MULTI-PLANE DEEP LEARNING AND DOMAIN-ADAPTED RADIOMICS FOR CT-BASED PNEUMONIA DIFFERENTIATION: FROM ALGORITHM DEVELOPMENT TO MULTI-CENTER GENERALIZATION

SONG LIMING

PhD

The Hong Kong Polytechnic University

2025

The Hong Kong Polytechnic University

Department of Health Technology and Informatics

Multi-Plane Deep Learning and Domain-Adapted Radiomics for CT-Based Pneumonia Differentiation: From Algorithm Development to Multi-Center Generalization

SONG Liming

A thesis submitted in partial fulfillment of the requirements

for the degree of Doctor of Philosophy

June 2025

CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgment has been made in the text.

SONG Liming

ABSTRACT

Background: Pneumonia is a global health challenge requiring timely and accurate diagnosis to guide effective treatment. Pathological diagnosis is invasive and slow, while computed tomography (CT)-based differentiation of bacterial pneumonia (BP) from non-bacterial pneumonia (NBP), or tuberculosis (TB) from fungal pneumonia (FP), is challenged by overlapping features and limited biomarker specificity, often leading to empirical treatment. While artificial intelligence (AI) is promising, existing models for pneumonia often focus on simpler tasks, lacking robustness across diverse clinical settings or interpretability for broader adoption.

Purpose: This thesis addresses these challenges by developing and validating two novel AI models for pneumonia differentiation on chest CT: a deep learning model for BP versus NBP classification and a domain-adapted radiomics model for TB versus FP distinction.

Methods and Materials: For BP vs. NBP differentiation, the MPMT-Pneumo model, a hybrid convolutional neural network (CNN)-Transformer, was developed using CT data and four inflammatory biomarkers from 384 patients across two hospitals; Poly focal loss addressed class imbalance. For TB vs. FP differentiation, a radiomics model with a novel multicenter distribution adaptation (MDA) framework was developed using CT data from 528 patients across four centers, involving automated segmentation and feature extraction from seven ROIs. Performance was evaluated by area under the curve (AUC), accuracy, and sensitivity, benchmarked against alternatives, with SHAP values for TB/FP model interpretability.

Results: MPMT-Pneumo achieved AUC 0.874, accuracy 0.852, and sensitivity 0.894 for BP vs. NBP, outperforming baseline models and matching experienced radiologists' sensitivity for BP. Ablation studies confirmed MPMT-Pneumo's multi-modal design benefits. The MDA-radiomics model for TB vs. FP achieved AUCs of 0.871-0.977 in external validations, outperforming traditional classifiers. SHAP analysis offered feature insights, and t-SNE confirmed MDA’s efficacy in reducing inter-center feature variability.

Conclusion: This thesis developed two clinically relevant, generalizable AI solutions for CT-based pneumonia differentiation. Both the MPMT-Pneumo model and the MDA radiomics framework show significant gains in diagnostic accuracy and robustness, promising more rational antibiotic use and informed clinical decisions in diverse healthcare settings.

PUBLICATIONS DURING PH.D. STUDY

Journal Articles:

1. **Song L***, Zhan Y*, Li L, Li X, Wu Y, Zhao M, Li Z, Ren G, Cai J. Differentiating Bacterial and Non-Bacterial Pneumonia on Chest CT Using Multi-Plane Features and Clinical Biomarkers. *Acad Radiol*, 2025 Jun 9. doi: 10.1016/j.acra.2025.05.047. *(IF: 3.8)*
2. **Song L***, Li L*, Li X*, Zhang Y, Wu Y, Zhao M, Li Z, Zhan Y, Ren G, Cai J. Feature-Level Domain Adaptation in CT Radiomics for Differentiating Tuberculous and Fungal Pneumonia: A Multicenter Study. *Insights Imaging*, *(IF: 4.1)* (**Under review**)
3. **Song L**, Sun H, Xiao H, Lam S. K, Zhan Y, Ren G, Cai J. Artificial intelligence for chest X-ray image enhancement. *Radiation Medicine and Protection*, 2025 Feb 28; 6(1), 61–68. doi: 10.1016/j.radmp.2024.12.0003. *(IF: N/A)*
4. Li L*, **Song L***, Liu Y, Ayoub M, Song Y, Shu Y, Liu X, Deng Y, Liu Y, Xia Y, Li H, Peng D. Combining static and dynamic functional connectivity analyses to identify male patients with obstructive sleep apnea and predict clinical symptoms. *Sleep Med*. 2025 Feb;126:136-147. doi: 10.1016/j.sleep.2024.12.013. *(IF: 3.8)*
5. Chen Z, Li Z, Huang YH, Teng X, Zhang J, Xiong T, Dong Y, **Song L,** Ren G, Cai J. Anatomy-wise lung ventilation imaging for precise functional lung avoidance radiation therapy. *Phys Med Biol*. 2025 Feb 12;70(4). doi: 10.1088/1361-6560/adb123. *(IF: 3.3)*
6. Zhao M, **Song L**, Zhu J, Zhou T, Zhang Y, Chen SC, Li H, Cao D, Jiang YQ, Ho W, Cai J, Ren G. Non-contrasted computed tomography (NCCT) based chronic thromboembolic pulmonary hypertension (CTEPH) automatic diagnosis using cascaded network with multiple instance learning. *Phys Med Biol*. 2024 Sep 13;69(18). doi: 10.1088/1361-6560/ad7455. *(IF:3.3)*

7. Liu X, Yang R, Xiong T, Yang X, Li W, **Song L**, Zhu J, Wang M, Cai J, Geng L. CBCT-to-CT Synthesis for Cervical Cancer Adaptive Radiotherapy via U-Net-Based Model Hierarchically Trained with Hybrid Dataset. *Cancers (Basel)*. 2023 Nov 20;15(22):5479. doi: 10.3390/cancers15225479. *(IF: 4.5)*
8. Sun H, Ren G, Teng X, **Song L**, Li K, Yang J, Hu X, Zhan Y, Wan SBN, Wong MFE, Chan KK, Tsang HCH, Xu L, Wu TC, Kong FS, Wang YXJ, Qin J, Chan WCL, Ying M, Cai J. Artificial intelligence-assisted multistrategy image enhancement of chest X-rays for COVID-19 classification. *Quant Imaging Med Surg*. 2023 Jan 1;13(1):394-416. doi: 10.21037/qims-22-610. *(IF:2.9)*
9. Lam NFD, Sun H, **Song L**, Yang D, Zhi S, Ren G, Chou PH, Wan SBN, Wong MFE, Chan KK, Tsang HCH, Kong FS, Wáng YXJ, Qin J, Chan LWC, Ying M, Cai J. Development and validation of bone-suppressed deep learning classification of COVID-19 presentation in chest radiographs. *Quant Imaging Med Surg*. 2022 Jul;12(7):3917-3931. 10.21037/qims-21-791 *(IF:2.9)*

Conference Abstracts:

1. **Song L**, Ren G, Wu Y, Zhan Y, Li L, Zhao M, Li X, Cai J. Mpdt: A Multi-View Pneumonia Diagnostic Transformer for Differentiating between Bacterial and Non-Bacterial Pneumonia in CT Images. In American Association of Physicists in Medicine (AAPM) Annual Meeting; 21-25 July 2024, Los Angeles, USA. (**Oral Presentation**)
2. Ni Y, **Song L**, Lee V, Chiang C, Cheung L, Teng X, Zhang J, Ren G, Cai J, Li T. Functional Dose-Volume Analysis Based on a Novel Image Biomarker Derived from DCE-MRI for Predicting Post-SBRT Liver Function Preservation in Hepatocellular Carcinoma Patients. In American Association of Physicists in Medicine (AAPM) Annual Meeting; 21-25 July 2024, Los Angeles, USA. (**Oral Presentation**)

3. Chen Z, Sun H, **Song L**, Ren G, Yang J, Cai J. Chest X-Ray Enhancement Based On Bone-Suppression Model Using Multi-Dilated-Rate Strategy. In AAPM Annual Meeting; 10-14 July 2022, Washington D.C., USA. (**Poster**)

ACKNOWLEDGEMENTS

I would like to express my heartfelt gratitude to my chief supervisor, Prof. CAI Jing, and my co-supervisor, Dr. REN Ge, for their invaluable guidance, encouragement, and support throughout my Ph.D. journey. Prof. CAI not only provided me with the opportunity to pursue doctoral studies but also offered patient mentorship and genuine care for my well-being in many aspects of daily life.

I am truly grateful to my master's supervisors, Prof. DONG Guoya, Prof. XIE Yaoqin, and Prof. QIN Wenjian, for introducing me to the field of medical image processing and encouraging my academic development.

I am truly grateful to Dr. LI Wen, who has been both a senior and a mentor to me during my master's and Ph.D. studies. His guidance, support, and friendship have meant a great deal to me, and I wish him every success in his future endeavors.

I am truly grateful to Dr. ZHAN Yuefu, Dr. ZHANG Yuanpeng and Dr. LI Tian for their support in my research projects.

I am truly grateful to my close friends, Mr. LI Kang, Mr. CAO Jin, Mr. WANG Weixu, and Mr. LI Zhichun, for their enduring friendship, encouragement, and for making this journey brighter and more enjoyable. I wish them all the best as they approach the completion of their Ph.D. studies.

I am truly grateful to my roommates, Dr. XIAO Haonan, Dr. LIU Chenyang, Mr. LIU Xi, and Mr. HUANG Chao, for their companionship during the pandemic and beyond, which made those times warmer and truly memorable.

I would also like to express my sincere gratitude to all the staff members in the Department of Health Technology and Informatics and The Hong Kong Polytechnic University for providing a supportive academic environment.

Finally, my deepest gratitude goes to my parents and my wife, Mrs. SUI Xinyu, for their unconditional love, understanding, and support throughout these years. Without them, this journey would not have been possible.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Abbreviations

AI	Artificial intelligence
AMR	Antimicrobial resistance
ANC	Absolute neutrophil count
AUC	Area under curve
BP	Bacterial pneumonia
CAP	Community-acquired pneumonia
CNN	Convolutional neural network
COVID	Coronavirus disease
CRP	C-reactive protein
CT	Computed tomography
DA	Domain adaptation
DL	Deep learning
DT	Decision tree
FP	Fungal pneumonia
FS	Feature selection
GAN	Generative adversarial networks
GB	Gradient boosting
GBD	Global Burden of Disease
GGO	Ground-glass opacity
GPU	Graphics Processing Unit

LLL	Left lower lobe
LR	Logistic regression
LUL	Left upper lobe
MCC	Matthews correlation coefficient
MDA	Multicenter distribution adaptation
ML	Machine learning
MMD	Maximum mean discrepancy
MRI	Magnetic resonance imaging
MPMT	Multi-plane and multi-modal transformer
NBP	Non-bacterial pneumonia
NP	No pneumonia
NTIL	Non-tuberculous infectious lesions
PCT	Procalcitonin
PCR	Polymerase chain reaction
PET	Positron emission tomography
PFL	Poly focal loss
RF	Random forests
RLL	Right lower lobe
RML	Right middle lobe
ROC	Receiver operating characteristic
ROI	Region of interest
RUL	Right upper lobe

SU	Supervised and unsupervised
SVM	Support vector machine
TB	Tuberculosis
ViT	Vision transformer
VP	Viral pneumonia
WBC	White blood cell

CHAPTER 1 Introduction

1.1 Pneumonia

1.1.1 Background

Pneumonia stands as a formidable global health concern, responsible for substantial morbidity and mortality across all age groups. According to the 2021 Global Burden of Disease (GBD) report [1], 2.1 million people died of pneumonia, including over 700,000 children under five and more than 1 million adults over 70 [2]. This positions pneumonia as one of a leading cause of death globally [3].

The challenge of pneumonia management is compounded by its diverse etiology. Bacteria, viruses, fungi, and atypical pathogens such as Mycoplasma pneumoniae and Legionella pneumophila can all cause pneumonia [4]. Accurate and timely etiological differentiation is therefore critical, as it determines the appropriate therapeutic approach [5]. For instance, bacterial pneumonia (BP) necessitates treatment with antibiotics, whereas viral or fungal pneumonia (FP) require specific antiviral or antifungal agents, respectively [6].

Inaccurate or delayed etiological diagnosis can lead to suboptimal or unnecessary treatment, which not only fails to address the underlying infection but also exposes patients to adverse drug reactions [7], increases the risk of serious complications (such as lung abscess, pleural effusion, or respiratory failure) [8], and, in severe cases, may result in death [9]. Furthermore, the empirical use of broad-spectrum antibiotics in the absence of a clear diagnosis drives the selection and spread of drug-resistant pathogens, contributing significantly to the global burden of antimicrobial resistance (AMR) [10]. According to the 2019 GBD report, 1.27 million deaths were attributable to bacterial AMR, with Streptococcus pneumoniae as a major contributor [11].

Therefore, improving diagnostic precision is essential for both optimal patient outcomes and the broader effort to combat AMR.

Furthermore, precise diagnosis optimizes the utilization of healthcare resources. By ensuring that patients receive the correct treatment promptly, accurate etiological differentiation can reduce the length of hospital stays, minimize the need for additional diagnostic tests or interventions, and decrease overall healthcare costs. This optimization of resources is particularly critical in low- and middle-income countries, which often bear the highest burden of pneumonia and operate under significant resource constraints. Delayed or inaccurate etiological diagnosis leads to inappropriate or delayed treatment, which in turn can result in poorer patient outcomes, an increased burden of AMR, and inefficient use of healthcare resources [12].

1.1.2 Challenges of Accurate Diagnosis

The diagnosis of pneumonia and the subsequent differentiation of its etiological agents have traditionally relied on a combination of clinical assessment, radiological imaging, and various laboratory tests [13]. Conventional diagnostic methods for identifying the causative pathogens include sputum and blood cultures, direct microscopy of respiratory specimens, polymerase chain reaction (PCR) assays, and antigen or antibody detection tests. While these methods can provide valuable information, they are beset by several inherent limitations that can impede timely and accurate etiological differentiation [14-16], as summarized in **Table 1.1**.

Table 1.1. Limitations of conventional pneumonia diagnostic methods.

Method	**Time**	**Sensitivity**	**Specificity**	**Limitations**
Sputum Culture	48-72 hours	Low to moderate	Moderate	Time-consuming; low yield for some pathogens; affected by sample quality; cannot distinguish infection from colonization
Blood Culture	24-48 hours	Very low	High when positive	Low yield, especially in non-severe cases; invasive procedure
Microscopy	Hours	Variable	Moderate to high	Operator-dependent; low sensitivity with low pathogen loads; some pathogens not easily visualized
PCR Tests	1–6 hours	Generally high	Moderate to high	May not distinguish colonization from infection; expensive; requires specialized labs; false positives/negatives possible
Antigen/ Antibody Tests	Minutes to hours	Variable	Variable	Often pathogen-specific; possible cross-reactivity; timing of sample collection is critical

One major limitation is the time required for results. Microbiological cultures, though considered the gold standard for bacterial identification, can take 48 to 72 hours or longer, which is problematic in acute settings where prompt decisions are essential [17]. Sensitivity and specificity are also important concerns. For example, sputum cultures often have low sensitivity, especially if the patient has received antibiotics prior to sample collection [18]. Sputum samples may also be contaminated with oropharyngeal flora, making it difficult to distinguish between true pathogens and colonizing organisms. This challenge extends to PCR-based detection, as the presence of bacterial DNA in nonsterile samples does not always indicate active infection, complicating clinical decision-making [17]. Blood cultures have a very low yield in non-severe pneumonia. Microscopic examination of stained smears can offer rapid results, but its sensitivity is often operator-dependent and can be hampered by low pathogen loads, as seen in non-HIV patients with Pneumocystis pneumonia [19].

