Faster R-CNN model encounters challenges when applied to lung nodule detection. One of the critical challenges of Faster R-CNN in detecting small objects is the limited spatial resolution of feature maps at higher convolutional layers. This can lead to difficulty in accurately capturing small objects' details and distinctive features, making their detection less reliable. One solution to this challenge is incorporating multi-scale features into the Faster R-CNN model. Combining feature maps from different convolutional layers, including those with finer spatial resolutions, the model can capture detailed information from lower layers and semantic context from higher layers. This allows for more effective detection of small objects by providing richer and more informative features. Additionally, feature pyramid networks [94] can handle objects at different scales more effectively.

In [123], our focus was on enhancing the efficiency of lung nodule detection using multi-scale CNNs based on Faster R-CNN. Specifically, we optimized feature extraction with the VGG16 model and improved region proposal accuracy through a region proposal network (RPN). This article proposes an improved Faster R-CNN model based on OMS-CNNs to enhance lung nodule detection sensitivity in CT scans. This current work extends that research by completing the remaining stages and significantly improving the performance of the Faster R-CNN algorithm for detecting nodules of various sizes. The primary contributions of this paper are as follows:

1. An improved Faster R-CNN framework has been developed by enhancing and optimizing the multi-scale CNNs feature extraction model. The system exhibits robust and

accurate identification of nodules across various sizes, particularly small ones. This capability holds promise for easing the burden on medical practitioners and lowering the likelihood of diagnostic errors.

2. The use of metaheuristic algorithmic techniques, specifically advanced PSF-HS and BAS optimization, has been suggested for determining the optimal number of composite layers and optimizing the initial weights and biases of these layers. These techniques improve the accuracy and efficiency of pinpointing prospective areas, thereby decreasing the number of redundant nodules and accelerating the detection process within the specified context.

3. A novel approach is proposed for the false positive reduction stage, which combines 3D deep CNNs and exhibits commendable performance. Unlike 2D convolutional neural networks, 3D CNNs utilize three-dimensional contextual data, capturing richer spatial intricacies and producing inherently more distinct features to characterize pulmonary nodules.

The remainder of this paper is structured as follows: Section 5.3 offers an overview of the original faster R-CNN model. Section 5.4 outlines the design framework for an automated pulmonary nodule detection system using OMS-CNN. Section 5.5 elaborates on the experimental outcomes and ensuing discussions. The concluding section encapsulates the key findings of this study.

## 5.3 Faster R-CNN Model

The Faster R-CNN model is an advanced approach to object detection in images, leveraging deep convolutional neural networks. In this model, input images undergo feature extraction via a CNN to generate a feature map. Subsequently, regions of interest (RoIs) likely to contain objects are automatically identified using an RPN.

A notable feature of this model is its utilization of an anchor-based strategy for generating proposed regions. A set of anchor boxes is obtained for each pixel in the feature map, effectively considering various points in the image and thereby adeptly identifying different objects. In the subsequent stage, these proposed regions are utilized for classification and regression tasks. Each RoI is transformed into a feature vector through another convolutional network, which is then used for classification and object localization through classification and regression.

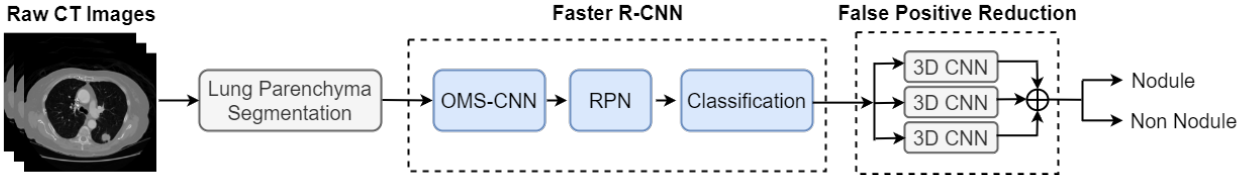


Figure 5.1 Overall framework of an automatic pulmonary nodule detection system based on OMS-CNN.

With this approach, Faster R-CNN demonstrates high proficiency in object detection in images. Combining deep convolutional networks with anchor-based methods achieves higher object detection accuracy and speed than previous approaches.

The Faster R-CNN approach is widely used for object detection in medical images. Therefore, this method is employed to detect lung nodules in CT images. However, one of the significant challenges in using this approach for lung nodule detection is achieving sufficient accuracy. Detecting small nodules is particularly important for early-stage lung cancer detection. While the convolutional neural networks used in Faster R-CNN are robust in feature extraction, they tend to abstract features of such small nodules excessively during the convolution process, leading to undesirable detection outcomes.

To address the challenge above, this paper enhances the initial model by focusing on the following aspects: First, the resolution of the input image is improved, and the lower dimensional features are concatenated with the higher-dimensional features by path augmentation [123]. Second, essential hyperparameters like the number of combined channels are fine-tuned using the sophisticated advanced PSF-HS optimization technique [6]. Third, in the conventional Faster R-CNN architecture, the weights and biases of combined kernel filters are initialized randomly before the training process, resulting in a decline in diagnostic precision. To counteract these challenges, the BAS optimization algorithm [124] is employed to initialize combined kernel filter weights and biases. Fourth, to enhance the accuracy of bounding box prediction, the generalized intersection over union (GIoU) [125] loss function is implemented instead of the intersection over union (IoU) loss function.

Fig. 5.1 illustrates an automated pulmonary nodule detection system based on OMS-CNN. This system takes three-dimensional CT scan images as input and outputs the position of nodules. The implementation of this system aims to achieve high sensitivity in nodule detection while reducing the average number of false positives per scan.

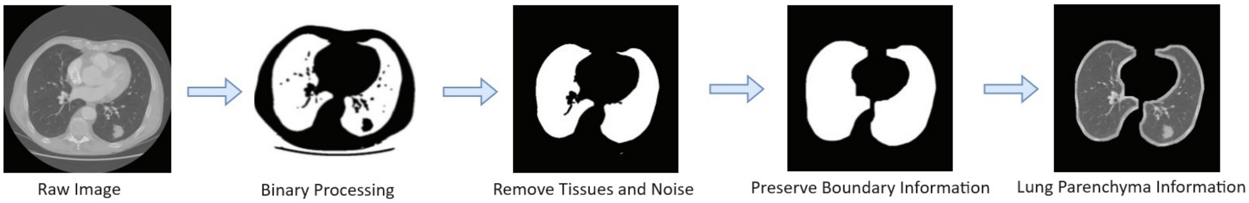


Figure 5.2 Segmentation process of lung parenchyma.

## 5.4 Design Framework

### 5.4.1 Lung Parenchyma Segmentation Strategy

Automated lung segmentation from CT chest scans plays a vital role in removing extraneous elements, thereby enhancing the accuracy of lung nodule detection. As depicted in Fig. 5.2, the raw images are initially binarized. In this process, pixel values are adjusted within the -1000 HU to 400 HU range and then normalized to the standard range of 0 to 1. Subsequently, a threshold segmentation is performed based on the mean pixel value of the CT scan image, dividing the chest region into external and internal compartments. To improve this stage, a 4-connected neighbourhood operator is utilized to eliminate noise and additional internal and external tissues. One of the challenges of automated lung segmentation is the exclusion of nodules attached to the walls. Morphological operations such as opening and closing are employed to address this issue. Finally, to enhance accuracy, small noisy regions are eliminated using the morphological operation of 'Binary Fill Holes' [93]. After completing the steps above, the mask is extracted, and the lung parenchyma is well-segmented.

### 5.4.2 Optimized Multi-Scale CNN (OMS-CNN)

The selection of an appropriate feature extraction architecture is a significant factor influencing the performance of modern convolutional neural networks for lung nodule detection. Various models with different feature extraction architectures exist, with the most prominent ones being DensNet, VGGNet, and ResNet. These models can extract object features from images with high accuracy. However, their performance diminishes when faced with small lung nodules. One of the most common feature extraction architectures is VGG16, characterized by compact 3x3 kernels and optimal layers that enable the detection of small nodules with higher accuracy. In [123], feature maps are generated by combining the final three layers of the VGG16 network to extract features from diverse nodule sizes, especially small ones.

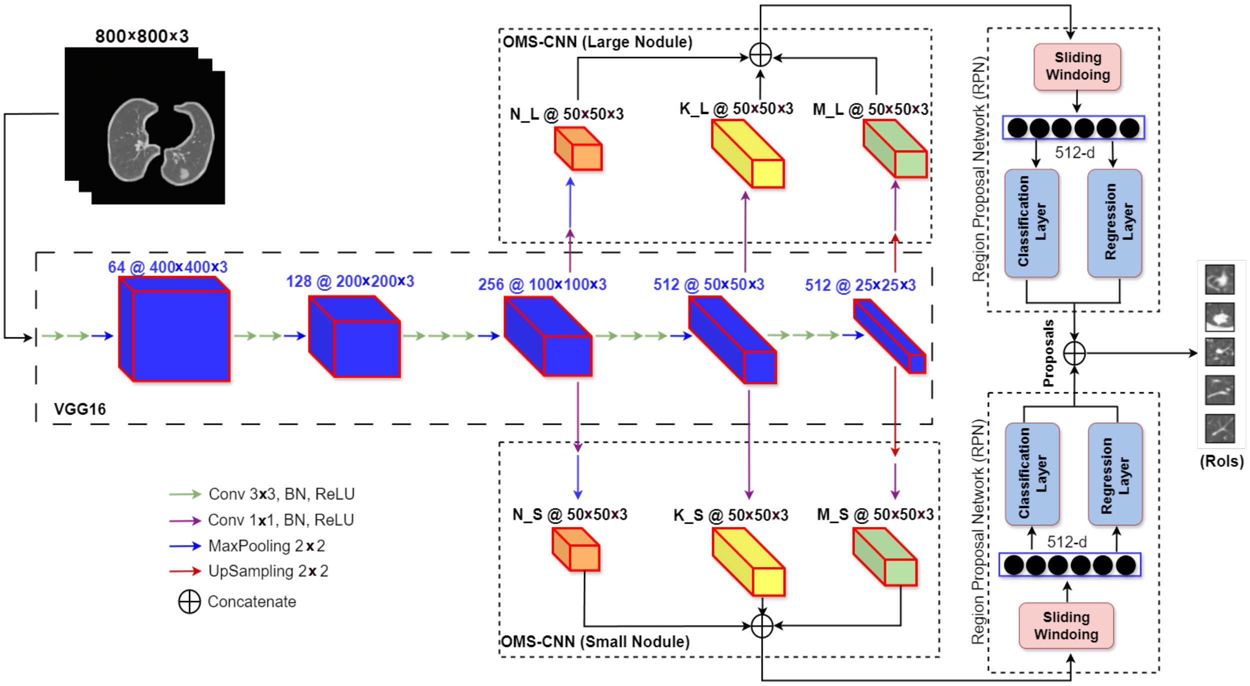


Figure 5.3 Overall framework of optimized multi-scale CNN model (OMS-CNN).

This paper proposes an efficient feature extractor structure called OMS-CNNs for extracting optimal RoIs, as depicted in Fig. 5.3. The backbone of this structure utilizes the VGG16 model and comprises two independent RPNs. To enhance the accuracy of RoI extraction, the advanced PSF-HS optimization method is employed to adjust the number of combined kernels [N, K, M] in the final layers of the VGG16. Additionally, the BAS optimization method is utilized instead of the random initialization approach to adjust the weights and biases of the composite layer filters.

### Input and Output

We opted for a 3D input data approach instead of 2D input data. Due to the substantial computational demands associated with using the original 3D volume of the CT scan as input for the nodule detection network, we resorted to employing axial slices as the input data. This was achieved by consolidating the primary CT scan slice housing the nodule with the adjacent upper and lower slices. 3D input data imparts a richer contextual backdrop and a more comprehensive portrayal of the nodule, aiding the model's differentiation between nodules and other structures or artifacts. Consequently, we extract the three contiguous slices for every axial slice in the CT images and convert this data into an $800 \times 800 \times 3$ image.

### Network Structure

The fundamental architecture of the feature extractor is based on VGG16, comprising five convolutional groups. The upper layers encompass a semantic feature map and get more intricate details from nodules. Conversely, the lower layers offer heightened resolution but cannot extract finer intricacies from nodules. To harness the advantages of detailed features with enhanced resolution, we concatenate the upper and lower layers of VGG16 (Fig. 5.3).

By adopting this configuration, we can selectively access features from feature maps of three distinct layers, encompassing heightened precision and resolution traits. In this structure, two identical configurations are employed. Each of these configurations possesses an independent RPN for extracting proposed regions. The number of kernels in the composite layers of each structure differs from one another. These kernel counts are adjusted within one structure to extract features of large nodules and within another to extract features of small nodules.

### Hyperparameter Tuning Stage

In neural networks, hyperparameters are parameters predetermined by individuals or set automatically through an external model mechanism. Hyperparameter optimization pertains to selecting the most suitable set of hyperparameters for a machine-learning algorithm. Techniques for hyperparameter optimization encompass grid search, Bayesian optimization, random search, and gradient-based optimization [126]. For CNNs, these hyperparameters involve configuring parameters like kernel size, stride, the number of channels, zero-padding, and more.

In this paper, we present an approach for fine-tuning the hyperparameters of the feature extraction phase in the multi-scale CNN model. This optimization is accomplished using an advanced PSF-HS algorithm [6]. In the context of the PSF-HS algorithm, we designate harmony as the variable of interest to be optimized. Our working hypothesis posits that we can achieve the optimal hyperparameter settings by configuring the hyperparameters as harmonies and employing the PSF-HS algorithm to adjust these harmonies iteratively.

Two crucial factors, harmony memory consideration rate (HMCR) and pitch adjustment rate (PAR), utilized in the harmony search (HS) algorithm significantly impact its effectiveness. These parameters are essential in determining whether to utilize the variable within the harmony memory (HM) with the best previously calculated values or to adjust and use it anew. The PSF-HS algorithm [127] operates with fixed HMCR and PAR values over a set number of iterations. However, this approach encounters a limitation in the PSF-HS method, where only one value is employed after a certain stage of the HS process. To address this, the advanced PSF-HS approach [6] has been introduced, which incorporates a maximal improvisation setup that adjusts both HMCR and PAR values.

The equations used to determine HMCR and PAR are as follows:

$$\text{HMCR} = 0.5 + 0.5 \times \textit{sigmoid}(10\frac{i}{n} - \frac{5}{\log(v)}) \tag{5.1}$$

$$\text{PAR} = \text{HMCR} \times \textit{sigmoid}(\frac{4}{v} - 2) \tag{5.2}$$

$$\textit{sigmoid}(x) = \frac{1}{1 + e^{-x}} \tag{5.3}$$

Here, $i$ signifies the current iteration, $n$ indicates the maximum number of iterations, and $v$ denotes the number of variables subject to modification within the HS algorithm.

Furthermore, the pitch adjustment bandwidth ($bw$) is used to set the upper limit for the adjustment magnitude, influenced by both the HMCR and PAR probabilities. This is calculated as:

$$bw = (U - L) \times \lambda \tag{5.4}$$

In Eq. (5.4), $U$ and $L$ represent the upper and lower bounds of the variable inputs and $\lambda$ is a constant parameter adjusted within the range of 0.01 to 0.1 [6].

Finally, to update the value of HM at each iteration, we use the following equations:

$$\text{HM} = \begin{cases} \text{HM} + \text{PAR} \times bw, & \textit{rand}() \geq 0.5 \\ \text{HM} - \text{PAR} \times bw, & \textit{rand}() < 0.5 \end{cases} \tag{5.5}$$

$$\text{HM}_{new} = \max(L, \min(U, \text{HM})) \tag{5.6}$$

In this paper, as illustrated in Fig. 5.3, the number of channels in the composite layers is treated as a hyperparameter for feature map creation. Therefore, the parameters $N$, $K$, and $M$ represent the number of channels in the third, fourth, and fifth layers of CNNs, respectively. By varying their values, one can determine the optimal combinations $[N_S, K_S, M_S]$ and $[N_L, K_L, M_L]$ for achieving high-precision detection of small and large lung nodules, respectively.

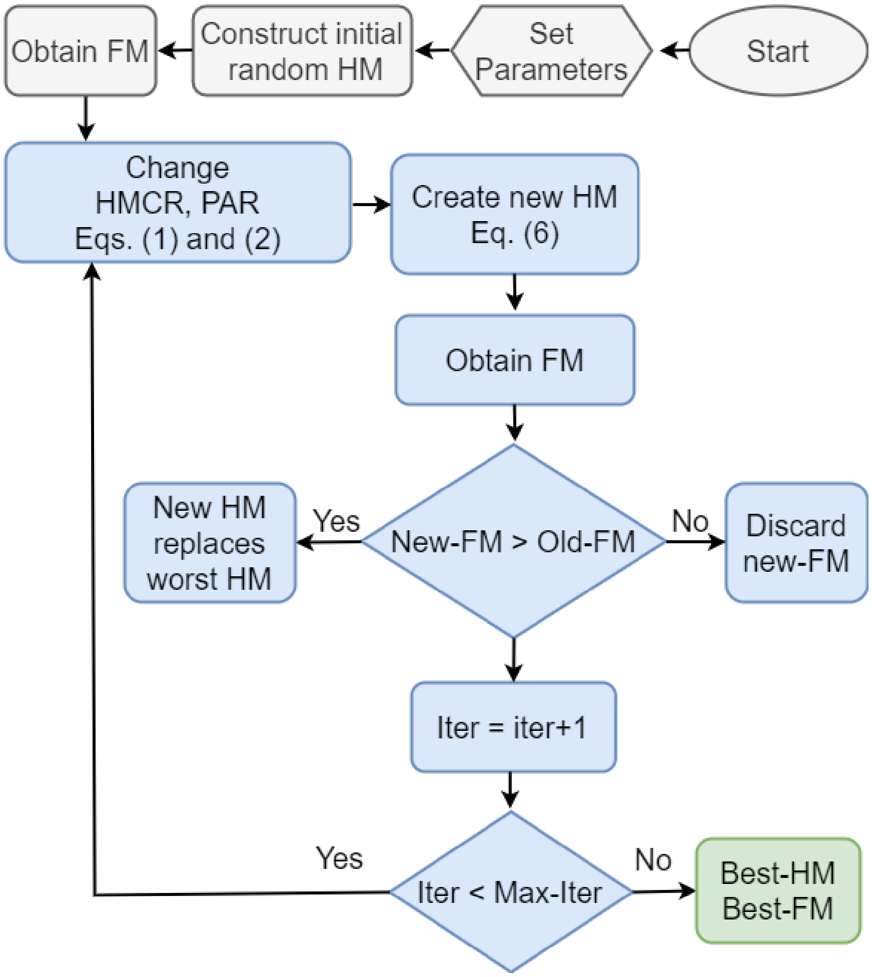


Figure 5.4 The advanced PSF-HS optimization algorithm.

As depicted in Fig. 5.4, the initial parameters — such as the harmony memory size (HMS), maximum number of iterations, and variable ranges — are initialized. Subsequently, the HM matrix is generated randomly within these variable ranges. To compute the fitness memory (FM) using the values $[N, K, M]$ obtained from HM, the RPN network is trained, and the accuracy of the trained model is recorded in FM. Subsequently, a new HM matrix is formed using the HMCR and PAR, and a corresponding new FM matrix is generated. If the new FM outperforms the previous one, the old HM matrix is replaced with the new one; otherwise, no changes are made. This process of generating new HMs is repeated at each iteration until the optimal HM matrix is obtained.

This paper aims to enhance the accuracy of lung nodule detection by utilizing two separate Region Proposal Networks (RPNs): one specialized for identifying small nodules and the other for large nodules. The advanced PSF-HS algorithm is applied to optimize this process

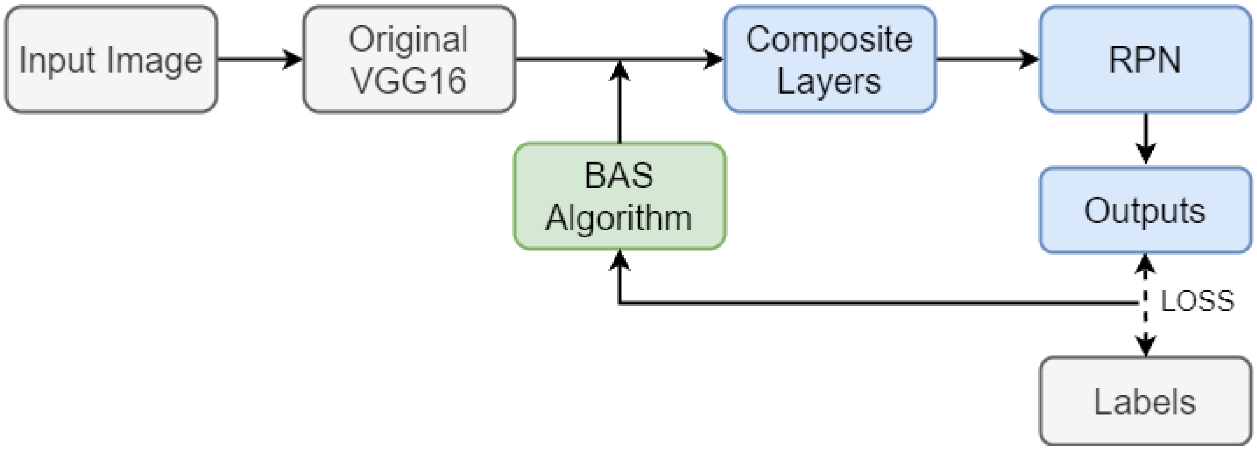


Figure 5.5 The model for initial weight and bias optimization using the BAS algorithm.

to select the most effective combination of VGG16 layers for each RPN, ensuring precise detection of both small and large lung nodules.

### Initial Weights and Biases Optimization Stage

The initial assignment of weights and biases in a CNN significantly influences the training pace and the accuracy of the region of interest extraction. CNNs learn by adjusting these parameters through the backpropagation of errors during training. This continual adjustment process, applied to the convolutional and fully connected layers, progressively leads to achieving the desired learning outcomes. Typically, in CNNs, these weights and biases are initially set randomly using various methods [128–130], resulting in unnecessary computational overhead during training. To overcome this challenge and improve training efficiency, we adopt a two-step approach: First, we initialize the weights and biases of all VGG16 layers by leveraging a pre-trained model for ImageNet classification [101]. Subsequently, we utilize the BAS optimization algorithm to fine-tune the weights and biases of additional composite layers before training. Inspired by beetle search principles, the BAS algorithm is a sophisticated metaheuristic optimization technique [7]. It offers a straightforward implementation and requires minimal computational resources. This strategy effectively addresses the drawbacks associated with CNNs stemming from the random initialization of weights and biases [131].

Concerning the optimized model illustrated in Fig. 5.5, the best configuration involves initializing the parameters of compound layers within a designated number of iterations before commencing model training. Implementing the BAS algorithm necessitates defining several parameters, such as the step size $\delta^t$, the maximum iteration count $N$, the frequency of step size adjustments $\eta$, the initial direction of the beetle represented as $b$, the initial spacing $d_0$

between the beetle's antennae, and the beetle's starting position $w^0$.

$$b = \frac{\text{rands}(k,1)}{||\text{rands}(k,1)||} \tag{5.7}$$

$$\begin{cases} w^{rt} = w^t + d_0 b/2 \\ w^{lt} = w^t - d_0 b/2 \end{cases} \tag{5.8}$$

The initial direction of the beetle is calculated by Eq. (5.7). In this equation, *rands* is a random function that generates a $k$-dimensional column vector comprising random numbers ranging from -1 to 1. Here, $k$ is chosen as the dimensions of the CNN weight matrix. Furthermore, the position of the beetle antennae must be determined based on the initial position of the beetle that is calculated in Eq. (5.8).

$$w^{t+1} = w^t - \delta^t b sign(f(w^{rt}) - f(w^{lt})) \tag{5.9}$$

$$f = \frac{1}{N} \sum_{x_i} -[y_i \log(a_i) + (1 - y_i) \log(1 - a_i)] \tag{5.10}$$

$$sign(w) = \begin{cases} 1, & w > 0 \\ 0, & w = 0 \\ -1, & w < 0 \end{cases} \tag{5.11}$$

$$\delta^{t+1} = \delta^t \times \eta \tag{5.12}$$

The beetle's location is updated regarding Eq. (5.9). In this equation, $f$ represents the algorithm's fitness function, calculated in Eq. (5.10). In this scenario, $x_i$ stands for the $i$th image instance in the dataset, with $y_i$ indicating the corresponding diagnostic label. The variable $a_i$ represents the output generated by processing image $x_i$ through the feature extractor. Over a set number of iterations ($N$), adjustments are made to the weights matrix until the algorithm reaches its optimal fitness level. In addition, the BAS algorithm enhances its search accuracy by Eq. (5.12), which updates the step size by its step update frequency.

In this paper, we employ the BAS optimization algorithm to discover improved weight and bias values for the filters in the new convolutional layers before training. This allows us to substitute the initial random initialization process with optimized parameters, thus creating an optimal model for RoI extraction.

### Loss Function

In deep learning networks, model parameters iteratively undergo training to minimize the loss function. The Faster R-CNN model uses two independent loss functions for classification and regression layers. Typically, the IoU metric is used during regression loss evaluation as a standard, measuring the overlap between the predicted bounding box and the ground-truth box. Eq. (5.13) quantifies the regression loss in this scenario.

$$\begin{cases} \mathrm{IoU} = \frac{|B^{\mathrm{GT}} \cap B^{\mathrm{pred}}|}{|B^{\mathrm{GT}} \cup B^{\mathrm{pred}}|} \\ \\ \mathrm{Loss}_{\mathrm{IoU}} = 1 - \mathrm{IoU} \end{cases} \tag{5.13}$$

where $B^{\mathrm{GT}}$ is the real bounding box, $B^{\mathrm{pred}}$ is the predicted bounding box.

The proposed method uses the Generalized IoU (GIoU) loss function instead of $IoU$ to determine the regression loss. As depicted in Eq. (5.14), when the predicted bounding box and the ground-truth box do not overlap, $GIoU$ ranges between -1 and 0, while IoU is 0. This advantage of $GIoU$ addresses the issue of gradient optimization instability [125].

$$\begin{cases} \mathrm{GIoU} = \mathrm{IoU} - \frac{|B - (B^{\mathrm{GT}} \cup B^{\mathrm{pred}})|}{|B|} \\ \\ \mathrm{Loss}_{\mathrm{GIoU}} = 1 - \mathrm{GIoU} \end{cases} \tag{5.14}$$

where $B$ represents the minimal bounding box that contains both $B^{GT}$ (ground-truth box) and $B^{pred}$ (prediction box).

### 5.4.3 Classification Stage

After combining the RoIs extracted from two RPNs and eliminating duplicate RoIs, a deep convolutional neural network (DCNN) is designed to make binary classifications for each RoI, determining whether it represents a nodule or not.

The positions of candidate nodules, referred to as regions of interest (ROIs), are predicted by the RPN regression layer. These values represent the candidate nodule's centre position and the width and height $(W, H)$ of ROI patches. Subsequently, three-dimensional patches are extracted from the feature map using the obtained values. The output values of the RPN classification layer are utilized to select appropriate patches for input into the classification network. These output values range between 0 and 1. Therefore, patches with a probability value greater than 0.5 are selected as nodule patches and inputted into the classification network [49].

An RoI pooling layer is initially employed to project each RoI onto a smaller feature map with a predetermined spatial dimension of $W \times H$ (specifically, $7 \times 7$ as outlined in this paper). The RoI pooling process involves dividing the RoI into a grid of sub-windows measuring $W \times H$ and performing max-pooling within each sub-window, resulting in values being mapped to their corresponding output grid cells. This pooling operation is carried out independently across each feature map channel, akin to standard max pooling procedures. After the RoI pooling layer, a fully connected network comprising two 4096-dimensional fully connected layers is utilized to transform the fixed-size feature map into a feature vector. Ultimately, a classifier is employed to predict confidence scores for potential candidates. The training of the binary classification model generally uses CrossEntropyLoss as the loss function to optimize the model.

### 5.4.4 False Positive Reduction

A combination of multiple 3D CNNs is utilized in the phase to reduce false positives. The final outcome is determined through a voting mechanism. Fig. 5.6 illustrates the passage of a 3D image patch through a sequence of three trained 3D CNN models, with the final outcome determined by a voting process.

#### Network Structure

We obtain 3D patches of size $(32 \times 32 \times 32)$ by extracting various slices to identify potential nodules. We chose this dimension not only because it encapsulates the majority of nodules but also because it provides ample contextual information, which is particularly valuable for analyzing smaller nodules. Subsequently, we employ flip and duplication methods to augment the data. The dataset is then partitioned into five subsets: three for training, one for validation, and one for testing.

The primary influencing factor in object detection is the substantial number of negative samples, comprising the majority of the total loss. Notably, many of these negative samples are easily classifiable [61]. This suggests the significance of hard mining in enhancing 3D CNN performance, emphasizing the necessity for the network to focus on more challenging samples. In line with this concept, the training dataset is refined to include the more intricate samples, which persist in the subsequent model training iterations, thereby augmenting the classification accuracy of each individual model. Each of these models is built upon the VGG-

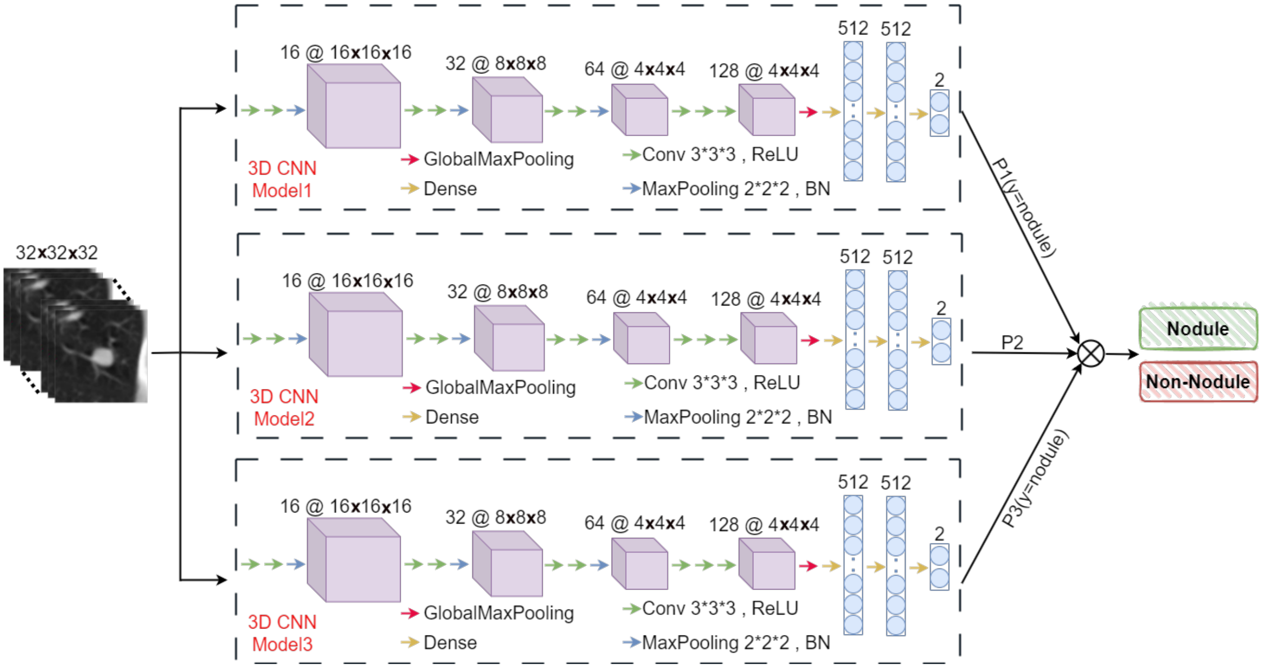


Figure 5.6 The framework of false positive reduction model.