Accessibility and cost are additional barriers. Sophisticated molecular assays are often expensive and may require specialized laboratory infrastructure and trained personnel, which are not universally available, particularly in resource-limited settings where pneumonia burden is highest [20]. This often forces clinicians to balance diagnostic accuracy with practical feasibility.

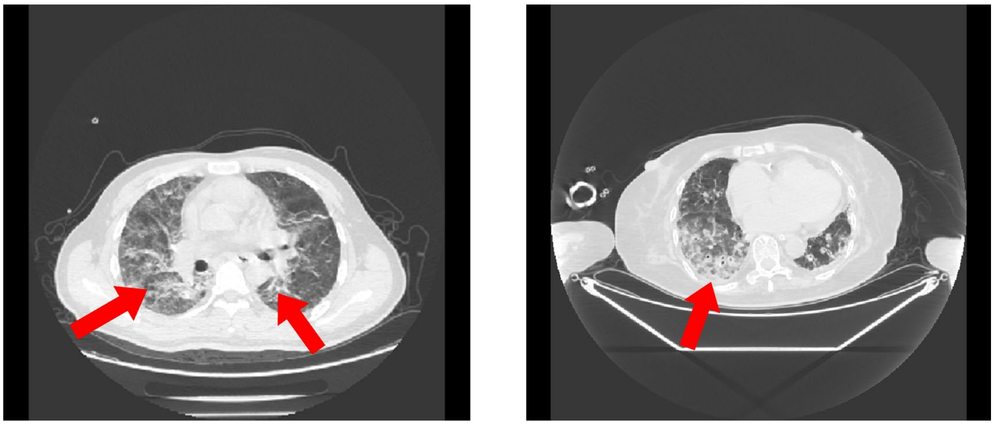
These limitations collectively create a "diagnostic gap." In the absence of rapid, reliable, and accessible diagnostic tools, clinicians often resort to empirical broad-spectrum antibiotics [18]. While this may be necessary to avoid treatment delays, it substantially contributes to antibiotic overuse and the development of AMR. This diagnostic gap underscores the urgent need for novel and efficient diagnostic strategies, such as advanced medical imaging and artificial intelligence (AI), to improve pneumonia management.

1.1.3 CT for Pneumonia Differentiation

Computed tomography (CT) has become a valuable imaging modality in the diagnostic pathway for pneumonia, particularly in complex or equivocal cases [21]. Its main advantage lies in providing detailed, cross-sectional images of the lungs with superior spatial and contrast resolution compared to conventional chest radiography [22]. This enhanced visualization allows for more accurate detection and characterization of pneumonic infiltrates, as well as assessment of disease extent and severity.

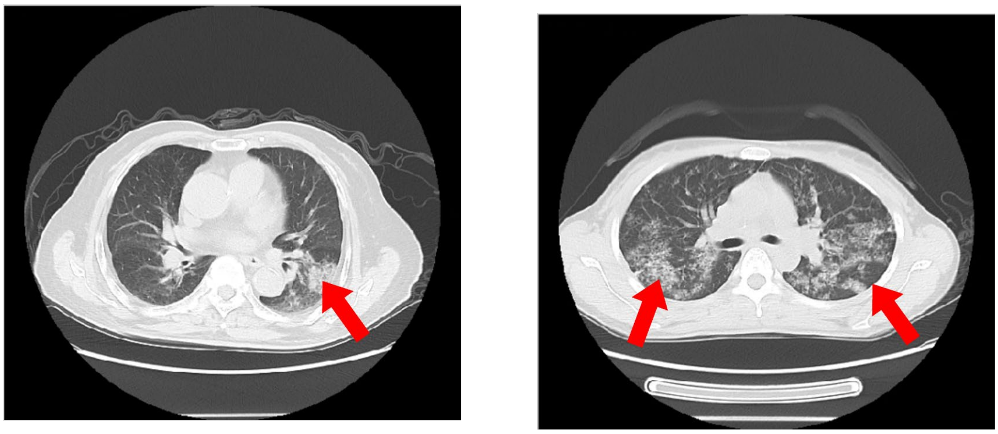
CT can reveal early or subtle pneumonia that may be missed on chest X-ray and is especially effective in detecting complications, such as abscesses, pleural effusions, or cavitation [23]. Studies report that CT sensitivity and specificity for pneumonia differentiation often exceed 95% and 90%, respectively [22]. This makes CT a crucial reference tool, especially when initial imaging is inconclusive or in patients with severe disease. For instance, early chest CT scans in emergency department patients with clinically suspected community-acquired pneumonia (CAP) have been shown to improve diagnostic certainty and influence therapeutic plans, such as decisions regarding antibiotic initiation or hospitalization [24]. Recent advances, including ultra-low-dose CT and AI-driven denoising, have further reduced radiation risks and broadened CT applicability, especially for radiosensitive groups or cases requiring serial imaging [25].

Despite the significant advantages of CT in detecting and characterizing pneumonia, a substantial challenge persists: the accurate differentiation of various pneumonia etiologies based solely on visual interpretation of CT images remains difficult for human observers [26]. The primary reason for this challenge lies in the phenomenon of overlapping imaging features [27]. Different causative pathogens—be they bacterial, viral, or fungal—can manifest with remarkably similar visual patterns on CT scans as shown in **Figure 1.1**. Common CT findings such as ground-glass opacities (GGOs), consolidations, and nodular patterns are not specific to any single etiology and can be observed across a spectrum of infectious processes. For example, distinguishing BP from non-bacterial pneumonia (NBP) (often viral) is notoriously challenging due to this overlap; features like lobar consolidation might suggest bacterial origin [28], while diffuse GGOs might point towards viral, but exceptions and mixed patterns are frequent [29, 30]. Similarly, differentiating tuberculous pneumonia from FP, especially in immunocompromised patients, can be exceedingly difficult, as both can present with nodules, consolidations, and cavitation, leading to diagnostic uncertainty[31, 32]. Studies have shown that even experienced radiologists exhibit poor agreement when trying to distinguish conditions like pulmonary tuberculosis (TB) from invasive pulmonary aspergillosis in transplant recipients based on CT findings alone[31]. Even within viral pneumonias (VP), differentiating specific viral agents based on CT patterns is often unreliable due to the diversity of presentations and similarities between them[33]. The challenge is further compounded in cases like Coronavirus disease (COVID)-19, where atypical CT findings can mimic other lung diseases [34].



(a) Bacterial Pneumonia

(b) Viral Pneumonia



(c) Fungal Pneumonia

(d) Tuberculosis

Figure 1.1. CT for lung imaging in patients with pneumonia.

Representative axial CT images displayed with a lung window setting (Window Width: 1500 HU, Window Level: -600 HU). **(a)** Bacterial Pneumonia. **(b)** Viral Pneumonia. **(c)** Fungal Pneumonia. **(d)** Tuberculosis

Besides, the inherent limitations of human visual interpretation contribute significantly to diagnostic uncertainty [35]. CT interpretation is a subjective process, susceptible to inter-observer variability, meaning that different radiologists, or even the same radiologist at different times, may interpret the same scan differently [23]. This variability can be substantial, even among experienced thoracic radiologists, particularly when assessing complex diffuse lung diseases or subtle patterns [36]. For instance, inter-observer agreement for identifying specific patterns like usual interstitial pneumonia in interstitial lung diseases has been reported as moderate [37], and diagnostic agreement for pneumonia on chest radiographs can also be moderate, with kappa values sometimes falling significantly in specific patient subgroups or for certain pathogens [38]. This subjectivity and variability make it hard to consistently identify the subtle signs that can distinguish different causes of pneumonia. The combination of overlapping imaging features and the limitation of human interpretation results in a diagnostic bottleneck, which means that doctors often have to use slower, sometimes less sensitive or specific, standard microbiological methods, or make treatment decisions based on experience [39]. This clearly defines an area where more objective and powerful analytical tools are needed.

1.2 AI for CT-Based Pneumonia Differentiation

The field of medical image analysis is undergoing a rapid transformation, primarily driven by the emergence of AI and its specialized subfields, machine learning (ML), deep learning (DL), and radiomics. These technologies enable the extraction of vast quantitative information from medical images, supporting more objective, accurate, and efficient diagnostic and prognostic assessments.

1.2.1 Background

AI in medical imaging has evolved from early expert systems, which codified medical knowledge into rule-based frameworks, to advanced ML and DL approaches [40]. Notably, DL models, especially convolutional neural networks (CNNs) [41], have shown outstanding performance in image-based tasks by learning hierarchical visual features directly from raw image data, eliminating the need for manual feature engineering [42]. More recently, vision transformers (ViTs) [43], adapted from natural language processing, have further advanced the field by capturing global context and long-range dependencies within images [44].

Concurrently, the field of radiomics has emerged as a distinct but related discipline. Radiomics focuses on high-throughput extraction of quantitative features, such as intensity, shape, texture, and wavelet characteristics, from CT, magnetic resonance imaging (MRI), or positron emission tomography (PET) images [45]. These features characterize tissue properties and generate a quantitative "imaging phenotype" for regions of interest, which can be correlated with clinical outcomes or used for predictive modeling in diagnosis and prognosis [46]. While traditional radiomics relies on "hand-crafted" features, DL techniques can also be employed within the radiomics workflow, for instance, for automated segmentation of regions of interest (ROIs) or even for learning "deep features" directly from the image data [47].

The rapid proliferation and advancement of these AI technologies in medical imaging have been significantly catalyzed by several factors. These include the increasing availability of large medical image datasets (though challenges in data sharing and annotation persist), substantial improvements in computational power, particularly through the use of Graphics Processing Units (GPUs) and cloud computing platforms, and the development of open-source software libraries and frameworks (e.g., TensorFlow, PyTorch, MONAI) [40]. This convergence of algorithmic innovation, data availability, and computational resources has enabled AI to tackle complex challenges in medical image analysis.

1.2.2 Advancements and Applications

AI-driven approaches, encompassing both DL and radiomics, hold considerable potential to address the inherent limitations in conventional CT interpretation for pneumonia differentiation. As previously discussed, human assessment of CT scans is often limited by overlapping imaging features and inter-observer variability. AI overcomes these challenges by analyzing complex image patterns that are often imperceptible to the human eye [48]. For example, DL models, particularly CNNs, can learn to identify subtle textural and morphological distinctions from vast amounts of image data, potentially discerning discriminative features that differentiate various pneumonia types even when they appear visually similar[49]. Concurrently, radiomics, through the extraction of a multitude of quantitative features, aims to capture an "occult" imaging phenotype, subtle characteristics of tissue heterogeneity and structure that are not typically assessed in routine visual inspection but may correlate with underlying pathology or etiology [47]. This detailed analysis is essential for distinguishing pneumonia etiologies that present similar visual characteristics.

The application of AI can enhance CT-based pneumonia differentiation in several keyways, leading to significant improvements in diagnostic capabilities. One major benefit is improved accuracy; by learning from large datasets and identifying complex patterns, AI models have the potential to achieve higher diagnostic accuracy in differentiating pneumonia etiologies compared to visual interpretation alone [48]. Studies on COVID-19, for example, have demonstrated AI systems capable of diagnosing the disease from CT scans with high accuracy and differentiating it from other common pneumonias [49].

Beyond accuracy, AI contributes to enhanced objectivity and consistency in diagnoses. Once trained, AI algorithms provide deterministic, reproducible results, minimizing the subjectivity and inter-observer variability that can affect traditional interpretation [48]. This leads to more reliable and standardized assessments. Furthermore, the use of AI can lead to increased efficiency in clinical workflows. Automated AI analysis streamlines processes, reduces image reading time, and alleviates radiologists' workload, which is particularly crucial in high-volume or crisis settings [48].

The integration of AI in this field is prompted by the pursuit of improved diagnostic accuracy, enhanced efficiency, and greater consistency, as well as the potential to extract more detailed information from medical images. Effective implementation of these technologies may promote more judicious use of antibiotics, thereby supporting global initiatives to address AMR. Additionally, such advancements can facilitate the optimal allocation of healthcare resources by enabling more precise and timely diagnoses.

CHAPTER 2 Research Gaps, Aims and Objectives

2.1 Research Gaps

While the potential of AI in medical imaging is vast, its application to specific, challenging clinical problems in pneumonia differentiation requires focused research and development to address persistent unmet needs. Current diagnostic pathways, even when incorporating advanced imaging like CT, face significant limitations in accurately and reliably differentiating key pneumonia etiologies. AI methodologies have not yet adequately targeted these specific, nuanced differentiation tasks with robust and generalizable solutions. Furthermore, a broader barrier to the clinical translation of many AI models is their failure to generalize reliably across diverse clinical centers and patient populations. This thesis identifies the following key research gaps:

2.1.1 Lack of Targeted and Effective AI-Driven Solutions for Differentiating Bacterial from Non-Bacterial Pneumonia

BP represents a highly common and clinically significant form of pneumonia encountered in practice, rendering its accurate and timely differentiation from NBP particularly crucial [50]. However, CT imaging features such as consolidation, ground-glass opacities, and nodules are often present in both BP and NBP, making reliable visual differentiation difficult [27]. Similarly, commonly used clinical biomarkers (e.g., C-reactive protein (CRP) and white blood cell (WBC) count) frequently lack sufficient specificity, as elevations can occur in various inflammatory states [51]. While DL has shown promise in general pneumonia differentiation, there is a significant gap in AI models specifically designed and validated for the critical binary task of BP versus NBP differentiation [52, 53]. Existing AI research has often focused on other classifications (e.g., COVID-19) [54, 55] or has not sufficiently explored the synergistic potential of integrating comprehensive multi-plane CT image analysis with relevant clinical biomarkers to achieve the accuracy and robustness needed for this specific differentiation. Consequently, a clear need exists for advanced AI methodologies that can effectively address this frequent and diagnostically ambiguous challenge to better guide antibiotic therapy.

2.1.2 Lack of Robust and Generalizable Non-Invasive Tools for Differentiating Tuberculous from Fungal Pneumonia

Distinguishing TB from FP is a persistent clinical challenge due to their overlapping and complex CT presentations, especially in vulnerable patient groups [56]. Given the distinct, prolonged treatment regimens required and the severe consequences of misdiagnosis, accurate differentiation is crucial. However, conventional methods, including visual CT assessment and microbiological tests, have notable limitations in providing timely and reliable diagnoses for these specific conditions [57]. While radiomics has indicated potential for characterizing infectious lesions [58], there is a distinct lack of AI-driven models, including those based on advanced radiomics or DL, that are specifically developed, validated, and proven to be generalizable across multiple centers for the explicit task of differentiating TB from FP. This gap leaves clinicians without effective non-invasive tools for this challenging differentiation.