Net, renowned for its strong performance in image classification tasks and widely adopted as a deep learning network for feature extraction [90]. The architecture comprises 12 layers, with $(32 \times 32 \times 32)$ patches serving as inputs. It entails a repetitive sequence of layers, each consisting of two convolutional layers with 16 kernels of size $(3 \times 3 \times 3)$, a max-pooling layer, and a batch normalization layer. This configuration iterates three times, with the number of kernels doubling in each iteration. Subsequently, a global max-pooling layer is employed to preserve the most salient features across the entire dataset. The concluding segment encompasses two fully connected layers and three dense layers.

The training dataset is divided into three segments, each dedicated to training the 3D CNN model separately. Initially, the first segment is utilized for training Model1. Subsequently, misclassified samples from both Model1 and the second segment are employed to train Model2 from scratch independently. Similarly, Model3 is independently trained using misclassified data from Model1, Model2, and the third segment. This iterative process involves utilizing misclassified samples from the previous round as training data for the next model, thus amplifying the importance of these erroneous samples. We can fine-tune the weight parameters during the training of the subsequent model, allowing it to learn more representative features and enhance its ability to differentiate challenging mimics. An overview of the proposed false positive reduction system is illustrated in Fig. 5.7.

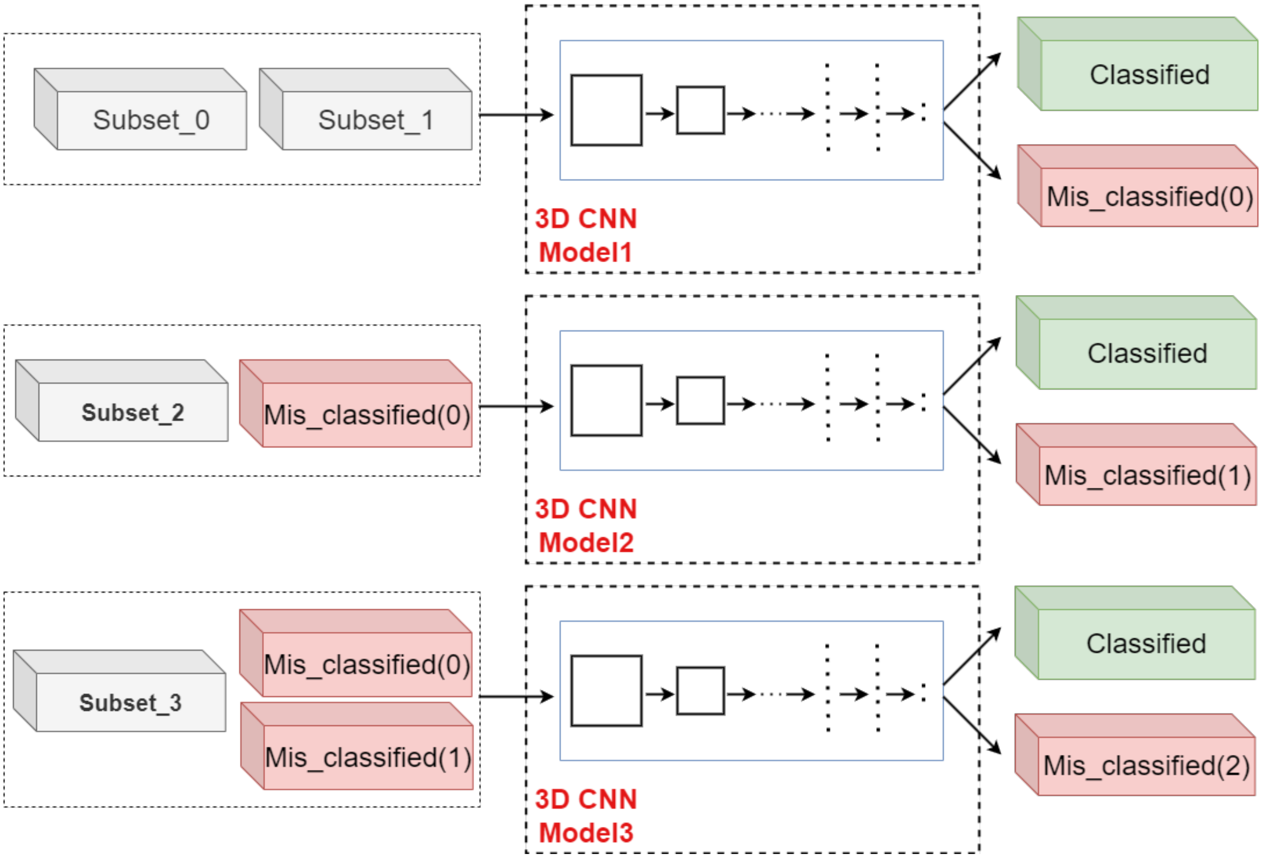


Figure 5.7 The proposed false positive reduction network model.

## 5.5 Experiment Results and Discussion

### 5.5.1 Datasets

Our study uses the LUNA16 dataset [97]. This dataset comprises 888 CT scans and 1186 lung nodules, each with a diameter greater than 3 millimetres. At least three radiologists have evaluated each nodule. The primary data format for CT scans is DICOM, consisting of 100 to 500 axial slices and the size of each slice is 512×512. The CT scans within this dataset are partitioned into three subsets: a training set (622 scans, 70%), a validation set (88 scans, 10%), and a test set (178 scans, 20%).

The primary focus of this research is the detection of lung nodules of various sizes. Therefore, the nodules of test data are categorized into three groups based on their size: non-small nodules ($nodule > 10mm$), small nodules ($6mm < nodule \leq 10mm$), and very small nodules ($3mm < nodule \leq 6mm$). Table 5.1 illustrates the number of test data in each category.

Table 5.1 The number of nodules of various sizes in the LUNA16

| Category | Number |
|---|---|
| Non-Small Nodules (NSN) | 70 |
| Small Nodules (SN) | 70 |
| Very Small Nodules (VSN) | 97 |

In this study, we utilize a distinct dataset named PN9 [114] to evaluate the generalizability of the proposed model. PN9 comprises 8,798 CT scans, featuring 40,439 annotated nodules. The CT scan images were collected from hospitals, where lung nodules were annotated by specialist physicians through a two-step process.

### 5.5.2 Evaluation Metrics

The performance of networks is frequently described using a confusion matrix. The proposed approach's performance is assessed through cross-validation, utilizing metrics such as the free-receiving operating curve (FROC) and the competition performance metric (CPM). Sensitivities are calculated at specific false positive rates (FPRs) per patient, including 1/8, 1/4, 1/2, 1, 2, 4, and 8 FPRs. The CPM for the system is derived by averaging sensitivities at these specific points. Recall(Sensitivity), Precision, and CPM are defined as follows:

$$\text{Recall(Sensitivity)} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{5.15}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{5.16}$$

$$\text{CPM} = \frac{1}{N} \sum_{i \in \text{I}} \text{Recall}_{\text{fpr}=i} \tag{5.17}$$

with $\text{I} = \{0.125, 0.25, 0.5, 1, 2, 4, 8\}$ and where the value of $N$ is set at seven, the variable $fpr$ represents the average number of false positives per scan, while $Recall_{fpr=i}$ signifies the recall rate associated with $fpr = i$.

Following the completion of model training and validation, the test images undergo the computation of both average precision and average recall. The average precision ($AP$) is determined using an IoU threshold of 0.5, while the average recall ($AR$) involves IoU thresholds ranging from 0.5 to 0.95 in 100 region proposals.

### 5.5.3 Detection of Pulmonary Nodules

In this paper, lung nodule detection is carried out in three stages: region proposal extraction, classification, and false positive reduction. In the first stage, the OMS-CNN structure is employed as a feature extraction, and the RPN is utilized for training. It also uses six anchors of different sizes, including 4 × 4, 6 × 6, 10 × 10, 16 × 16, 22 × 22 and 32 × 32 for each sliding window, as in [100]. In this phase, We perform 10-fold cross-validation to assess the system's performance and employ stochastic gradient descent (SGD) optimization with a momentum factor of 0.9. We incorporate a weight decay rate of 0.00001 and set the base learning rate to 0.0001. The network model trains in the 2 V100 GPU environment, and the memory is 192 GB.

In the context of the classification network, it is imperative to address the challenge of class imbalance. This issue is solved by equilibrating the quantities of negative and positive patches, leveraging the output from the trained RPN. Within this approach, in conjunction with the ground truth, we identify suggested region proposals with an intersection over union (IoU) exceeding 0.7 as positive patches and select an equivalent number of suggested regions with an IoU less than 0.1 as negative patches in a randomized manner. The execution of this technique serves to achieve not only class balance but also an augmentation in the quantity of positive patches. Furthermore, the model's parameter updates are facilitated by the utilization of the Adam optimizer. An initial learning rate of 0.001 is employed, and the learning rate is adjusted by applying the cosine decay function.

In the false positive reduction (FPR) stage, the primary data augmentation strategy involves flipping each positive candidate patch along three orthogonal dimensions (coronal, sagittal, and axial positions). In the subsequent stage, the determination of a candidate's positivity or negativity hinges upon whether the geometric centre of the candidate resides within a nodule. Positive patches are replicated eight fold to balance the count of positive and negative patches within the training set. Furthermore, all three network models employ the Adam optimizer, with the learning rate, momentum, and batch size set to 0.0001, 0.4, and 16, respectively.

Fig. 5.8 illustrates the process of lung nodule detection by the proposed method at each stage. The detection trend encompasses nodules of various sizes. The number of region proposals is 100 in this experiment, considering an IoU of 0.5. The results depict the count of classified regions in the classification stage and the number of false positives in the FPR stage.

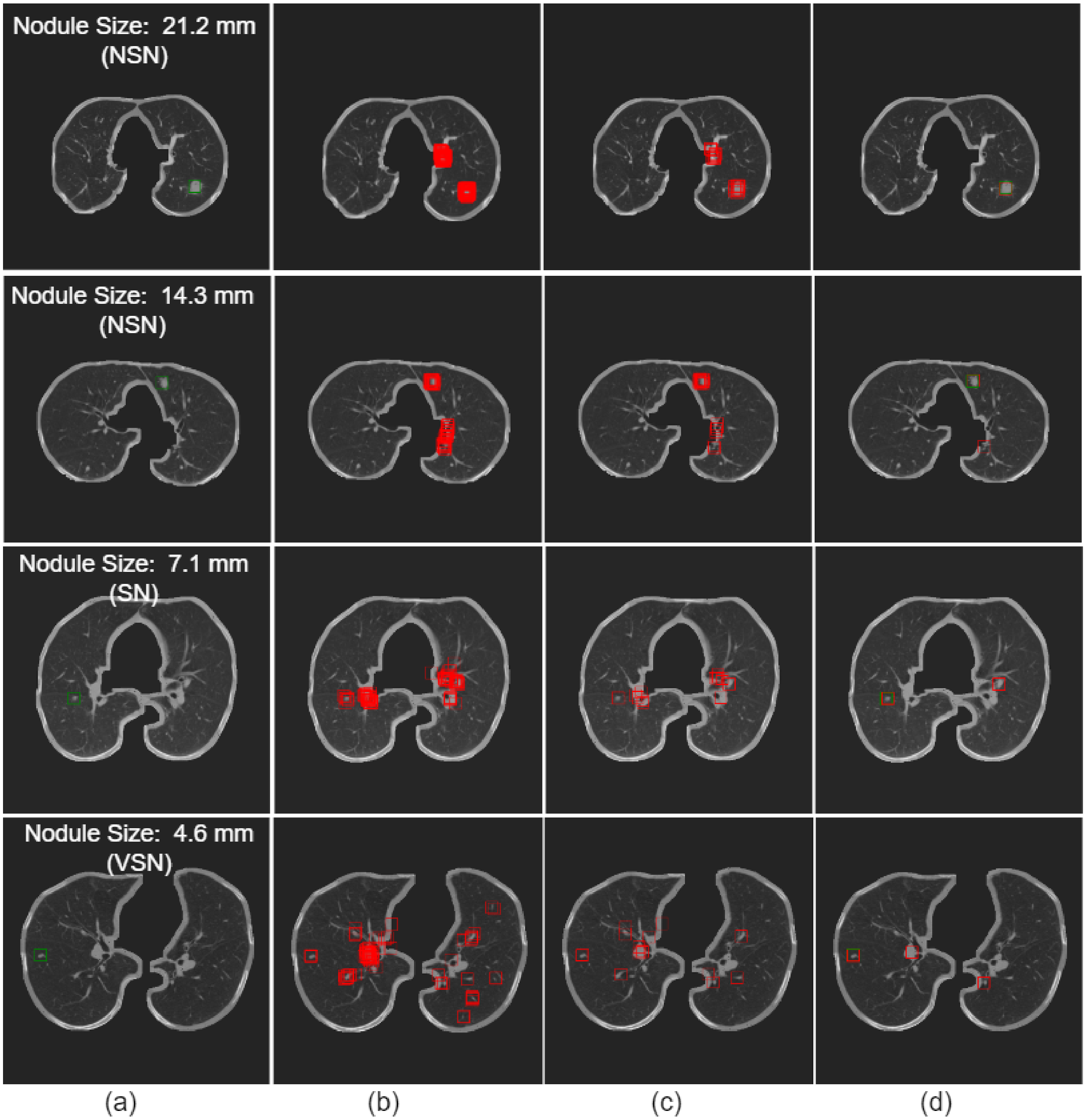


Figure 5.8 The lung nodule detection process with different nodule sizes on LUNA16, (a) The CT image after lung parenchyma segmentation stage, (b) The location of 100 region proposal after RPN stage, (c) The location of classified region proposal after classification stage, (d) The location of lung nodule after false positive reduction stage.

### 5.5.4 Ablation Study

We have designed an ablation experiment based on the Faster R-CNN method to verify the effectiveness of different components in the proposed architecture, employing its foundational

VGG16 backbone model. The results obtained pertain to three categories of data: non-small nodules (NSN), small nodules (SN), and very small nodules (VSN). In addition, the normal CNN (N-CNN) model signifies the Faster R-CNN method, using the last deconvolution layer of VGG16 as the feature map [100]. The multi-scale CNN (MS-CNN) model combines the last three layers of VGG16 as the feature map [123]. In the tuned multi-scale CNN (TMS-CNN) model, the number of combined channels in the last three layers of VGG16 is adjusted as hyperparameters to optimize the accuracy of detecting both small and large lung nodules using the advanced PSF-HS optimization algorithm. Lastly, in the optimised multi-scale CNN (OMS-CNN) model, the weights and biases of the filters from the fine-tuned last three layers are optimized using the BAS optimization algorithm before training.

### The Effectiveness of Multi-Scale CNN

In the MS-CNN approach, a feature map is constructed by combining the last layer, which contains semantic features, with the third and fourth layers of VGG16, which contain representative features. This feature map extracts high-resolution and detailed features of small and very small nodules. By comparing the N-CNN and MS-CNN methods in Fig. 5.9, it is evident that the MS-CNN method shows an increase in recall values for small and very small nodules when the number of proposed regions is set to 1000. On average, this increase amounts to 10.91% for the SN and VSN datasets. Conversely, the average increase for non-small nodules is around 2.12%. Fig. 5.10 displays the FROC curves of various proposed feature extractor structures. The average CPM scores for N-CNN and MS-CNN are 0.785 and 0.808, respectively. The first and second experiments in Table 5.2 demonstrate that the average recall values for SN and VSN have increased by 18.15% and 19.76%, respectively, with only a very slight increase observed for NSN. Furthermore, the MS-CNN method demonstrates a greater increase in average precision and sensitivity for small nodules (SN) and very small nodules (VSN) compared to non-small nodules (NSN), indicating its superior performance in detecting small nodules compared to the N-CNN method.

Table 5.2 Experimental results of various lung nodule detection models on LUNA16.

| | NSN | | | SN | | | VSN | | |
|---|---|---|---|---|---|---|---|---|---|
| Model | AP | AR | Sensitivity | AP | AR | Sensitivity | AP | AR | Sensitivity |
| N-CNN | 87.79% | 70.43% | 95.01% | 73.38% | 41.08% | 85.45% | 52.50% | 20.54% | 53.71% |
| MS-CNN | 89.14% | 70.58% | 96.84% | 78.00% | 48.54% | 91.25% | 56.81% | 24.60% | 62.32% |
| TMS-CNN | 95.48% | 75.27% | 99.17% | 87.58% | 65.91% | 94.26% | 70.84% | 30.08% | 74.13% |
| OMS-CNN | **96.14%** | **77.83%** | **100%** | **90.75%** | **73.19%** | **97.41%** | **75.35%** | **34.49%** | **85.12%** |

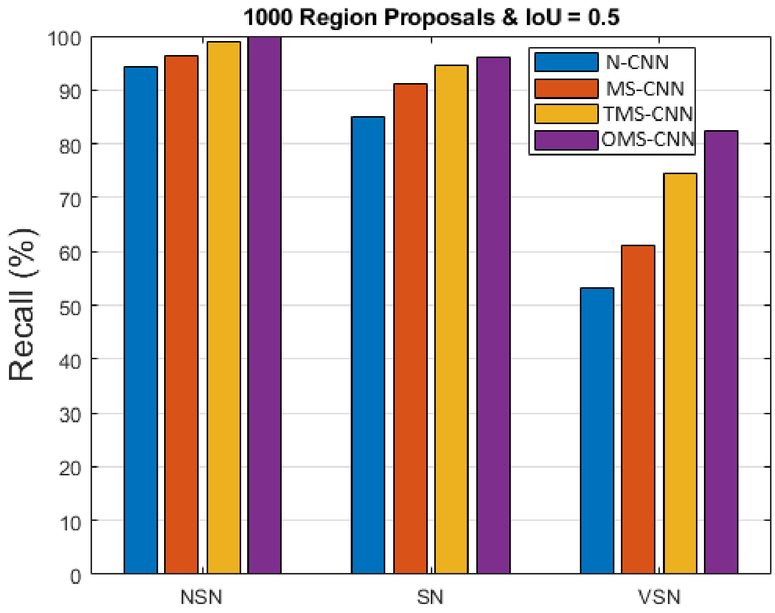


Figure 5.9 The recall of non-small (NSN), small (SN) and very small (VSN) nodules with various feature extractor based on VGG16 after RPN stage on LUNA16.

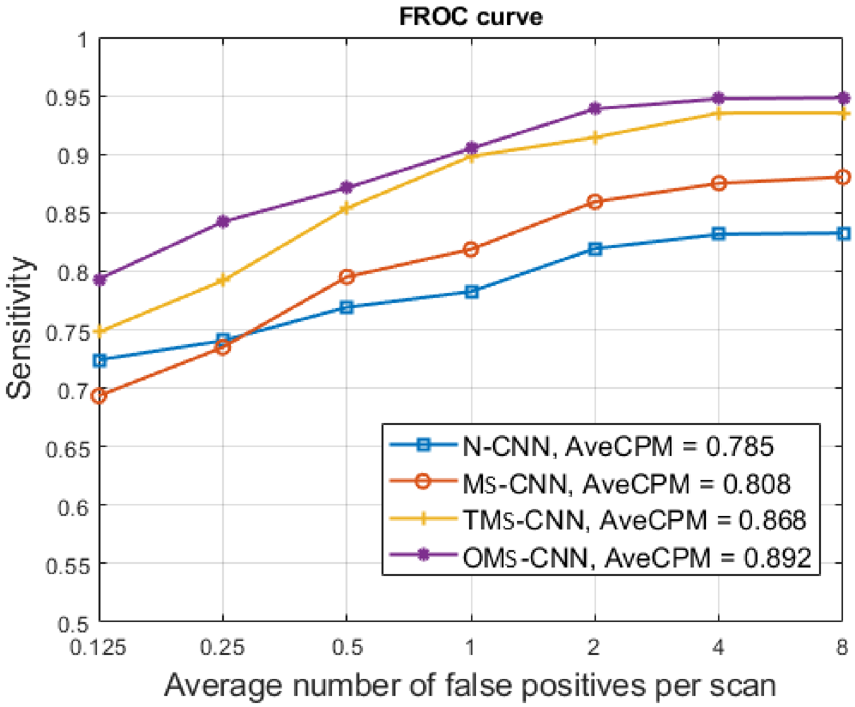


Figure 5.10 FROC curves of different proposed models on LUNA16.

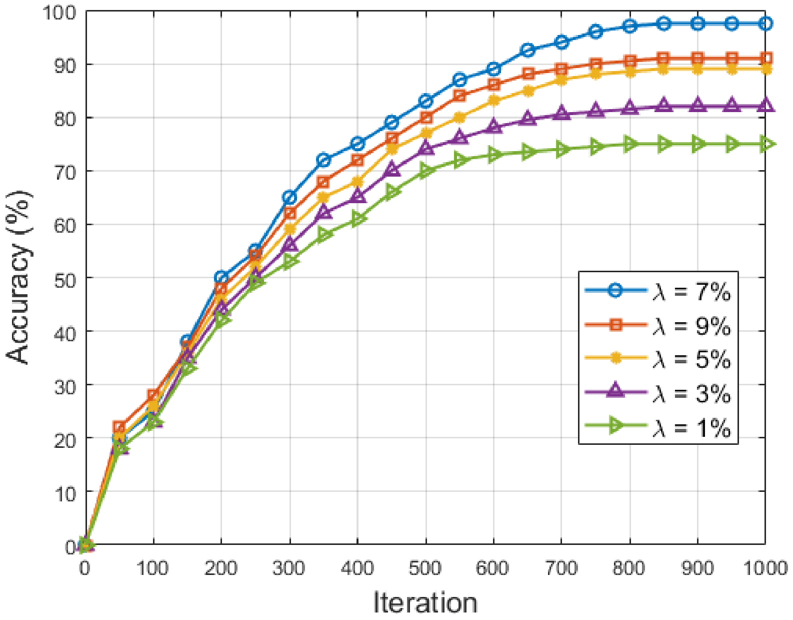


Figure 5.11 Accuracy of lung nodule detection with various pitch adjustment bandwidth on LUNA16.

### The Effectiveness of Tuned Multi-Scale CNN

The number of channels in the composite layers in the MS-CNN method significantly influences the accuracy of the created feature map. In this paper, $N$ represents the number of channels in the third layer of VGG16, while $K$ and $M$ denote the number of channels in the fourth and fifth layers, respectively. Therefore, the combinations $[N_S, K_S, M_S]$ and $[N_L, K_L, M_L]$ are considered hyperparameters that are adjusted before training to detect small and large lung nodules separately by two independent RPNs. The feature map extracted from the upper CNN layers is suitable for detecting larger lung nodules, while the feature map extracted from the lower CNN layers is appropriate for detecting smaller lung nodules. Initial experiments reveal that the optimal range for varying the number of channels in each layer for the detection of different nodule sizes includes $N = [1, 20]$, $K = [500, 510]$, and $M = [1, 20]$. This paper employs the advanced PSF-HS algorithm, a metaheuristic optimization, to precisely adjust the number of channels in each layer. HMS is set to 5 in this experiment for sufficient harmony memory considerations. Each hyperparameter is randomly generated under conditions where the number of channels in the composite layers $[N, K, M]$ is an integer within the specified range. HMCR and PAR are calculated for each hyperparameter using Eq. (5.1) and Eq. (5.2). Fig. 5.11 illustrates that varying the constant $\lambda$ in Eq. (5.4), along with setting the pitch adjustment bandwidth, impacts the accuracy of lung nodule detection. This approach produced good results at 7% pitch adjusting bandwidth over the entire input range in 1000 iterations. After the initial settings, the advanced PSF-HS algorithm is applied. The fitness function value for tuning $[N_S, K_S, M_S]$ is determined by the average accuracy of detecting small and very small nodules, while for tuning $[N_L, K_L, M_L]$, it is equivalent to the accuracy of detecting large nodules. Approximately after 700 iterations, all harmony vectors converge to a vector. The simulation results reveal that the optimal hyperparameter values for detecting small nodules converge to $N_S = 6$, $K_S = 501$, and $M_S = 15$, while for large nodules, they converge to $N_L = 4$, $K_L = 510$, and $M_L = 16$.

As depicted in Fig. 5.9, in the TMS-CNN method, the recall value for various nodule sizes has increased compared to the MS-CNN method, with the highest increase observed for VSN at 22.01%. Additionally, the sensitivity values for TMS-CNN and MS-CNN methods at four FPs per scan are 0.875 and 0.831, respectively (Fig. 5.10). As indicated in Table 5.2, in the third experiment, compared to the second experiment, the average precision (AP) has increased by 24.69%, 11.53%, and 7.12% for the VSN, SN, and NSN datasets, respectively. Thus, in the TMS-CNN approach, through the adjustment of kernel numbers within the composite layers of each RPN, more efficient features are extracted to detect nodules of various sizes.

### The Effectiveness of Optimized Multi-Scale CNN

After determining the values of $[N_S, K_S, M_S]$ and $[N_L, K_L, M_L]$ in the preceding step, the dimensions of the weight and bias matrices for the composite layers are specified. In conventional approaches, the initial weight and bias values of these layers are randomly acquired. This paper employs the BAS optimization method to derive the initial weight and bias values in order to enhance lung nodule detection accuracy. In this study, we set the initial step size $\delta^0$ to 0.8, the step update frequency $\eta$ to 0.95, and the initial distance between the beetle's left and right antennae $d_0$ to 0.5 [124]. In addition, a maximum of 120 iterations is considered, during which the weight and bias values are updated in each iteration based on the equations of the BAS algorithm. Fig. 5.12 illustrates the changes in the optimal loss value during the optimization process. It is observed that after approximately 100 iterations, the optimal values for the initial weights and biases of the composite layers are reached, leading to the attainment of the global optimum of the function.

As indicated in Table 5.2, the AP, AR, and sensitivity values for the OMS-CNN method have increased compared to the TMS-CNN method across various nodule sizes. The highest increases are observed for SN and VSN. In Fig. 5.10, the proposed method achieves the highest CPM score of 0.892 compared to previous methods. Furthermore, the sensitivity reaches 90.48% and 94.73% at one and four FPs per scan, respectively.

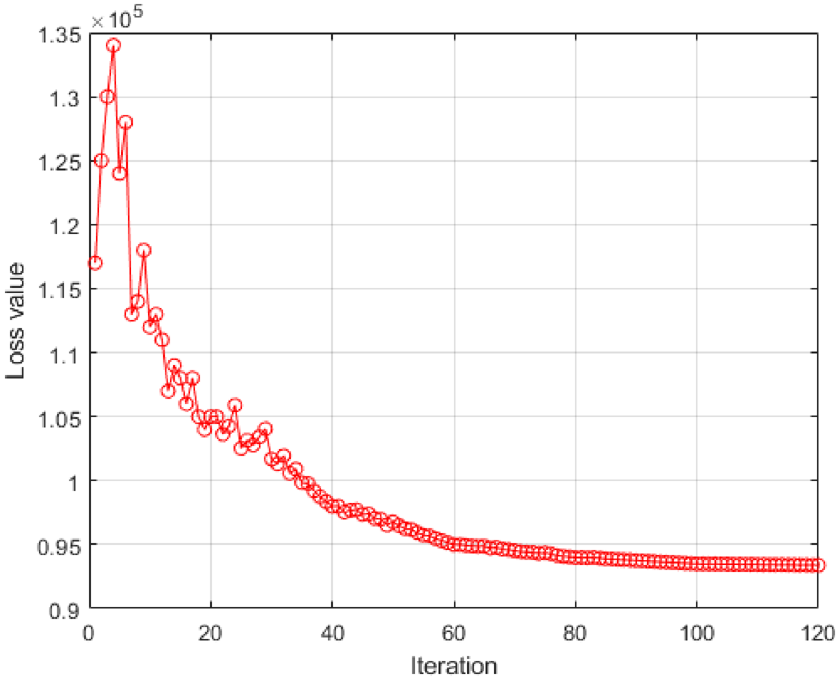


Figure 5.12 Convergence of the minimum loss value by BAS algorithm.

### 5.5.5 Experimental Comparison

To further assess the efficacy of the proposed nodule candidate detection network, the detection outcomes presented in this study on LUNA16 are compared with those of other established methodologies, employing the CPM score for comparison. The quantitative outcomes are detailed in Table 5.3. The tabulated results reveal that our proposed detection network attains the highest CPM score of 0.839.

Table 5.3 Comparison of the proposed candidate nodule detection network with other methods on LUNA16.

| CAD Method | Year | 0.125 | 0.25 | 0.5 | 1.0 | 2.0 | 4.0 | 8.0 | CPM |
|---|---|---|---|---|---|---|---|---|---|
| Dou et al. [132] | (2017) | 0.6590 | 0.7540 | 0.8190 | 0.8650 | 0.9060 | 0.9330 | 0.9460 | 0.8390 |
| Gu et al. [133] | (2018) | 0.4801 | 0.6495 | 0.7920 | 0.8794 | 0.9163 | 0.9293 | 0.9301 | 0.7967 |
| Pezeshk et al. [112] | (2018) | 0.6370 | 0.7230 | 0.8040 | 0.8650 | 0.9070 | 0.9380 | 0.9520 | 0.8320 |
| Xie et al. [111] | (2019) | 0.4390 | 0.6880 | 0.7960 | 0.8520 | 0.8640 | 0.8640 | 0.8640 | 0.7750 |
| OMS-CNN | | 0.7215 | 0.7357 | 0.7993 | 0.8521 | 0.9162 | 0.9243 | 0.9283 | **0.8396** |

Table 5.4 Performance comparison of different methods for false positive reduce on LUNA16.

| CAD Method | Year | 0.125 | 0.25 | 0.5 | 1.0 | 2.0 | 4.0 | 8.0 | CPM |
|---|---|---|---|---|---|---|---|---|---|
| OverFeat [110] | (2015) | 0.6000 | 0.6020 | 0.7100 | 0.7300 | 0.7600 | 0.7700 | 0.7700 | 0.7143 |
| Nodule ResNet [134] | (2017) | 0.5170 | 0.6020 | 0.7200 | 0.7880 | 0.8220 | 0.8390 | 0.8560 | 0.7350 |
| 3D Faster R-CNN [4] | (2018) | 0.6620 | 0.7460 | 0.8150 | 0.8640 | 0.9020 | 0.9180 | 0.9320 | 0.8340 |
| Leaky Noisy-OR [33] | (2019) | 0.5938 | 0.7266 | 0.7813 | 0.8438 | 0.8750 | 0.8906 | 0.8984 | 0.8013 |
| Xie et al. [111] | (2019) | 0.7340 | 0.7440 | 0.7630 | 0.7960 | 0.8240 | 0.8320 | 0.8340 | 0.7900 |
| DeepSEED [113] | (2020) | 0.7390 | 0.8030 | 0.8580 | 0.8880 | 0.9070 | 0.9160 | 0.9200 | 0.8620 |
| Zeo et al [53] | (2020) | 0.6300 | 0.7530 | 0.8190 | 0.8690 | 0.9030 | 0.9150 | 0.9200 | 0.8300 |
| CBAM [54] | (2021) | 0.4670 | 0.6020 | 0.7300 | 0.812 | 0.8770 | 0.9150 | 0.9310 | 0.7620 |
| I3DR-Net [55] | (2022) | 0.6356 | 0.7131 | 0.7984 | 0.8527 | 0.8760 | 0.8992 | 0.9147 | 0.8128 |
| MSM-CNN [49] | (2022) | 0.6770 | 0.7410 | 0.8160 | 0.8500 | 0.8900 | 0.9050 | 0.9250 | 0.8290 |
| MS-3DCNN [48] | (2023) | 0.7280 | 0.7990 | 0.860 | 0.8080 | 0.9260 | 0.9410 | 0.9560 | 0.8730 |
| AttentNet [135] | (2024) | 0.7520 | 0.8170 | 0.8570 | 0.8850 | 0.9200 | 0.9330 | 0.9330 | 0.8710 |
| MK-3DCNN [56] | (2024) | 0.7099 | 0.7723 | 0.8356 | 0.8836 | 0.9174 | 0.9384 | **0.9562** | 0.8591 |
| TED [51] | (2024) | 0.7619 | 0.8222 | **0.8736** | **0.9069** | 0.9302 | 0.9443 | 0.9530 | 0.8846 |
| TMS-CNN |  | 0.7479 | 0.7918 | 0.8538 | 0.8981 | 0.9143 | 0.9352 | 0.9352 | 0.8778 |
| OMS-CNN |  | **0.7932** | **0.8421** | 0.8712 | 0.9048 | **0.9387** | **0.9473** | 0.9481 | **0.8922** |

The false positive reduction network categorizes candidate nodules acquired in the preceding stage, thereby eliminating false positive instances to enhance the accuracy of the detection outcome. To evaluate the efficiency of the suggested automated system for detecting pulmonary nodules, we benchmark our results against other leading methodologies using the LUNA16 dataset. In comparison to the MSM-3DCNN [48], which enhances the performance of Faster R-CNN by utilizing a combination of multiscale feature maps and the K-means++ clustering method to improve the scale and proportion of anchor boxes in the RPN, our proposed method integrates multiscale feature maps and optimizes the number of combined channels and their initial weights and biases using advanced PSF-HS and BAS algorithms, respectively. In comparison to AttentNet [135], which enhances 3D lung nodule detection by introducing a 3D cross-channel and cross-sectional spatial attention unit and utilizing a fully convolutional network to efficiently apply attention with richer spatial descriptors, our approach improves the feature extraction structure in the RPN and combines multiple 3D CNNs for false positive reduction, resulting in superior performance in lung nodule detection. In comparison to the TED [51], which employs a transformer encoder-decoder to capture long-range dependencies and provide a global description, our method demonstrates superior performance, notably achieving a 3.13% improvement in sensitivity at 0.125 false positives per scan. These findings validate that our OMS-CNN method outperforms TED [51] and also exhibits a high detection rate under conditions of lower false positives per scan. Table 5.4 presents the detection sensitivities at seven different false positives per scan (FPs/Scan) and the CPM score. Our proposed detection system achieves the highest CPM score of 0.892. The sensitivities at 0.125, 0.25, 2, and 4 FPs/scan are 0.793, 0.842, 0.938,

Table 5.5 The sensitivity and CPM score compared with other methods on PN9.

| CAD Method | Year | 0.125 | 0.25 | 0.5 | 1.0 | 2.0 | 4.0 | 8.0 | CPM |
|---|---|---|---|---|---|---|---|---|---|
| SSD512 [121] | (2016) | 0.0462 | 0.0848 | 0.1476 | 0.2506 | 0.4032 | 0.5727 | 0.7080 | 0.3161 |
| RetinaNet [61] | (2017) | 0.0260 | 0.0556 | 0.1095 | 0.1925 | 0.2929 | 0.4049 | 0.5105 | 0.2274 |
| NoduleNet [136] | (2019) | 0.2117 | 0.3023 | 0.4038 | 0.5102 | 0.6129 | 0.7070 | 0.7693 | 0.5025 |
| SA-Net [114] | (2021) | 0.2672 | 0.3603 | 0.4746 | 0.5699 | 0.6635 | 0.7352 | 0.7832 | 0.5506 |
| I3DR-Net [55] | (2022) | 0.1564 | 0.2313 | 0.3700 | 0.5154 | 0.6454 | 0.7291 | 0.7753 | 0.4890 |
| OMS-CNN | | 0.2865 | 0.3841 | 0.4775 | 0.5907 | 0.6974 | 0.7853 | 0.8432 | **0.5807** |

and 0.947, respectively, surpassing those of the best-performing method presented.

Additionally, to demonstrate the efficacy, generality, and robustness of our approach, we evaluate the proposed trained model on the PN9 database. Subsequently, we compare the obtained results with several advanced methods, including NoduleNet [136], I3DR-Net [55], and SA-Net [114]. Table 5.5 illustrates that our OMS-CNN method attains a CPM score of 0.580, indicating noteworthy performance in comparison to other methods. Thus, the method proposed in this study demonstrates superiority and significant clinical value.

Table 5.6 Comparison of Training and Inference Times for Various Methods of the Proposed Faster R-CNN on LUNA16.

| Proposed Faster R-CNN | | | | |
|---|---|---|---|---|
| Without Optimization | Advanced PSF-HS | BAS Algorithm | Training Time/fold (h) | Test Time/CT (s) |
| ✓ | – | – | 18.46 | 2.95 |
| – | ✓ | – | 16.73 | 2.76 |
| – | ✓ | ✓ | 14.25 | 2.72 |

Table 5.6 illustrates the comparison of training and inference times for various methods of the proposed Faster R-CNN across 10-fold cross-validation. This model without optimization attained a training time per fold of 18.46 hours and a test time per CT of 2.95 seconds. By applying the advanced PSF-HS algorithm for hyperparameter tuning and the BAS algorithm for weight and bias optimizations, the training time and inference time decreased by 4.21 hours and 0.23 seconds, respectively. Therefore, the proposed Faster R-CNN with these optimizations helps improve both training and inference times by finding the most efficient parameter values, potentially resulting in a more accurate model that requires fewer complex operations.

Table 5.7 presents the results of experiments conducted on various proposed feature extraction network based on Faster R-CNN for the entire test dataset. As indicated, the OMS-CNN achieves a recall rate of 61.83%, demonstrating a significant improvement of 40.49% compared to the N-CNN model. Additionally, the proposed method achieves a precision rate of 87.41%. The proposed method for detecting potential nodules demonstrates a sensitivity of 94.89%.

Table 5.7 Experimental results of different proposed feature extraction networks based on Faster R-CNN with LUNA16.

| Model | AP | AR |
|---|---|---|
| N-CNN | 71.22% | 44.01% |
| MS-CNN | 72.98% | 47.91% |
| TMS-CNN | 84.63% | 57.08% |
| OMS-CNN | **87.41%** | **61.83%** |

On average, there are 9.25 candidates per scan.

## 5.6 Conclusion

In this paper, we applied a deep learning-based lung nodule detection algorithm to CT images and proposed an optimized multi-scale CNN feature extraction method within the Faster R-CNN algorithm to address certain challenges that arise in lung nodule detection. First, we extracted the lung parenchyma using simple image processing methods. Second, we combined features from the last three layers of the VGG16 structure to create a feature map with higher precision. Third, we adjusted the number of combined channels using the advanced PFS-HS optimization algorithm to obtain the best feature map for accurately detecting small and large lung nodules. Fourth, rather than determining the initial weight and bias of the combined channel filters randomly, we utilized the BAS optimization algorithm to enhance these filters' weight and bias. Furthermore, we propose a combined 3D CNN model for the false positive reduction stage, aimed at removing false positive samples to enhance the accuracy of the detection results.

After several experiments, it has been demonstrated that the proposed model outperforms the recent computer-aided detection (CAD) model in lung nodule detection on LUNA16 and PN9 datasets. Incorporating a deep CNN with a dual attention mechanism in the classification stage of the Faster R-CNN algorithm could further enhance sensitivity, particularly in identifying small and hard-to-detect nodules. Further improvements in other components of the Faster R-CNN algorithm and refinements to the false positive reduction process are anticipated to boost the overall precision of lung nodule detection in future research.

# CHAPTER 6 ARTICLE 3: DA OMS-CNN: DUAL ATTENTION OMS-CNN WITH 3D SWIN TRANSFORMER FOR EARLY-STAGE LUNG CANCER DETECTION

**Preface:** This chapter introduces an advanced lung nodule detection method based on the Faster R-CNN framework, enhanced by the integration of a dual-attention optimized multi-scale CNN (DA OMS-CNN), dual-attention RoIPooling, and 3D Swin Transformers. These components collectively improve the model's sensitivity and reduce false-positive detections in early-stage lung cancer screening. This work has been peer-reviewed and was published as an open-access article in *Informatics*, Volume 12, Issue 3, on July 7, 2025.

**Contributions:** This research was conducted during my doctoral studies at Polytechnique Montréal. I initiated and led the conceptual development of the dual-attention OMS-CNN framework, designed the DA-RoIPooling module for region-wise refinement, and integrated an ensemble of 3D Swin Transformers for false-positive reduction. I also performed the experimental evaluations on LUNA16 and PN9 datasets and authored the manuscript. My co-authors contributed to architectural insights, evaluation design, and extensive feedback on manuscript revisions.

**Full Citation:** Yadollah Zamanidoost, Matis Rivron, Tarek Ould-Bachir, and Sylvain Martel, *"DA OMS-CNN: Dual-Attention OMS-CNN with 3D Swin Transformer for Early-Stage Lung Cancer Detection," Informatics*, vol. 12, no. 3, article 65, July 7, 2025.

**DOI**: 10.3390/informatics12030065

## 6.1 Abstract

Lung cancer is one of the most prevalent and deadly forms of cancer, accounting for a significant portion of cancer-related deaths worldwide. It typically originates in the lung tissues, particularly in the cells lining the airways, and early detection is crucial for improving patient survival rates. Computed tomography (CT) imaging has become a standard tool for lung cancer screening, providing detailed insights into lung structures and facilitating the early identification of cancerous nodules. In this study, an improved Faster R-CNN model is employed to detect early-stage lung cancer. To enhance the performance of Faster R-CNN,

a novel dual-attention optimized multi-scale CNN (DA OMS-CNN) architecture is used to extract representative features of nodules at different sizes. Additionally, dual-attention RoIPooling (DA-RoIpooling) is applied in the classification stage to increase the model's sensitivity. In the false-positive reduction stage, a combination of multiple 3D shift window transformers (3D SwinT) is designed to reduce false-positive nodules. The proposed model was evaluated on the LUNA16 and PN9 datasets. The results demonstrate that integrating DA OMS-CNN, DA-RoIPooling, and 3D SwinT into the improved Faster R-CNN framework achieves a sensitivity of 96.93% and a CPM score of 0.911. Comprehensive experiments demonstrate that the proposed approach not only increases the sensitivity of lung cancer detection but also significantly reduces the number of false-positive nodules. Therefore, the proposed method can serve as a valuable reference for clinical applications.

## 6.2 Introduction

The pursuit of technological advancements in healthcare remains a continuous and pressing endeavor, especially in light of the critical need to mitigate the devastating effects of serious illnesses such as cancer [137–139]. Among the myriad forms of this disease, lung cancer represents a particularly formidable global threat, claiming countless lives with little warning. Data from the world health organization (WHO) [8] starkly illustrate the scale of this issue, with 2.21 million new lung cancer cases reported in 2020, constituting 11.4% of all cancer diagnoses worldwide. Furthermore, the estimated 1.8 million deaths attributed to lung cancer that year reaffirm its status as the primary cause of cancer-related mortality on a global level. Despite the differences in lung cancer prevalence across regions, demographics, and age groups, there is an unwavering need for early-stage detection to improve patient outcomes. Early diagnosis is widely recognized as a critical factor in enhancing the success of treatment interventions and increasing survival rates.

In response to this urgent health challenge, medical researchers and technology experts have joined forces to explore innovative strategies that can potentially transform the approach to lung cancer diagnosis and therapy. One of the most promising areas of advancement involves the application of deep learning (DL) algorithms to the identification of lung nodules within diagnostic imaging modalities such as X-rays [139], computed tomography (CT) scans [138], and magnetic resonance imaging (MRI) [140]. In particular, the Faster R-CNN algorithm has emerged as a prominent tool for the early detection of lung cancer.

The integration of advanced technologies like Faster R-CNN into the diagnostic process marks a significant step forward in equipping healthcare professionals with powerful tools for more accurate detection and treatment of lung cancer [141]. This synergy between medical imaging and DL algorithms offers a beacon of hope in the fight against lung cancer, as the remarkable capabilities of Faster R-CNN for accurate identification of cancerous nodules open up new possibilities for early intervention. As efforts continue globally to address this pressing health issue, there is renewed optimism for a future where early lung cancer diagnosis can become a standard, potentially saving numerous lives and providing hope to those affected by this devastating disease. Faster R-CNN operates as a two-stage, region-based detection system that excels in extracting significant information from medical images, including MRIs and CT scans. Its detection methodology initiates with the generation of a comprehensive set of candidate regions, followed by classification and refinement using convolutional neural networks (CNNs).

The Faster R-CNN method is extensively utilized for detecting objects in medical imaging, particularly for identifying lung nodules in CT scans. A significant obstacle encountered when employing this technique for lung nodule detection is achieving adequate accuracy. The precise identification of small nodules is essential for the early detection of lung cancer. Although the CNNs that are part of the Faster R-CNN framework excel at feature extraction, they often overly generalize the attributes of these small nodules during the convolutional process, which can lead to less-than-optimal detection results.

In ref. [50], we introduced a novel architecture, OMS-CNN, which employs VGG16 as its backbone to enhance feature extraction. This architecture improves feature map representation by integrating the final layers of VGG16 and optimizing the number of merged channels to facilitate the detection of both large and small lung nodules. To further refine this process, the advanced PSF-HS optimization algorithm [6] is applied for channel selection, while the BAS optimization algorithm [7] is utilized for initializing the weights and biases of the merged layers. Additionally, an ensemble of multiple 3D CNNs is incorporated to mitigate false-positive detections. The overall framework of this approach is depicted in Figure 6.1. In this study, we propose an enhanced Faster R-CNN model based on DA OMS-CNN, which improves the sensitivity of early-stage lung nodule detection.

Although Faster R-CNN has demonstrated considerable success in object detection tasks within medical imaging, its effectiveness in identifying small and morphologically diverse lung nodules remains limited [142]. This shortcoming is primarily due to the constraints of traditional CNN-based feature extractors and standard RoIPooling methods, which often struggle to capture subtle spatial details and contextual variations critical for early-stage nodule detection. To overcome these challenges, we enhance the OMS-CNN architecture with a dual-attention mechanism designed to strengthen the model's capacity to emphasize both spatially significant regions and channel-specific features within the input data. Additionally,

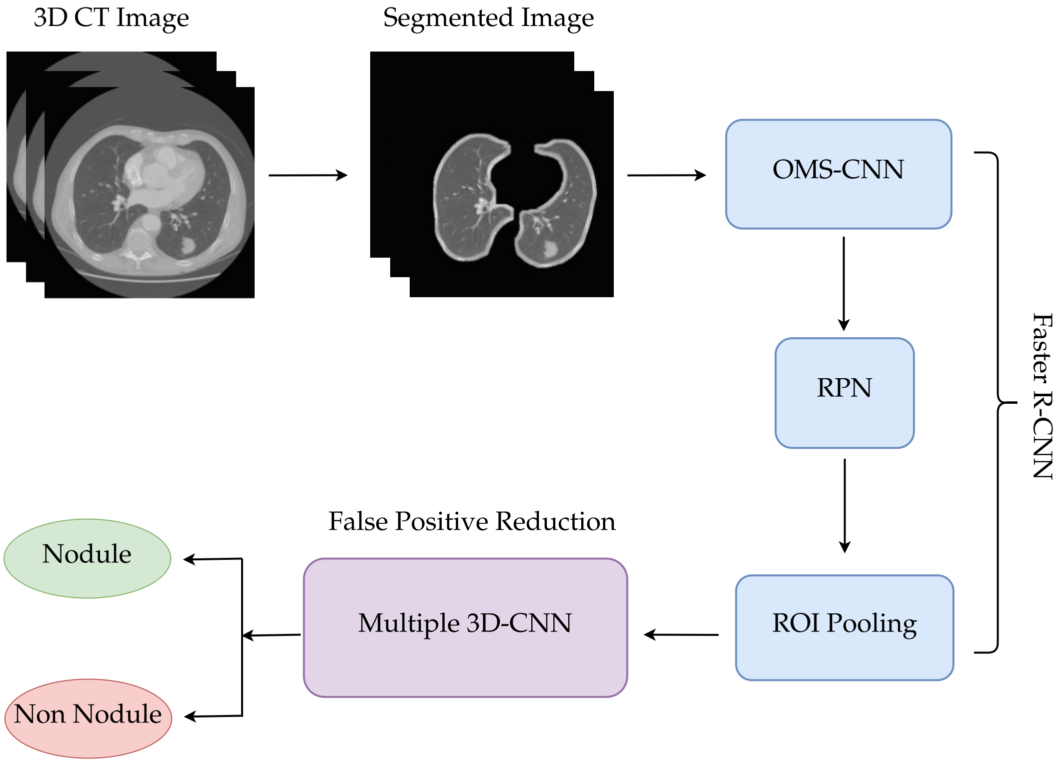


Figure 6.1 Overall framework of an automatic pulmonary nodule detection system based on OMS-CNN [50].

we propose a novel dual-attention RoIPooling (DA-RoIPooling) module that integrates attention into the region of interest feature extraction process. This approach enables the model to better isolate and utilize the most salient features within each region, thereby improving its ability to discriminate true nodules from benign structures or artifacts. Collectively, these methodological advancements aim to address key limitations of existing Faster R-CNN-based approaches by improving detection sensitivity and substantially reducing false-positive rates. The main contributions of this paper diverge from the existing literature in several key aspects:

- The first contribution of this study is the integration of a dual-attention mechanism into the final layers of the OMS-CNN. The dual-attention mechanism enhances the network's ability to capture both spatial and channel-wise dependencies within the feature maps. By incorporating both spatial attention, which emphasizes important regions in the image, and channel attention, which focuses on relevant feature channels, the DA OMS-CNN achieves improved sensitivity in detecting small lung nodules. This approach ensures that critical regions and fine-grained details in the input data are highlighted, leading to more accurate and robust feature extraction.

- The second contribution is the introduction of the dual-attention RoIPooling (DA-RoIPooling) mechanism at the classification stage of the framework. DA-RoIPooling applies spatial and channel-wise attention to the pooled features, enabling the model to focus on the most relevant features within each region of interest (RoI). This dual-attention mechanism ensures that the classification network emphasizes the key characteristics of the nodules while suppressing irrelevant background information. By refining the feature representation within the RoIs, DA-RoIPooling improves the overall classification accuracy, particularly in distinguishing true nodules from false positives. This innovation significantly enhances the performance of the Faster R-CNN framework by reducing misclassifications and improving sensitivity and precision, particularly for challenging cases.

- The third contribution involves the utilization of three distinct 3D Swin Transformers for the false-positive reduction stage. This approach leverages the powerful feature representation capabilities of the 3D Swin Transformer, which uses hierarchical feature extraction and self-attention mechanisms across spatial and temporal dimensions. By combining three separate 3D Swin Transformers, the proposed framework effectively processes volumetric data from different perspectives, ensuring a more comprehensive analysis of nodule candidates. This ensemble strategy reduces false positives by capturing subtle variations and dependencies in the 3D CT data, improving the model's ability to differentiate between true nodules and irrelevant structures. The use of 3D Swin Transformers in this stage not only enhances the overall detection accuracy but also strengthens the robustness of the proposed framework in clinical scenarios.

## 6.3 Related Works

Zamanidoost et al. [123] present a study focused on improving the detection of lung cancer nodules by enhancing feature extraction in convolutional networks. The research addresses the limitations of standard models like VGGNet and ResNet in detecting small objects, such as lung nodules, due to their feature extraction limitations. The authors propose a modified approach using the VGG16 network, known for its 3 × 3 kernels and optimal layer configuration, which can effectively capture features of small objects. Their method involves combining the feature maps from the last three layers of VGG16 to create a comprehensive representation of nodules of varying sizes. A region proposal network (RPN) is used to evaluate the proposed feature map's accuracy compared to the original VGG16. Results indicate that the proposed feature map outperforms traditional VGG16 layers in capturing nodule features and maintains higher recall stability when the number of region proposals is reduced. This approach highlights the potential benefits of optimizing feature extraction strategies in convolutional networks for lung nodule detection.

Zamanidoost et al. [50] propose an enhanced lung nodule detection method by introducing an optimized multi-scale CNN (OMS-CNN) within the Faster R-CNN framework to address the challenges of detecting nodules of varying sizes, particularly small ones. Their approach combines feature maps from the last three layers of the VGG16 architecture to create a detailed representation of nodules, which is further optimized using advanced metaheuristic algorithms, specifically the PSF-HS and BAS. These algorithms fine-tune the number of combined channels and initialize filter weights and biases, significantly enhancing feature extraction precision and efficiency. The optimized OMS-CNN effectively captures multi-scale features, improving the detection sensitivity and robustness of the model. Furthermore, a novel 3D CNN model is employed in the false-positive reduction stage, utilizing three-dimensional contextual data to refine the detection process. Experimental results on the LUNA16 and PN9 datasets demonstrate the effectiveness of the OMS-CNN in achieving higher sensitivity, reducing false positives, and achieving superior CPM scores compared to existing models, highlighting its potential for clinical application in lung nodule detection.

Tan et al. [48] proposed a multi-scale 3D CNN to improve lung nodule detection accuracy while reducing false positives. Their model integrates a 3D UNet++ architecture with a region proposal network and employs cross-layer feature fusion for enhanced feature learning. Using multiple input sizes and residual connections, the model achieves an average sensitivity of 87.3% on the LUNA16 dataset, outperforming UNet++ by 7.8% and VGG16 by 8.1%, demonstrating its effectiveness for clinical applications.

Recent studies highlight the role of attention mechanisms in improving lung nodule detection. Traditional 2D modules like SE and CBAM enhance feature extraction but are computationally expensive for 3D imaging. To address this, Almahasneh et al. [135] introduce AttentNet, a 3D fully convolutional attention mechanism that reduces computational load while preserving feature quality. Evaluations on the LUNA16 dataset demonstrate its efficiency in candidate proposal and false-positive reduction, making it a suitable approach for 3D medical imaging.

Wu et al. [56] propose the multi-kernel driven 3D CNN (MK-3DCNN) to enhance lung nodule detection in CT scans. Their model integrates a residual encoder-decoder structure with a multi-kernel joint learning block to capture multi-scale spatial features. Additionally, a mixed pooling strategy improves feature representation. Experiments on the LUNA16 dataset show superior performance over existing methods, with further validation on the CQUCH-LND clinical dataset demonstrating its practical applicability.

Lung cancer is the leading cause of cancer-related deaths worldwide, emphasizing the need for early detection to improve survival rates. Deep learning has shown great potential in medical imaging, particularly for lung cancer identification in CT scans. Srivastava et al. [143] introduced the hybridized Faster R-CNN (HFRCNN), a two-stage model that generates and refines region proposals using a CNN. Trained on diverse datasets, HFRCNN achieves over 97% detection accuracy, outperforming many existing methods and highlighting the transformative role of deep learning in lung cancer diagnosis.

Ma et al. [51] propose TiCNet, a transformer-enhanced 3D CNN designed for early lung cancer detection. By integrating transformers with CNNs, TiCNet captures both short- and long-range dependencies, improving nodule characterization. The model incorporates attention blocks, multi-scale skip pathways, and a two-head detector to enhance sensitivity and specificity. Evaluations on LUNA16 and PN9 datasets show that TiCNet outperforms existing methods, demonstrating its potential for improving lung cancer screening.

Sun et al. [144] explored the use of the Swin Transformer model for lung cancer detection, demonstrating its potential to improve diagnostic accuracy for radiologists. Their study showed that the pre-trained Swin-B model achieved a top-1 accuracy of 82.26%, surpassing the Vision Transformer (ViT) by 2.529%. In segmentation tasks, the Swin-S model outperformed traditional methods, showing significant improvements in mean intersection over union (mIoU). This research highlights the effectiveness of pre-trained transformers in enhancing medical imaging performance, advancing reliable diagnostic tools for lung cancer detection.

These contributions illustrate the wide array of deep learning approaches that have been explored to enhance lung cancer detection, ranging from modifications of classical CNN architectures such as VGG16 and ResNet to the integration of attention mechanisms and transformer-based models. While these methods have undoubtedly advanced the field, they also exhibit several limitations. Many CNN-based models struggle to selectively focus on the most informative regions or features, thereby limiting their sensitivity, particularly for small or ambiguous nodules. Attention mechanisms and transformer models, though promising, often suffer from high computational costs or are insufficiently integrated into multi-stage detection pipelines. Moreover, few approaches effectively combine attention mechanisms at both the feature extraction and region pooling stages, or leverage 3D context in the false-positive reduction phase. These gaps highlight the need for a unified framework that integrates spatial and channel-wise attention, multi-scale feature learning, and volumetric

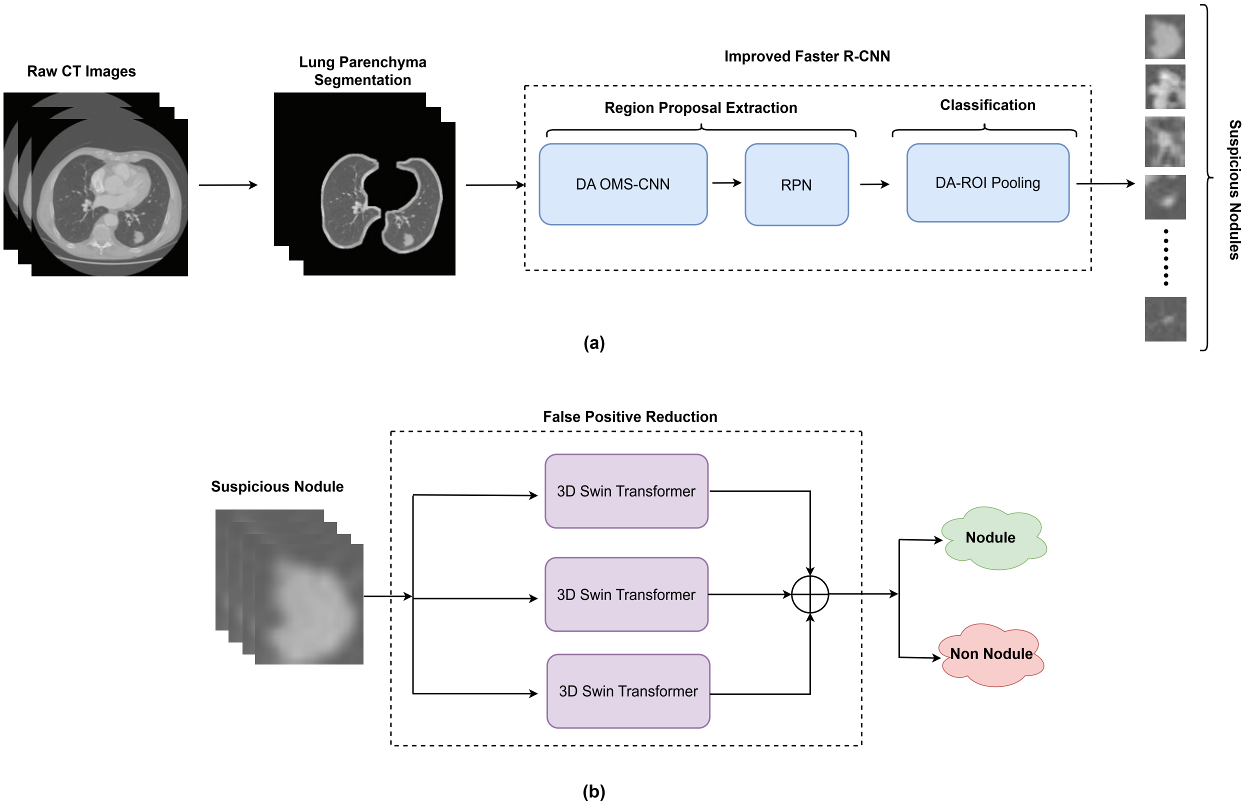


Figure 6.2 Overall framework of an early-stage lung cancer detection system based on DA OMS-CNN: (**a**) the proposed lung nodule detection framework; (**b**) the proposed false-positive reduction framework.

analysis to improve both sensitivity and specificity. In response to these challenges, our work proposes a novel architecture that incorporates dual-attention-enhanced feature extraction, dual-attention RoIPooling, and a 3D Swin Transformer ensemble to provide a comprehensive solution for early-stage lung cancer detection.

## 6.4 Materials and Methods

Figure 6.2 illustrates an automated pulmonary nodule detection system based on the DA OMS-CNN. This system processes three-dimensional CT scan images as input and outputs the positions of detected nodules. The implementation is designed to achieve high sensitivity in nodule detection while minimizing the average number of false positives per scan.

The process begins with image preprocessing, where CT scans are segmented and normalized to enhance data quality. Feature extraction is performed using the DA OMS-CNN, which incorporates spatial and channel attention mechanisms to focus on discriminative regions within the scans. Subsequently, the RPN identifies candidate regions of interest (RoIs) that may contain lung nodules. To refine these candidate regions, the DA-RoI Pooling mechanism is employed, ensuring that the most relevant features are extracted for further analysis. The refined RoIs are then processed through the stack of 3D SwinT blocks, which are used for false-positive reduction. These blocks effectively capture both global and local dependencies in the 3D space of CT scans. The outputs of the 3D SwinT networks are aggregated to deliver accurate predictions, enabling the robust identification of pulmonary nodules.

### 6.4.1 Dataset and Preprocessing

To develop and evaluate the proposed framework for lung nodule detection, two datasets were utilized: the LUNA16 [28] and PN9 [114] datasets. These datasets were chosen due to their high-quality annotations and complementary characteristics. The LUNA16 dataset was used for training, validation, and testing purposes, while the PN9 dataset was employed to evaluate the model's generalization capability. This two-pronged approach ensures that the model not only performs well on the training data but also generalizes effectively to unseen data from different clinical sources.

#### LUNA16

The LUNA16 dataset, derived from the LIDC-IDRI, is a widely used benchmark dataset in lung disease research. It consists of 888 low-dose thoracic CT scans with detailed annotations by multiple expert radiologists. The dataset includes nodules with varying sizes, shapes, and malignancy probabilities, offering a diverse set of examples for model training and evaluation. Each nodule is annotated with descriptors such as size (ranging from 3 mm to 30 mm) and shape (e.g., round, irregular, lobulated), providing a comprehensive representation of nodular characteristics. The dataset's thin-slice CT scans, with slice thickness ranging from 0.4 mm to 2.5 mm and pixel spacing between 0.310 mm and 1.091 mm, ensure high resolution for precise nodule detection.

For this study, the dataset was divided into three subsets: 70% for training (622 scans), 10% for validation (88 scans), and 20% for testing (178 scans), ensuring a balanced distribution for robust performance evaluation. Each CT scan, stored in DICOM format, comprises 100 to 500 axial slices with a resolution of 512 × 512 pixels. To prepare the data for input into the model, we extracted three contiguous slices for each axial slice and stacked them to form a three-channel image. These were then resized to a uniform shape of $800 \times 800 \times 3$ to ensure consistency across inputs and facilitate multi-scale feature learning during training.