2.1.3 Insufficient Generalizability of AI Models for CT-based Pneumonia Differentiation Across Diverse Clinical Centers

A major overarching challenge hindering the widespread clinical adoption of AI tools in pneumonia differentiation, is the issue of "domain shift" [59]. AI models, whether based on DL or radiomics, frequently exhibit a significant degradation in performance when applied to data from different hospitals, CT scanners, or patient populations than those on which they were initially developed and validated [60]. This lack of robustness and generalizability is a critical gap. There is a specific need for the development and rigorous validation of effective strategies, such as domain adaptation (DA) techniques, which can enhance the resilience of AI models to such inter-center variability in the context of CT-based pneumonia differentiation.

2.2 Research Aim

The aim of this study is to develop and rigorously validate novel AI methodologies for CT-based differentiation of key pneumonia types. Specifically, we seek to achieve accurate, robust, and generalizable distinction between BP and NBP, as well as TB and FP, by leveraging multi-plane DL with biomarker integration and domain-adapted radiomics. These AI tools are intended to enhance clinical decision-making, promote rational antimicrobial use, and improve patient outcomes.

2.3 Research Objectives

To achieve the overall aim, this thesis pursues the following specific and measurable objectives, each aligned with the core research questions:

1. To develop and evaluate a multi-plane and multi-modal transformer (MPMT) model that synergistically integrates imaging features extracted from multiple CT views (axial, coronal, and sagittal planes) with routine clinical inflammatory biomarkers for the improved differentiation of BP from NBP.
2. To develop and validate a robust CT-based radiomics pipeline, incorporating a novel feature-level DA framework, termed multicenter distribution adaptation (MDA), designed to enhance the differentiation of TB from FP with improved generalizability across diverse, multi-center clinical datasets.
3. To comprehensively benchmark the diagnostic performance of the developed MPMT-Pneumo model and MDA radiomics framework against relevant alternative computational approaches and/or the interpretative performance of clinical radiologists.

CHAPTER 3 Literature Review of AI-based Pneumonia Differentiation on CT Images

This chapter provides a comprehensive review of the existing literature relevant to the core themes of this PhD thesis: the application of advanced AI techniques, specifically DL, radiomics, and DA, for the etiological differentiation of pneumonia using CT imaging. As established in Chapter 1, differentiating pneumonia etiologies is crucial but challenging with conventional methods and visual CT interpretation alone. This review will examine the current state of AI in addressing these challenges.

The purpose of this literature review is twofold. Firstly, it aims to establish a thorough understanding of the current state of knowledge regarding AI, DL, and radiomics as applied to pulmonary image analysis and pneumonia differentiation. Secondly, and critically, this review seeks to identify the existing gaps, limitations, and unanswered questions within the current body of research. By pinpointing these deficiencies, this chapter will provide the scholarly context and robust justification for the specific research questions, aims, and methodologies adopted in this thesis, thereby underscoring the novelty and significance of the proposed work.

3.1 Deep Learning-based Methods

The emergence of DL as a powerful tool for image analysis has opened new avenues for addressing the previously outlined challenges in CT-based pneumonia differentiation. DL models, by learning intricate patterns directly from image data, offer the potential to overcome the limitations of human visual interpretation, such as subjectivity and difficulty with overlapping features, and the constraints of traditional computer-aided diagnosis systems.

3.1.1 Key Deep Learning Architectures

DL, a subfield of ML, utilizes artificial neural networks with multiple layers to learn hierarchical representations of data [61]. In the context of pulmonary CT imaging, several DL architectures have gained prominence.

CNNs are the cornerstone of DL for image analysis [42]. Their architecture, comprising convolutional layers, pooling layers, and fully connected layers, is inherently suited for processing grid-like data such as images. Convolutional layers apply learnable filters to input images, extracting features like edges, textures, and more complex shapes at various levels of abstraction. Pooling layers then reduce dimensionality while retaining essential information, making the models more robust to variations in object position and scale. CNNs have been successfully applied to a wide range of tasks in pulmonary CT, including lung nodule detection and classification, segmentation of anatomical structures (e.g., lungs, airways, lobes) and pathological findings (e.g., lesions, tumors), and general disease classification [42].

Transformers, initially achieving state-of-the-art results in natural language processing, have more recently been adapted for computer vision tasks, including medical image analysis [62]. ViTs treat image patches as sequences, analogous to words in a sentence, and use self-attention mechanisms to model global relationships between these patches. This allows them to capture long-range dependencies within images, which can be advantageous for understanding complex and diffuse patterns seen in many lung diseases.

Hybrid Models that combine elements of CNNs and transformers, or other neural network architectures like Recurrent Neural Networks [63] for sequential data or Generative Adversarial Networks (GANs) [64] for image synthesis and augmentation, are also being explored. These hybrid approaches aim to leverage the strengths of different architectures; for example, CNNs for robust local feature extraction and transformers for modeling global context.

The broad success of DL in diverse pulmonary CT applications, such as accurately detecting lung nodules or segmenting complex lung structures [65, 66], provides a strong foundation and compelling proof-of-concept for its potential application to the more nuanced and challenging task of pneumonia differentiation. The ability of DL models to learn discriminative features directly from image data, without reliance on the manual feature engineering that characterized traditional ML, enables them to analyze highly complex image patterns that may be subtle or invisible to the human eye. This inherent capability is particularly relevant for distinguishing between pneumonia etiologies that present overlapping visual features on CT scans, as established in Chapter 1. The continuous evolution from CNNs to transformers and hybrid architecture reflects an ongoing effort to develop models that can more effectively capture the multifaceted nature of medical image data, paving the way for more sophisticated diagnostic tools for pneumonia differentiation.

3.1.2 Applications of Deep Learning in Pneumonia Differentiation

Building on the general capabilities of DL in pulmonary CT imaging, numerous studies have specifically focused on applying DL models for the detection and classification of pneumonia. Initial efforts often concentrated on general pneumonia detection, confirming the presence or absence of pneumonic infiltrates on CT scans. However, the field has rapidly advanced towards the more complex task of differentiating between various pneumonia etiologies.

A significant body of research emerged during the COVID-19 pandemic, with many DL models developed to detect COVID-19 pneumonia and distinguish it from non-COVID-19 pneumonia, including BP, or from normal lung CTs [49]. For instance, some studies employed ResNet-based architectures to classify lesion patches from COVID-19 and BP patients, achieving high accuracies (e.g., 91.2% in one study) by identifying distinct lesion patterns such as GGOs in peripheral lungs for COVID-19 versus crazy-paving patterns for BP [67]. Other models, like COVNet, were developed to extract visual features from volumetric chest CTs for COVID-19 detection, demonstrating strong performance in differentiating COVID-19 from CAP and other lung conditions [68].

Beyond COVID-19, a growing body of literature demonstrates the use of DL to differentiate other clinically critical pneumonia types. Some studies have explored the classification of bacterial versus VP directly. For example, the PneumoFusion-Net, a multimodal framework incorporating CT images with clinical and laboratory data, reported very high accuracy (98.96%) in distinguishing BP from VP [69]. More recently, efforts have been directed towards multicategory pneumonia differentiation. One such study developed a DL model using DenseNet architecture to classify CT scans into VP, BP, FP, pulmonary TB, and no pneumonia (NP) categories, achieving promising F1-scores and area under curve (AUC) values on both internal and external test sets [70]. This indicates a trend towards developing more comprehensive DL tools that can address a wider range of differential diagnoses in pneumonia.

The performance of these DL models in pneumonia differentiation tasks has often been shown to meet or even exceed that of human radiologists, those with less experience. Several studies have benchmarked their AI systems against radiologists and found comparable or superior performance, especially when AI assistance was provided to clinicians [52]. This suggests a significant potential for DL to serve as a valuable assistive tool in clinical practice, aiming to improve diagnostic accuracy and consistency. The availability of CT datasets, even if initially from single centers or of modest size, combined with powerful DL algorithms, has demonstrated the fundamental feasibility of developing models for these complex pneumonia differentiation tasks, setting the stage for more refined and robust solutions.

3.1.3 Key Methodology for Pneumonia Differentiation

To further enhance the capabilities of DL models in the complex task of pneumonia CT analysis and differentiation, researchers have been exploring several methodological advancements. Two such advancements particularly relevant to this thesis are multi-plane feature integration and multi-modal DL. multi-plane feature integration and multi-modal DL.

CT scans inherently provide volumetric data of the lungs. While early DL approaches often processed this data as a series of 2D axial slices, there is a growing recognition that incorporating information from multiple anatomical planes (axial, coronal, and sagittal) or utilizing full 3D volumetric data can lead to improved spatial context and diagnostic performance [61]. Pneumonia, especially when presented as diffuse or multifocal disease, has complex 3D morphology. Analyzing features from orthogonal planes can provide complementary information about the distribution, shape, and relationships of lesions that might be missed or be less apparent in a single plane. Some studies have developed 3D CNNs that directly process volumetric CT data, inherently capturing 3D spatial relationships [71]. Others have explored multi-view fusion techniques, where features from 2D slices in axial, coronal, and sagittal views are extracted and then fused at a later stage in the network [72]. This multi-view approach can offer a balance between capturing comprehensive spatial information and managing computational complexity, as full 3D CNNs can be very resource intensive. The rationale is that a more holistic understanding of 3D lesion architecture can lead to better discrimination between pneumonia etiologies that may have subtle differences in their spatial presentation.

Clinical diagnosis rarely relies on imaging in isolation; clinicians typically integrate imaging findings with patient history, symptoms, physical examination, and laboratory test results. Multi-modal DL aims to emulate this comprehensive diagnostic process by integrating imaging data (e.g., CT scans) with other types of data [73]. For pneumonia differentiation, this could involve combining CT image features with clinical information, such as patient demographics (age, sex), symptoms (fever, cough, dyspnea), and medical history (comorbidities, immune status), or laboratory biomarkers. Laboratory biomarkers include inflammatory markers like WBC, absolute neutrophil count (ANC), CRP, and procalcitonin (PCT), which can provide insights into the type and severity of infection [73]. PneumoFusion-Net [69] is an example of such a system, integrating CT images, clinical text, numerical lab results, and radiology reports to classify pneumonia types. ML models, in general, have demonstrated superior performance when they can leverage large amounts of diverse data, such as those available in Electronic Patient Records, which often include biomarkers and clinical notes[73]. The integration of these non-imaging data sources can provide valuable contextual information that helps to disambiguate imaging findings and improve the accuracy and robustness of pneumonia differentiation models.

3.1.4 Limitations

DL models have demonstrated considerable promise in the realm of CT-based pneumonia differentiation, with numerous studies reporting high performance metrics. For instance, models like COVNet achieved a sensitivity of 90% and specificity of 96% for COVID-19 detection against other pneumonias and normal CTs [68]. The multimodal PneumoFusion-Net reported an accuracy of 98.96% for distinguishing BP from VP [69]. A multicategory DL model for classifying viral, bacterial, fungal, tuberculous, and no pneumonia conditions on CT showed weighted F1-averages of 0.883 on an internal test set and 0.846 on an external test set, with AUCs for individual classes often exceeding 0.90 [70]. Another study focusing on lesion patterns to differentiate COVID-19 from BP achieved 91.2% accuracy [67].

These results are often comparable, and sometimes superior, to the performance of human radiologists, especially those with less experience [52].

Despite these encouraging performance figures, a critical analysis of the current DL literature reveals several significant limitations and persistent gaps that need to be addressed for these technologies to translate effectively into routine clinical practice:

A primary cluster of limitations relates to data dependency and its impact on model generalizability. Many studies rely on small and/or single-center datasets [61]; models trained on such limited and potentially homogeneous data may struggle to generalize well to new, unseen data from different patient populations, imaging equipment, or clinical settings, raising concerns about their true real-world performance. Closely intertwined with this is the frequent lack of rigorous external validation on diverse, independent test sets from multiple institutions [71]. High performance on an internal test set (drawn from the same data distribution as the training set) does not guarantee comparable performance on external data, yet robust external validation is a critical step for assessing model robustness and generalizability. Furthermore, the efficacy of supervised DL models is heavily contingent upon the quality and quantity of labeled training data. The process of obtaining large, accurately annotated medical image datasets is labor-intensive and costly, often requiring expert radiological input, thereby posing a significant barrier to developing broadly applicable and validated models.

Beyond data-centric issues, DL models face intrinsic challenges related to their analytical capabilities and practical implementation hurdles. For instance, DL shows promise, accurately differentiating pneumonia types with very subtle or highly overlapping imaging features—such as distinguishing specific viral etiologies from one another or reliably separating certain bacterial from non-bacterial presentations in all cases—remains a significant hurdle. The specific challenge of BP vs. NBP differentiation, a central focus of this thesis, falls squarely into this category. Another major barrier is model interpretability, often referred to as the "black box" problem. Many DL models, especially complex architectures like deep CNNs or transformers, provide predictions without clear insight into their reasoning [61]. This lack of transparency can impede clinical trust and adoption, as clinicians need to understand and verify the basis of AI-generated diagnostic suggestions. Additionally, the computational demands for training very deep or complex models, particularly those utilizing 3D volumetric data or extensive multi-modal inputs, can be substantial, requiring specialized hardware (e.g., GPUs) and significant training time. These resource requirements can limit accessibility for some research groups and present deployment challenges in resource-constrained clinical environments.

Finally, prevailing research practices can affect the objective assessment and comparability of DL approaches. The variability in datasets, model architectures, training protocols, and evaluation metrics used across different studies makes direct comparison of various DL models challenging [74]. The absence of standardized benchmarks and consistent reporting practices further hinders the systematic evaluation of progress in the field.

These limitations highlight that while reported accuracy in specific studies can be high, they should be interpreted with caution, considering the context of the data and validation methods used. The "performance inflation" sometimes observed in AI research underscores this point. The identified gaps directly motivate the methodological choices and focus areas of this thesis. Specifically, the persistent challenge in robustly differentiating BP vs. NBP using CT, which often involves subtle or overlapping features, is a key gap the MPMT-Pneumo model developed in this thesis (Objective 1) aims to address through multi-plane and multi-modal DL. The broader issues of data diversity, robust validation, and generalizability, while critical for all AI models including DL, are also central themes, with generalizability being particularly tackled by Objective 2 (in the context of radiomics) and through comprehensive benchmarking in Objective 3. Addressing these limitations, particularly concerning data diversity, robust validation, generalizability, and interpretability, is crucial for advancing DL in pneumonia CT analysis from a promising research avenue to a clinically impactful and trustworthy diagnostic tool.

To provide a clearer overview of the current landscape, **Table 3.1** summarizes key DL studies focused on CT-based pneumonia differentiation, highlighting their methodologies, performance, and limitations.