### PN9

The PN9 dataset serves as a benchmark for assessing the generalization capabilities of the proposed lung nodule detection framework. This dataset comprises CT scans collected from two major hospitals, representing diverse clinical scenarios such as outpatient visits, hospitalizations, and physical examinations. The scans were acquired over the period from 2015 to 2019, ensuring a wide temporal distribution. To guarantee data quality, the initial CT images underwent a meticulous validation process, focusing on compliance with DICOM standards. Scans containing significant respiratory motion artifacts or other disruptive interferences were excluded. Additionally, all sensitive patient information embedded within the DICOM headers—such as patient identifiers, institutional details, and referring physician names—was securely anonymized through data masking techniques.

Pulmonary nodules in the PN9 dataset were annotated using a two-step process. In the first stage, each CT scan was independently reviewed by a physician, who generated an initial medical report detailing the type, size, and approximate location of detected nodules. These reports were then cross-validated by a second physician to ensure accuracy. In the second stage, nodules were annotated in a detailed, slice-by-slice manner, with physicians referencing the corresponding medical reports to ensure consistency. For each nodule, bounding boxes and classification labels were stored in structured XML files. The nodules were categorized into nine distinct groups based on their size and type, following established medical standards. To enhance the reliability of these annotations, a second physician reviewed the outputs, and any discrepancies were resolved collaboratively.

### Data Augmentation

In scenarios where data imbalance poses a challenge, augmentation becomes an essential strategy to enhance dataset diversity and improve model performance. Given the inherent imbalance in our dataset, we employed manual augmentation techniques to address this issue effectively. By rotating images in multiple directions and generating additional variations from different angles [145], we created a more diverse representation of the original data. These transformations mitigate the class imbalance problem and ensure a more robust learning process. Additionally, advanced augmentation methods, such as zooming in and out, applying various shear ranges, and flipping images, were utilized to further enrich the dataset. These techniques not only introduce variability but also enable the model to interpret data from multiple perspectives, ultimately improving its generalization capabilities.

### 6.4.2 Lung Parenchyma Segmentation

Accurate segmentation of the lung parenchyma from chest CT scans is a critical preprocessing step for isolating relevant anatomical structures and enhancing the precision of nodule detection. In our approach, this task is accomplished through a series of image processing techniques, starting with intensity normalization and binarization.

Raw CT images are first clipped to a predefined Hounsfield unit (HU) range of $[-1000, 400]$, which effectively captures the range of lung tissue densities. The pixel intensities are then normalized to the interval $[0, 1]$ using the following linear transformation:

$$I_{\text{norm}}(x, y) = \frac{\min(\max(I(x, y), -1000), 400) + 1000}{1400} \tag{6.1}$$

where $I(x, y)$ represents the original intensity at pixel $(x, y)$, and $I_{\text{norm}}$ is the normalized value. This scaling ensures that lung tissue contrasts are preserved while suppressing irrelevant high-density regions such as bone.

To isolate the internal thoracic region, a global threshold $T$ is computed as the mean of all normalized pixel intensities:

$$T = \frac{1}{N} \sum_{x=1}^{H} \sum_{y=1}^{W} I_{\text{norm}}(x, y) \tag{6.2}$$

Pixels with values below $T$ are set to 1 (foreground), and others to 0 (background), generating an initial binary lung mask $B(x, y)$:

$$B(x, y) = \begin{cases} 1, & \text{if } I_{\text{norm}}(x, y) < T \\ 0, & \text{otherwise} \end{cases} \tag{6.3}$$

To remove irrelevant components such as bones and air outside the lungs, a four-connected component labeling algorithm is applied. Only the two largest connected regions (corresponding to left and right lungs) are retained. All other regions are discarded as noise.

Morphological operations are used to enhance the binary mask. A morphological opening operation removes small artifacts using a circular structuring element, and a morphological closing operation is then used to fill small gaps near lung boundaries. Internal voids and gaps within the segmented lung areas are removed using a hole-filling operation [93].

The final binary mask provides a clean, contiguous representation of the lung parenchyma. This mask is subsequently used to crop the original CT images and suppress non-lung regions, ensuring that subsequent processing stages, such as nodule detection and classification, operate exclusively within the anatomically relevant domain.

### 6.4.3 Lung Nodule Detection

Detecting lung nodules, particularly small ones, is a complex challenge due to their subtle appearance and varying sizes. To address this, we developed an improved Faster R-CNN framework, enhanced with several innovative techniques designed to optimize sensitivity and accuracy for small nodule detection. Our method integrates a DA OMS-CNN to better capture multiscale features, an advanced RPN for generating accurate region proposals, and a DA-RoIPooling mechanism to enhance the classification process.

#### Dual-Attention Optimized Multi-Scale CNN (DA OMS-CNN)

Choosing the right feature extraction architecture plays a crucial role in determining the effectiveness of modern convolutional neural networks for detecting lung nodules. Several architectures, such as DenseNet, VGGNet, and ResNet, are widely used due to their ability to extract object features from images with remarkable precision. Despite their strengths, these models often struggle with accurately identifying small lung nodules. Among these architectures, VGG16 stands out for its compact $3 \times 3$ convolutional kernels and well-optimized layers, which enhance its capability to detect small nodules with improved precision. In ref. [123], the final three layers of the VGG16 network are merged to create feature maps, enabling the extraction of features from nodules of varying sizes, with a particular emphasis on smaller nodules.

This study proposes a dual-attention OMS-CNN designed for optimal RoI extraction, as illustrated in Figure 6.3. At this stage, fully convolutional dual-attention blocks are integrated to enhance the network's ability to focus on critical features across both channel and spatial dimensions. To achieve this, the dual-attention mechanism is incorporated at three key stages of the architecture: (1) the fourth layer of the VGG16 backbone, (2) the fifth layer of the VGG16 backbone, and (3) the concatenated feature map layer, where outputs from the last three convolutional layers are fused. The spatial attention component dynamically highlights spatial regions of interest by weighting feature maps based on their positional importance. Simultaneously, the channel attention component amplifies feature maps that contain the most discriminative information for nodule detection. This mechanism strengthens the model's focus on small nodules that may otherwise be overlooked, particularly in high-dimensional feature spaces.

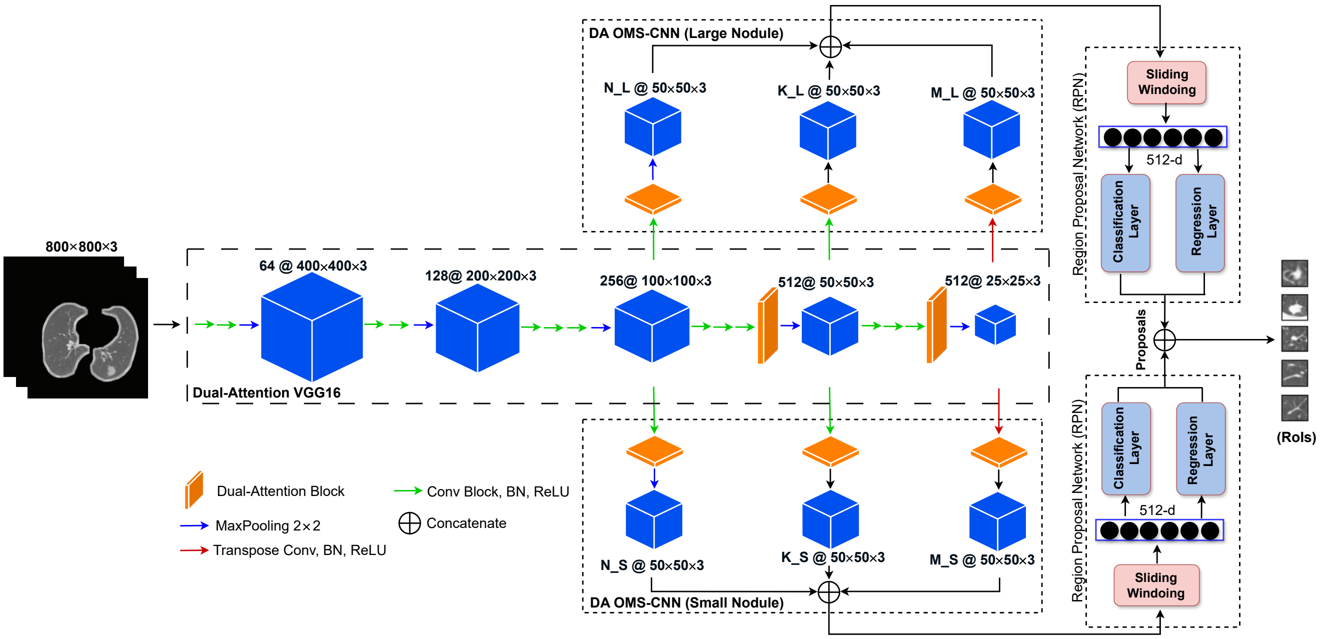


Figure 6.3 Overall framework of dual-attention OMS-CNN model (DA OMS-CNN).

To tailor the convolutional capacity of the network to different nodule scales, we define two sets of kernel configurations—$[N_S, K_S, M_S]$ for small nodules and $[N_L, K_L, M_L]$ for large nodules—corresponding to the number of output channels in the final three convolutional layers. The optimal values for these configurations are not selected heuristically but are instead determined through a principled optimization process. Specifically, we utilize the advanced parameter-setting-free harmony search (PSF-HS) algorithm [6], which formulates the search for kernel parameters as a global optimization problem. In this context, each candidate configuration is treated as a "harmony" whose fitness is evaluated based on the model's sensitivity in detecting annotated nodules within the training set. Through adaptive control of exploration and exploitation, PSF-HS iteratively refines candidate solutions and converges to high-performing configurations. This optimization strategy eliminates the need for manual tuning and improves the generalizability of the learned representations to nodules of varying shapes and sizes. For additional technical details and mathematical formulations, readers are referred to our earlier work [50]. Furthermore, to improve convergence stability and initialization quality, we adopt the BAS optimization algorithm [7], which replaces conventional random initialization for the convolutional layer filters by providing a more robust search mechanism for optimal weights and biases.

The dual-attention blocks are added only in the deeper layers of the VGG16 backbone and the concatenated feature map layer for specific reasons. In the lower layers of the network,

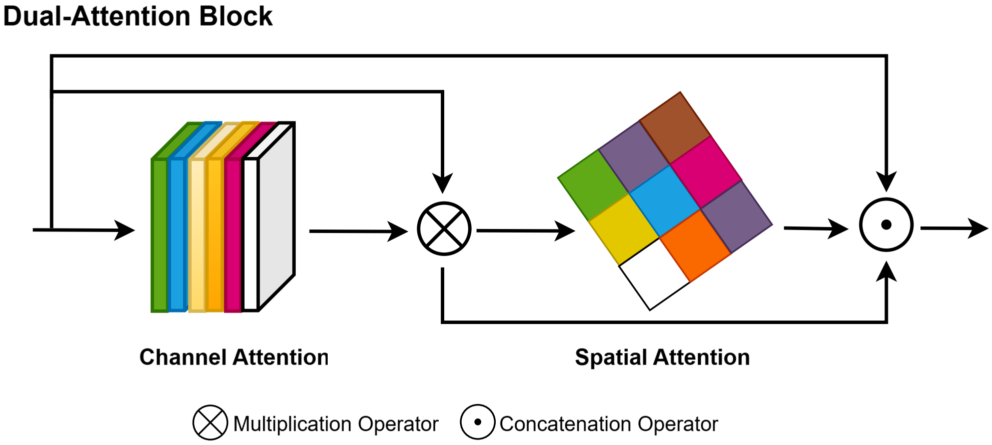


Figure 6.4 Proposed dual-attention block.

the extracted features are low-level representations, primarily capturing general patterns such as edges, textures, and basic shapes. Applying attention mechanisms at these stages could lead to a loss of critical general-purpose information needed for constructing high-level features. Instead, the dual attention is integrated into the deeper layers (fourth and fifth) of the backbone where the feature maps are more abstract, containing high-level semantic information critical for identifying lung nodules. These layers are better suited for attention mechanisms as they focus on more meaningful regions in the image.

Additionally, the concatenated feature map layer combines outputs from the last three convolutional layers, offering a multi-scale feature representation that captures nodules of varying sizes. By applying dual attention at this stage, the network selectively emphasizes the most relevant features across scales, improving its sensitivity to small nodules. Incorporating dual attention into this layer enhances the hierarchical understanding of multi-scale features while suppressing redundant or irrelevant information. This approach ensures that the network retains the ability to detect small and subtle nodules with high precision.

By avoiding the application of dual attention in the lower layers, the architecture balances the need for preserving low-level feature diversity with the refinement of high-level features. This strategic design significantly enhances the network's ability to detect nodules of varying sizes while maintaining robustness against background noise and irrelevant details, thus improving the overall sensitivity and performance of the system.

The dual-attention mechanism is designed with a sequential structure where channel attention is applied first, followed by spatial attention, as illustrated in Figure 6.4. This order ensures an optimal refinement of feature representations by addressing the "what" and "where" aspects of attention hierarchically [122]. By applying channel attention first, the model identifies and amplifies the most informative feature channels, effectively enhancing the global semantic understanding of the input. This step prioritizes features that are most relevant for identifying lung nodules, reducing noise at the channel level. The channel attention mechanism begins by extracting global information from the input feature map using both average pooling and max pooling operations. These operations yield two distinct descriptors, denoted as $\chi^C_{\text{AUG}}$ and $\chi^C_{\text{Max}}$. These descriptors are then processed through a scale network, which generates a channel attention map, represented as $M_C \in \mathbb{R}^{C/2G\times1\times1}$. The channel attention map is subsequently used to modulate the input feature map $\chi$, enabling element-wise summation with the corresponding sub-feature. Following this, average pooling and max pooling operations are applied to both branches of each sub-feature $\chi_K$. The resulting feature vectors are combined using element-wise summation, producing the final output $\chi_C \in \mathbb{R}^{C/2G\times1\times1}$. The process can be mathematically expressed as:

$$M_C(\chi) = \text{MLP}(\text{AVG}(\chi)) + \text{MLP}(\text{MAXPool}(\chi)) \tag{6.4}$$

$$\chi = \text{MLP}(\chi_{K1} + \chi_{K2}) + \text{MAXPool}(\chi_{K1} + \chi_{K2}) + M_C(\chi) \tag{6.5}$$

To complement the channel attention mechanism, a compact feature representation is created to enable precise and adaptive selection. This is achieved using a straightforward gating mechanism with a sigmoid activation function. The final output of the channel attention is computed as:

$$\chi_C = \delta(F_C(\chi)) \cdot \chi_{K1} = \delta(W_C\chi + b_C) \cdot \chi_{K1} \tag{6.6}$$

Subsequently, the output of the channel attention block is passed to the spatial attention block, which focuses on localizing the critical regions of interest within the feature maps. This sequential process ensures that spatial attention operates on already refined feature maps, making it more effective at highlighting the precise locations of nodules. To compute spatial attention, we apply group normalization (GN) to the $\chi_{K1}$ and $\chi_{K2}$ branches. This approach reduces computational complexity while ensuring that spatial information is effectively utilized, providing more accurate data to the feature extraction network. The calculation for spatial attention is expressed as:

$$\chi_S = \delta(W_S \cdot (\text{GN}(\chi_{K2}) + \text{GN}(\chi_{K1})) + b_S) \cdot \chi_{K2} \tag{6.7}$$

Here, $W_S$ and $b_S$ are parameters with a shape of $\mathbb{R}^{C/2G\times1\times1}$. The $\chi_{K1}$ and $\chi_{K2}$ branches are subsequently combined to align the number of channels with the input dimensions. This integration allows spatial attention to improve the representation of the feature map effectively.

By separating these two stages and processing channel importance before spatial localization, the network achieves a better balance between global feature importance and local feature refinement, improving detection accuracy and robustness.

### Region Proposal Network (RPN)

The RPN processes an input image and generates a set of rectangular proposals, each associated with an objectness score. The RPN is implemented as a fully convolutional network, designed to operate efficiently on feature maps produced by the last convolutional layer of the feature extraction network. A small network, which slides over the input feature map using a 3 × 3 spatial window, serves as the core of the RPN. Each sliding window extracts a feature vector (512 dimensions in the case of DA OMS-CNN), which is then passed through a box-classification layer to predict objectness scores and a box-regression layer to estimate the bounding box coordinates. To address the detection of both small and large lung nodules, two distinct RPNs are utilized within the framework. These RPNs are specifically designed to leverage different perspectives and extract complementary information, which enhances the overall proposal generation process. The two networks are integrated with the DA OMS-CNN backbone, with one RPN tailored for small nodules and the other for large nodules (Figure 6.3). These networks operate on feature maps of the same dimensions, ensuring seamless integration with the backbone architecture. To accommodate the varying sizes of lung nodules, seven anchor boxes with different scales are employed: $4 \times 4$, $6 \times 6$, $10 \times 10$, $16 \times 16$, $22 \times 22$, and $32 \times 32$, as in [50]. This multi-scale anchor design is particularly effective in capturing nodules of diverse sizes, enabling the framework to improve detection accuracy for both small and large nodules. With these definitions, the multi-task loss function for an image is expressed as:

$$L(p_i, t_i, p_{kj}, t_{kj}) = \sum_i L_1(p_i, t_i) + \sum_{k=1}^{2} \sum_j L_2(p_{kj}, t_{kj}) \tag{6.8}$$

Here, $L_1$ and $L_2$ are defined as:

$$L_1(p_i, t_i) = L_{\text{cls}}(p_i, p_i^*) + \lambda \times p_i^* \times L_{\text{reg}}(t_i, t_i^*) \tag{6.9}$$

$$L_2(p_{kj}, t_{kj}) = \frac{1}{N_{\text{cls}}} \times L_{\text{cls}}(p_j, p_j^*) + \frac{\lambda}{N_{\text{reg}}} \times p_j^* \times L_{\text{reg}}(t_{kj}, t_j^*) \tag{6.10}$$

The regression loss is defined as:

$$L_{\text{reg}}(t_i, t_i^*) = R(t_i - t_i^*) \tag{6.11}$$

$$R(x) = \begin{cases} 0.5x^2 & \text{if } |x| < 1, \\ |x| - 0.5 & \text{otherwise.} \end{cases} \tag{6.12}$$

In these equations:

- $i$ represents the index of the proposals generated by the region proposal networks.
- $j$ identifies an anchor selected.
- $k$ refers to one of the two region proposal networks.
- $p_i$ is the predicted probability of proposal $i$ being a nodule.
- $p_i^*$ is the ground truth label, where $p_i^* = 1$ if the proposal is positive, otherwise $p_i^* = 0$.
- $t_i$ and $t_i^*$ are the predicted and ground truth bounding box regression parameters, respectively.
- $L_{cls}$ is a binary cross-entropy loss.
- $L_{reg}$ represents the regression loss.
- $\lambda$ is a balancing factor between the classification and regression losses.
- $N_{cls}$ and $N_{reg}$ are normalization terms for classification and regression, respectively.
- $R$ is the smooth $L_1$ function.

### Classification Stage

After obtaining the RoIs predicted by the RPN and removing duplicates, a deep convolutional neural network (DCNN) is employed to classify each RoI, determining whether it corresponds to a nodule or not. The RPN regression layer generates candidate nodule positions, specifying

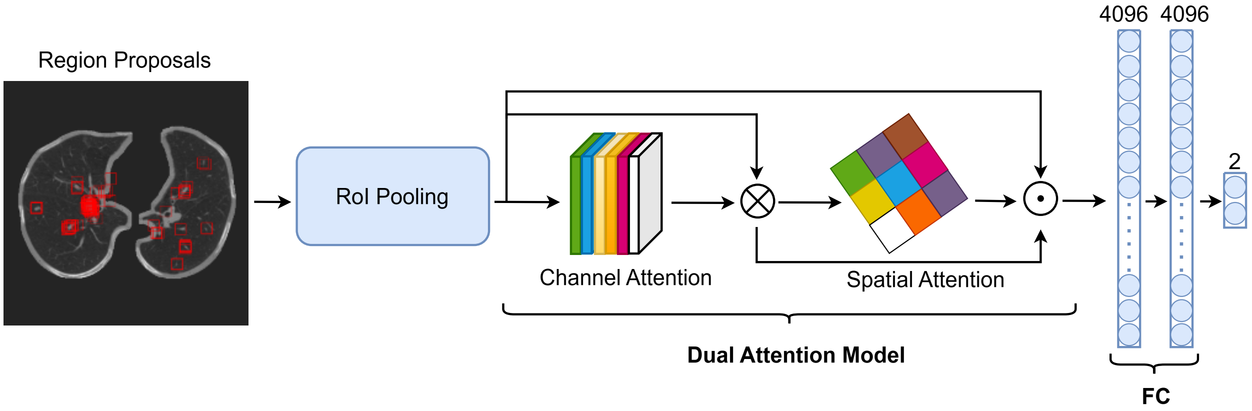


Figure 6.5 Overall framework of classification stage (DA RoIPooling).

the center coordinates as well as the width and height $(W, H)$ of each RoI. These values are used to extract patches from the feature map, which serve as input to the classification network. The RPN classification layer provides a probability score for each patch, ranging between 0 and 1. Patches with scores exceeding a threshold of 0.5 are considered nodule candidates and forwarded to the classification stage for further analysis [49].

In the proposed method, we introduce a dual-attention mechanism after the RPN stage and before the fully connected layers in the RoIPooling structure, as shown in Figure 6.5. An RoIPooling layer is employed to project each RoI onto a smaller feature map with a predetermined spatial dimension of $W \times H$ (specifically, $7 \times 7$ as outlined in this paper). The RoIPooling process involves dividing the RoI into a grid of sub-windows measuring $W \times H$ and performing max-pooling within each sub-window, resulting in values being mapped to their corresponding output grid cells. This pooling operation is carried out independently across each feature map channel, akin to standard max pooling procedures.

Following the RoI pooling operation, the dual-attention mechanism is integrated to enhance the extracted feature representations. The channel attention block selectively emphasizes informative channels while suppressing less relevant ones, ensuring that critical features for nodule classification are highlighted. The output of the channel attention block is then passed through the spatial attention block, which focuses on relevant spatial regions within each feature map. This combination allows the network to refine RoI feature maps by simultaneously considering channel-level and spatial-level dependencies. By applying dual attention at this stage, we aim to better capture subtle and discriminative features critical for accurate classification, especially in challenging cases.

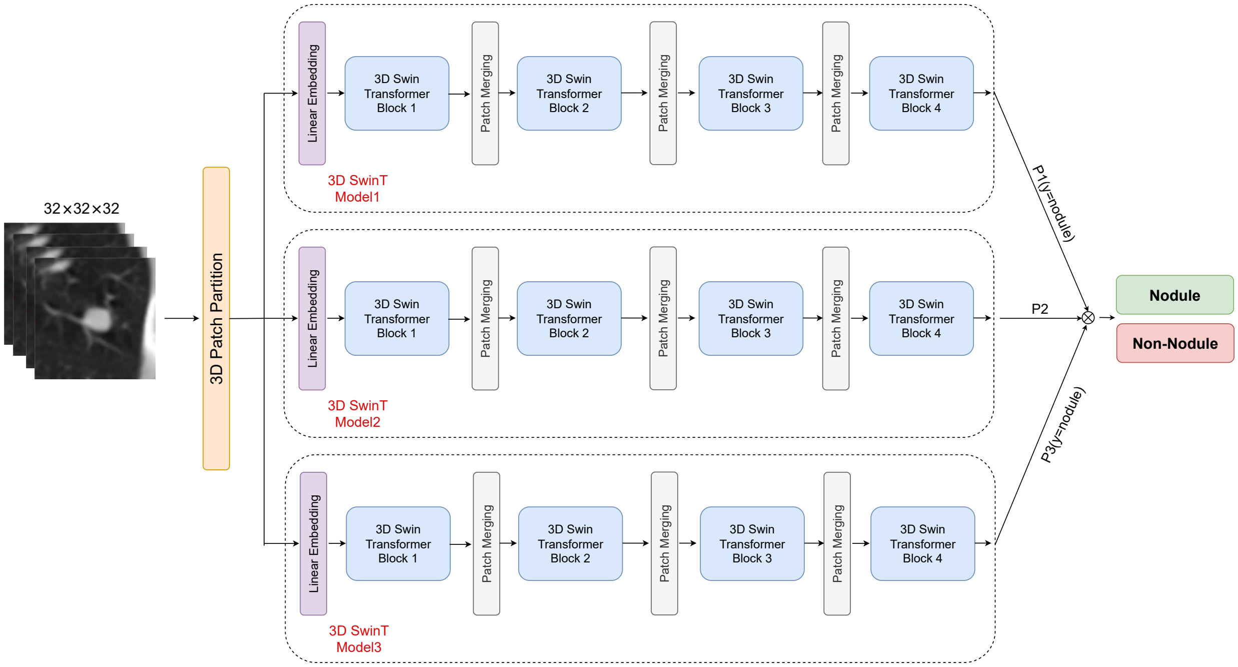


Figure 6.6 The structure of the proposed false-positive reduction model.

After dual-attention processing, a fully connected network comprising two 4096-dimensional fully connected layers is employed to transform the fixed-size feature map into a feature vector. Finally, a binary classifier predicts confidence scores for potential candidates. The training of the classification model utilizes CrossEntropyLoss as the loss function to optimize the network. This enhanced architecture aims to reduce false positives and improve the overall sensitivity and specificity of the nodule detection pipeline.

### 6.4.4 False Positive Reduction

In the false-positive reduction phase, we employ a sequence of 3D Swin Transformer models to enhance classification accuracy and reduce false positives. The pipeline processes 3D image patches, where each patch is passed through multiple trained 3D SwinT models, as shown in Figure 6.6. The outputs from these models are combined using a voting mechanism to determine the final classification as either “Nodule” or “Non-Nodule.” This approach leverages the hierarchical structure and self-attention mechanism of the 3D SwinT, enabling the extraction of both local and global features from volumetric data for robust decision making.

One of the main challenges in object detection is the overwhelming number of negative samples, which dominate the total loss. Many of these samples are relatively easy to classify, highlighting the importance of hard sample mining to improve performance. Following this concept, the training data is curated to emphasize more difficult samples, which persist through subsequent training iterations, enhancing the classification accuracy of each individual model.

The 3D SwinT, chosen for this phase, is designed to process volumetric data effectively by leveraging shifted window-based multi-head self-attention mechanisms. The hierarchical structure of the transformer enables the model to capture both global and local spatial relationships within the patches, offering improved performance compared to traditional convolutional models. Each model is initialized and trained independently using a specific subset of the data, focusing on misclassified samples from previous iterations [50]. Initially, the first subset is used to train Model1. Misclassified samples from both Model1 and the second subset are subsequently used to train Model2. Similarly, Model3 is trained using misclassified samples from the first two models and the third subset. This iterative training process ensures that challenging examples are emphasized, enabling the models to learn robust and discriminative features. The models are fine-tuned during successive iterations, refining their weight parameters to enhance their ability to classify difficult samples.

This iterative approach, combined with the 3D SwinT's ability to effectively capture spatial dependencies and represent complex patterns, significantly improves the classification accuracy of the false-positive reduction system. To further enhance the performance of the false-positive reduction phase, we employ various patch augmentation techniques and leverage the advanced hierarchical design of the 3D SwinT. These approaches are discussed in detail in the subsequent subsections.

### 3D Swin Transformer

The Swin Transformer (SwinT) is a hierarchical transformer that efficiently generates multiscale feature maps by integrating neighboring patches and employing a window partition mechanism. This approach ensures linear computational complexity relative to image size, which is particularly advantageous for dense prediction tasks and processing high-resolution images. To adapt this architecture for the 3D characteristics of CT images, we extend SwinT into a 3D structure (3D SwinT), enabling it to capture detailed spatial and volumetric information. The architecture of 3D SwinT, illustrated in Figure 6.7, differs from the standard SwinT in several key aspects:

- CT images are represented as $H \times W \times D$, where $D$ refers to the depth, and $H$ and $W$ denote the image's height and width, respectively.

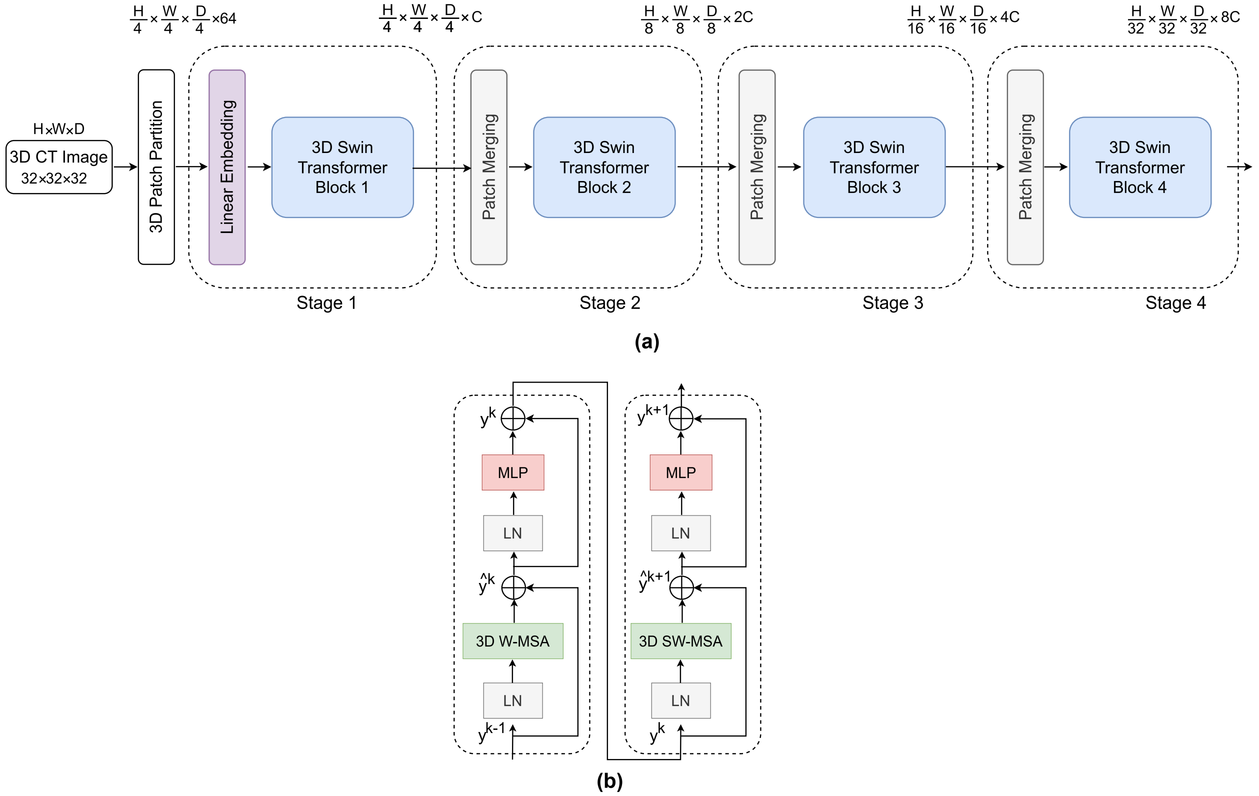


Figure 6.7 Overall architecture of 3DSwinT: (**a**) network architecture; (**b**) two consecutive 3DSwinT blocks.