Table 3.1. Summary of key deep learning studies for CT-based pneumonia differentiation.

Study	**Task**	**Model(s)**	**Dataset Size(s)**	**Evaluation Metrics**	**Identified Limitations/Gaps**
Li L et al., 2020 (COVNet) [68]	COVID-19 / CAP / NP	ResNet50-based (COVNet)	4,352 CT scans (3,322 patients) from 6 hospitals	COVID-19 vs. others: AUC 0.960; CAP vs. others: AUC 0.950	Potential selection bias, focus primarily on COVID-19.
Zhang K et al., 2020 [48]	COVID-19 / Common Pneumonia / Normal	Segmentation network + 3D Classification network	3,777 CT scans	NCP vs. others: AUC 0.979	Focus on COVID-19, interpretability of complex model.
Han Z et al., 2020 [67]	COVID-19 / BP	ResNet50	73 COVID-19, 97 BP patients (single center)	Acc 0.912	Small single-center dataset.
Fu et al., 2023 [70]	VP, BP, FP, PTB, NP	DenseNet	1,611 CT scans (2 hospitals) for training; 165 cases (3 hospitals) for external test	AUC >0.90	Small samples for BP/FP.
Li Y et al., 2025 (PneumoFusion-Net) [69]	BP vs. VP	ResNet-GCSA, BiLSTM, MLP, Swin transformer	10,095 CT scans	Acc 0.989	Data for non-imaging modalities partly synthetic/literature-based for some conditions.

Abbreviations: VP, viral pneumonia; BP, bacterial pneumonia; FP, fungal pneumonia; PTB, pulmonary tuberculosis; NP, non-pneumonia; CAP, community-acquired pneumonia; NCP, novel coronavirus pneumonia; Acc, accuracy; AUC, area under curve.

3.2 Radiomics-based Methods

In addition to DL, radiomics, the process of extracting and analyzing large amounts of quantitative features from medical image, offers another promising avenue for enhancing CT-based pneumonia differentiation. By converting images into quantitative, high-dimensional data, radiomics aims to uncover subtle patterns and signatures that may not be perceptible to the human eye but are indicative of underlying pathophysiology, potentially offering objective biomarkers for disease characterization.

3.2.1 Overview of the Radiomics Workflow

Radiomics is founded on the premise that medical images contain rich, quantifiable information that reflect tissue characteristics, including heterogeneity, morphology, and texture, which can be correlated with clinical, pathological, or genomic data [60]. The overarching goal is to develop non-invasive imaging biomarkers for improved diagnosis, prognosis, and treatment response prediction. The radiomics process is a meticulous, multi-stage workflow, where each step critically influences the final outcome and requires careful consideration to ensure robust and reproducible results.

The workflow typically commences with Image Acquisition, the initial step of obtaining medical images (e.g., CT, MRI, PET). While standardized acquisition protocols are crucial for ensuring feature robustness and reproducibility, achieving this presents a common challenge in retrospective studies where data often originate from diverse sources with varying scanner parameters and protocols [47]. Following acquisition is the crucial step of Segmenting the ROI. This involves delineating the specific anatomical area or pathological lesion to be analyzed on the image. Segmentation can be performed manually by expert radiologists, semi-automatically with user guidance, or, with increasing frequency, automatically using DL-based segmentation algorithms. The accuracy and consistency of this segmentation are paramount, as the values of extracted features are highly dependent on the precise boundaries of the defined ROI.

Once the ROI is defined, an extensive array of quantitative imaging features is computationally extracted—often hundreds to thousands in number. This Feature Extraction stage yields various categories of features. First-order statistics (or intensity/histogram-based features) describe the distribution of voxel intensities within the ROI (e.g., mean, median, skewness, kurtosis) without considering spatial relationships. Shape-based features quantify the geometric properties of the ROI, such as volume, surface area, sphericity, and compactness. Second-order statistics, commonly known as texture features, characterize the spatial arrangement and interrelationships of voxel intensities, thereby providing information about tissue heterogeneity; common methods for deriving these include the Gray Level Co-occurrence Matrix, Gray Level Run Length Matrix, Gray Level Size Zone Matrix, and Gray Level Dependence Matrix. Finally, higher-order statistics, or filter-based features, involve applying mathematical filters (e.g., wavelet transforms, Laplacian of Gaussian) to the image prior to feature calculation, aiming to capture patterns at different scales or enhance specific image characteristics.

Given the high dimensionality of the extracted feature set, particularly in studies with limited sample sizes, Feature selection (FS) and reduction is a vital subsequent stage. This step aims to identify the most informative, robust, and non-redundant features, which is essential for preventing model overfitting and enhancing the generalizability of the resultant radiomic signature. Common techniques include assessing feature stability (e.g., against image perturbations or across different segmentations), removing highly correlated features, and applying various statistical or ML-based algorithms for optimal subset selection.

The final stage in the core workflow is Model Building and Validation. Here, the selected radiomic features are used to train a predictive or prognostic model, typically employing ML algorithms such as logistic regression (LR), support vector machines (SVMs), random forests (RF), and neural networks. Rigorous validation, encompassing internal cross-validation and, crucially, external validation on independent datasets, is essential to reliably assess model performance and its potential for generalization.

The output of this process is often a "radiomic signature" or an "imaging phenotype" that can provide insights beyond conventional visual image assessment, potentially aiding in more precise and personalized clinical decision-making [47]. The transformation of qualitative visual interpretation into quantitative, objective analysis is a key strength of radiomics, offering a way to standardize image assessment and potentially uncover subtle disease characteristics. However, the multi-step nature of the radiomics workflow means that each stage can introduce variability, underscoring the need for meticulous methodology, standardization efforts, and thorough validation to ensure reliable and reproducible results [75].

3.2.2 Applications for Radiomics in Pneumonia Differentiation

CT-based radiomics has been increasingly applied to a wide spectrum of pulmonary diseases, demonstrating its versatility as a quantitative image analysis tool [76-78]. Much of the initial research focused on oncological applications, such as characterizing lung nodules (e.g., differentiating benign from malignant), estimating tumor malignancy, predicting treatment response in lung cancer, or identifying underlying genetic mutations from imaging features [79, 80].

The application of radiomics to infectious lung diseases, particularly for the etiological differentiation of pneumonia, is a more recent but rapidly growing area of investigation [81]. This development is driven by the need to overcome the limitations of visual CT interpretation, especially where imaging features overlap significantly between different types of infections. Notable applications in pneumonia differentiation include the challenge of distinguishing TB from Non-tuberculous Infectious Lesions (NTIL). Several studies have investigated the utility of CT radiomics for this purpose, with NTIL representing a category that can encompass bacterial and fungal infections presenting as solid pulmonary nodules or masses [82, 83] . These studies often report promising diagnostic performance, with AUC values frequently exceeding 0.8 or 0.9. Some research also suggests that contrast-enhanced CT radiomics may offer superior discriminative ability compared to non-contrast CT radiomics in this context, potentially due to enhancement patterns reflecting lesion vascularity and activity [82]. The ability to non-invasively differentiate TB is crucial given its public health importance and specific treatment requirements.

Another key application area is the differentiation of BP from Mycoplasma pneumoniae Pneumonia. Mycoplasma pneumoniae is an atypical bacterium, and its differentiation from typical BP can significantly influence antibiotic choice. Radiomics based on non-contrast chest CT has shown good performance in distinguishing pleural effusions associated with BP from those associated with Mycoplasma pneumoniae pneumonia in pediatric populations, achieving AUCs greater than 0.9 in some studies [84]. This highlights the potential of radiomics to analyze not just the parenchymal infiltrates but also associated complications like pleural effusions for diagnostic clues.

These examples illustrate that radiomics is often targeted at specific, challenging clinical scenarios within pneumonia differentiation where conventional imaging interpretation is limited. By extracting a rich set of quantitative features, radiomics aims to capture subtle differences in lesion texture, shape, and intensity that are not readily apparent to the human eye, thereby providing a more objective basis for differentiating between etiologies that may otherwise appear similar on CT scans. This quantitative approach holds the promise of improving diagnostic accuracy and supporting more tailored treatment decisions in complex pneumonia cases.

3.2.3 Challenges and Limitations of Radiomics Approaches

Despite its promise of improving scans and distinguishing lung diseases like pneumonia, radiomics still face many linked hurdles that slow their routine clinical use. A major group to data integrity, methodological consistency, and the subsequent impact on model generalizability. A critical concern is featuring robustness, reproducibility, and stability. Radiomic features are known to be highly sensitive to variations in image acquisition parameters and image preprocessing steps. Even minor variations can lead to substantial differences in extracted feature values, making it difficult to ensure that features are stable and reproducible across different imaging sites or time points. This lack of robustness is a primary barrier to developing generalizable radiomic models. Compounding this, segmentation variability in delineating the ROI—whether manual, semi-automated, or automated—can introduce inconsistencies, as feature calculations are confined to this area and are highly dependent on its precise definition. Manual segmentation, while often a reference, is time-consuming and prone to inter-observer and intra-observer variability. The predominance of retrospective and single-center studies in many radiomics research, including pneumonia studies, further limits the generalizability of findings [85]. This is exacerbated by considerable methodological variability across studies in terms of patient cohort selection, CT acquisition parameters, ROI segmentation techniques, feature extraction sets, selection algorithms, and classification models, making direct comparisons and evidence synthesis difficult. Consequently, there is a pressing need for robust external validation on independent, multi-center datasets, and ideally, prospective validation, to truly ascertain clinical utility.

Another set of significant hurdles involves analytical challenges inherent in the radiomics approach and difficulties in clinical interpretation. Radiomics typically extracts a very large number of features, often from small sample sizes. This "curse of dimensionality", a high number of features relative to patient samples, creates a substantial risk of model overfitting, where models learn noise or training data idiosyncrasies rather than true underlying patterns, leading to poor generalization. Effective FS to identify a small, optimal subset of informative and generalizable features is therefore a complex but crucial task. Beyond these analytical issues, the interpretability of radiomic models and the biological correlation of their features remain major concerns. Many radiomic features, especially complex texture features and wavelet features, lack a clear and intuitive biological meaning, making it difficult for clinicians to understand what these features represent in terms of underlying pathophysiology. This "black box" nature can hinder clinical trust and the integration of radiomic outputs into decision-making [86].

The literature review identified key gaps that directly motivate the work in this thesis. These gaps include the need for more multi-center studies with standardized methodologies, a deeper investigation into the biological basis of radiomic features, and, crucially, the development of robust solutions to address domain shift for real-world applicability. The challenge of differentiating TB from FP, for example, benefits significantly from radiomic approaches that can maintain performance across diverse clinical settings. This thesis, particularly through Objective 2, aims to address some of these critical gaps by developing and validating a novel MDA framework for CT-radiomics. This framework is specifically designed to enhance the differentiation of TB from FP pneumonia by improving feature stability and model generalizability across multiple centers.

Table 3.2 provides a summary of key radiomics studies relevant to CT-based pneumonia differentiation, outlining their approaches, performance, and limitations, thereby contextualizing the contributions of this thesis.

Table 3.2. Summary of key radiomics studies for CT-based pneumonia differentiation.

Study	**Task**	**Model (s)**	**Dataset Size(s)**	**Evaluation Metrics**	**Identified Limitations/Gaps**
Li X et al., 2022 [82]	TB / Non-TB	LR	101 lesions (95 patients), single center	AUC 0.833	Small sample, single CT phase.
Yan L et al., 2024 [87]	Non-TB / TB	LR, KNN, SVM	179 patients (NTB: 85, TB: 94), single center	AUC 0.89	Single center, only consolidation analyzed.
Li Z et al., 2025 [84]	BP / Mycoplasma Pneumoniae Parapneumonic Effusion (MPPE)	LR	249 children (BP: 160, MPPE: 89), single center	AUC 0.917	Single center, focused on PE not parenchyma.

Abbreviations: LR, logistic regression; KNN, K-nearest neighbors; SVM, support vector machine; AUC, area under curve.

3.3 Domain Adaptation

A key challenge in the routine clinical application of AI models, including those for DL and radiomics, is ensuring their robust and generalizable performance across diverse healthcare settings. AI Models that perform exceptionally well on data from the hospital or scanner where they were created often do not perform as well when used with data from new, unseen places. This problem, known as domain shift, makes it hard to trust and rely on AI in medical imaging. This section will explore the problem of domain shift and review DA strategies that try to lessen its effects, focusing on how these apply to CT imaging of lung diseases.

3.3.1 The Problem of Domain Shift and an Overview of Adaptation Techniques

Domain shift happens when the data used to train an AI model (the source data) is different in important ways from the data the model encounters later in a new setting (the target data) [60]. In medical imaging, these differences, or shifts, are very common and can happen for many reasons. For example, different hospitals might use different imaging equipment; variations between scanner brands, models, or even software versions can change how images look, affecting things like image noise, sharpness, and contrast [60]. Similarly, acquisition protocols, meaning the specific settings used during a scan like slice thickness, how images are reconstructed, radiation dose for CT, and how contrast dye is given, can also vary and add to data differences. Patient populations themselves can differ too; factors like age, sex, ethnicity, how common or severe a disease is, and other health problems can vary between hospitals or regions, causing shifts in the appearance of medical conditions. Finally, post-processing steps, such as how images are normalized, how areas of interest are segmented, or how features are extracted at different centers, can also create domain shifts.

When an AI model trained on specific source data is then used with target data that has different characteristics, the patterns it learned may no longer be accurate, leading to worse performance. This drop in performance is a serious problem because it makes AI models less reliable when used in various real-world clinics, where they need to work well even with data they haven't seen during training. The effect of scanner differences on AI model performance has been studied, and even though CT scans are generally more standardized than images from some other methods like MRI, domain shift is still a significant concern for CT-based AI models [88]. Therefore, finding ways to handle domain shifts is crucial for creating AI tools that can be widely used, trusted, and effective across many different medical centers. Many promising AI models that work well in research studies fail when faced with diverse 'real-world' data, showing that domain shift is a major roadblock for bringing medical AI into regular clinical use.

To address the problem of domain shift, researchers use techniques called DA. DA includes a set of ML methods that help a model trained with source data to still work well when it sees new, different target data [89]. As a special kind of transfer learning, the model tries to do the same job, but the source data and the new data look different. There are several ways to understand and categorize these DA strategies.

One important way to group DA methods is by looking at how much labeled data is available from the new target setting when the model is being adapted. If there's a small amount of labeled target data, this is called supervised DA. This new labeled data can facilitate the adjustment (or fine-tuning) of a model pre-trained on source data, or aid in learning features common to both the source and target data. Sometimes, there might be only a few labeled examples from the target setting but many unlabeled ones; this approach is known as semi-supervised DA. The extra unlabeled data can help the model get a better idea of what the new target data looks like overall. The most common situation, and often the hardest, is unsupervised DA. Here, the model only has labeled source data and completely unlabeled target data to work with during the adaptation phase. The aim is to build a model that can learn from the labeled old data and still make good predictions on the new, unlabeled data.