- The patch partitioning process in SwinT divides the input into $(H/4) \times (W/4)$ patches, each sized $4 \times 4$. In contrast, 3D SwinT utilizes 3D cubes of size $4 \times 4 \times 4$, producing $(H/4) \times (W/4) \times (D/4)$ patches. These patches, with a feature dimension of 64, are projected into an arbitrary dimension $C$ via a linear embedding layer. Following this, the neighboring patches are combined during the patch merging stage, where the spatial and depth resolution decrease progressively (4, 8, 16, 32).

- The main distinction between the SwinT and 3D SwinT blocks lies in the multi-head self-attention mechanism. For 3D SwinT, the window-based multi-head self-attention (W-MSA) is extended into a 3D version (3D W-MSA), incorporating the volumetric information. This is achieved using 3D windows sized $P \times M \times M$, where $P$ represents the depth dimension, instead of the 2D $M \times M$ windows used in SwinT. Additionally, the window shifting mechanism in 3D SwinT introduces shifts of $(P/2, M/2, M/2)$ patches along the depth, height, and width dimensions, enhancing inter-window information interaction.

The 3D SwinT architecture comprises four stages. Each stage includes a patch merging module and multiple 3D SwinT blocks (except Stage 1). The patch merging module aggregates neighboring $2 \times 2 \times 2$ patches into larger patches, effectively reducing the spatial resolution to a quarter of its original size. A linear layer then projects the concatenated feature dimensions to half their size. The 3D SwinT blocks in each stage extract self-attention features while preserving the input resolution. Consequently, the feature map sizes at different stages are $(H/4) \times (W/4) \times (D/4) \times C$ (Stage 1), $(H/8) \times (W/8) \times (D/8) \times 2C$ (Stage 2), and so forth. Compared to standard SwinT blocks, 3D SwinT employs 3D window-based multi-head self-attention (3D W-MSA) to capture both spatial and volumetric information. Other architectural components, such as the multilayer perceptron (MLP), layer normalization (LN), and residual connections, remain unchanged from SwinT. Figure 6.7b depicts two adjacent 3D SwinT blocks within each stage, which can be represented by following the equation:

$$\begin{cases} \hat{y}^k = \text{3D W-MSA}(LN(y^{(k-1)})) + y^{(k-1)} \\ y^k = \text{MLP}(LN(\hat{y}^k)) + \hat{y}^k \\ \hat{y}^{k+1} = \text{3D SW-MSA}(LN(y^k)) + y^k \\ y^{k+1} = \text{MLP}(LN(\hat{y}^{k+1})) + \hat{y}^{k+1} \end{cases} \tag{6.13}$$

where 3D W-MSA and 3D SW-MSA represent the 3D window-based and shifted W-MSA mechanisms, respectively, and $\hat{y}^k$ and $y^k$ are the outputs of 3D (S)W-MSA and MLP in block $K$, respectively.

### 6.4.5 Evaluation Metrics

To comprehensively assess the performance of the proposed model, two key evaluation metrics—recall (sensitivity) and competition performance metric (CPM)—were employed. These metrics are widely utilized in the field of computer-aided detection (CAD) to evaluate the accuracy and robustness of nodule detection systems.

Recall, or sensitivity, quantifies the model's ability to correctly identify all existing nodules within the annotated dataset. Specifically, it measures the proportion of true nodules (ground truth) that are successfully detected by the model. This metric is especially critical in medical imaging applications, where missing even a single malignant nodule can delay diagnosis and significantly affect patient outcomes. In the context of lung cancer screening, a high recall is imperative to minimize false negatives and ensure that potential cancerous regions are not overlooked.

The recall metric is mathematically defined as:

$$\text{Recall} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}} \tag{6.14}$$

where $TP$ denotes the number of correctly detected nodules, and $FN$ represents the number of nodules present in the dataset but missed by the model. A higher recall value indicates stronger detection sensitivity and reduced clinical risk, which is particularly important in early-stage cancer detection when nodules are small and harder to detect.

The competition performance metric (CPM) measures the average sensitivity of the model across a range of false-positive rates (typically 1/8, 1/4, 1/2, 1, 2, 4, and 8 false positives per scan). The CPM provides a holistic assessment of the model's performance, balancing its sensitivity and specificity at varying levels of false positives.

$$\text{CPM} = \frac{\sum_{i=1}^{n} \text{Sensitivity at } FP_i}{n} \tag{6.15}$$

where $n$ is the number of predefined false-positive thresholds ($FP_i$).

## 6.5 Experimental Results and Discussion

In this section, we present a comprehensive evaluation of the proposed DA OMS-CNN framework. The results are structured into three key subsections: (1) implementation details and training setup, (2) an ablation study to assess the individual contributions of each proposed module, and (3) experimental comparisons with state-of-the-art lung nodule detection methods on both the LUNA16 and PN9 datasets. These analyses collectively demonstrate the effectiveness and generalization capabilities of our proposed approach.

### 6.5.1 Implementation

This study approaches lung nodule detection through three key stages: region proposal extraction, classification, and false-positive reduction. Initially, the DA OMS-CNN architecture is utilized for feature extraction, while the RPN is employed for training. The hyperparameters $[N_S, K_S, M_S]$ for small nodules and $[N_L, K_L, M_L]$ for large nodules are tuned prior to training, using two distinct RPNs for each category. After optimization, the values for small nodules are found to be $N_S = 8$, $K_S = 505$, and $M_S = 14$, while for large nodules, the values converge to $N_L = 3$, $K_L = 512$, and $M_L = 16$. A 10-fold cross-validation strategy is implemented to evaluate the system's performance, with stochastic gradient descent (SGD) optimization applied using a momentum factor of 0.9. Additionally, a weight decay of 0.00001 is incorporated, and the base learning rate is set at 0.0001. The training process is conducted in a computing environment equipped with two V100 GPUs and 192 GB of memory.

In the classification stage, addressing class imbalance is a crucial aspect of the classification network. This challenge is tackled by ensuring an equal distribution of positive and negative patches, utilizing the output from the trained RPN. In this method, region proposals with an intersection over union (IoU) greater than 0.7, along with the ground truth, are designated as positive patches, while an equal number of randomly selected proposals with an IoU below 0.1 are considered negative patches. This approach not only balances the classes but also increases the number of positive samples. During training, key hyperparameters were set, including an initial learning rate of 0.01 and a maximum of 150 epochs. To prevent overfitting, a weight decay of $1 \times 10^{-4}$ was applied. The learning rate was adjusted at specific checkpoints: it was reduced to 0.001 after 50% of the epochs, further decreased to 0.0001 after 75%, and finally set to 0.00001 after 90% of the epochs. These modifications contributed to a more effective training process. Additionally, stochastic gradient descent (SGD) with a momentum of 0.9 was employed to enhance model performance.

In the false-positive reduction (FPR) phase, as mentioned in the classification section, nodule and non-nodule patches are first generated, with a size of $32 \times 32 \times 32$, and then augmented using patch augmentation techniques. The architectural hyperparameters for all three 3D SwinT models are set as follows: $C = 96$, and the layer configurations are 2, 2, 6, and 2. Furthermore, the number of multi-head self-attention heads per stage is set to 3, 6, 12, and 24, respectively. The models are trained for 100 epochs using a fivefold cross-validation approach to assess performance. All three models utilize the AdamW optimizer, with the learning rate, momentum, batch size, and weight decay values set to 0.001, 0.6, 16, and $1 \times 10^{-5}$, respectively. A warm-up cosine annealing learning rate schedule is applied, with the warm-up phase lasting for 30 steps. Figure 6.8 compares pulmonary nodules as detected by the proposed network against their correspondent ground-truth locations.

The entire model was implemented using the PyTorch deep learning framework (version 1.13). All experiments were conducted on a server equipped with two NVIDIA V100 GPUs and 192 GB of RAM.

### 6.5.2 Ablation Study

To evaluate the effectiveness of the proposed model, we conducted ablation studies under identical conditions using the LUNA16 dataset with tenfold cross-validation, as depicted in Figures 6.9 and 6.10. The ablation experiments were performed on four different configurations: (1) OMS-CNN, (2) DA OMS-CNN, (3) DA OMS-CNN with DA-RoIPooling, and

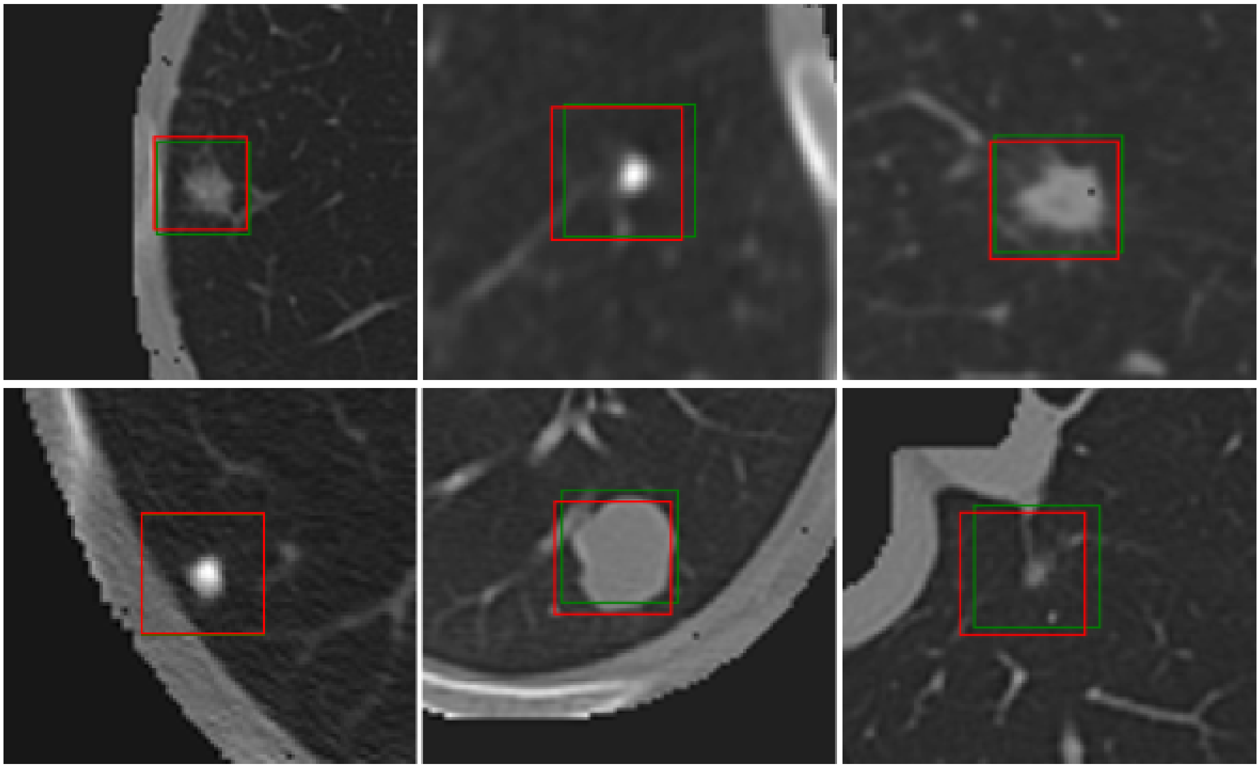


Figure 6.8 Pulmonary nodules detected by DA OMS-CNN (red) and their corresponding ground-truth boxes (green).

(4) DA OMS-CNN with DA-RoIPooling and the proposed FPR module. This analysis helps to assess the contribution of each component to the overall model performance.

In the DA OMS-CNN approach, a dual-attention mechanism is incorporated into the final layers of OMS-CNN to enhance feature representation, as illustrated in Figure 6.3. This enhancement enables the extraction of high-resolution, fine-grained features, which are particularly beneficial for the early detection of lung nodules. A comparative analysis between OMS-CNN and DA OMS-CNN, presented in Figure 6.9, demonstrates that integrating the dual-attention mechanism into the final layers of OMS-CNN increases the average recall for 1000 region proposals by 1.3%. Additionally, as shown in Figure 6.10, this modification improves the CPM score from 0.839 in OMS-CNN to 0.849, further highlighting its effectiveness.

In our second contribution, we refine the classification stage by replacing RoIPooling with DA-RoIPooling. This modification enhances the model's ability to capture both spatial and channel-wise dependencies, leading to more discriminative feature representations and, ultimately, improved accuracy in lung nodule detection. To assess the effectiveness of this enhancement, Figure 6.9 shows that the average recall of DA OMS-CNN with DA-RoIPooling is 4.2% higher than that of OMS-CNN and 3.1% higher than DA OMS-CNN. Additionally,

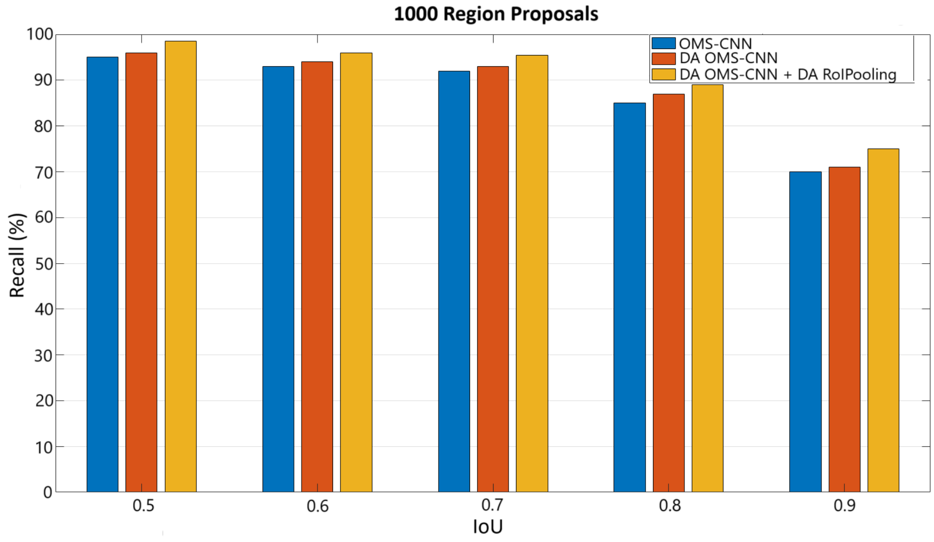


Figure 6.9 Recall vs. IoU overlap ration.

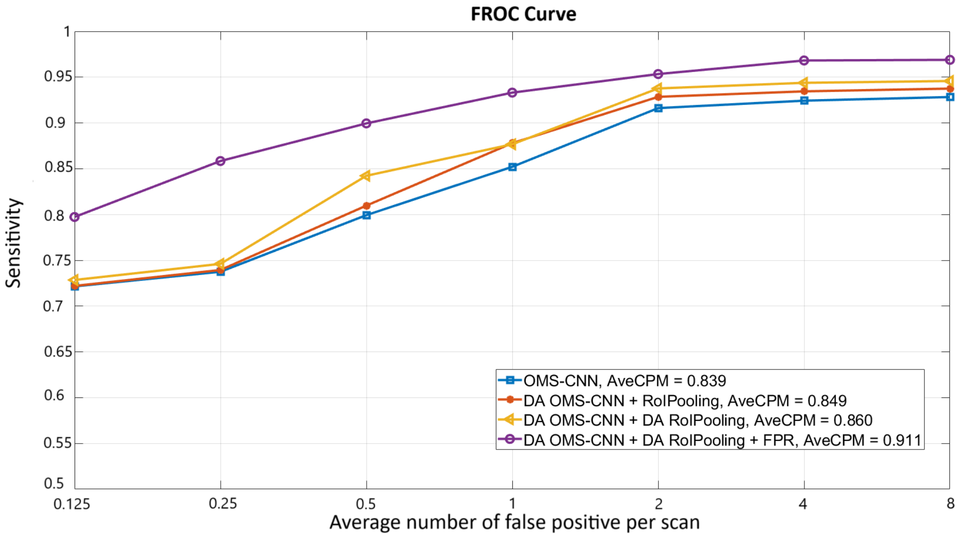


Figure 6.10 FROC curves of different proposed models on LUNA16.

as depicted in Figure 6.10, this method achieves a CPM score of 0.86, reflecting a 2.5% improvement over the OMS-CNN approach. In the final stage, we utilize an ensemble of three 3D SwinT models to reduce false-positive nodules. As illustrated in Figure 6.10, the proposed method improves the CPM score by 8.5%, 7.3%, and 5.9% compared to OMS-CNN, DA OMS-CNN, and DA OMS-CNN with DA-RoIPooling, respectively.

To further clarify the impact of each proposed module, Table 6.1 summarizes the key results of the ablation study in a tabular format. It presents the CPM scores and sensitivity values at 1.0 false positive per scan for different configurations of our model. This complementary table enables a more intuitive comparison of performance gains achieved through the integration of dual attention mechanisms, DA-RoIPooling, and the final false-positive reduction (FPR) module. The results demonstrate the incremental improvements in both sensitivity and overall CPM, highlighting the contribution of each component to the final detection performance.

Table 6.1 Ablation study: performance comparison of different model configurations on LUNA16.

| **Model Configuration** | **CPM Score** | **Sensitivity at 1.0 FP/scan** |
|---|---|---|
| OMS-CNN | 0.839 | 0.8521 |
| DA OMS-CNN | 0.849 | 0.8967 |
| DA OMS-CNN + DA-RoIPooling | 0.860 | 0.9331 |
| DA OMS-CNN + DA-RoIPooling + FPR | 0.911 | 0.9601 |

### 6.5.3 Experimental Comparison

This section presents the performance evaluation of the proposed lung nodule detection framework using different experimental settings. The results are reported in three tables: Table 6.2 shows the performance of the proposed candidate nodule detection network before false-positive reduction on the LUNA16 dataset, Table 6.3 presents the results after applying false-positive reduction using the LUNA16 dataset, and Table 6.4 demonstrates the generalization capability of the proposed method by evaluating it on the PN9 dataset.

Table 6.2 provides a comparative analysis of the proposed method against existing candidate nodule detection methods on the LUNA16 dataset. The comparison is based on the competition performance metric (CPM) score at different sensitivity thresholds. The results indicate that the proposed DA OMS-CNN method achieves the highest CPM score of 0.8601, outperforming the baseline OMS-CNN, which achieves a CPM of 0.8396. Compared to other state-of-the-art methods, such as Dou et al. [132] and Gu et al. [133], the proposed method consistently achieves higher detection performance across all sensitivity thresholds. This improvement can be attributed to the integration of domain adaptation and optimized feature extraction techniques, as well as the addition of a dual-attention mechanism in the last layers of OMS-CNN [50], which enhances the network's ability to identify candidate nodules more effectively. Moreover, our model shows notable improvements particularly at mid-to-low false-positive rates (0.5–2.0 FP/scan), which are critical operating points in clinical screen-

Table 6.2 Comparison of the proposed candidate nodule detection network with other methods on LUNA16.

| **CAD Method** | **Year** | **0.125** | **0.25** | **0.5** | **1.0** | **2.0** | **4.0** | **8.0** | **CPM** |
|---|---|---|---|---|---|---|---|---|---|
| Dou et al. [132] | (2017) | 0.6590 | 0.7540 | 0.8190 | 0.8650 | 0.9060 | 0.9330 | 0.9460 | 0.8390 |
| Gu et al. [133] | (2018) | 0.4801 | 0.6495 | 0.7920 | 0.8794 | 0.9163 | 0.9293 | 0.9301 | 0.7967 |
| Pezeshk et al. [112] | (2018) | 0.6370 | 0.7230 | 0.8040 | 0.8650 | 0.9070 | 0.9380 | 0.9520 | 0.8320 |
| Xie et al. [111] | (2019) | 0.4390 | 0.6880 | 0.7960 | 0.8520 | 0.8640 | 0.8640 | 0.8640 | 0.7750 |
| OMS-CNN [50] | (2024) | 0.7215 | 0.7357 | 0.7993 | 0.8521 | 0.9162 | 0.9243 | 0.9283 | 0.8396 |
| DA OMS-CNN | | 0.7285 | 0.7461 | 0.8223 | 0.8967 | 0.9377 | 0.9438 | 0.9458 | 0.8601 |

Table 6.3 Performance comparison of different methods for false-positive reduction on LUNA16.

| **CAD Method** | **Year** | **0.125** | **0.25** | **0.5** | **1.0** | **2.0** | **4.0** | **8.0** | **CPM** |
|---|---|---|---|---|---|---|---|---|---|
| Zeo et al. [53] | (2020) | 0.6300 | 0.7530 | 0.8190 | 0.8690 | 0.9030 | 0.9150 | 0.9200 | 0.8300 |
| CBAM [54] | (2021) | 0.4670 | 0.6020 | 0.7300 | 0.812 | 0.8770 | 0.9150 | 0.9310 | 0.7620 |
| I3DR-Net [55] | (2022) | 0.6356 | 0.7131 | 0.7984 | 0.8527 | 0.8760 | 0.8992 | 0.9147 | 0.8128 |
| MSM-CNN [49] | (2022) | 0.6770 | 0.7410 | 0.8160 | 0.8500 | 0.8900 | 0.9050 | 0.9250 | 0.8290 |
| MS-3DCNN [48] | (2023) | 0.7280 | 0.7990 | 0.860 | 0.8080 | 0.9260 | 0.9410 | 0.9560 | 0.8730 |
| AttentNet [135] | (2024) | 0.7520 | 0.8170 | 0.8570 | 0.8850 | 0.9200 | 0.9330 | 0.9330 | 0.8710 |
| MK-3DCNN [56] | (2024) | 0.7099 | 0.7723 | 0.8356 | 0.8836 | 0.9174 | 0.9384 | 0.9562 | 0.8591 |
| TED [51] | (2024) | 0.7619 | 0.8222 | 0.8736 | 0.9069 | 0.9302 | 0.9443 | 0.9530 | 0.8846 |
| OMS-CNN [50] | (2024) | 0.7932 | 0.8421 | 0.8712 | 0.9048 | 0.9387 | 0.9473 | 0.9481 | 0.8922 |
| DA OMS-CNN | | 0.7973 | 0.8584 | 0.8995 | 0.9331 | 0.9534 | 0.9682 | 0.9689 | 0.9112 |

ing scenarios. For example, at 1.0 FP/scan, the DA OMS-CNN achieves a sensitivity of 0.8967, significantly higher than the 0.8650 reported by Dou et al. [132] and the 0.8521 of the baseline OMS-CNN. This enhanced detection capability is mainly due to the effective integration of the dual-attention mechanism and domain adaptation strategies, which allow the model to better focus on relevant features and reduce noise from surrounding anatomical structures. These improvements contribute to the overall 2.1% increase in CPM score compared to OMS-CNN, demonstrating the practical benefits of the proposed enhancements.

Table 6.3 evaluates the impact of false-positive reduction using different methods on the LUNA16 dataset. The results show that the proposed DA OMS-CNN achieves the highest CPM score of 0.9112, surpassing other state-of-the-art approaches, including TED [51] (CPM = 0.8846) and MK-3DCNN [56] (CPM = 0.8591). The improvement in performance highlights the effectiveness of the false-positive reduction strategy employed in the proposed method, which incorporates an ensemble of three 3D SwinT models. This ensemble learning approach refines the detection process, effectively reducing the number of false positives while maintaining high sensitivity for true-positive nodules. The sensitivities at 0.125, 0.25, 2, and 4 FPs/scan are 0.797, 0.858, 0.953, and 0.968, respectively, surpassing those of the

Table 6.4 The sensitivity and CPM score compared with other methods on PN9.

| **CAD Method** | **Year** | **0.125** | **0.25** | **0.5** | **1.0** | **2.0** | **4.0** | **8.0** | **CPM** |
|---|---|---|---|---|---|---|---|---|---|
| SSD512 [121] | (2016) | 0.0462 | 0.0848 | 0.1476 | 0.2506 | 0.4032 | 0.5727 | 0.7080 | 0.3161 |
| RetinaNet [61] | (2017) | 0.0260 | 0.0556 | 0.1095 | 0.1925 | 0.2929 | 0.4049 | 0.5105 | 0.2274 |
| NoduleNet [136] | (2019) | 0.2117 | 0.3023 | 0.4038 | 0.5102 | 0.6129 | 0.7070 | 0.7693 | 0.5025 |
| SA-Net [114] | (2021) | 0.2672 | 0.3603 | 0.4746 | 0.5699 | 0.6635 | 0.7352 | 0.7832 | 0.5506 |
| I3DR-Net [55] | (2022) | 0.1564 | 0.2313 | 0.3700 | 0.5154 | 0.6454 | 0.7291 | 0.7753 | 0.4890 |
| OMS-CNN [50] | (2024) | 0.2865 | 0.3841 | 0.4775 | 0.5907 | 0.6974 | 0.7853 | 0.8432 | 0.5807 |
| DA OMS-CNN | | 0.3015 | 0.3952 | 0.4978 | 0.6221 | 0.7205 | 0.8241 | 0.8629 | 0.6034 |

best-performing method presented. Furthermore, compared to the baseline OMS-CNN [50], which achieves a CPM of 0.8922, DA OMS-CNN provides a 2.1% increase in detection accuracy, further demonstrating its robustness in distinguishing true nodules from non-nodular structures. The proposed method for detecting potential nodules demonstrates a sensitivity of 96.93%. On average, there are 9.38 candidates per scan. These results underline the significant advantage of the ensemble-based false-positive reduction strategy in balancing sensitivity and specificity. Notably, the DA OMS-CNN maintains superior sensitivity even at very low false-positive rates, which is critical for clinical usability to minimize unnecessary follow-ups. The integration of the three 3D SwinT models contributes to capturing diverse contextual features, thereby effectively filtering out false positives without compromising the true-positive detection rate. This comprehensive improvement emphasizes the robustness and practicality of the proposed approach in real-world screening settings.

To evaluate the generalization capability of the proposed method, we conducted an experiment using the PN9 dataset, and the results are presented in Table 6.4. The performance of the proposed approach is compared with several existing methods, including SSD512 [121], RetinaNet [61], and NoduleNet [136]. The results indicate that the DA OMS-CNN model achieves a CPM score of 0.6034, outperforming the baseline OMS-CNN (0.5807) and other existing methods such as SA-Net [114] (CPM = 0.5506) and I3DR-Net [55] (CPM = 0.4890). The consistent improvement across different sensitivity thresholds suggests that the proposed method generalizes well to unseen datasets, making it a promising approach for real-world clinical applications. These findings demonstrate the strong generalization ability of the DA OMS-CNN framework beyond the primary training domain, which is essential for clinical translation, where data variability is common. The steady increase in CPM and sensitivity across various false-positive rates indicates robustness to domain shifts and dataset heterogeneity. This suggests that the combined use of dual attention mechanisms and the 3D Swin Transformer architecture effectively captures invariant and discriminative features, enabling reliable detection performance even on previously unseen datasets such as PN9.

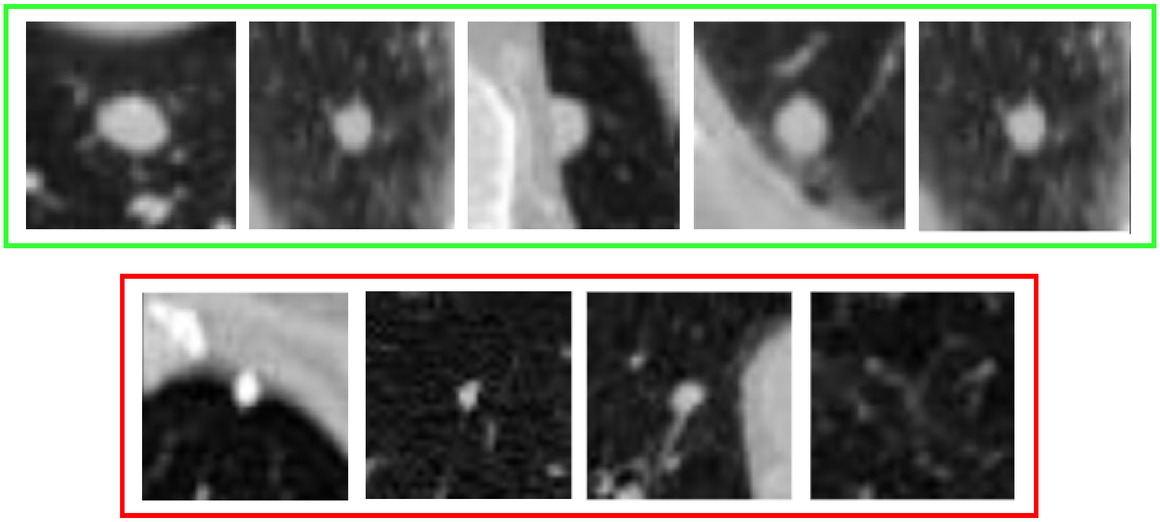


Figure 6.11 Examples of qualitative detection results by the proposed DA OMS-CNN. Nodules outlined in green represent correctly detected cases, while those in red indicate missed nodules.

To further understand the behavior of the proposed DA OMS-CNN model, we conducted a qualitative analysis of both successful and failed detection cases. In successful cases, the model accurately identified nodules with clear boundaries, moderate size, and strong contrast from surrounding tissues. These nodules typically appeared in central lung regions with less anatomical noise. However, the model showed reduced sensitivity in detecting extremely small nodules (less than 3mm), nodules located near complex anatomical structures such as blood vessels or the pleural wall, and in scans with low image quality or artifacts. In such cases, misclassification often resulted from insufficient contrast or structural ambiguity.

Figure 6.11 illustrates representative examples of both detected (green box) and missed (red box) nodules. As shown, successfully detected nodules tend to be well isolated and exhibit clearer margins, while missed cases often involve small or low-contrast nodules embedded within complex anatomical surroundings. This qualitative evidence supports our earlier quantitative findings and further highlights the strengths and current limitations of the proposed framework.

The experimental results highlight the superior performance of the proposed DA OMS-CNN framework in lung nodule detection. The candidate nodule detection stage achieves a higher CPM score compared to existing methods, demonstrating the effectiveness of the proposed feature extraction and detection strategies. The integration of an ensemble-based false-positive reduction approach significantly enhances detection accuracy, reducing false positives while maintaining high sensitivity. Finally, the generalization experiment on the PN9 dataset further validates the robustness of the proposed method, confirming its capability to perform well on different datasets.

## 6.6 Conclusions

In this study, we presented an improved Faster R-CNN model for early-stage lung cancer detection, which integrates a novel dual-attention optimized multi-scale CNN (DA OMS-CNN) architecture and a dual-attention RoIPooling (DA-RoIPooling) technique to enhance the model's sensitivity. The DA OMS-CNN effectively captures representative features of nodules at varying sizes, while the DA-RoIPooling method further refines classification accuracy, ensuring a higher detection rate. Additionally, the incorporation of an ensemble of three 3D Swin Transformer (3D SwinT) models for false-positive reduction significantly improves the precision of the detection system. Our model demonstrated superior performance on the LUNA16 and PN9 datasets. The experimental results validate the effectiveness of the integrated DA OMS-CNN and DA-RoIPooling techniques in improving the sensitivity of lung cancer detection, while also reducing the occurrence of false-positive nodules. This advancement marks a significant step forward in the development of more accurate and reliable lung nodule detection systems, with potential applications in clinical practice.

As a future direction, we aim to enhance the clinical applicability of our system by improving its transparency and reliability. To this end, we are investigating explainable AI strategies that allow the model's decisions to be more interpretable for clinicians, helping bridge the gap between automated predictions and clinical trust. This will support the development of more user-centric and deployable CAD systems for lung cancer diagnosis.