Another way to classify DA techniques is by where or how the adaptation happens in the AI model's process. Some methods try image-level adaptation, which aims to change the appearance of the images from one domain to make them look as if they came from another domain. For example, tools like GANs, including one called CycleGAN [90], can be used to make images from different scanners look more similar before the main AI model even sees them. Other methods focus on feature-level adaptation. The idea here is to teach the model to find and use features or characteristics in the images that don't change much between the old (source) and new (target) data. This often involves trying to make the patterns the model learns look similar across both the source and target data, perhaps by using special training techniques like adversarial training or by mathematically matching how features are distributed. It's also possible to do model-level adaptation, which involves changing the main AI model itself, like adjusting its internal settings or how it makes final decisions, to better fit the new target data. This might include giving more importance to certain data examples or directly tweaking the part of the model that classifies things.

Beyond these main types, it's also useful to know about domain generalization. This is a slightly different but related goal where the aim is to train a model on data from one or more source settings so well that it can make good predictions on data from completely new, unseen settings, even without seeing any examples from those new settings beforehand during training [91]. Most DA methods today, whether they work with simpler, pre-defined features or are built into complex DL models, are chosen based on the specific medical problem, how much the data differs between settings, and what kind of data is available.

The main goal of all these DA techniques is to build AI models that are more dependable and perform consistently well, even when faced with the variety of medical data found in real-world clinical practice, thus making them more useful and trustworthy for doctors and patients.

3.3.2 Applications for Domain Adaptation in Medical Imaging

Researchers are increasingly using DA methods for CT images analysis of lung diseases [92], including pneumonia [93, 94]. This helps AI models work more reliably across different scanners, hospitals, and groups of patients. Because the way CT images look can change a lot based on scanner settings and how images are made, DA techniques are very important for creating AI models that can be used with confidence in many different clinics.

While there are not many comprehensive reviews looking specifically at DA applications for all types of pneumonia on CT [95], work in related areas of lung CT analysis and broader medical imaging shows how these methods can be helpful [96, 97]. The COVID-19 Domain Adaptation Challenge, which encouraged developing models that could work with data from different hospitals, whether the data was labeled or not [98]. Scientists explored ways like using specific techniques (e.g., Monte Carlo Dropout) to see how sure the model was about its answers and to help decide which new, unlabeled data needed to be checked by experts.

Many DA techniques first created for general medical image analysis can also be used for lung CTs. Since labeling medical images takes a lot of time and effort, Unsupervised DA methods, which don't need labeled data from the new (target) setting, are especially useful. In the specific area of lung image analysis, studies on finding lung nodules, understanding interstitial lung disease, and other lung conditions using CT have started to use DA. For instance, a method called contrastive DA with consistency matching has been suggested to help transfer knowledge from existing datasets to new, smaller, unlabeled datasets for telling different types of pneumonia apart using chest X-rays [99]. This method considers both the differences between the data sources and the lack of labeled new data. Although this example is for X-rays, the basic ideas of reducing differences between datasets and using related data can also be applied to CT scans.

Developing DA techniques for CT-based pneumonia differentiation is very important because signs of pneumonia can be subtle, and their appearance can easily change with different CT scanner settings. Feature-level DA methods, which try to make feature patterns similar across different hospitals or scanners, seem particularly promising. These methods aim to create a common way of seeing features so that data from different places look the same to the model, allowing one classifier to work well everywhere. This is especially helpful for radiomics, where the numbers calculated from features can be very sensitive to scanner settings. By making these features less affected by such shifts, their utility in multi-center studies can be significantly enhanced.

3.3.3 Limitations of Current Domain Adaptation Strategies

Although DA methods are helping AI work better with medical images, there are still several limitations and unanswered questions. These challenges are especially important when trying to use DA for complicated tasks like telling different types of pneumonia apart from CT scans from many medical centers.

One set of problems concerns how well these DA methods actually work and how reliable they are. The effectiveness of different DA techniques can change a lot depending on the specific medical task, how much the new data differs from the old data, the type of medical images used, and how much data is available. It's often hard to find studies that compare different DA methods on the same datasets for specific tasks like pneumonia differentiation. This makes it difficult to know which DA method is best for a particular situation. Sometimes, trying to adapt a model can even make things worse. This is called "negative transfer," and it can happen if the original (source) data is too different from or not relevant to the new (target) data. In such cases, the DA process might actually lower the model's performance on the new data compared to not using DA at all.

Another group of challenges relates to the data needed for DA and practical issues in using these methods. Most DA techniques, even those called "unsupervised," still need to see at least some unlabeled data from the new (target) hospital or scanner during the training or adaptation phase. In real clinical situations, it might not always be possible to get this new data before the AI model needs to be used. This difficulty has led to more interest in Domain Generalization, which aims to train models that are naturally strong enough to work well on new, unseen data without any prior look at it. Additionally, some advanced DA methods can be very complex to set up and fine-tune, especially if they involve complicated training steps or major changes to the AI model. Also, when DA methods change image features or how the model learns, it can become harder to understand why the model makes certain decisions, which is a concern for clinical use.

Finally, there are significant gaps in how DA models are tested and proven ready for real-world clinical tasks. It is difficult to truly know how well a DA model generalizes without testing it thoroughly on data from many different hospitals, with diverse patient groups, and on data it has never seen before. Many studies only test their DA models on a few new data sources, or on new data that is still quite similar to the original training data. Therefore, a key gap in the field is the need for DA strategies that are practical, truly effective, and well-tested for complex, real-world clinical jobs like telling pneumonia types apart from using CT scans from multiple centers. This means we need DA methods that are not too complicated, can handle different types and amounts of data differences well, and whose benefits can be clearly shown through detailed testing across many medical centers. The development of the MDA framework in this thesis aims to help fill this gap, especially for using CT radiomics to differentiate between TB and FP.

3.4 Summary

The comprehensive review of AI applications in CT-based pneumonia differentiation has surveyed existing methodologies and, critically, highlighted persistent limitations and unanswered questions within the field. This summary now serves to connect these literature-derived insights directly to the research objectives of this PhD thesis, demonstrating how this study intends to address these identified shortcomings and contribute meaningfully to current knowledge.

One significant challenge underscored by the review of DL methods is the ongoing difficulty in achieving accurate and reliable differentiation between BP and NBP using CT imaging. The literature suggests a specific gap in effectively integrating CT information from multiple viewing planes with readily available clinical biomarkers, a synergy whose full potential has not yet been clearly established for this vital diagnostic task. In direct response to this literature-confirmed need, the first objective of this thesis is to develop and evaluate the MPMT-Pneumo model. This novel multi-plane, multi-modal DL approach is specifically designed to enhance BP versus NBP differentiation by leveraging this combined information.

Furthermore, the examination of radiomics-based approaches confirmed that while promising for characterizing infections like TB, a major impediment is the limited generalizability of these models when applied to data from different hospitals or CT scanners. This issue is particularly acute for the difficult task of distinguishing TB from FP, where subtle CT distinctions require highly dependable AI models. While DA techniques offer potential solutions, the literature also points to limitations in current strategies. Consequently, the second objective of this work is to develop and validate a robust CT-based radiomics pipeline featuring a novel feature-level DA framework, termed MDA. This framework is engineered to improve the differentiation of TB from FP with enhanced reliability across diverse, multi-center clinical datasets, addressing a critical need for robust and generalizable radiomic solutions.

Finally, a consistent theme emerging from the wider AI literature reviewed is the crucial requirement for more thorough benchmarking and transparent, robust validation of AI models against established clinical practices and alternative computational methods. Many studies lack comprehensive multi-center validation, which is essential for translating AI tools into real-world clinical settings. The third objective of this thesis directly addresses this by committing to a comprehensive benchmarking of the developed MPMT-Pneumo model and the MDA-enhanced radiomics framework, ensuring their performance is rigorously assessed.

In essence, this thesis endeavors to advance the field by developing AI solutions specifically tailored to overcome these literature-identified challenges in pneumonia differentiation. By focusing on innovative approaches like multi-plane DL with biomarker integration and domain-adapted radiomics, and by prioritizing robust multi-center validation, this research aims to produce AI tools that are not only accurate but also dependable and broadly applicable in diverse healthcare environments, ultimately contributing to improved clinical decision-making and patient care.

CHAPTER 4 Development of the MPMT-Pneumo Model: Integrating Multi-Plane CT Features and Clinical Biomarkers for Differentiating Bacterial and Non-Bacterial Pneumonia

4.1 Introduction

BP continues to present a frequent diagnostic challenge in clinical practice, particularly when it is necessary to distinguish it from NBP during the early stages of disease [100, 101]. This differentiation directly impacts decisions regarding antibiotic therapy, which is often initiated empirically due to overlapping clinical symptoms and delayed microbiological confirmation [102]. However, empirical antibiotic use carries risks, including AMR and unnecessary adverse effects, especially when applied without clear bacterial evidence [103].

CT plays a valuable role in the evaluation of pneumonia, offering superior anatomic detail compared to chest radiographs [104]. Nevertheless, CT findings alone are often insufficient for reliable etiological classification. Radiologic features such as consolidation or ground-glass opacities are not specific to bacterial infection [105], and considerable interobserver variability exists in image interpretation [106]. Microbiologic testing remains the reference standard but is frequently inconclusive or delayed [107]. These limitations have led to increased interest in AI-based tools that combine imaging and clinical data to improve diagnostic accuracy in pneumonia [108, 109].

Recent advances in DL have shown promising results in image-based pneumonia differentiation [53]. CNNs and ViTs have been applied to tasks including COVID-19 detection and general pneumonia subtyping [52, 110, 111]. However, many existing models are limited to multiclass classification schemes or focus on image-only inputs[112-114], and few studies address the binary classification of BP versus NBP, which is critical for guiding antibiotic therapy. Moreover, the integration of routinely available inflammatory biomarkers, such as WBC, ANC, CRP, and PCT, remains underexplored, despite their established role in infection assessment.

To address these limitations, we developed the multi-plane and multi-modal transformer (MPMT), a DL model designed to differentiate BP from NBP using non-contrast chest CT and four inflammatory biomarkers. The model combines representations from axial, coronal, and sagittal planes to simulate radiologists' multi-plane interpretation approach and incorporates biomarker information through a joint transformer-based fusion module [115]. The model was trained and validated on a dual-center dataset with microbiologically confirmed cases, enabling assessment across diverse clinical environments.

This study aims to demonstrate the potential of combining multi-plane imaging features with clinical biomarkers to improve the accuracy and robustness of BP differentiation and support more informed antibiotic decision-making in real-world settings.

4.2 Materials and Methods

4.2.1 Imaging Acquisition

This retrospective study was approved by the Institutional Review Board. The requirement of informed consent was waived. A total of 507 patients diagnosed with community-acquired pneumonia were screened at two institutions: Center 1 and Center 2, from December 2021 to November 2023. Inclusion criteria were as follows: (1) pneumonia etiology confirmed by microbiological tests, including sputum culture, bronchoalveolar lavage, or polymerase chain reaction for respiratory pathogens. Only patients with a definitive pathogen identified and consistent clinical–radiological correlation were included; (2) chest CT performed within two days of hospital admission; and (3) availability of laboratory test results for inflammatory biomarkers: WBC, ANC, CRP and PCT within 48 hours after CT acquisition. Patients were excluded if they were under 18 years of age, had evidence of co-infection, or had incomplete imaging or laboratory records. A total of 384 patients met the inclusion criteria, comprising 239 BP cases and 145 NBP cases (including viral, fungal, and mycoplasma pneumonia). **Figure 4.1** shows the distribution of the two types of pneumonia at two centers. The allocation of data sets by institution resulted in a training set (n=309) and two test sets (n=28, 47). The average age of patients was 61.2 ± 16.3. A detailed summary of patient characteristics is presented in **Table 4.1**.

Table 4.1. Summary of clinical characteristics of patients at Center 1 and Center 2.

Characteristics	**Center 1 (n=144)**		***p*-value**	**Center 2 (n=240)**		***p*-value**
	B (n=62)	**NB (n=82)**		**B (n=177)**	**NB (n=63)**	
Age (years)	70.5±16.5	67.0±19.0	0.247	56.3 ±12.8	58.1±15.0	0.424
Gender (%)			0.003			0.901
Male (%)	49 (79.0%)	44 (53.7%)		110 (62.3%)	38 (60.3%)	
Female (%)	13 (21.0%)	38 (46.3%)		67 (37.7%)	25 (39.7%)	

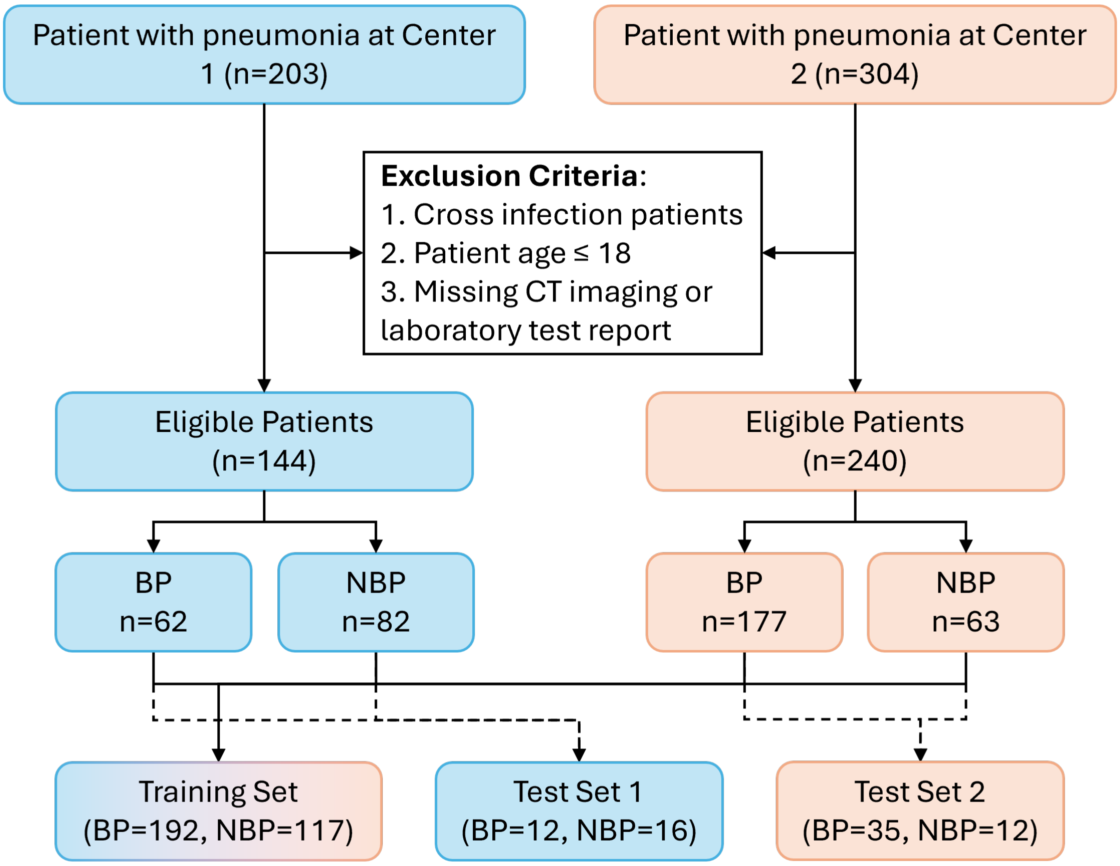


Figure 4.1. Flowchart illustrating patient enrollment and dataset division.