# CHAPTER 7 ARTICLE 4: IMPROVED 3D SWINT-CNN: A HYBRID AND INTERPRETABLE DEEP LEARNING MODEL FOR LUNG NODULE DIAGNOSIS

**Preface:** This chapter presents a hybrid and interpretable deep learning model (SwinT-CNN) for lung nodule diagnosis. The proposed architecture combines a 3D CNN and a 3D Swin Transformer within a dual-path framework, enhanced by anatomical attention gates (AAG) informed by a pretrained U-Net segmentation model. The method is designed to improve both diagnostic accuracy and model transparency, making it suitable for clinical applications in early-stage lung cancer screening. The full manuscript was submitted for peer review to the *IEEE Journal of Biomedical and Health Informatics* on August 13, 2025.
**Contributions:** This research was conducted during my doctoral studies at Polytechnique Montréal. I conceptualized and implemented the SwinT-CNN model, integrated anatomical priors via AAG, and designed the interpretability evaluation framework using Grad-CAM and entropy metrics. I performed all experiments on the LIDC-IDRI and PN9 datasets, analyzed the results, and led the manuscript writing. My co-authors contributed to theoretical modeling, experimental design, and critical feedback during the revision process.
**Manuscript Title:** Yadollah Zamanidoost, Tarek Ould-Bachir, and Sylvain Martel, *"Improved 3D SwinT-CNN: A Hybrid and Interpretable Deep Learning Model for Lung Nodule Diagnosis"*, submitted to *IEEE Journal of Biomedical and Health Informatics*, August 13, 2025.
**Submission Status:** Under peer review (submission date: August 13, 2025).

## 7.1 Abstract

Lung cancer is a leading cause of cancer deaths worldwide, and early detection improves outcomes. Detecting pulmonary nodules in CT scans is a time-consuming process that relies on expert interpretation. The clinical use of deep learning-based computer-aided diagnosis (CAD) systems is limited due to their poor interpretability. This research introduces SwinT-CNN, a hybrid end-to-end deep learning system that combines a 3D CNN and a 3D Swin Transformer. It uses a dual-path architecture to capture both fine local features and broad global relationships. The CNN branch receives anatomical priors from a pretrained 3D U-Net segmentation model through an anatomical attention gate (AAG) to enhance interpretability.

The network design enables it to concentrate on clinically meaningful areas, which prior voxel-level segmentation identifies. The model receives end-to-end training from LIDC-IDRI data and its performance evaluation occurs on the external PN9 dataset to test its ability to handle different imaging scenarios. The proposed method achieves 97.3% accuracy in distinguishing between benign and malignant nodules, surpassing existing baseline models. Interpretability assessment using Grad-CAM heatmaps, sensitivity, and entropy metrics shows the model reliably focuses on nodule areas. The enhanced 3D SwinT-CNN combines high accuracy and explainability, making it a promising tool for the early detection of lung cancer.

## 7.2 Introduction

cancer stands as one of the deadliest cancers globally because pulmonary nodules frequently appear as the first signs of the disease. Early detection of these nodules through accurate methods leads to better treatment results and longer patient survival times [103, 146]. CT imaging stands out as essential for both lung nodule screening and diagnosis among all available imaging techniques [147,148]. The increasing number of CT scans in clinical practice presents a growing challenge for radiologists, as their diagnostic abilities become vulnerable to workload and fatigue, as well as inter-observer differences [149]. The introduction of computer-aided diagnosis (CAD) systems utilizing deep learning technology enables clinicians to achieve more accurate and reliable lung nodule identification and classification [81].

The promising results of deep learning (DL) models in medical image analysis face an essential challenge because they lack transparency [150]. The healthcare sector requires artificial intelligence systems to deliver precise predictions while providing medical professionals with explanations they can trust and understand [151]. The "black box" nature of conventional DL models prevents clinicians from understanding the diagnostic process and from confirming which medical features the model uses [73]. The lack of transparency in AI systems reduces trust in their decision-making assistance and prevents their adoption in evidence-based clinical practices [152]. Research now focuses on developing explainable AI (XAI) methods which improve interpretability through visual explanations, feature attribution, and model transparency [74,153]. The goal of these efforts is to provide both the rationale behind a prediction and the location of the model's attention, so that AI-based diagnostic tools become more reliable and clinically useful [154].

The classification of pulmonary nodules in CAD systems relies on CNNs because these networks learn hierarchical features directly from raw CT images in recent years [33, 123]. The development of CNN-based architectures for nodule malignancy detection includes multi-view analysis [155] and optimized multi-scale CNN (OMS-CNN) [50] and attention mechanisms [52, 57] to boost diagnostic accuracy. The ability of traditional CNNs to detect fine-grained local structures does not translate into effective modeling of long-range dependencies and contextual information needed for precise malignancy assessment.

Medical imaging applications now use transformer-based models to solve the limitations of traditional CNNs. The Swin Transformer (SwinT) stands out among vision transformers because it uses a hierarchical structure to extract both local and global features through shifted window attention [156]. SwinT and other vision transformers demonstrate better performance than standard CNNs in medical imaging applications such as tumor classification and segmentation [144, 157] according to preliminary research. The combination of convolutional and transformer-based models in AI-driven lung cancer diagnosis is showing increasing interest in improving both reliability and interpretability.

The research introduces an enhanced 3D SwinT-CNN architecture, which combines CNNs and Swin Transformers through a parallel dual-path architecture for the diagnosis of lung nodules. The architecture combines simultaneous local structural feature extraction with global contextual dependency analysis to overcome the limitations of single-stream models when dealing with the heterogeneous characteristics of pulmonary nodules. The model incorporates anatomical priors from a 3D U-Net segmentation model, which was trained separately to enhance interpretability and guide attention to critical clinical areas. The AAG method enables the classification model to focus on nodule-specific regions with greater precision by injecting encoder features from U-Net into the CNN pathway. The proposed method achieves better accuracy and transparency in distinguishing between benign and malignant nodules by evaluating publicly available benchmark datasets.

The main contributions of this work are summarized as follows:

- A dual-path hybrid architecture combines CNNs and Swin Transformers to extract local and global features from CT scans. CNNs capture fine structural details, while Swin Transformers analyze distant spatial patterns. This approach enhances understanding of pulmonary nodules for more accurate classification.

- The interpretability of the CNN pathway improves with anatomical information from a U-Net segmentation model. The AAG integrates U-Net encoder features, rich in spatial and semantic context, into the CNN branch to enhance nodule features and suppress background noise, enabling more focused and meaningful decisions.

- The model's interpretability is evaluated both qualitatively and quantitatively using attention-based visualizations. Heatmaps show consistent focus on pulmonary nodules during predictions. This transparency allows radiologists to verify AI reasoning and fosters trust in the system's diagnostic decisions.

The structure of this paper is organized as follows. Section 7.3 reviews prior studies relevant to lung nodule classification and model interpretability. Section 7.4 details the architectural design of the proposed interpretable SwinT-CNN framework. Section 7.5 presents the experimental results along with a comprehensive analysis. Finally, the conclusions and key contributions are summarized in the last section.

## 7.3 Related Works

The integration of XAI techniques into deep learning frameworks has garnered growing attention in recent years, particularly in high-stakes fields such as medical imaging. Deep neural networks achieve remarkable prediction results, but their black-box operation creates transparency and trust issues in clinical settings. The lack of transparency in model decision-making poses a significant concern in lung cancer screening, as it directly impacts patient outcomes. The medical field now requires diagnostic models that achieve high accuracy while providing interpretable results, which clinicians can use to make decisions.

Wang et al. [5] developed ExPN-Net as a multi-task, explainable deep learning model that predicts malignancy while simultaneously detecting and localizing specific nodule characteristics. The model achieves accurate classification through its soft activation module and segmentation-derived attention maps, which enhance interpretability. The authors demonstrate the effectiveness of attribute-level guidance through their results on both public and private datasets, which improve clinical understanding and diagnostic accuracy.

The authors Fu et al. [78] developed an attention-based multi-task CNN to evaluate multiple visual attributes of lung nodules, thereby solving the problem of inter-observer variability. The model uses slice-level, cross-attribute and attribute-specific attention to eliminate irrelevant features and highlight clinically meaningful patterns. The framework validated on LIDC-IDRI improves attribute scoring and enables malignancy classification with interpretable outputs.

Gu et al. [158] created VINet as an interpretable CAD system which combines classification with visual attention to show important diagnostic areas. The method utilizes feature destruction to reduce noise while enhancing attention maps, resulting in both high accuracy and visual transparency. The LUNA16 dataset tests show that VINet has strong potential for trustworthy clinical diagnosis.

Recent advances in lung nodule diagnosis with deep learning show promise, but most rely solely on convolutional architectures. The limited receptive field of CNNs hinders their ability to capture long-range dependencies and holistic anatomical context, reducing accuracy and interpretability for subtle malignancy indicators. Our method combines convolutional and transformer-based models in a dual-path architecture, improving diagnostic precision by integrating local and global features along with segmentation-based anatomical priors, enhancing both transparency and clinical relevance.

## 7.4 Design Framework

The proposed method framework appears in Fig. 7.1. The pipeline consists of four main stages: image preprocessing, feature extraction and classification, explanation map generation, and explanation evaluation. The initial step of the process involves performing multiple preprocessing operations on raw input images to improve their quality before analysis. The extracted discriminative features enable the classification of pulmonary nodules after preprocessing. The model provides transparency through explainability maps, which show the specific areas that drive the prediction results. The final stage involves a systematic evaluation of the generated explanations to assess their quality and reliability.

### 7.4.1 Datasets

The proposed SwinT-CNN framework for lung nodule classification and segmentation was evaluated using two publicly available chest CT datasets: LIDC-IDRI [95] and PN9 [114]. The selected datasets contained high-quality annotations and clinical relevance, as well as complementary characteristics, which enabled a comprehensive evaluation of both diagnostic accuracy and model generalizability.

#### LIDC-IDRI

The LIDC-IDRI dataset contains thoracic CT scans from 1,018 patients, which up to four expert radiologists annotated. The average score of multiple expert radiologists was used to evaluate nodule malignancy on a scale from 1 (benign) to 5 (malignant). The nodule classification system used benign for scores below 3 and malignant for scores above 3, while excluding nodules with a score of exactly 3 to prevent classification uncertainty [159]. The study included only nodules that exceeded 3 mm in size. The *pylidc* toolkit [160] processed all volumes before they were resampled to $64 \times 64 \times 32$ voxels. The dataset served both classification and segmentation purposes through a 60% training set, a 20% validation set,

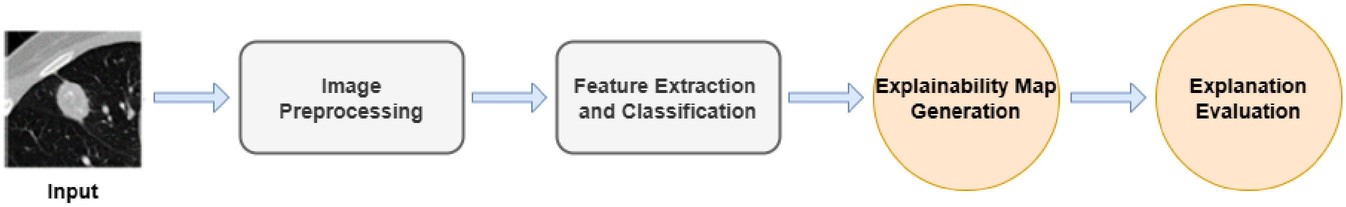

Input
Image Preprocessing
Feature Extraction and Classification
Explainability Map Generation
Explanation Evaluation

Figure 7.1 Workflow of proposed approach for interpretable lung nodule diagnosis

and 20% testing set distribution.

### PN9

The PN9 dataset contains CT scan images of pulmonary nodules which were obtained from nine medical centers to create a diverse population for external validation purposes. The clinical diagnosis determines whether each nodule receives a benign or malignant binary label. The PN9 dataset lacks voxel-level segmentation masks and multi-rater malignancy scores which distinguishes it from LIDC-IDRI. The dataset provides excellent generalizability assessment because it contains diverse acquisition settings and well-defined diagnostic labels.

### Data Preprocessing

The LIDC-IDRI and PN9 datasets underwent unified preprocessing operations. The CT scans received an isotropic voxel spacing transformation to $0.7 \times 0.7 \times 1.25$ mm. The extraction of each nodule involved obtaining a fixed-size volume of $64 \times 64 \times 32$ voxels, which centred on the annotated location and added zero-padding when needed. The voxel intensity values received clipping to the range $[-1000, 400]$ followed by z-score standardization based on training set statistical data. The training process involved real-time data augmentation through 3D rotations (90°, 180°, 270°), horizontal and vertical flipping, and z-axis slice reversal to enhance model robustness. The segmentation task used the same LIDC-IDRI dataset, extracting nodule-centered volumes and binary masks from radiologist annotations. Identical transformations were applied to both inputs and masks for data augmentation, preserving spatial alignment.

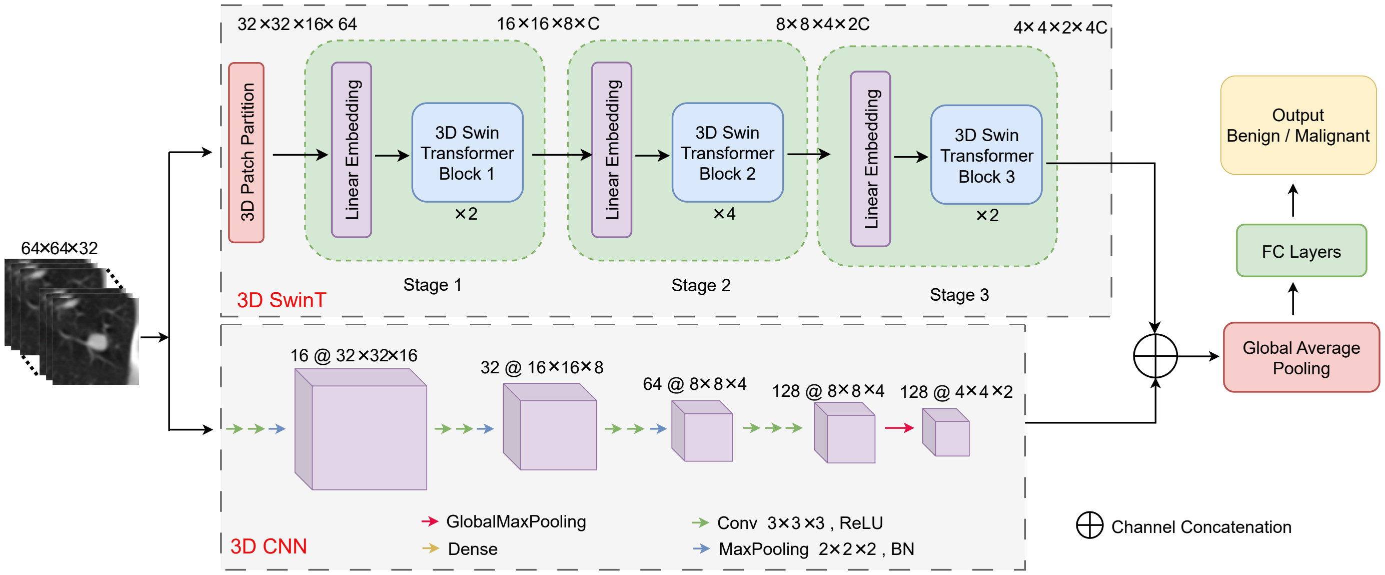


Figure 7.2 The backbone structure of the 3D SwinT-CNN model.

## 7.4.2 Dual-Path Feature Extraction and Classification

### Backbone Structure

The proposed backbone network consists of two parallel pathways: a convolutional stream based on a 3D VGG-style architecture [90], and a transformer stream constructed using a hierarchical 3D Swin Transformer [156]. The input CT scan has a spatial resolution of $64 \times 64 \times 32$. The dual-path architecture layout is shown in Fig. 7.2 which demonstrates the sequential operations in both branches before their integration for classification.

The lower branch consists of multiple 3D convolutional layers, which increase channel depth while decreasing spatial resolution. The convolutional blocks in this network consist of $3{\times}3{\times}3$ kernels followed by batch normalization, ReLU activation, and 3D max pooling. The network design enables the detection of detailed local patterns, including nodule boundaries, textures, and edge variations, while maintaining volumetric context [90]. The simple hierarchical structure of this branch, combined with its lack of skip connections, makes it suitable for interpretability because it enables better localization of image regions that influence classification results.

The upper branch utilizes a 3D Swin Transformer backbone, which transforms the input volume into non-overlapping 3D patches that are embedded as tokens in a sequence. The Swin Transformer blocks (as illustrated in Fig. 7.3) operate in stages to process the tokens through two consecutive self-attention stages, which start with W-MSA followed by SW-MSA [156].

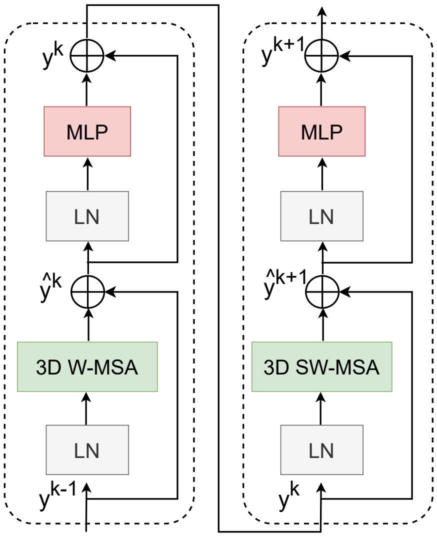


Figure 7.3 Swin transformer block.

Each attention block includes residual connections and layer normalization (LN), followed by a multi-layer perceptron (MLP) for transforming the representation. SW-MSA improves global context modeling by shifting window partitions to enable cross-window information exchange. A Swin Transformer block operates using the following sequence:

$$\begin{cases} \hat{y}^k = \text{3D W-MSA}(LN(y^{(k-1)})) + y^{(k-1)} \\ y^k = \text{MLP}(LN(\hat{y}^k)) + \hat{y}^k \\ \hat{y}^{k+1} = \text{3D SW-MSA}(LN(y^k)) + y^k \\ y^{k+1} = \text{MLP}(LN(\hat{y}^{k+1})) + \hat{y}^{k+1} \end{cases} \tag{7.1}$$

Where $\hat{y}^k$ and $y^k$ are the outputs of 3D (S)W-MSA and MLP in block $K$, respectively.

The two representations perform best when their feature maps are spatially aligned and fused after the final encoder stages. This fusion combines local detail with global context into a unified representation. A global average pooling layer condenses spatial data into a compact feature vector, which is then fed to fully connected layers for classification. Pooling reduces feature dimensions while preserving key global activations, thereby improving generalization. The backbone preserves each branch's strengths while supporting accurate, interpretable lung nodule diagnosis.

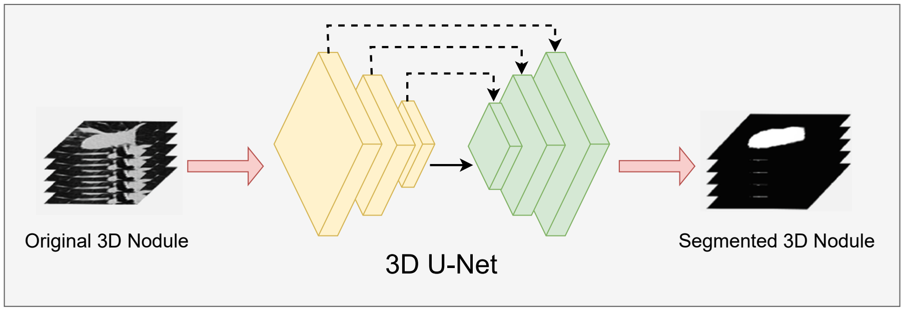


Figure 7.4 3D U-Net segmentation pipeline.

### 3D U-Net-Based Segmentation Branch

The proposed framework employs a 3D U-Net architecture [35,161,162] to segment pulmonary nodules from chest CT scans. This segmentation branch enhances the classification model's interpretability by explicitly detecting nodule regions to guide CNN feature extraction. As shown in Fig. 7.4, the workflow transforms the raw 3D CT volume into a voxel-wise probability map, which is then converted into a binary segmentation mask highlighting the target nodules. The 3D U-Net model uses a modified encoder–decoder architecture for volumetric medical imaging. The encoder extracts abstract features from broader spatial contexts using convolutional layers and downsampling [35]. Each encoder block includes 3D convolutions with $3 \times 3 \times 3$ kernels, ReLU activation, and 3D max pooling.

The network uses symmetric upsampling via transposed convolutions in the decoder to gradually restore spatial resolution. Decoder layers receive encoder feature maps through skip connections, preserving spatial details lost in downsampling [35]. This fusion enables accurate nodule localization while maintaining semantic abstraction.

### Overall Structure of the SwinT-CNN Model

The model gains spatial focus through anatomical prior embedding in the convolutional branch as shown in Fig. 7.5. The 3D U-Net model receives pretraining to extract voxel-wise segmentation features, which are then integrated into the CNN pathway through an AAG [163]. The network design enables the model to focus on clinically meaningful areas, such as pulmonary nodules, during classification without needing extra supervision or retraining of the segmentation network.

The 3D U-Net encoder trained independently for nodule segmentation remains fixed during end-to-end classification training to maintain its anatomical priors. The AAG modules feed

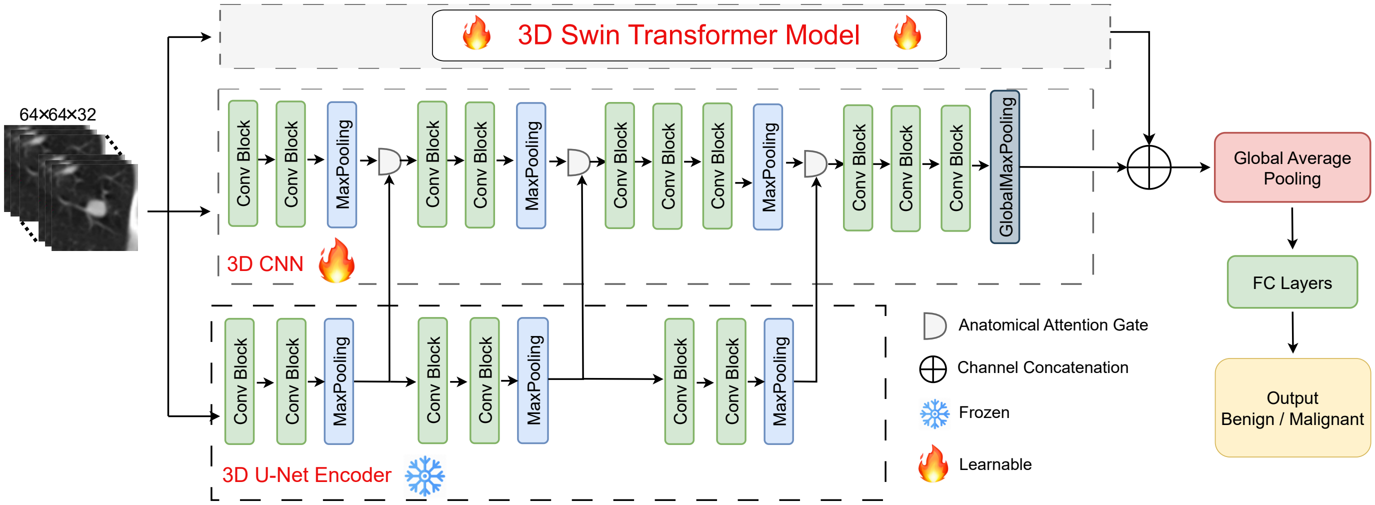


Figure 7.5 Overall Structure of the proposed improved 3D SwinT-CNN model.

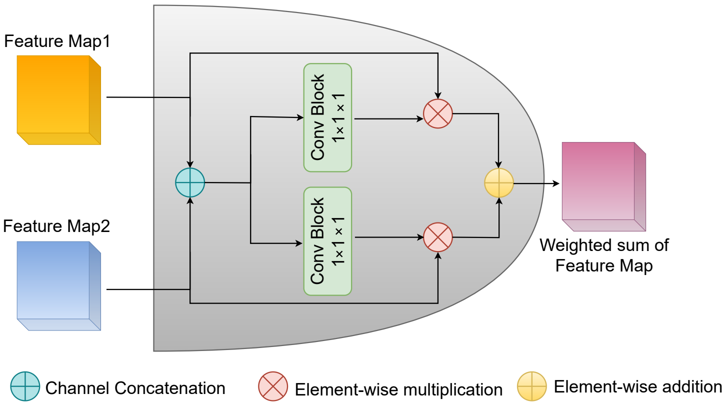


Figure 7.6 Overview of the Anatomical Attention Gate (AAG).

intermediate feature maps containing nodule structural, semantic, and spatial information to corresponding CNN layers. The AAG modules combine the CNN's $s$-th convolutional layer feature map $f_i^s$ with anatomical feature map $f_a^s$ extracted from the 3D U-Net encoder as shown in Fig. 7.6. The two feature maps are first combined by stacking them along the channel axis.

$$f_{concat}^s = [f_i^s, f_a^s] \tag{7.2}$$

The combined feature map passes through two parallel $1 \times 1 \times 1$ convolutional layers with sigmoid activations to learn attention weights $O_i^s$ and $O_a^s$ for each stream:

$$\begin{cases} O_i^s = \delta(W_i^s \times f_{concat}^s + b_i^s) \\ O_a^s = \delta(W_a^s \times f_{concat}^s + b_a^s) \end{cases} \tag{7.3}$$

The 3D convolution operation is denoted by ×, and $\delta$ represents the sigmoid function, while $W_i^s$, $W_a^s$, $b_i^s$, and $b_a^s$ are learnable parameters of the gating layers. The attention maps perform element-wise multiplication to weight the respective input features:

$$\begin{cases} f_i^{s1} = O_i^s \times f_i^s \\ f_a^{s1} = O_a^s \times f_a^s \end{cases} \tag{7.4}$$

Finally, the two weighted maps are combined via element-wise addition to yield the output of the gate:

$$f_o^s = f_i^{s1} + f_a^{s1} \tag{7.5}$$

which is then passed as input to the next convolutional block in the CNN stream.

The Swin Transformer branch runs in parallel, enabling joint optimization during end-to-end training. A gating mechanism allows the CNN to adjust its attention based on spatial regions identified by the segmentation model, thereby integrating anatomical priors into the classification process. AAG modules require no additional supervision, as the U-Net encoder remains fixed, thereby preserving anatomical consistency. The network merges outputs from both branches before applying global average pooling and fully connected layers for prediction.

The proposed architecture achieves both explicit anatomical localization interpretability and end-to-end deep classification model adaptability through frozen segmentation features that utilize anatomically guided attention gates. The proposed design enables precise and dependable predictions which match the requirements for trustworthy AI applications in lung nodule diagnosis.

### 7.4.3 Explainability Map Generation

The SwinT-CNN architecture requires visual explanation methods to improve transparency by showing which CT regions influence classification. Its dual-path structure calls for saliency maps reflecting the combined decision process, not separate branches. Fused Grad-CAM maps better align with the final output by highlighting the joint impact of local and global features.

The 3D extension of gradient-weighted class activation mapping (Grad-CAM) [74] serves as our chosen explainability approach, which operates on the fused feature representation before the global pooling layer. The selection of Grad-CAM for this approach stems from three essential factors: (1) Grad-CAM functions well with convolutional structures and works with 3D volume data; (2) The method works directly with activation maps to provide spatially-resolved explanations that match the original image structure; and (3) The approach enables model variant comparison through consistent analysis between models with and without segmentation guidance [164].

The Grad-CAM computation begins by locating the final convolutional block in the fused representation which unites features from both CNN and Swin Transformer branches. The backward pass computes the gradient of the predicted class score relative to each feature channel in this layer. The importance weights for each channel are obtained by averaging the gradients across all channels before applying a weighted sum with the corresponding feature maps. The final output of the saliency map emerges after applying ReLU activation:

$$L_{\text{Grad-CAM}} = \text{ReLU}\left(\sum_k W_k \cdot A^k\right) \tag{7.6}$$

where $A^k$ represents the $k$-th channel of the fused feature map, and $W_k$ is the corresponding importance weight derived from the gradient signal. The resulting map is upsampled to match the original input dimensions and visualized as a volumetric heatmap overlaid on the CT scan.

The strategy offers a unified interpretability framework compatible with any model, enabling version comparisons. We apply it to two configurations: (1) The baseline SwinT-CNN without segmentation; and (2) The enhanced version with anatomical priors from a pretrained U-Net integrated via anatomical attention gates. Heatmap analysis reveals how anatomical priors and hierarchical attention influence model focus and diagnostic reasoning.

## 7.5 Experiment Results and Discussion

### 7.5.1 Experimental Setting

#### Segmentation Experimental Setting

The nnU-Net framework [162] served as our method for pulmonary nodule segmentation because it automatically adjusts network architecture and training strategy based on input data characteristics. The training process utilizes a composite loss function which combines

Dice loss ($\mathcal{L}_{\text{Dice}}$) with cross-entropy loss ($\mathcal{L}_{\text{CE}}$):

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{Dice}} + w \cdot \mathcal{L}_{\text{CE}} \tag{7.7}$$

The weighting factor $w$ was set to 1 in our experiments. The training configuration is summarized in Table 7.1.

Table 7.1 Training parameters using nnU-Net.

| **Parameter** | **Configuration** |
|---|---|
| Optimizer | Stochastic Gradient Descent |
| Learning rate | 0.01 |
| Momentum | 0.99 |
| Weight decay | $3 \times 10^{-5}$ |
| Batch size | 14 |
| Number of epochs | 200 |

The training process employed a five-fold cross-validation approach to achieve robustness and generalizability. The five models were combined into an ensemble to generate the final segmentation output.

### Classification Experimental Setting

The SwinT-CNN model received end-to-end training on $64 \times 64 \times 32$ volumetric CT patches through PyTorch as its deep learning framework. The training process took place on two NVIDIA V100 GPUs, each with 192 GB of system memory. The training process used AdamW optimizer with $10^{-3}$ learning rate and $5 \times 10^{-5}$ weight decay and 16 batch size for 250 epochs. The training process used a cosine annealing learning rate scheduler to decrease the learning rate to 1% of its initial value during the training period.

The architecture follows a dual-path design, combining a 3D convolutional stream and a 3D Swin Transformer branch. The transformer branch adopts a hierarchical structure with three stages, where each stage includes 3D W-MSA and SW-MSA blocks. The network uses a hidden embedding size (C) of 96, a window size of $4 \times 4 \times 4$, and a stage-wise configuration of `[2, 4, 2]` transformer layers with corresponding attention heads set to `[3, 6, 12]`. The classification objective was treated as a binary classification problem to distinguish between benign and malignant nodules. The model was trained using the binary cross-entropy loss:

$$\mathcal{L}_{\text{Cls}} = -\left[y \cdot \log(\hat{y}) + (1-y) \cdot \log(1-\hat{y})\right] \tag{7.8}$$

The CNN and SwinT, along with fusion layers and classification head, received joint updates through backpropagation while the U-Net encoder remained frozen to maintain anatomical prior integrity.

Table 7.2 Ablation study of classification performance on the LIDC and PN9 datasets.

| **Dataset** | **Model** | **Accuracy (%)** | **Sensitivity (%)** | **Specificity (%)** | **Precision(%)** | **AUC** |
|---|---|---|---|---|---|---|
| LIDC | 3D SwinT | 94.1<br>(93.3-94.9) | 90.0<br>(88.9-91.1) | 88.8<br>(87.3-90.3) | 90.2<br>(89.3-91.1) | 0.976<br>(0.973-0.979) |
| | 3D SwinT-CNN | 95.4<br>(94.1-96.7) | 92.8<br>(91.7-93.9) | 91.7<br>(90.2-93.2) | 92.3<br>(91.5-93.1) | 0.986<br>(0.981-0.991) |
| | Improved 3D SwinT-CNN | 97.3<br>(96.4-98.2) | 96.7<br>(95.5-97.9) | 95.1<br>(93.5-96.7) | 95.8<br>(94.5-97.1) | 0.993<br>(0.991-0.995) |
| PN9 | 3D SwinT | 86.3<br>(83.6-89.0) | 81.7<br>(79.6-83.8) | 84.6<br>(82.7-86.5) | 83.9<br>(81.7-86.1) | 0.931<br>(0.923-0.939) |
| | 3D SwinT-CNN | 89.7<br>(87.3-92.1) | 86.6<br>(84.0-89.2) | 88.4<br>(86.3-90.5) | 87.2<br>(85.3-89.1) | 0.954<br>(0.947-0.961) |
| | Improved 3D SwinT-CNN | 93.8<br>(92.7-94.9) | 91.5<br>(90.2-92.8) | 89.9<br>(88.4-91.4) | 91.0<br>(89.6-92.4) | 0.963<br>(0.956-0.970) |

### 7.5.2 Ablation Study

The evaluation of each component in the proposed improved SwinT-CNN framework required running ablation experiments on LIDC-IDRI and PN9 datasets. The experiments evaluated how each architectural enhancement specifically contributed to the model's classification performance by adding a CNN branch and implementing anatomical attention. Table 7.2 shows the quantitative results from different configurations. The evaluation of interpretability included Grad-CAM saliency maps in addition to classification metrics. The attention maps from SwinT-CNN and improved SwinT-CNN models are visually compared in Fig. 7.7 for three representative nodule samples.

### 7.5.3 Classification Evaluation

The proposed model's ability to distinguish between malignant (positive class) and benign (negative class) pulmonary nodules through binary classification is evaluated using sensitivity, specificity, precision, and accuracy metrics. The confusion matrix components, true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), are used to derive these metrics.

$$\text{Sensitivity} = \frac{TP}{TP + FN}, \quad \text{Specificity} = \frac{TN}{TN + FP} \tag{7.9}$$

$$\text{Precision} = \frac{TP}{TP + FP}, \tag{7.10}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{7.11}$$

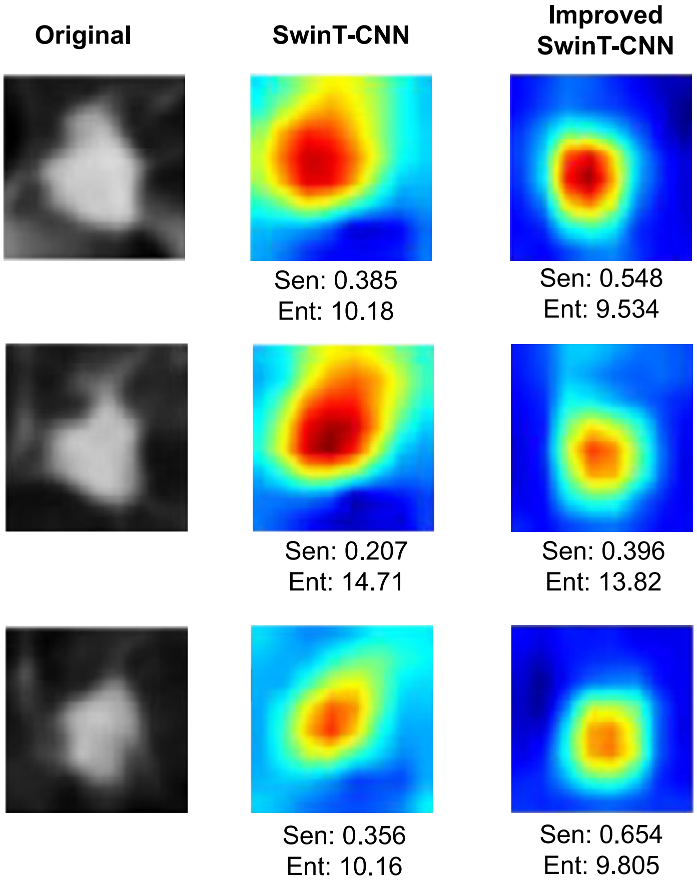


Figure 7.7 Comparison of Grad-CAM heatmaps from SwinT-CNN and improved SwinT-CNN for three nodules.

These indicators show how well the model detects cancer (sensitivity), avoids false alarms (specificity), makes reliable positive predictions (precision), and maintains overall correctness (accuracy).

### 7.5.4 Explanation Evaluation

The interpretability of Grad-CAM heatmaps generated by SwinT-CNN model variants is measured through sensitivity and entropy metrics. The metrics assess nodules to determine how well the model focuses its attention and how clearly it distinguishes between clinical categories.

**Sensitivity (Heatmap-ROI Overlap)** The metric evaluates the spatial relationship between the saliency map and the annotated nodule region. The proportion of attention energy contained within the ground truth mask defines this metric.

$$\text{Sensitivity} = \frac{\sum_{i \in R} H_i}{\sum_i H_i} \tag{7.12}$$

where $H_i$ denotes the Grad-CAM intensity at voxel $i$, and $R$ is the set of voxels within the annotated nodule region. Higher values indicate that the model focuses more precisely on the relevant anatomical area.

**Entropy (Spatial Concentration)** The normalized Shannon entropy calculation of the saliency distribution determines the spatial concentration or diffusion of the attention map.

$$\text{Entropy} = -\sum_i p_i \log(p_i), \quad p_i = \frac{H_i}{\sum_j H_j} \tag{7.13}$$

The level of entropy determines how compact and easy to understand heatmaps are, as lower entropy values create more compact heatmaps. In comparison, higher entropy values result in heatmaps with dispersed attention.

### 7.5.5 Evaluation Results

#### 3D SwinT (Baseline)

The 3D SwinT model served as the base configuration to extract global contextual dependencies from volumetric CT data. The hierarchical window-based self-attention mechanism in this architecture successfully models both long-range interactions and high-level semantics, which are essential for malignancy assessment. The 3D SwinT model demonstrated excellent performance on the LIDC dataset, achieving 94.1% accuracy, 90.0% sensitivity, 88.8% specificity, and 90.2% precision. The model showed strong generalization capabilities on the external PN9 dataset, achieving 86.3% accuracy and an AUC of 0.931.

#### 3D SwinT-CNN (Enhancing Local Representations)

SwinT gains enhanced global representation by integrating a parallel 3D CNN branch that captures nodule shape, boundary texture, and internal intensity features. SwinT-CNN merges both streams at the feature level, allowing simultaneous use of local and global information. As shown in Table 7.2, this fusion improved LIDC performance, increasing sensitivity by 2.8% and specificity by 2.9%. On the PN9 test set, adding local context increased classification accuracy to 89.7% and sensitivity to 86.6%, thereby enhancing the model's robustness and generalization.

The hybrid structure of the model enhanced its interpretability, as demonstrated by the Grad-CAM visualizations (Fig. 7.7). The saliency maps generated by SwinT-CNN showed moderate alignment with the annotated nodules and achieved sensitivity values of 0.385, 0.207, and 0.356 across the three samples. The entropy values (0.948, 0.897, 0.821) indicate that the model paid attention to the region, but the attention was spread out spatially, which suggests that there is room for improvement in localization precision.

### Improved 3D SwinT-CNN (Integrating Anatomical Priors)

The CNN stream is guided toward clinically relevant areas by integrating anatomical priors via AAG, which injects features from a pretrained 3D U-Net encoder. This enhancement enables the model to focus on nodule-related structures using spatial cues from segmentation. As shown in Table 7.2, the improved architecture achieved the highest classification results on both datasets: 97.3% accuracy, 96.7% sensitivity, 95.1% specificity, and 95.8% precision on LIDC. The AUC improved from 0.976 (baseline SwinT) to 0.993 (SwinT-CNN). On the external PN9 dataset, the model reached 93.8% accuracy and 0.963 AUC, confirming gains in diagnostic precision and generalization.

The improved SwinT-CNN achieved better accuracy results while providing enhanced interpretability features. The Grad-CAM heatmaps of this model achieved higher sensitivity values (0.548, 0.396, 0.654) than those of SwinT-CNN, indicating better alignment between the attention map and the actual lesion. The entropy values remained consistently low at (0.678, 0.595, 0.483), which suggests that the model focuses on clinically significant areas. The results demonstrate that adding anatomical priors enhances both diagnostic accuracy and model transparency, which is crucial for clinical use.

### 7.5.6 Comparison Results and Discussion

Research has introduced multiple deep learning frameworks for pulmonary nodule classification, which strive to achieve both high diagnostic precision and model interpretability. The earlier approaches, including MC-CNN [165], HSCNN [77], and MTMR-Net [45], investigated multi-task or attribute-driven strategies. The MTMR-Net model provided valuable insights into inter-attribute relationships but required additional classifiers and complex post-processing analysis methods, which made the model more complicated and limited its ability to function as a comprehensive system.

The research of Fu et al. [78] introduced two new attention-based methods that improved diagnostic region detection through slice-level weighting and cross-attribute attention mechanisms. The technique achieved better performance in specific cases, yet its interpretability remained focused on particular areas without providing a comprehensive understanding of the complete spatial relationships. ExPN-Net [5] achieved competitive results by integrating segmentation priors with classification through anatomical attention and soft activation maps. The methods required multiple training stages and separate attention modules that operated independently.

Table 7.3 Performance comparison of benign-malignant classification on LIDC dataset.

| **Model** | **Year** | **Accuracy (%)** | **Sensitivity (%)** | **Specificity (%)** | **Precision(%)** | **AUC** |
|---|---|---|---|---|---|---|
| MC-CNN [165] | 2017 | 87.1 | 77.0 | 93.0 | - | 0.930 |
| Song et al. [166] | 2017 | 84.2 | 84.0 | 84.3 | - | - |
| HSCNN [77] | 2019 | 84.2 | 70.5 | 88.9 | - | 0.856 |
| Xie et al. [167] | 2019 | 91.6 | 86.5 | 94.0 | 87.8 | 0.957 |
| MTMR-Net [45] | 2019 | 93.5 | 83.0 | 89.4 | - | 0.979 |
| MSCS-DeepLN [168] | 2020 | 92.7 | 85.6 | 94.9 | 90.4 | 0.940 |
| Fu et al. [78] | 2022 | 94.7 | 96.2 | 82.9 | **97.8** | 0.959 |
| Swin-T [169] | 2024 | 93.0 | 86.0 | 85.3 | 87.7 | 0.960 |
| ExPN-Net [5] | 2024 | 95.5 | **1.00** | 94.7 | 78.0 | 0.992 |
| Improved SwinT-CNN (Ours) |  | **97.3** | 96.7 | **95.1** | 95.8 | **0.993** |

This study presents an end-to-end, unified framework that simultaneously captures global semantics, local features, and anatomical relevance for the classification of lung nodules. By leveraging anatomical attention gates (AAG), the model integrates segmentation-derived priors into parallel CNN and transformer-based pathways. This fusion enables adaptive, context-aware attention to clinically meaningful regions, eliminating the need for handcrafted attribute hierarchies or auxiliary modules.

Compared to all state-of-the-art approaches (Table 7.3), the proposed SwinT-CNN model demonstrates superior performance, achieving the highest accuracy (97.3%), specificity (95.1%), and AUC (0.993) on the LIDC dataset. These results confirm that the joint modeling of spatial priors and hierarchical features leads to enhanced diagnostic precision and interpretability. Notably, the framework aligns with clinical diagnostic practices by producing anatomically interpretable decisions that closely reflect radiological reasoning. Grad-CAM-based heatmaps confirm improved attention localization and reduced entropy, indicating more focused and reliable diagnostic inference. These properties address key limitations of previous CAD systems, paving the way for more trustworthy AI-assisted clinical workflows.

## 7.6 Conclusion

The research presented improved 3D SwinT-CNN as an end-to-end hybrid architecture which unites 3D Swin Transformers with CNNs and segmentation-derived anatomical priors for lung nodule classification. The model achieved strong classification performance and reliable interpretability through its ability to jointly model global context and local features and spatial attention through anatomical attention gates. The model demonstrated robustness and generalization capabilities through experiments conducted on LIDC-IDRI and PN9 datasets across different imaging conditions. Our future research will focus on enhancing both accuracy and interpretability through the integration of SwinT-UNet modules into the transformer stream to achieve better anatomical structure and global semantic fusion. This research aims to motivate additional studies that unite model performance improvement with interpretability enhancement to advance AI adoption in medical imaging.

# CHAPTER 8 GENERAL DISCUSSION

This chapter provides a comprehensive discussion of the main contributions of this thesis in light of the three primary research objectives: (1) Enhancing sensitivity in detecting small pulmonary nodules; (2) Reducing false positives to improve diagnostic precision; (3) Improving computational efficiency and real-world applicability; and (4) Promoting interpretability and clinical transparency. Drawing from four peer-reviewed research articles, the discussion integrates quantitative evaluations, ablation studies, explainability analyses, and experimental comparisons across two benchmark datasets (LUNA16 and PN9), highlighting how each work package contributes to the overarching goals of early-stage lung cancer detection and its adoption in clinical settings.

## 8.1 Sensitivity Improvement

Improving sensitivity—particularly for small and early-stage nodules—was a central goal throughout this thesis. Accurate detection of such nodules is crucial in clinical settings, where missed detections may lead to delayed diagnoses and reduced patient survival rates.

The first research contribution, presented in Chapter 4, addressed the challenge of low sensitivity in conventional CNN architectures by enhancing feature extraction using a modified VGG16 model. By combining the last three convolutional layers into a composite feature map, the Region Proposal Network (RPN) generated more fine-grained proposals. This significantly improved recall for small nodules, especially at low IoU thresholds. The proposed approach maintained high performance even when the number of region proposals was reduced from 2000 to 300.

In Chapter 5, sensitivity was further improved via the Optimized Multi-Scale CNN (OMS-CNN), which used metaheuristic algorithms—Parameter Setting-Free Harmony Search (PSF-HS) and Beetle Antenna Search (BAS)—to optimize the feature extraction layers. This led to a CPM score of 0.8922 on LUNA16, surpassing baseline Faster R-CNN configurations and other models.

Chapter 6 introduced the DA OMS-CNN model, which integrates dual-attention mechanisms and DA-RoIPooling. With a sensitivity of 96.93% and a CPM score of 0.9112 on LUNA16, the model achieved state-of-the-art performance. Notably, it reached 0.9331 sensitivity at 1 FP/scan, an essential metric for clinical application.

Ablation studies confirmed that each component (dual attention, DA-RoIPooling, and FPR) contributed incrementally to the performance improvements. These enhancements were crucial for detecting subtle and small nodules, thereby fulfilling Objectives A1 and A2. In addition, the improved SwinT-CNN model in Chapter 7 preserved high sensitivity (96.7%) while enhancing interpretability, demonstrating that accurate detection and transparency can be achieved simultaneously.

## 8.2 False Positive Reduction

Reducing false positives is essential to avoid unnecessary clinical interventions and improve trust in CAD systems. Each of the four work packages tackled this challenge through architectural, algorithmic, and interpretability-driven strategies.

Chapter 5 introduced hybrid FPR using 3D CNNs, which effectively filtered non-nodular candidates. Chapter 6 further enhanced this stage through ensemble-based learning with three 3D Swin Transformer (3D SwinT) models. These models leveraged attention mechanisms to suppress noise and irrelevant anatomical patterns.

On LUNA16, the proposed approach achieved a CPM of 0.9112, outperforming leading models such as TED (0.8846) and MK-3DCNN (0.8591). On PN9, it maintained strong performance with a CPM of 0.6034, demonstrating its ability to generalize across datasets. The system showed robustness even at very low FP rates, a key requirement for clinical deployment.

Qualitative analysis further confirmed the model's effectiveness in distinguishing ambiguous structures near pleural walls or blood vessels. This highlights the value of ensemble attention and domain-invariant feature learning.

Chapter 7 also contributed to false positive reduction by using anatomical attention gates to suppress irrelevant activations, thereby improving both classification precision and the spatial focus of the network.

## 8.3 Efficiency and Clinical Applicability

Efficiency and deployability were central to the design of all models in this thesis. Deep learning in medical imaging requires not just accuracy but also speed, memory efficiency, and interpretability.

Chapter 5 optimized training complexity using PSF-HS and BAS, allowing for automated hyperparameter tuning. Experiments used scalable infrastructure (V100 GPUs, 192 GB RAM), warm-up cosine learning rate schedules, and optimizers such as SGD and AdamW.

Chapter 7 maintained a lightweight dual-path design and leveraged pretrained segmentation priors, reducing the need for retraining auxiliary modules and improving training stability.

Clinical relevance was evaluated using the PN9 dataset, which includes different acquisition settings and patient variability. DA OMS-CNN outperformed methods such as SSD512 and RetinaNet, maintaining a CPM above 0.6 and demonstrating its robustness in unseen domains.

Moreover, the pipeline's modular design allows for seamless integration into existing CAD workflows. The candidate detection stage can act as a triage tool, while the FPR module supports decision-making. These features point to real-world deployability and fulfill Objective C1.

## 8.4 Interpretability and Clinical Transparency

Interpretability plays a crucial role in gaining clinical trust and facilitating the real-world deployment of AI-based CAD systems. While prior work focused on improving accuracy and reducing false positives, Chapter 7 addressed the challenge of model transparency by introducing an interpretable hybrid architecture.

The proposed improved SwinT-CNN model integrates a dual-path structure that combines 3D CNNs and 3D Swin Transformers, along with anatomical attention gates (AAGs) that inject voxel-level priors from a pre-trained U-Net segmentation model into the CNN branch. This design helps the model focus on clinically meaningful regions during classification.

Quantitative evaluation using Grad-CAM showed enhanced alignment between the model's attention and actual nodule locations. The improved model achieved higher sensitivity values (0.548, 0.396, 0.654) and lower entropy scores (0.678, 0.595, 0.483), indicating more precise and focused visual explanations. These interpretability gains were achieved without compromising performance, as the model reached 97.3% accuracy and 0.993 AUC on LIDC.

Together, these results fulfill Objectives D1 and D2, demonstrating that incorporating anatomical priors and attention mechanisms can simultaneously improve diagnostic performance and model explainability—key requirements for clinical integration.

## 8.5 Summary

This chapter has synthesized the key findings of the thesis in alignment with the four core research objectives: enhancing sensitivity, reducing false positives, improving efficiency and clinical applicability, and promoting interpretability and clinical transparency. Each of the four research articles contributed distinct architectural and algorithmic innovations that together addressed the major limitations in early-stage lung nodule detection using deep learning.

The first objective, sensitivity improvement, was achieved through progressive enhancements in feature representation, starting from a modified VGG16 backbone to dual-attention mechanisms and adaptive pooling techniques. These improvements significantly increased recall rates, especially for small nodules, without compromising efficiency. The second objective focused on reducing false positives through the use of 3D CNNs and attention-based Swin Transformers. These models improved specificity by distinguishing true nodules from anatomical lookalikes across diverse patient scans.

The third objective emphasized computational practicality and real-world integration. Techniques such as metaheuristic optimization and modular pipeline design ensured that the proposed methods remain scalable, generalizable, and compatible with existing CAD systems.

Finally, the fourth objective addressed the growing demand for model interpretability in clinical workflows. By incorporating anatomical priors and attention-guided saliency mapping, the proposed SwinT-CNN framework offered high diagnostic accuracy alongside transparent decision-making. This integration supports clinical trust and positions the model for deployment in explainable AI-based medical imaging systems.

Overall, this body of work presents a robust and clinically meaningful framework for early-stage lung cancer detection, setting the stage for future advancements in explainable and deployable AI tools in medical imaging.

# CHAPTER 9 CONCLUSION

The research conducted in this thesis contributes to the advancement of early-stage lung cancer detection using deep learning, with a particular emphasis on improving sensitivity to small nodules, reducing false positives, enhancing clinical applicability, and promoting interpretability and trust in clinical practice. This work was structured around four primary research objectives, which were addressed through a series of interrelated studies. Each study proposed novel methods and architectures grounded in convolutional neural networks, metaheuristic optimization, vision transformers, and anatomical attention mechanisms to tackle key limitations of existing computer-aided detection (CAD) systems.

## 9.1 Insights and Reflections

Throughout this doctoral research, the overarching theme has been the design and refinement of a robust, interpretable, and clinically relevant framework for lung nodule detection. This journey began with the generation of enhanced region proposals using a modified VGG16 backbone, continued with metaheuristically optimized multi-scale CNNs, and advanced to a dual-attention architecture incorporating 3D Swin Transformers for reducing false positives. Finally, the thesis culminated in a hybrid model that integrates anatomical priors to enhance both diagnostic accuracy and interpretability.

A key takeaway from this trajectory is the importance of domain-specific architectural adaptations for medical imaging tasks. While standard object detection frameworks like Faster R-CNN provide a solid foundation, their performance on subtle and variable clinical targets—such as small pulmonary nodules—can be significantly improved through tailored enhancements. This thesis demonstrates how components such as DA-RoIPooling, ensemble attention mechanisms, and anatomical attention gates can bridge the gap between generic deep learning methods and domain-specific demands.

Another core insight is the value of interpretability in clinical AI. Beyond raw accuracy, the ability of a model to explain its decisions visually and anatomically is essential for building clinical trust. The final study in this thesis integrated segmentation-derived priors and saliency-based evaluations (e.g., Grad-CAM, sensitivity, entropy) to ensure that predictions were not only accurate but also explainable and focused on relevant structures.

Equally important was the emphasis on reproducibility and generalizability. The use of benchmark datasets (LUNA16 and PN9), along with systematic experimental design and ablation studies, helped ensure that the proposed methods are scientifically rigorous and clinically transferable. Each contribution built upon the previous one, forming a coherent methodological pipeline that elevates both the performance and practical utility of CAD systems.

From a research perspective, this work has provided extensive experience in designing deep learning systems, experimental tuning, and integrating optimization techniques. From a broader standpoint, it reinforced the importance of aligning technical innovation with clinical needs, emphasizing that trust, transparency, and interpretability are no longer optional—they are foundational for the adoption of AI in healthcare.

## 9.2 Future Research Directions

Building upon the foundations established in this thesis, several promising directions can be pursued to enhance further the impact, interpretability, and clinical viability of AI-driven lung nodule detection systems:

- **Advancing Explainable AI for Clinical Trust:** While this thesis introduced anatomical attention and Grad-CAM-based visual explanations, future work could focus on integrating multi-level interpretability frameworks. These may include concept-based explanations, clinician-aligned attribution methods, and post-hoc analysis tools such as SHAP and LIME. The goal is to align model reasoning with radiological decision-making and support transparent clinical deployment.

- **Multimodal and Longitudinal Data Integration:** Current methods rely solely on single-phase CT imaging. Incorporating multimodal sources—such as PET scans, EHRs, or longitudinal CT studies—could provide more context for malignancy assessment and temporal tracking, enabling models to predict nodule behavior over time.

- **Model Compression and Deployment Efficiency:** Although the proposed models demonstrate high diagnostic performance, their computational demands may limit real-world use. Future research could explore techniques such as neural architecture search (NAS), pruning, quantization, and knowledge distillation to reduce inference time and memory footprint.

- **Progressive Nodule Modeling:** Developing predictive models that track morphological changes in pulmonary nodules across timepoints can enhance malignancy risk stratification and assist in personalized treatment planning. Temporal modeling using recurrent or transformer-based architectures could be explored.

- **Clinical Trials and User-Centered Validation:** Ultimately, AI tools must be validated in clinical settings. Future efforts should involve prospective clinical trials, real-time radiologist feedback, and clinician-in-the-loop systems to assess usability, interpretability, and diagnostic value under real-world constraints.

In conclusion, this thesis offers a comprehensive contribution to the field of AI-based medical image analysis by proposing novel, interpretable, and clinically relevant solutions for early lung cancer detection. The developed models not only surpass existing benchmarks in terms of accuracy and robustness, but also promote transparency and clinical trust—key enablers for real-world adoption. Grounded in rigorous experimentation and aligned with evolving healthcare needs, these contributions lay a strong foundation for future innovations at the intersection of artificial intelligence and clinical decision-making.

# REFERENCES

[1] S. Qiu, D. Wen, Y. Cui, and J. Feng, “Lung nodules detection in ct images using gestalt-based algorithm,” *Chinese Journal of Electronics*, vol. 25, no. 4, pp. 711–718, 2016.

[2] S. Dai, K. Lu, J. Dong, Y. Zhang, and Y. Chen, “A novel approach of lung segmentation on chest ct images using graph cuts,” *Neurocomputing*, vol. 168, pp. 799–807, 2015.

[3] P. H. Viale, “The american cancer society’s facts & figures: 2020 edition,” *Journal of the advanced practitioner in oncology*, vol. 11, no. 2, p. 135, 2020.

[4] W. Zhu, C. Liu, W. Fan, and X. Xie, “Deeplung: Deep 3d dual path nets for automated pulmonary nodule detection and classification,” in *2018 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 2018, pp. 673–681.

[5] C. Wang, Y. Liu, F. Wang, C. Zhang, Y. Wang, M. Yuan, and G. Yang, “Towards reliable and explainable ai model for pulmonary nodule diagnosis,” *Biomedical Signal Processing and Control*, vol. 88, p. 105646, 2024.

[6] Y.-W. Jeong, S.-M. Park, Z. W. Geem, and K.-B. Sim, “Advanced parameter-setting-free harmony search algorithm,” *Applied Sciences*, vol. 10, no. 7, p. 2586, 2020.

[7] Q. Wu, Z. Ma, G. Xu, S. Li, and D. Chen, “A novel neural network classifier using beetle antennae search algorithm for pattern classification,” *IEEE access*, vol. 7, pp. 64 686–64 696, 2019.

[8] C. O. WHO, “World health organization,” *Air Quality Guidelines for Europe*, no. 91, 2020.

[9] National Centre for Disease Informatics and Research, “Clinicopathological profile of cancers in india: A report of the hospital based cancer registries,” https://ncdirindia.org/all_reports/hbcr_2021/, 2021, accessed: July 7, 2025.

[10] S. Cressman, S. J. Peacock, M. C. Tammemägi, W. K. Evans, N. B. Leighl, J. R. Goffin, A. Tremblay, G. Liu, D. Manos, P. MacEachern *et al.*, “The cost-effectiveness of high-risk lung cancer screening and drivers of program efficiency,” *Journal of Thoracic Oncology*, vol. 12, no. 8, pp. 1210–1222, 2017.

[11] W. Cao, R. Wu, G. Cao, and Z. He, “A comprehensive review of computer-aided diagnosis of pulmonary nodules based on computed tomography scans,” *IEEE Access*, vol. 8, pp. 154 007–154 023, 2020.

[12] A. Snoeckx, P. Reyntiens, D. Desbuquoit, M. J. Spinhoven, P. E. Van Schil, J. P. van Meerbeeck, and P. M. Parizel, “Evaluation of the solitary pulmonary nodule: size matters, but do not ignore the power of morphology,” *Insights into imaging*, vol. 9, pp. 73–86, 2018.

[13] D. Ost, A. M. Fein, and S. H. Feinsilver, “The solitary pulmonary nodule,” *New England Journal of Medicine*, vol. 348, no. 25, pp. 2535–2542, 2003.

[14] A. Shankar, D. Saini, A. Dubey, S. Roy, S. J. Bharati, N. Singh, M. Khanna, C. P. Prasad, M. Singh, S. Kumar *et al.*, “Feasibility of lung cancer screening in developing countries: challenges, opportunities and way forward,” *Translational lung cancer research*, vol. 8, no. Suppl 1, p. S106, 2019.

[15] L. Chen, H. Sun, and Y. Huang, “Pet-ct principles and applications in lung cancer management,” in *Medical Imaging-Principles and Applications.* IntechOpen, 2019.

[16] J. Liang, G. Ye, J. Guo, Q. Huang, and S. Zhang, “Reducing false-positives in lung nodules detection using balanced datasets,” *Frontiers in public health*, vol. 9, p. 671070, 2021.

[17] H. Zhang, Y. Peng, and Y. Guo, “Pulmonary nodules detection based on multi-scale attention networks,” *Scientific Reports*, vol. 12, no. 1, p. 1466, 2022.

[18] F. Shariaty and M. Mousavi, “Application of cad systems for the automatic detection of lung nodules,” *Informatics in Medicine Unlocked*, vol. 15, p. 100173, 2019.

[19] J. Juan, E. Monsó, C. Lozano, M. Cufí, P. Subías-Beltrán, L. Ruiz-Dern, X. Rafael-Palou, M. Andreu, E. Castañer, X. Gallardo *et al.*, “Computer-assisted diagnosis for an early identification of lung cancer in chest x rays,” *Scientific Reports*, vol. 13, no. 1, p. 7720, 2023.

[20] M. Liu, J. Wu, N. Wang, X. Zhang, Y. Bai, J. Guo, L. Zhang, S. Liu, and K. Tao, “The value of artificial intelligence in the diagnosis of lung cancer: A systematic review and meta-analysis,” *PLoS One*, vol. 18, no. 3, p. e0273445, 2023.

[21] G. Li, W. Zhou, W. Chen, F. Sun, Y. Fu, F. Gong, and H. Zhang, “Study on the detection of pulmonary nodules in CT images based on deep learning,” *IEEE Access*, vol. 8, pp. 67 300–67 309, 2020.

[22] H. Tang, D. R. Kim, and X. Xie, “Automated pulmonary nodule detection using 3D deep convolutional neural networks,” in *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*. IEEE, 2018, pp. 523–526.

[23] R. Sharma, M. Saqib, C.-T. Lin, and M. Blumenstein, “A survey on object instance segmentation,” *SN Computer Science*, vol. 3, no. 6, p. 499, 2022.

[24] E. Elyan, P. Vuttipittayamongkol, P. Johnston, K. Martin, K. McPherson, C. F. Moreno-García, C. Jayne, and M. M. K. Sarker, “Computer vision and machine learning for medical image analysis: recent advances, challenges, and way forward,” *Artificial Intelligence Surgery*, vol. 2, no. 1, pp. 24–45, 2022.

[25] H. Lindroth, K. Nalaie, R. Raghu, I. N. Ayala, C. Busch, A. Bhattacharyya, P. Moreno Franco, D. A. Diedrich, B. W. Pickering, and V. Herasevich, “Applied artificial intelligence in healthcare: a review of computer vision technology application in hospital settings,” *Journal of Imaging*, vol. 10, no. 4, p. 81, 2024.

[26] The Cancer Imaging Archive, “Downloading TCIA images,” https://wiki.cancerimagingarchive.net/display/NBIA/Downloading+TCIA+Images, 2021, accessed: July 2025.

[27] Lung Image Database Consortium, “LIDC dataset description,” https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=1966254, 2021, accessed: July 2025.

[28] LUNA Challenge Team, “Lung nodule analysis 2016 (LUNA) dataset,” https://academictorrents.com/details/58b053204337ca75f7c2e699082baeb57aa08578, 2016, accessed: July 2025.

[29] National Cancer Institute, “National lung screening trial (NLST) dataset,” https://cdas.cancer.gov/nlst/, 2011, accessed: July 2025.