The study flow diagram presented the screening and categorization of patients with pneumonia at Center 1 and Center 2. The process involved exclusion criteria application, resulting in 384 eligible patients. Data were subdivided into one training set and two test set based on data source for bacterial pneumonia (BP) and non-bacterial pneumonia (NBP) classification.

4.2.2 Imaging Preprocessing

Chest CT examinations were acquired using multidetector CT scanners at both participating centers, following institutional standard protocols for non-contrast chest imaging. The scanning parameters included a tube voltage of 100–120 kVp and a slice thickness ranging from 1.25 to 5 mm. All CT scans were obtained in supine position during the end inspiratory breath hold.

To ensure data consistency and minimize interinstitutional variability, several image preprocessing steps were applied as shown in **Figure 4.2**. First, all CT volumes were resampled at an isotropic voxel size of 1.25 mm using bicubic interpolation [116]. The lung and infection regions were segmented using publicly available U-Net [117] and custom-trained nnU-Net [66], respectively. The nnU-Net was trained on a public dataset [118], achieving a Dice coefficient of 0.9386 on its validation set.

Segmentation quality control was conducted through systematic review of all 384 infection masks by two board-certified thoracic radiologists (with 5 and 20 years of experience). The reviewers confirmed that each mask appropriately captured pneumonia-related parenchymal abnormalities—including instances of multifocal disease, lesions with irregular contours, and those adjacent to major anatomical structures—such as consolidation, ground-glass opacities, or mixed attenuation patterns, while excluding unrelated findings. Regarding cavities, those clearly integral to the acute pneumonic process were included in the segmentation, whereas large, chronic-appearing cavities or those unrelated to the primary pneumonia were excluded. Similarly, the segmentation focused on parenchymal pneumonia; direct infection of the trachea (tracheitis) without parenchymal lung involvement was outside the scope of the infection masks. Minor manual corrections were required in 17 cases (4.4%), with a median adjustment time of 25 seconds (maximum < 90 seconds). No cases were excluded.

This dual-radiologist verification process ensured high-quality, anatomically accurate segmentation across the dataset. The visualization of the axial view of infection masks of different pneumonia are also shown in the dashed box in **Figure 4.2.** The lung region was then cropped using the lung mask to reduce background noise and computational burden. Window settings were standardized to lung window (L/W: -212 HU / 1624 HU), corresponding to a Hounsfield unit range of −1024 HU to 600 HU, and pixel intensities were normalized across scans. For the inflammatory biomarkers (WBC, ANC, CRP, PCT), Z-score standardization was applied to account for scale differences and ensure compatibility with DL model input. For this study, specific volumetric measurements of the pneumonia (e.g., total volume, volume of consolidation, or ground-glass opacity) were not performed, nor were these specific characteristics explicitly quantified for each scan as separate input features for the model. Our model was designed to learn relevant visual features directly from the pixel data within the provided expert-verified infection masks. The radiologists who verified the infection masks confirmed that these segmented regions corresponded to areas clinically and radiologically consistent with pneumonia, encompassing patterns such as consolidation and ground-glass opacities as commonly identified in clinical practice.

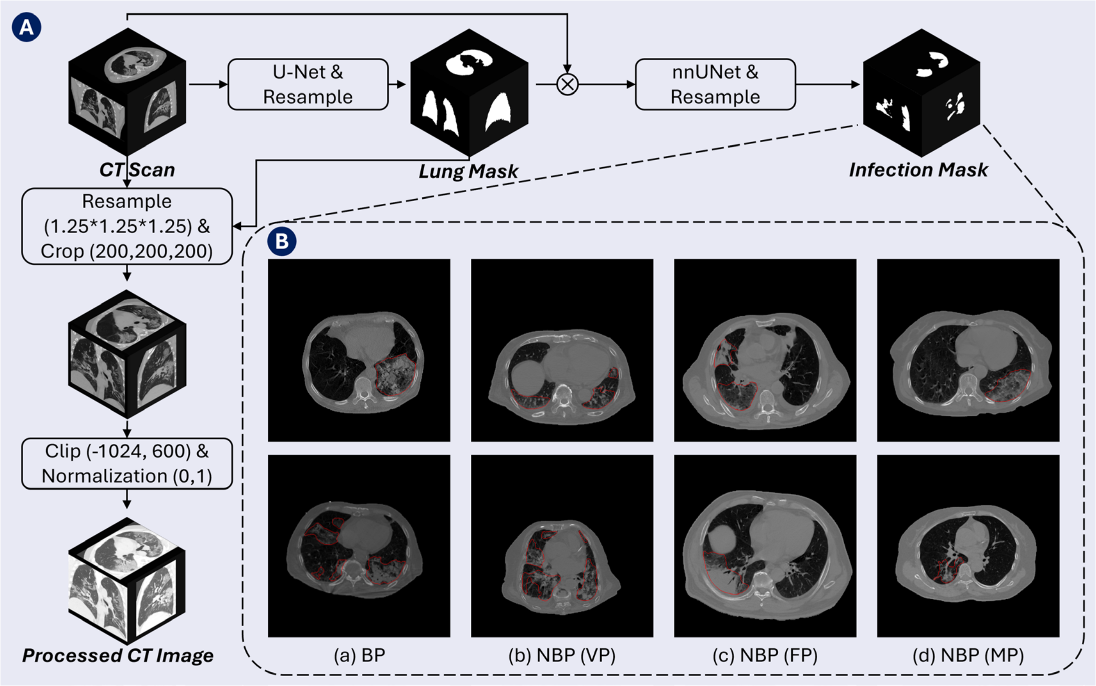


Figure 4.2. Image preprocessing pipeline and examples of segmented infection areas. Preprocessing Pipeline Diagram.

(A): This diagram illustrates the image processing workflow. (1) Input CTs are resampled to 1.25 mm isotropic voxels and cropped. (2) A U-Net model generates lung masks from the resampled CTs. (3) These lung masks isolate lung parenchyma, which serves as input for (4) a custom-trained nnU-Net model that segments infection masks. (5) Separately, the resampled and cropped CT images are windowed (L/W: -212 HU / 1624 HU, corresponding to a -1024 to 600 HU range) and normalized to create the 'Processed CT Image'. Both this 'Processed CT Image' and the radiologist-verified 'Infection Mask' from step (4) are inputs to the MPMT-Pneumo model. Examples of Segmented Infection Areas **(B)**: Representative axial CT slices displaying radiologist-verified infection masks (red contours) for various pneumonia types. These examples highlight the method's application to diverse presentations, including cases with multifocal involvement and irregular lesion borders (particularly visible in examples **a**, **b**, and **c**). BP: bacterial pneumonia; NBP: non-bacterial pneumonia; VP: viral pneumonia; FP: fungal pneumonia; MP: mycoplasma pneumonia.

4.2.3 Model Development and Training

We developed a novel multi-plane and multi-modal transformer (MPMT-Pneumo) to differentiate BP from NBP using chest CT images and inflammatory biomarker data. The model was inspired by the M3T [119], combining CNN for local spatial feature extraction with ViT for global contextual representation. The MPMT-Pneumo consists of three parts with three inputs: chest CT images, corresponding pneumonia infection masks and four clinical biomarkers. An overview of the model architecture is shown in **Figure 4.3.**

4.2.3.1 Multi-Plane Visual Feature Extraction

The 3D chest CT volume, combined with the pneumonia lesion mask, is first processed by a 3D convolutional encoder consisting of two convolutional layers with 5×5×5 kernels. Each subsequent 1×1×1 convolution followed by batch normalization and ReLU activation compresses the output into compact 3D features. These features are then sliced along three anatomical planes—axial, coronal, and sagittal—to simulate the routine multi-plane interpretation performed by radiologists. The resulting 2D slices are fed into a ResNet-50 [120] network pretrained on the MedicalNet [121] to extract high-level imaging features.

4.2.3.2 Multi-modal Feature Alignment

The image features extracted from the ResNet backbone are first projected into a 256-dimensional space using a lightweight two-layer neural network. These features are then grouped according to anatomical plane and arranged sequentially for subsequent integration. Special tokens are introduced to guide the model's attention: a token marks the start of the sequence, and additional tokens indicate boundaries between different anatomical planes. The four standardized inflammatory biomarkers are encoded into a single 256-dimensional vector using a multilayer perception, forming a dedicated clinical token. Plane-specific embeddings are added to each token to preserve anatomical context within the sequence.

All clinical and image features are concatenated into a unified token sequence. A special [CLS] token is prepended to serve as an aggregate representation for classification. Separator [SEP] tokens are inserted between the multi-plane image features, while a [CLI] token denotes the clinical biomarker data, ensuring the model can distinguish between different data modalities.

4.2.3.3 Transformer Encoder and Prediction

The unified sequence of imaging and clinical tokens is processed by a transformer encoder composed of eight layers. This architecture is designed to model both within-plane and cross-plane dependencies, as well as the interactions between imaging features and clinical data. Each layer in the encoder consists of a multi-head self-attention mechanism (with 8 heads and a 256-dimensional embedding) and a position-wise feed-forward network (with a 1024-dimensional inner layer). Pre-layer normalization is applied in every layer to improve training stability and convergence.

After passing through all transformer layers, the output corresponding to the classification token is extracted and passed through a fully connected layer to generate the final prediction of BP versus NBP

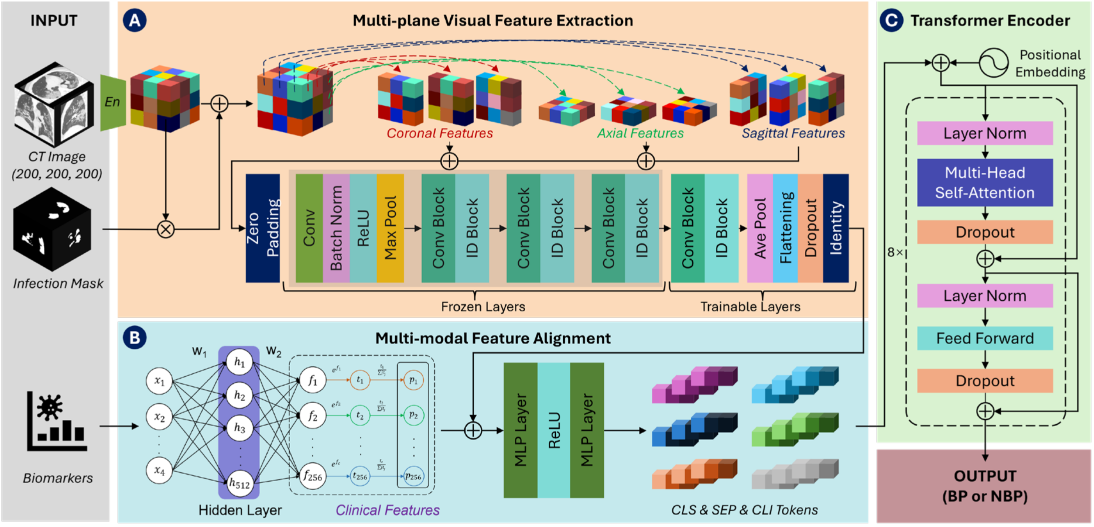


Figure 4.3. Overview of the MPMT-Pneumo network architecture for pneumonia differentiation.

The diagram illustrates the model's process for integrating multi-modal inputs to differentiate bacterial pneumonia from non-bacterial pneumonia. Input Section (Left): Shows the three primary inputs: 3D CT images, corresponding infection masks, and clinical biomarkers. Multi-plane Visual Feature Extraction **(A)**: Details the processing of CT images using a CNN architecture to extract features from coronal, axial, and sagittal planes. Multi-modal Feature Alignment **(B)**: Illustrates the processing of clinical biomarkers through an MLP (Hidden Layer) and the subsequent alignment and fusion of these clinical features with the extracted multi-plane visual features, generating combined tokens (e.g., CLS, SEP, CLI tokens). Transformer Encoder and Output **(C)**: Depicts the transformer block that processes the fused feature sequence for global context aggregation, leading to the final classification output (BP or NBP). En, encoder; BP, bacterial pneumonia; NBP, non-bacterial pneumonia.

4.3 Loss Function

To improve classification performance and address class imbalance in the dataset, we adopt Poly Focal Loss (PFL) [122] as the training objective. This loss function extends the standard Focal Loss [123], which down-weights well-classified examples to emphasize hard, misclassified cases. Focal Loss is defined as:

$$Focal\ Loss = -\alpha_t(1 - p_t)^{\gamma}\ log(p_t) \tag{4.1}$$

where p_t denotes the model's predicted probability for the correct class, γ is the focusing parameter, and α_t balances positive and negative samples. PFL introduces an additional polynomial term to stabilize optimization and accelerate convergence, particularly useful for subtle pneumonia distinctions. The loss is defined as:

$$Poly\ Focal\ Loss = \alpha_t \cdot FL(p_t) + \epsilon(1 - p_t)^{\gamma+1} \tag{4.2}$$

where ϵ controls the influence of the polynomial term. In our experiments, we set $\alpha_t = 0.40$, $\gamma = 2.0$, and $\epsilon = 1$.

4.4 Experimental Environment and Evaluation Methods

All experiments were conducted on a workstation equipped with an NVIDIA A6000 GPU (48 GB) running Ubuntu 22.04. The model was implemented using the PyTorch DL framework. Runtime was evaluated on the same workstation by timing the complete inference process across all 75 test cases. Training was performed for 200 epochs with a batch size of 1, using the Adam optimizer with an initial learning rate of 0.0001. Data augmentation was applied only during training using the TorchIO library. Each input scan underwent normalization to a canonical orientation, followed by probabilistic augmentations: affine transformation (40% probability), Gaussian noise addition (40%), and identity mapping (20%). To optimize hyperparameters, 10% of the training data was reserved as a validation set. For a fair comparison, all baseline models were re-implemented and trained under identical conditions. MPMT-Pneumo was further evaluated on separate test sets from each participating center to assess cross-domain generalizability. Lastly, ablation studies were conducted to evaluate the individual contribution of each model component.

4.4.1.1 Biomarker-only Statistical Analysis

Four conventional machine-learning classifiers— gradient boosting (GB, 500 estimators), LR, RF (100 trees), and SVM (RBF kernel)—were trained exclusively on the four inflammatory biomarkers (WBC, ANC, CRP, PCT). Hyper-parameters were tuned on the 10 % validation split via 5-fold grid search optimizing Matthews correlation coefficient (MCC) [124]. Class weights were set inversely proportional to class frequency to mitigate imbalance. Performance metrics were computed on the same test dataset.