[30] Mei et al., “PN9: A large-scale pulmonary nodule dataset with 8,798 CT scans and 40,439 annotated nodules,” https://github.com/mj129/SANet, 2021, accessed: July 2025.

[31] S. G. Armato III and W. F. Sensakovic, “Automated lung segmentation for thoracic CT: impact on computer-aided diagnosis1,” *Academic radiology*, vol. 11, no. 9, pp. 1011–1021, 2004.

[32] I. Sluimer, A. Schilham, M. Prokop, and B. Van Ginneken, “Computer analysis of computed tomography scans of the lung: a survey,” *IEEE transactions on medical imaging*, vol. 25, no. 4, pp. 385–405, 2006.

[33] F. Liao, M. Liang, Z. Li, X. Hu, and S. Song, “Evaluate the malignancy of pulmonary nodules using the 3-d deep leaky noisy-or network,” *IEEE transactions on neural networks and learning systems*, vol. 30, no. 11, pp. 3484–3495, 2019.

[34] S. M. Naqi, M. Sharif, and I. U. Lali, “A 3D nodule candidate detection method supported by hybrid features to reduce false positives in lung nodule detection,” *Multimedia Tools and Applications*, vol. 78, pp. 26 287–26 311, 2019.

[35] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, “3D U-Net: learning dense volumetric segmentation from sparse annotation,” in *International conference on medical image computing and computer-assisted intervention.* Springer, 2016, pp. 424–432.

[36] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, “Unet++: A nested U-Net architecture for medical image segmentation,” in *Deep learning in medical image analysis and multimodal learning for clinical decision support: 4th international workshop, DLMIA 2018, and 8th international workshop, ML-CDS 2018, held in conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, proceedings 4.* Springer, 2018, pp. 3–11.

[37] M. Z. Alom, C. Yakopcic, M. Hasan, T. M. Taha, and V. K. Asari, “Recurrent residual U-Net for medical image segmentation,” *Journal of medical imaging*, vol. 6, no. 1, pp. 014 006–014 006, 2019.

[38] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz *et al.*, “Attention U-Net: Learning where to look for the pancreas,” *arXiv preprint arXiv:1804.03999*, 2018.

[39] G. Savitha and P. Jidesh, “A fully-automated system for identification and classification of subsolid nodules in lung computed tomographic scans,” *Biomedical Signal Processing and Control*, vol. 53, p. 101586, 2019.

[40] S. A. Khan, S. Hussain, S. Yang, and K. Iqbal, “Effective and reliable framework for lung nodules detection from CT scan images,” *Scientific reports*, vol. 9, no. 1, p. 4989, 2019.

[41] S. A. El-Regaily, M. A. M. Salem, M. H. A. Aziz, and M. I. Roushdy, “Multi-view convolutional neural network for lung nodule false positive reduction,” *Expert systems with applications*, vol. 162, p. 113017, 2020.

[42] B. K. Veronica, “An effective neural network model for lung nodule detection in CT images with optimal fuzzy model,” *Multimedia Tools and Applications*, vol. 79, no. 19, pp. 14 291–14 311, 2020.

[43] S. V. Fotin, D. F. Yankelevitz, C. I. Henschke, and A. P. Reeves, “A multiscale laplacian of gaussian (LoG) filtering approach to pulmonary nodule detection from whole-lung CT scans,” *arXiv preprint arXiv:1907.08328*, 2019.

[44] H. Cheng, Y. Zhu, and H. Pan, “Modified U-Net block network for lung nodule detection,” in *2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference (ITAIC).* IEEE, 2019, pp. 599–605.

[45] L. Liu, Q. Dou, H. Chen, J. Qin, and P.-A. Heng, “Multi-task deep model with margin ranking loss for lung nodule analysis,” *IEEE transactions on medical imaging*, vol. 39, no. 3, pp. 718–728, 2019.

[46] J. Wang, J. Wang, Y. Wen, H. Lu, T. Niu, J. Pan, and D. Qian, “Pulmonary nodule detection in volumetric chest CT scans using CNNs-based nodule-size-adaptive detection and classification,” *IEEE Access*, vol. 7, pp. 46 033–46 044, 2019.

[47] J. Liu, L. Cao, O. Akin, and Y. Tian, “Accurate and robust pulmonary nodule detection by 3D feature pyramid network with self-supervised feature learning,” *arXiv preprint arXiv:1907.11704*, 2019.

[48] Y. Tan, X. Fu, J. Zhu, and L. Chen, “A improved detection method for lung nodule based on multi-scale 3d convolutional neural network,” *Concurrency and Computation: Practice and Experience*, vol. 35, no. 13, p. e7034, 2023.

[49] Y. Zhao, Z. Wang, X. Liu, Q. Chen, C. Li, H. Zhao, and Z. Wang, “Pulmonary nodule detection based on multiscale feature fusion,” *Computational And Mathematical Methods In Medicine*, vol. 2022, no. 1, p. 8903037, 2022.

[50] Y. Zamanidoost, T. Ould-Bachir, and S. Martel, “OMS-CNN: Optimized multi-scale CNN for lung nodule detection based on faster R-CNN,” *IEEE Journal of Biomedical and Health Informatics*, 2024.

[51] L. Ma, G. Li, X. Feng, Q. Fan, and L. Liu, “Ticnet: Transformer in convolutional neural network for pulmonary nodule detection on ct images,” *Journal of Imaging Informatics in Medicine*, vol. 37, no. 1, pp. 196–208, 2024.

[52] Y. Zamanidoost, M. Rivron, T. Ould-Bachir, and S. Martel, “DA OMS-CNN: Dual-attention OMS-CNN with 3D swin transformer for early-stage lung cancer detection,” in *Informatics*, vol. 12, no. 3. MDPI, 2025, p. 65.

[53] W. Zuo, F. Zhou, and Y. He, “An embedded multi-branch 3d convolution neural network for false positive reduction in lung nodule detection,” *Journal of digital imaging*, vol. 33, no. 4, pp. 846–857, 2020.

[54] L. Sun, Z. Wang, H. Pu, G. Yuan, L. Guo, T. Pu, and Z. Peng, “Attention-embedded complementary-stream cnn for false positive reduction in pulmonary nodule detection,” *Computers in Biology and Medicine*, vol. 133, p. 104357, 2021.

[55] I. W. Harsono, S. Liawatimena, and T. W. Cenggoro, “Lung nodule detection and classification from thorax ct-scan using retinanet with transfer learning,” *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no. 3, pp. 567–577, 2022.

[56] R. Wu, C. Liang, J. Zhang, Q. Tan, and H. Huang, “Multi-kernel driven 3d convolutional neural network for automated detection of lung nodules in chest ct scans,” *Biomedical Optics Express*, vol. 15, no. 2, pp. 1195–1218, 2024.

[57] M. Almahasneh, X. Xie, and A. Paiement, “Attentnet: Fully convolutional 3d attention for lung nodule detection,” *SN Computer Science*, vol. 6, no. 3, p. 292, 2025.

[58] M. S. Brown, P. Lo, J. G. Goldin, E. Barnoy, G. H. J. Kim, M. F. McNitt-Gray, and D. R. Aberle, “Toward clinically usable CAD for lung cancer screening with computed tomography,” *European radiology*, vol. 24, pp. 2719–2728, 2014.

[59] C. Jacobs, E. M. Van Rikxoort, T. Twellmann, E. T. Scholten, P. A. De Jong, J.-M. Kuhnigk, M. Oudkerk, H. J. De Koning, M. Prokop, C. Schaefer-Prokop *et al.*, “Automatic detection of subsolid pulmonary nodules in thoracic computed tomography images,” *Medical image analysis*, vol. 18, no. 2, pp. 374–384, 2014.

[60] A. Teramoto and H. Fujita, “Automated lung nodule detection using positron emission tomography/computed tomography,” *Artificial intelligence in decision support systems for diagnosis in medical imaging*, pp. 87–110, 2018.

[61] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[62] Z. Zhou, V. Sodha, M. M. Rahman Siddiquee, R. Feng, N. Tajbakhsh, M. B. Gotway, and J. Liang, "Models genesis: Generic autodidactic models for 3D medical image analysis," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part IV 22*. Springer, 2019, pp. 384–393.

[63] Z. Shi, H. Hao, M. Zhao, Y. Feng, L. He, Y. Wang, and K. Suzuki, "A deep CNN based transfer learning method for false positive reduction," *Multimedia Tools and Applications*, vol. 78, pp. 1017–1033, 2019.

[64] Z. Xiao, N. Du, L. Geng, F. Zhang, J. Wu, and Y. Liu, "Multi-scale heterogeneous 3D CNN for false-positive reduction in pulmonary nodule detection, based on chest CT images," *Applied Sciences*, vol. 9, no. 16, p. 3261, 2019.

[65] F. Ciompi, K. Chung, S. J. Van Riel, A. A. A. Setio, P. K. Gerke, C. Jacobs, E. T. Scholten, C. Schaefer-Prokop, M. M. Wille, A. Marchiano *et al.*, "Towards automatic pulmonary nodule management in lung cancer screening with deep learning," *Scientific reports*, vol. 7, no. 1, p. 46479, 2017.

[66] Ö. Günaydin, M. Günay, and Ö. Şengel, "Comparison of lung cancer detection algorithms," in *2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT)*. IEEE, 2019, pp. 1–4.

[67] H. Shakir, H. Rasheed, and T. M. Rasool Khan, "Radiomic feature selection for lung cancer classifiers," *Journal of Intelligent & Fuzzy Systems*, vol. 38, no. 5, pp. 5847–5855, 2020.

[68] M. Al-Shabi, B. L. Lan, W. Y. Chan, K.-H. Ng, and M. Tan, "Lung nodule classification using deep local–global networks," *International journal of computer assisted radiology and surgery*, vol. 14, pp. 1815–1819, 2019.

[69] I. D. Apostolopoulos, "Experimenting with convolutional neural network architectures for the automatic characterization of solitary pulmonary nodules' malignancy rating," *arXiv preprint arXiv:2003.06801*, 2020.

[70] J. L. Causey, J. Zhang, S. Ma, B. Jiang, J. A. Qualls, D. G. Politte, F. Prior, S. Zhang, and X. Huang, "Highly accurate model for prediction of lung nodule malignancy with CT scans," *Scientific reports*, vol. 8, no. 1, p. 9286, 2018.

[71] S. Hussein, P. Kandel, C. W. Bolan, M. B. Wallace, and U. Bagci, "Lung and pancreatic tumor characterization in the deep learning era: novel supervised and unsupervised learning approaches," *IEEE transactions on medical imaging*, vol. 38, no. 8, pp. 1777–1787, 2019.

[72] J. Yang, H. Deng, X. Huang, B. Ni, and Y. Xu, "Relational learning between multiple pulmonary nodules via deep set attention transformers," in *2020 IEEE 17th international symposium on biomedical imaging (ISBI)*. IEEE, 2020, pp. 1875–1878.

[73] E. Tjoa and C. Guan, "A survey on explainable artificial intelligence (xai): Toward medical xai," *IEEE transactions on neural networks and learning systems*, vol. 32, no. 11, pp. 4793–4813, 2020.

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

[75] M. T. Ribeiro, S. Singh, and C. Guestrin, "" why should i trust you?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[76] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[77] S. Shen, S. X. Han, D. R. Aberle, A. A. Bui, and W. Hsu, "An interpretable deep hierarchical semantic convolutional neural network for lung nodule malignancy classification," *Expert systems with applications*, vol. 128, pp. 84–95, 2019.

[78] X. Fu, L. Bi, A. Kumar, M. Fulham, and J. Kim, "An attention-enhanced cross-task network to analyse lung nodule attributes in ct images," *Pattern Recognition*, vol. 126, p. 108576, 2022.

[79] K. V. Aishwarya and A. Asuntha, "A survey on comparative study of lung nodules applying machine learning and deep learning techniques," *Multimedia Tools and Applications*, vol. 84, no. 5, pp. 2127–2181, 2025.

[80] R. Javed, T. Abbas, A. H. Khan, A. Daud, A. Bukhari, and R. Alharbey, "Deep learning for lungs cancer detection: a review," *Artificial Intelligence Review*, vol. 57, no. 8, p. 197, 2024.

[81] H. Jin, C. Yu, Z. Gong, R. Zheng, Y. Zhao, and Q. Fu, "Machine learning techniques for pulmonary nodule computer-aided diagnosis using ct images: A systematic review," *Biomedical Signal Processing and Control*, vol. 79, p. 104104, 2023.

[82] S. H. Hosseini, R. Monsefi, and S. Shadroo, "Deep learning applications for lung cancer diagnosis: a systematic review," *Multimedia Tools and Applications*, vol. 83, no. 5, pp. 14 305–14 335, 2024.

[83] L. A. Torre, F. Bray, R. L. Siegel, J. Ferlay, J. Lortet-Tieulent, and A. Jemal, "Global cancer statistics, 2012," *CA: a cancer journal for clinicians*, vol. 65, no. 2, pp. 87–108, 2015.

[84] M. Malvezzi, G. Carioli, P. Bertuccio, P. Boffetta, F. Levi, C. La Vecchia, and E. Negri, "European cancer mortality predictions for the year 2019 with focus on breast cancer," *Annals of Oncology*, vol. 30, no. 5, pp. 781–787, 2019.

[85] K. Yasufuku, "Early diagnosis of lung cancer," *Clinics in chest medicine*, vol. 31, no. 1, pp. 39–47, 2010.

[86] S. Blandin Knight, P. A. Crosbie, H. Balata, J. Chudziak, T. Hussell, and C. Dive, "Progress and prospects of early detection in lung cancer," *Open biology*, vol. 7, no. 9, p. 170070, 2017.

[87] K. Awai, K. Murao, A. Ozawa, M. Komi, H. Hayakawa, S. Hori, and Y. Nishimura, "Pulmonary nodules at chest CT: effect of computer-aided diagnosis on radiologists' detection performance," *Radiology*, vol. 230, no. 2, pp. 347–352, 2004.

[88] J. R. Uijlings, K. E. Van De Sande, T. Gevers, and A. W. Smeulders, "Selective search for object recognition," *International journal of computer vision*, vol. 104, no. 2, pp. 154–171, 2013.

[89] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[90] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[91] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[92] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[93] C. Bauckhage, "Numpy/scipy recipes for image processing: Binary images and morphological operations," *B-IT, Univ. Bonn, Fraunhofer IAIS, Sankt Augustin, Germany, Tech. Rep*, 2017.

[94] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[95] S. G. Armato III, 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 *et al.*, "The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans," *Medical physics*, vol. 38, no. 2, pp. 915–931, 2011.

[96] J. Wang, H. Ma, C.-J. Ni, J.-K. He, H.-T. Ma, and J.-F. Ge, "Clinical characteristics and prognosis of ground-glass opacity nodules in young patients," *Journal of Thoracic Disease*, vol. 11, no. 2, p. 557, 2019.

[97] A. A. A. Setio, A. Traverso, T. De Bel, M. S. Berens, C. Van Den Bogaard, P. Cerello, H. Chen, Q. Dou, M. E. Fantacci, B. Geurts *et al.*, "Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the LUNA16 challenge," *Medical image analysis*, vol. 42, pp. 1–13, 2017.

[98] H. Shaziya, K. Shyamala, and R. Zaheer, "Automatic lung segmentation on thoracic CT scans using U-Net convolutional network," in *2018 International conference on communication and signal processing (ICCSP)*. IEEE, 2018, pp. 0643–0647.

[99] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015.

[100] J. Ding, A. Li, Z. Hu, and L. Wang, “Accurate pulmonary nodule detection in computed tomography images using deep convolutional neural networks,” in *Medical Image Computing and Computer Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part III 20.* Springer, 2017, pp. 559–567.

[101] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, “Imagenet large scale visual recognition challenge,” *International journal of computer vision*, vol. 115, pp. 211–252, 2015.

[102] M. Sudhamani *et al.*, “Techniques for detection of solitary pulmonary nodules in human lung and their classifications-a survey,” *International Journal on Cybernetics & Informatics (IJCI)*, vol. 4, no. 1, p. 27, 2015.

[103] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, “Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries,” *CA: a cancer journal for clinicians*, vol. 71, no. 3, pp. 209–249, 2021.

[104] Z. Zhou, F. Gou, Y. Tan, and J. Wu, “A cascaded multi-stage framework for automatic detection and segmentation of pulmonary nodules in developing countries,” *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 11, pp. 5619–5630, 2022.

[105] H. Huang, R. Wu, Y. Li, and C. Peng, “Self-supervised transfer learning based on domain adaptation for benign-malignant lung nodule classification on thoracic ct,” *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 8, pp. 3860–3871, 2022.

[106] H. Mkindu, L. Wu, and Y. Zhao, “Lung nodule detection of ct images based on combining 3d-cnn and squeeze-and-excitation networks,” *Multimedia Tools and Applications*, vol. 82, no. 17, pp. 25 747–25 760, 2023.

[107] S. A. Agnes, J. Anitha, and A. A. Solomon, “Two-stage lung nodule detection framework using enhanced unet and convolutional lstm networks in ct images,” *Computers in Biology and Medicine*, vol. 149, p. 106059, 2022.

[108] L. Zhu, H. Zhu, S. Yang, P. Wang, and H. Huang, “Pulmonary nodule detection based on hierarchical-split hrnet and feature pyramid network with atrous convolution,” *Biomedical Signal Processing and Control*, vol. 85, p. 105024, 2023.

[109] D. Zhao, Y. Liu, H. Yin, and Z. Wang, "An attentive and adaptive 3d cnn for automatic pulmonary nodule detection in ct image," *Expert Systems with Applications*, vol. 211, p. 118672, 2023.

[110] 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 *2015 IEEE 12th International symposium on biomedical imaging (ISBI)*. IEEE, 2015, pp. 286–289.

[111] H. Xie, D. Yang, N. Sun, Z. Chen, and Y. Zhang, "Automated pulmonary nodule detection in ct images using deep convolutional neural networks," *Pattern Recognition*, vol. 85, pp. 109–119, 2019.

[112] A. Pezeshk, S. Hamidian, N. Petrick, and B. Sahiner, "3-d convolutional neural networks for automatic detection of pulmonary nodules in chest ct," *IEEE journal of biomedical and health informatics*, vol. 23, no. 5, pp. 2080–2090, 2018.

[113] Y. Li and Y. Fan, "Deepseed: 3d squeeze-and-excitation encoder-decoder convolutional neural networks for pulmonary nodule detection," in *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2020, pp. 1866–1869.

[114] J. Mei, M.-M. Cheng, G. Xu, L.-R. Wan, and H. Zhang, "Sanet: A slice-aware network for pulmonary nodule detection," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 8, pp. 4374–4387, 2021.

[115] Y.-S. Huang, P.-R. Chou, H.-M. Chen, Y.-C. Chang, and R.-F. Chang, "One-stage pulmonary nodule detection using 3-d dcnn with feature fusion and attention mechanism in ct image," *Computer Methods and Programs in Biomedicine*, vol. 220, p. 106786, 2022.

[116] W. Zhang, X. Wang, X. Li, and J. Chen, "3d skeletonization feature based computer-aided detection system for pulmonary nodules in ct datasets," *Computers in biology and medicine*, vol. 92, pp. 64–72, 2018.

[117] J. Xu, H. Ren, S. Cai, and X. Zhang, "An improved faster r-cnn algorithm for assisted detection of lung nodules," *Computers In Biology And Medicine*, vol. 153, p. 106470, 2023.

[118] Z. Cai and N. Vasconcelos, "Cascade r-cnn: Delving into high quality object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6154–6162.

[119] X. Han, J. Chang, and K. Wang, "You only look once: unified, real-time object detection," *Procedia Computer Science*, vol. 183, no. 1, pp. 61–72, 2021.

[120] X. Wu, H. Zhang, J. Sun, S. Wang, and Y. Zhang, "Yolo-msrf for lung nodule detection," *Biomedical Signal Processing and Control*, vol. 94, p. 106318, 2024.

[121] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*. Springer, 2016, pp. 21–37.

[122] Z. UrRehman, Y. Qiang, L. Wang, Y. Shi, Q. Yang, S. U. Khattak, R. Aftab, and J. Zhao, "Effective lung nodule detection using deep cnn with dual attention mechanisms," *Scientific Reports*, vol. 14, no. 1, p. 3934, 2024.

[123] Y. Zamanidoost, N. Alami-Chentoufi, T. Ould-Bachir, and S. Martel, "Efficient region proposal extraction of small lung nodules using enhanced vgg16 network model," in *2023 IEEE 36th International Symposium on Computer-Based Medical Systems (CBMS)*. IEEE, 2023, pp. 483–488.

[124] D. Chen, X. Li, and S. Li, "A novel convolutional neural network model based on beetle antennae search optimization algorithm for computerized tomography diagnosis," *IEEE transactions on neural networks and learning systems*, 2021.

[125] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 658–666.

[126] L. Zahedi, F. G. Mohammadi, S. Rezapour, M. W. Ohland, and M. H. Amini, "Search algorithms for automated hyper-parameter tuning," *arXiv preprint arXiv:2104.14677*, 2021.

[127] Z. W. Geem and K.-B. Sim, "Parameter-setting-free harmony search algorithm," *Applied Mathematics and Computation*, vol. 217, no. 8, pp. 3881–3889, 2010.

[128] S. Yuan, J. Liu, S. Wang, T. Wang, and P. Shi, "Seismic waveform classification and first-break picking using convolution neural networks," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 2, pp. 272–276, 2018.

[129] Y. Li, W. Xie, and H. Li, "Hyperspectral image reconstruction by deep convolutional neural network for classification," *Pattern Recognition*, vol. 63, pp. 371–383, 2017.

[130] Y. Dai, D. Liu, and F. Wu, "A convolutional neural network approach for post-processing in hevc intra coding," in *MultiMedia Modeling: 23rd International Conference, MMM 2017, Reykjavik, Iceland, January 4-6, 2017, Proceedings, Part I 23.* Springer, 2017, pp. 28–39.

[131] Z. Hu, I. Tereykovskiy, Y. Zorin, L. Tereykovska, and A. Zhibek, "Optimization of convolutional neural network structure for biometric authentication by face geometry," in *Advances in Computer Science for Engineering and Education 13.* Springer, 2019, pp. 567–577.

[132] Q. Dou, H. Chen, Y. Jin, H. Lin, J. Qin, and P.-A. Heng, "Automated pulmonary nodule detection via 3d convnets with online sample filtering and hybrid-loss residual learning," in *Medical Image Computing and Computer Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part III 20.* Springer, 2017, pp. 630–638.

[133] Y. Gu, X. Lu, L. Yang, B. Zhang, D. Yu, Y. Zhao, L. Gao, L. Wu, and T. Zhou, "Automatic lung nodule detection using a 3d deep convolutional neural network combined with a multi-scale prediction strategy in chest CTs," *Computers in biology and medicine*, vol. 103, pp. 220–231, 2018.

[134] A. Dobrenkii, R. Kuleev, A. Khan, A. R. Rivera, and A. M. Khattak, "Large residual multiple view 3d cnn for false positive reduction in pulmonary nodule detection," in *2017 IEEE conference on computational intelligence in bioinformatics and computational biology (CIBCB).* IEEE, 2017, pp. 1–6.

[135] M. Almahasneh, X. Xie, and A. Paiement, "Attentnet: Fully convolutional 3d attention for lung nodule detection," *arXiv preprint arXiv:2407.14464*, 2024.

[136] H. Tang, C. Zhang, and X. Xie, "Nodulenet: Decoupled false positive reduction for pulmonary nodule detection and segmentation," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part VI 22.* Springer, 2019, pp. 266–274.

[137] A. K. Balyan, S. Ahuja, U. K. Lilhore, S. K. Sharma, P. Manoharan, A. D. Algarni, H. Elmannai, and K. Raahemifar, "A hybrid intrusion detection model using ega-pso and improved random forest method," *Sensors*, vol. 22, no. 16, p. 5986, 2022.

[138] K. Barbouchi, D. El Hamdi, I. Elouedi, T. B. Aicha, A. K. Echi, and I. Slim, "A transformer-based deep neural network for detection and classification of lung cancer via PET/CT images," *International Journal of Imaging Systems and Technology*, vol. 33, no. 4, pp. 1383–1395, 2023.

[139] S. Bharati, M. R. H. Mondal, and P. Podder, "A review on explainable artificial intelligence for healthcare: Why, how, and when?" *IEEE Transactions on Artificial Intelligence*, vol. 5, no. 4, pp. 1429–1442, 2023.

[140] P. Dhiman, V. Kukreja, P. Manoharan, A. Kaur, M. Kamruzzaman, I. B. Dhaou, and C. Iwendi, "A novel deep learning model for detection of severity level of the disease in citrus fruits," *Electronics*, vol. 11, no. 3, p. 495, 2022.

[141] D. Dai, Y. Sun, C. Dong, Q. Yan, Z. Li, and S. Xu, "Effectively fusing clinical knowledge and AI knowledge for reliable lung nodule diagnosis," *Expert Systems with Applications*, vol. 230, p. 120634, 2023.

[142] C. Gao, L. Wu, W. Wu, Y. Huang, X. Wang, Z. Sun, M. Xu, and C. Gao, "Deep learning in pulmonary nodule detection and segmentation: a systematic review," *European radiology*, vol. 35, no. 1, pp. 255–266, 2025.

[143] D. Srivastava, S. K. Srivastava, S. B. Khan, H. R. Singh, S. K. Maakar, A. K. Agarwal, A. A. Malibari, and E. Albalawi, "Early detection of lung nodules using a revolutionized deep learning model," *Diagnostics*, vol. 13, no. 22, p. 3485, 2023.

[144] R. Sun, Y. Pang, and W. Li, "Efficient lung cancer image classification and segmentation algorithm based on an improved swin transformer," *Electronics*, vol. 12, no. 4, p. 1024, 2023.

[145] P. Chlap, H. Min, N. Vandenberg, J. Dowling, L. Holloway, and A. Haworth, "A review of medical image data augmentation techniques for deep learning applications," *Journal of medical imaging and radiation oncology*, vol. 65, no. 5, pp. 545–563, 2021.

[146] S. L. Tan, G. Selvachandran, R. Paramesran, and W. Ding, "Lung cancer detection systems applied to medical images: a state-of-the-art survey," *Archives of Computational Methods in Engineering*, vol. 32, no. 1, pp. 343–380, 2025.

[147] L. J. Crasta, R. Neema, and A. R. Pais, "A novel deep learning architecture for lung cancer detection and diagnosis from computed tomography image analysis," *Healthcare Analytics*, vol. 5, p. 100316, 2024.

[148] X. Zhu, L. Zhu, D. Song, D. Wang, F. Wu, and J. Wu, "Comparison of single-and dual-energy ct combined with artificial intelligence for the diagnosis of pulmonary nodules," *Clinical Radiology*, vol. 78, no. 2, pp. e99–e105, 2023.

[149] R. Alexander, S. Waite, M. A. Bruno, E. A. Krupinski, L. Berlin, S. Macknik, and S. Martinez-Conde, "Mandating limits on workload, duty, and speed in radiology," *Radiology*, vol. 304, no. 2, pp. 274–282, 2022.

[150] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical image analysis*, vol. 42, pp. 60–88, 2017.

[151] A. Holzinger, C. Biemann, C. S. Pattichis, and D. B. Kell, "What do we need to build explainable ai systems for the medical domain?" *arXiv preprint arXiv:1712.09923*, 2017.

[152] J. Amann, A. Blasimme, E. Vayena, D. Frey, V. I. Madai, and P. Consortium, "Explainability for artificial intelligence in healthcare: a multidisciplinary perspective," *BMC medical informatics and decision making*, vol. 20, pp. 1–9, 2020.

[153] G. Montavon, W. Samek, and K.-R. Müller, "Methods for interpreting and understanding deep neural networks," *Digital signal processing*, vol. 73, pp. 1–15, 2018.

[154] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2921–2929.

[155] Y. Yang, X. Li, J. Fu, Z. Han, and B. Gao, "3D multi-view squeeze-and-excitation convolutional neural network for lung nodule classification," *Medical Physics*, vol. 50, no. 3, pp. 1905–1916, 2023.

[156] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[157] S. Takahashi, Y. Sakaguchi, N. Kouno, K. Takasawa, K. Ishizu, Y. Akagi, R. Aoyama, N. Teraya, A. Bolatkan, N. Shinkai *et al.*, "Comparison of vision transformers and convolutional neural networks in medical image analysis: a systematic review," *Journal of Medical Systems*, vol. 48, no. 1, p. 84, 2024.

[158] D. Gu, Y. Li, F. Jiang, Z. Wen, S. Liu, W. Shi, G. Lu, and C. Zhou, “Vinet: A visually interpretable image diagnosis network,” *IEEE Transactions on Multimedia*, vol. 22, no. 7, pp. 1720–1729, 2020.

[159] Y. Lei, Y. Tian, H. Shan, J. Zhang, G. Wang, and M. K. Kalra, “Shape and margin-aware lung nodule classification in low-dose ct images via soft activation mapping,” *Medical Image Analysis*, vol. 60, p. 101628, 2020.

[160] M. C. Hancock and J. F. Magnan, “Lung nodule malignancy classification using only radiologist-quantified image features as inputs to statistical learning algorithms: probing the lung image database consortium dataset with two statistical learning methods,” *Journal of Medical Imaging*, vol. 3, no. 4, pp. 044 504–044 504, 2016.

[161] H. Rikhari, E. Baidya Kayal, S. Ganguly, A. Sasi, S. Sharma, A. Antony, K. Rangarajan, S. Bakhshi, D. Kandasamy, and A. Mehndiratta, “Improving lung nodule segmentation in thoracic ct scans through the ensemble of 3D U-Net models,” *International journal of computer assisted radiology and surgery*, vol. 19, no. 10, pp. 2089–2099, 2024.

[162] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, “nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation,” *Nature methods*, vol. 18, no. 2, pp. 203–211, 2021.

[163] L. Sun, W. Shao, D. Zhang, and M. Liu, “Anatomical attention guided deep networks for roi segmentation of brain MR images,” *IEEE transactions on medical imaging*, vol. 39, no. 6, pp. 2000–2012, 2019.

[164] S. Stassin, V. Corduant, S. A. Mahmoudi, and X. Siebert, “Explainability and evaluation of vision transformers: An in-depth experimental study,” *Electronics*, vol. 13, no. 1, p. 175, 2023.

[165] 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.

[166] Q. Song, L. Zhao, X. Luo, and X. Dou, “Using deep learning for classification of lung nodules on computed tomography images,” *Journal of healthcare engineering*, vol. 2017, no. 1, p. 8314740, 2017.

[167] Y. Xie, Y. Xia, J. Zhang, Y. Song, D. Feng, M. Fulham, and W. Cai, “Knowledge-based collaborative deep learning for benign-malignant lung nodule classification on chest ct,” *IEEE transactions on medical imaging*, vol. 38, no. 4, pp. 991–1004, 2018.

[168] X. Xu, C. Wang, J. Guo, Y. Gan, J. Wang, H. Bai, L. Zhang, W. Li, and Z. Yi, "MSCS-DeepLN: Evaluating lung nodule malignancy using multi-scale cost-sensitive neural networks," *Medical Image Analysis*, vol. 65, p. 101772, 2020.

[169] X. Zhao, J. Xu, Z. Lin, and X. Xue, "Bicformer: Swin transformer based model for classification of benign and malignant pulmonary nodules," *Measurement Science and Technology*, vol. 35, no. 7, p. 075402, 2024.