4.4.1.2 Radiologist Interpretations Study

To compared MPMT-Pneumo model with clinical radiologist interpretations, three thoracic radiologists (reader 1, 1 year; reader 2, 10 years; reader 3, 22 years of experience) independently reviewed the test set. CT scans were provided in DICOM format and loaded using ITK-SNAP (lung window −1024/600 HU). Each radiologist was instructed to classify each case as either BP or NBP, without access to clinical data or prior model outputs. No quantitative measurement or segmentation was required. Class labels were recorded in a blind spreadsheet.

4.4.1.3 Statistical Analysis

Model performance was assessed using seven evaluation metrics commonly applied in binary classification tasks: accuracy, sensitivity, specificity, precision, F1-score, and AUC.

Additionally, we report on the MCC, which is particularly informative under class imbalance conditions. MCC is calculated as:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{4.3}$$

where TP, TN, FP, and FN represent the numbers of true positives, true negatives, false positives, and false negatives, respectively. The MCC value ranges from −1 to +1, with 0 indicating a random prediction. Unlike metrics that rely solely on positive class performance, MCC offers a more balanced assessment of overall model performance.

4.4.1.4 Interpretability Analysis

To visualize how MPMT-Pneumo make the diagnostic decisions, we performed interpretability analysis at both the intermediate convolutional layer and the final transformer decision layer. Although the model processes full 3D CT volumes, interpretability was applied to representative axial slices extracted from these volumes. Grad-CAM was used to highlight spatial features contributing to the output of the ResNet-50 encoder, reflecting intermediate-level feature sensitivity. For the transformer module, we applied attention rollout to aggregate multi-head self-attention weights across all layers, identifying the relative influence of each 14 × 14 image patch on the final classification. Resulting attention maps were projected back onto the original 2D slice grid to visualize decision focus.

In addition, we examined attention weights assigned to the four clinical biomarkers (WBC, ANC, CRP, PCT) by averaging across all heads and layers. All visualizations were produced using PyTorch-Grad-CAM and Matplotlib. Although the model operates on full 3D CT volumes, interpretability results are presented in 2D views for clarity.

4.5 Results

4.5.1 Overall Performance on the Full Test Set

We first assessed the classification performance of MPMT-Pneumo on a combined test set consisting of 75 held-out patients from both Center 1 and Center 2, including 47 BP and 28 NBP cases. Inference was performed on an NVIDIA A6000 GPU. While this configuration was used for benchmarking, hardware with at least 21 GB of memory is sufficient for deployment. When processing the test set in batch mode, segmentation of a single case required approximately 72 seconds, and classification took less than 1.6 seconds. The total processing time per case, including both segmentation and prediction, was 73 ± 1.4 seconds. Four baseline models were included for comparison: ResNet50 and ViT, as canonical CNN and transformer architectures; Pneumonia-plus [52], a recent state-of-the-art model designed for tri-class pneumonia differentiation; and MMI [53], a multimodal integration framework incorporating CT and clinical biomarkers.

As shown in **Table 4.2**, MPMT-Pneumo achieved the best overall performance across all evaluation metrics, including an accuracy of 0.852, a sensitivity of 0.894, an F1-score of 0.883, and an AUC of 0.874. Importantly, it also obtained the highest MCC (0.682), indicating balanced performance on both BP and NBP classes. MMI yielded competitive results, achieving the highest specificity (0.821) and strong scores in precision and F1-score, though it fell slightly in AUC (0.801) and MCC (0.617).

Table 4.2. Comparative performance of various architectures in differentiating bacterial from non-bacterial pneumonia.

Model	Accuracy (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	Precision (95% CI)	F1-Score (95% CI)	MCC (95% CI)	AUC (95% CI)
Resnet50	0.614 (0.507-0.720)	0.532 (0.395-0.674)	0.750 (0.592-0.912)	0.786 (0.636-0.923)	0.632 (0.500-0.747)	0.279 (0.076-0.479)	0.647 (0.520-0.774)
ViT	0.547 (0.440-0.653)	0.596 (0.458-0.740)	0.464 (0.273-0.654)	0.652 (0.521-0.796)	0.619 (0.500-0.729)	0.060 (-0.173-0.276)	0.545 (0.406-0.676)
Pneumonia-plus	0.722 (0.613-0.827)	0.681 (0.551-0.813)	0.786 (0.636-0.929)	0.845 (0.714-0.946)	0.753 (0.646-0.851)	0.456 (0.238-0.647)	0.703 (0.579-0.823)
MMI	0.814 (0.720-0.893)	0.809 (0.694-0.917)	0.821 (0.667-0.957)	0.883 (0.775-0.974)	0.843 (0.750-0.915)	0.617 (0.427-0.781)	0.801 (0.693-0.904)
MPMT-Pneumo	**0.852 (0.760-0.933)**	**0.894 (0.795-0.976)**	**0.786 (0.607-0.929)**	**0.874 (0.771-0.961)**	**0.883 (0.809-0.944)**	**0.682 (0.482-0.840)**	**0.874 (0.781-0.956)**

These performance differences are further visualized in **Figure 4.4**, where the ROC curves show a clear separation between MPMT-Pneumo and all baselines across the entire false positive rate spectrum. MPMT-Pneumo's superior ROC trajectory supports its robust discriminative ability under variable classification thresholds. MMI yielded the closest ROC performance but fell short in sensitivity and MCC. Importantly, both ResNet50 and ViT were included as structural baselines, as MPMT-Pneumo integrates convolutional networks for local feature extraction and transformer modules for global context modeling. The model’s advantage over both isolated backbones underscores the benefit of this hybrid architecture.

The confusion matrices shown in **Figure 4.5** illustrate model-specific prediction tendencies. MPMT-Pneumo correctly identified 89.4% of BP cases and 78.6% of NBP cases, striking a good balance between sensitivity and specificity. MMI showed a slight bias toward BP detection, while ViT misclassified more than half of the NBP cases. These findings highlight MPMT-Pneumo’s robustness and balanced classification, especially when considering class imbalance and cross-domain input sources.

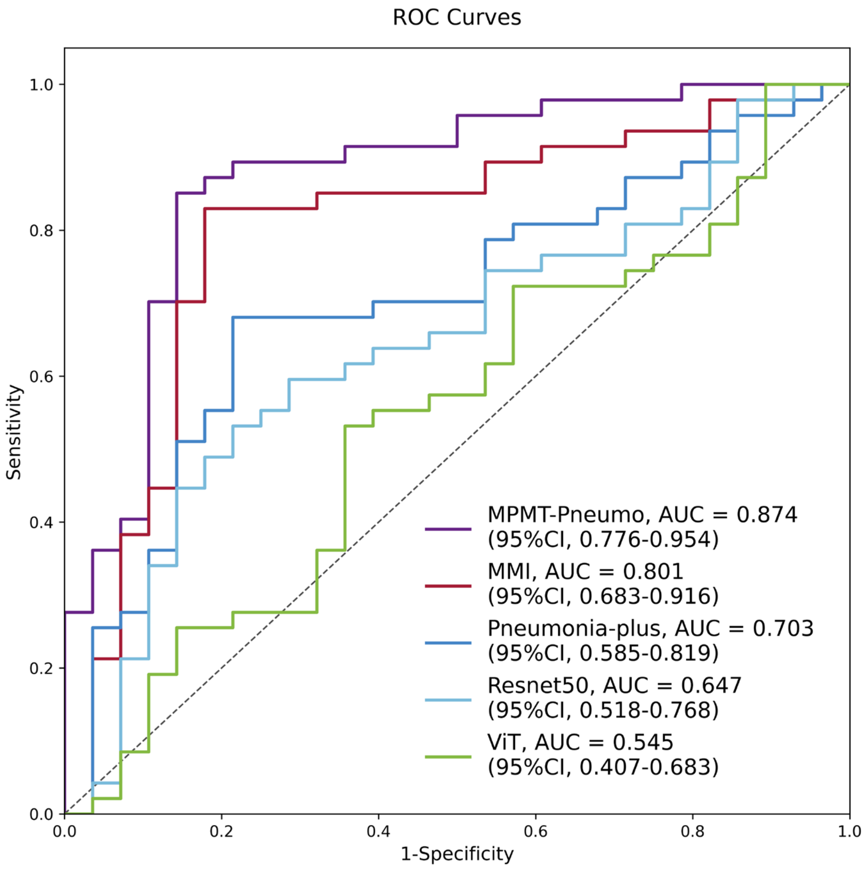


Figure 4.4. ROC curves of MPMT-Pneumo and four architectures in differentiating bacterial pneumonia from non-bacterial pneumonia.

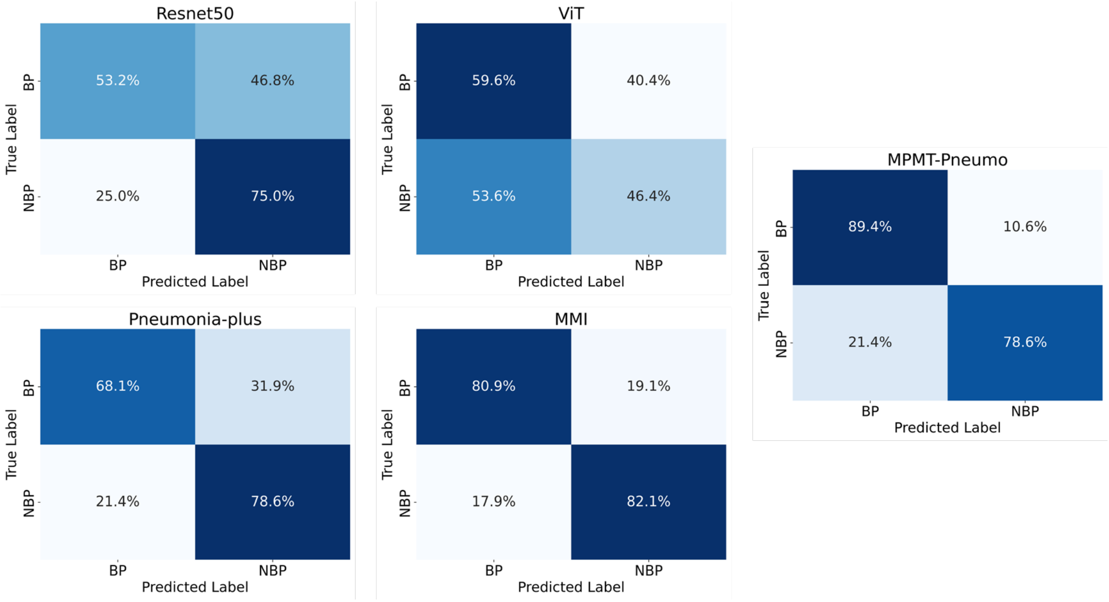


Figure 4.5. Confusion matrices of MPMT-Pneumo and four architectures in differentiating bacterial pneumonia from non-bacterial pneumonia.

Each matrix displays the percentage of correctly and incorrectly classified samples across bacterial pneumonia (BP) and non-bacterial pneumonia (NBP) classes.

4.5.2 Comparison with Biomarker-Only Models and Radiologists

We compared MPMT-Pneumo with conventional biomarker-only classifiers and expert radiologist interpretations. **Table 4.3** presents the performance of MPMT-Pneumo, four biomarker-only classifiers, and three radiologists on the 75-case test set. Among the conventional models, RF achieved the highest accuracy (0.613) and sensitivity (0.484), followed by GB, SVM, and LR. However, all biomarker-based models demonstrated limited sensitivity (< 0.5) and lower overall discriminative power, highlighting the insufficiency of laboratory values alone for pneumonia subtyping.

The results of radiologist interpretations shown the diagnostic accuracy increased with experience: Reader 3 (22 years) achieved the highest accuracy (0.733), sensitivity (0.654), and MCC (0.422), followed by Reader 2 (10 years) and Reader 1 (1 year). A consistent tendency among less experienced readers to misclassify NBP cases, particularly viral as BP. In contrast, MPMT-Pneumo demonstrated superior performance across all metrics. Statistical comparison using McNemar’s test confirmed that MPMT-Pneumo yielded significantly higher sensitivity for BP than both Reader 1 ($p = 0.011$) and Reader 2 ($p = 0.046$), while performing comparably to Reader 3 ($p = 0.053$). These results underscore the model’s clinical potential as a second-reader system capable of consistent, high-accuracy diagnosis across variable radiologist expertise and limited laboratory inputs.

Table 4.3. Comparative performance of statistical models based on four inflammatory biomarkers, three radiologists using CT alone, and the proposed MPMT-Pneumo model.

Method	**Accuracy (95% CI)**	**Sensitivity (95% CI)**	**Specificity (95% CI)**	**Precision (95% CI)**	**F1-Score (95% CI)**	**MCC (95% CI)**
GB	0.587 (0.474-0.691)	0.468 (0.333-0.608)	0.786 (0.605-0.898)	0.786 (0.605-0.898)	0.587 (0.487-0.687)	0.254 (0.154-0.354)
LR	0.467 (0.358-0.578)	0.350 (0.221-0.505)	0.600 (0.436-0.744)	0.500 (0.326-0.674)	0.412 (0.312-0.512)	0.052 (-0.152-0.048)
RF	0.613 (0.500-0.715)	0.484 (0.320-0.652)	0.705 (0.558-0.818)	0.536 (0.358-0.705)	0.508 (0.408-0.608)	0.192 (0.092-0.292)
SVM	0.573 (0.461-0.679)	0.423 (0.255-0.611)	0.653 (0.513-0.771)	0.393 (0.236-0.576)	0.407 (0.307-0.507)	0.075 (-0.025-0.175)
Radiologist 1	0.560 (0.447-0.667)	0.439 (0.299-0.590)	0.706 (0.538-0.832)	0.643 (0.458-0.793)	0.522 (0.422-0.622)	0.149 (0.049-0.249)
Radiologist 2	0.680 (0.568-0.775)	0.562 (0.393-0.718)	0.767 (0.623-0.868)	0.643 (0.458-0.793)	0.600 (0.500-0.700)	0.337 (0.237-0.437)
Radiologist 3	0.733 (0.624-0.820)	0.654 (0.462-0.806)	0.776 (0.641-0.870)	0.607 (0.424-0.764)	0.630 (0.530-0.730)	0.422 (0.322-0.522)
MPMT-Pneumo	0.852 (0.760-0.933)	0.894 (0.795-0.976)	0.786 (0.607-0.929)	0.874 (0.771-0.961)	0.883 (0.809-0.944)	0.682 (0.482-0.840)

Abbreviations: GB, gradient boosting; LR, logistic regression; RF, random forest; SVM, support vector machine

4.5.3 Generalization Across Medical Centers

We tested MPMT-Pneumo on two separate test sets from different hospitals. Test Set 1 came from Center 1 and included 12 BP and 16 NBP cases. Test Set 2 came from Center 2 and included 35 BP and 12 NBP cases. The two sets differed in sample size and class distribution. As shown in **Table 4.4**, the model correctly identified 83.3% of BP cases from Center 1 and 91.4% from Center 2. For NBP, the correct rates were 81.2% (Center 1) and 75.0% (Center 2). The AUC was 0.741 and 0.820 for Center 1 and Center 2, separately. **Figure 4.6** shows the confusion matrices for each center. The results in both hospitals show that the model had higher sensitivity for BP and slightly lower specificity for NBP. This pattern was seen across both centers. Despite different scanners and patient groups, the prediction trend stayed similar. These results show that the model performs consistently for each pneumonia type across hospitals. The differences in imaging protocols and patient profiles did not lead to major changes in accuracy. This suggests the model can generalize well in real clinical settings.

Table 4.4. Comparative performance of MPMT-Pneumo at Center 1 and Center 2.

Model	**Accuracy (95% CI)**	**Sensitivity (95% CI)**	**Specificity (95% CI)**	**Precision (95% CI)**	**F1-Score (95% CI)**	**MCC (95% CI)**	**AUC (95% CI)**
Center 1	0.819 (0.679-0.964)	0.833 (0.583-1.000)	0.812 (0.611-1.000)	0.767 (0.533-1.000)	0.789 (0.588-0.947)	0.634 (0.338-0.923)	0.741 (0.517-0.938)
Center 2	0.873 (0.766-0.957)	0.914 (0.816-1.000)	0.750 (0.461-1.000)	0.912 (0.806-1.000)	0.913 (0.841-0.973)	0.662 (0.397-0.888)	0.820 (0.668-0.949)

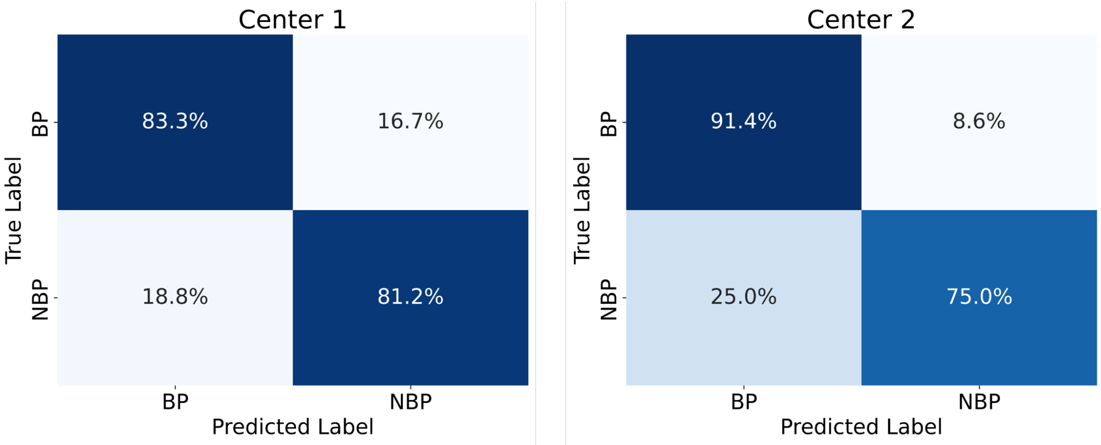
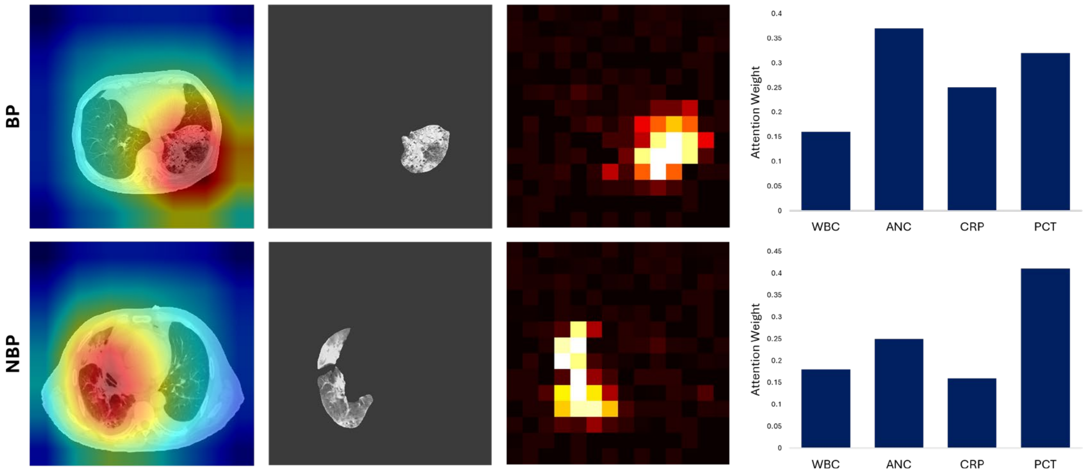


Figure 4.6. Confusion matrices of MPMT-Pneumo performance at Center 1 and Center 2. Each matrix displays the percentage of correctly and incorrectly classified samples across bacterial pneumonia (BP) and non-bacterial pneumonia (NBP) classes.

4.5.4 Model Interpretability Findings

We presented representative attention maps for BP and NBP cases in **Figure 4.7**. In the BP case (top row), the Grad-CAM heatmap from the CNN encoder (**Fig. 4.7-a**, top row) highlights a broad consolidation region in the right lower lobe. In contrast, the transformer attention map (**Fig. 4.7-c**) is focused within the infected region outlined by the ground-truth mask (**Fig. 4.7-b**), indicating better lesion localization at the final decision stage. Similarly, in the NBP case (bottom row), attention is distributed over the ground-glass opacities, but the transformer map again shows more specific localization to diseased areas compared to the broader CNN response. These findings suggest that while the CNN encoder captures early regional features, the transformer module refines this information to highlight the most diagnostically relevant areas.

The Bar plots in **Fig. 4.7-d** show the average attention weights for the four clinical biomarkers. Across both BP and NBP, the model consistently prioritized ANC and PCT over CRP and WBC. In BP cases, ANC received higher relative weight (0.37 vs. 0.25), while in NBP cases, PCT was more heavily weighted (0.41 vs. 0.32). These patterns suggest that the model leverages ANC to identify neutrophil-driven bacterial processes and recognizes the broader inflammatory relevance of PCT in atypical or VP. The consistent elevation of attention on ANC and PCT across diagnostic groups supports the clinical plausibility of the model's multi-modal feature integration.



(a) Grad-CAM Overlay on the Axial Slice (b) Infection Mask (c) Transformer Patch-level Attention (d) Biomarker-token Attention Weights

Figure 4.7. Qualitative interpretability examples for representative bacterial pneumonia and non-bacterial pneumonia cases from the test set.

Grad-CAM Overlay on the Axial Slice **(a)**: Heatmaps overlaid on the original CT scan, highlighting the discriminative image regions identified by the ResNet-50 encoder. Infection Mask **(b)**: The corresponding radiologist-verified infection mask for the slice shown in **(a)**, provided for anatomical reference and comparison with model attention areas. Transformer Patch-level Attention **(c)**: Visualization of aggregated self-attention weights from the transformer encoder, projected onto the input image patches (14×14 patches). Biomarker-token Attention Weights **(d)**: Bar plots showing the average attention weights assigned by the transformer to each of the four clinical biomarkers. BP, bacterial pneumonia; NBP, non-bacterial pneumonia; WBC, white blood cell count; ANC, absolute neutrophil count; CRP, C-reactive protein; PCT, procalcitonin.

4.5.5 Ablation Studies

To assess the individual and combined contributions of the loss function, multi-plane image features, and clinical biomarkers, we constructed six ablation variants. As shown in **Table 4.5**, each model had specific components removed or modified. Performance consistently declined with the removal of individual or paired components, particularly when both multi-plane imaging and clinical data were excluded (Study 5, MCC: 0.321). Below is a summary of each variant.

- Study 1: Loss function modification

Replacing PFL with CE loss resulted in decreased specificity (0.393) and a drop in MCC to 0.452 (baseline: 0.682), despite maintaining high sensitivity (0.957). This highlights the role of PFL in handling class imbalance.

- Study 2: Multi-plane removal

Using repeated axial image features instead of multi-plane inputs led to reduced performance (AUC: 0.731; MCC: 0.419), supporting the value of cross-sectional anatomical information in pneumonia detection.

- Study 3: Biomarker removal

Excluding the four inflammatory biomarkers (WBC, ANC, CRP, and PCT) resulted in lower specificity (0.643) and MCC (0.462), although overall discrimination remained good (AUC: 0.729; F1-score: 0.801). This suggests clinical data help balance classification outcomes.

- Study 4: Loss function modification and multi-plane removal

Removing both PFL and multi-plane input—retaining only axial features and clinical data—caused a notable performance drop (AUC: 0.685; MCC: 0.331). This underscores the combined contribution of imaging structure and loss function design.

- Study 5: Multi-plane and biomarker removal

Retaining PFL but removing multi-plane inputs and biomarkers yielded the lowest MCC (0.321; AUC: 0.688), indicating that imaging alone, even with optimized loss, is insufficient for robust classification.

- Study 6: Loss function modification and biomarker removal

Replacing PFL with CE loss and removing biomarkers—while keeping multi-plane views—also degraded performance (AUC: 0.654; MCC: 0.385), showing that both components contribute jointly to model robustness.

Table 4.5. Comparative performance of MPMT-Pneumo variants in ablation studies.

Study	**w/o PFL**	**w/o multi-plane**	**w/o biomarkers**	**Accuracy (95% CI)**	**Sensitivity (95% CI)**	**Specificity (95% CI)**	**Precision (95% CI)**	**F1-Score (95% CI)**	**MCC (95% CI)**	**AUC (95% CI)**
1	✓			0.747 (0.638, 0.831)	0.957 (0.858, 0.988)	0.393 (0.236, 0.576)	0.729 (0.672, 0.793)	0.827 (0.778, 0.876)	0.452 (0.264, 0.636)	0.704 (0.555-0.830)
2		✓		0.733 (0.624, 0.820)	0.830 (0.699, 0.911)	0.571 (0.391, 0.735)	0.765 (0.690, 0.844)	0.797 (0.720, 0.865)	0.419 (0.204, 0.624)	0.731 (0.602-0.843)
3			✓	0.747 (0.638, 0.831)	0.809 (0.675, 0.896)	0.643 (0.458, 0.793)	0.795 (0.711, 0.881)	0.801 (0.716, 0.878)	0.462 (0.248, 0.658)	0.729 (0.598-0.843)
4	✓	✓		0.707 (0.596, 0.798)	0.915 (0.801, 0.966)	0.357 (0.207, 0.542)	0.704 (0.647, 0.768)	0.795 (0.738, 0.852)	0.331 (0.103, 0.550)	0.685 (0.550-0.806)
5		✓	✓	0.680 (0.568, 0.775)	0.766 (0.628, 0.864)	0.536 (0.358, 0.705)	0.741 (0.659, 0.826)	0.753 (0.667, 0.835)	0.321 (0.088, 0.538)	0.688 (0.544-0.816)
6	✓		✓	0.720 (0.610, 0.809)	0.957 (0.858, 0.988)	0.321 (0.179, 0.507)	0.705 (0.657, 0.767)	0.812 (0.766, 0.860)	0.385 (0.180, 0.579)	0.654 (0.501-0.798)

4.6 Discussion

MPMT-Pneumo demonstrated high diagnostic performance in differentiating BP from NBP, with balanced sensitivity and specificity across two distinct clinical sites. This performance was achieved without requiring contrast enhancement or handcrafted radiologic features, indicating the feasibility of applying the model to routine non-contrast CT studies as part of early pneumonia triage. In head-to-head testing, MPMT-Pneumo also surpassed four biomarker-only classifiers and three thoracic radiologists, including a reader with 22 years of experience, confirming that multi-modal fusion adds diagnostic value beyond either modality alone. Average inference time was 73 ± 1.4 s per case on an A6000 GPU, which is compatible with real-time triage in routine practice. Notably, MPMT-Pneumo achieved comparable or superior performance to several established AI models across most evaluation metrics, particularly AUC and MCC, highlighting their robustness in imbalanced clinical settings.

The strength of MPMT-Pneumo stems from its alignment with diagnostic reasoning rather than purely statistical pattern recognition. By incorporating axial, coronal, and sagittal views, the model captures spatial heterogeneity and subtle signs of pneumonia that may be present differently across planes. This mirrors radiologists' workflow and enhances the recognition of infection distribution and anatomical relationships [125], particularly in early or atypical presentations.

Our ablation studies showed that removing the multi-plane input alone (Study 2) reduced the MCC from 0.682 to 0.419. Similarly, removing biomarkers (Study 3) led to a drop to 0.462. When both the anatomical views and clinical data were excluded (Study 5), the performance declined the most (MCC: 0.321), suggesting that these two parts works together to support the model. Study 4, which removed both the loss function and multi-plane input, also showed a clear drop in performance (MCC: 0.331), highlighting the combined value of proper loss design and diverse imaging input. In contrast, Study 6 kept the multi-plane input and performed better (MCC: 0.385), even though both the loss function and biomarkers were missing. This suggests that having information from different image planes is especially important for stable model performance. Overall, imaging structure, loss function, and clinical data each play a role, and their effects are stronger when used together.

Interpretability analysis provided additional insights. Transformer attention consistently assigned higher weights to ANC and PCT than to CRP and WBC. In bacterial cases, ANC predominated (0.37 vs 0.25), whereas PCT was emphasized in non-bacterial cases (0.41 vs 0.32). These patterns suggest that the model leverages ANC to identify neutrophil-driven BP and relies on PCT's broader inflammatory signal to flag NBP, supporting the clinical plausibility of its feature integration. Integration of laboratory markers therefore enabled more nuanced decisions, particularly where imaging findings were equivocal [126]. This finding is consistent with growing evidence that hybrid models combining imaging and clinical biomarkers outperform image-only frameworks [53].

MPMT-Pneumo also proved resilient to scanner hardware, acquisition parameters, and cohort differences, suggesting that site-specific calibration may not be necessary. Such cross-domain robustness is noteworthy given prior reports of performance drift due to technical variability [127, 128]. Clinically, the model could facilitate early empirical antibiotic selection and prioritize infectious-disease consultation. Automated infection masks and standard-resolution CT enable seamless integration into existing workflows, and high bacterial sensitivity may reduce missed infections while preserving specificity to avoid over-treatment.

Despite its promise, several limitations should be acknowledged. This study focused on binary differentiation of BP versus NBP, which is clinically meaningful but does not capture the full spectrum of pneumonia etiologies. Future development should expand the model toward multi-class classification to distinguish between specific non-bacterial subtypes such as viral, fungal, and atypical pneumonia. In addition, missing biomarker values were handled through simple exclusion without imputation, which may limit reproducibility in real-world settings with incomplete laboratory data. The model was evaluated using an internal split across two medical centers but was not tested on a fully independent external dataset. Multi-institutional external validation will be necessary to confirm generalizability.

Future directions include extension to multiclass pneumonia differentiation, improved handling of incomplete clinical data, and external validation across diverse healthcare systems. Given its strong cross-center performance and reliance on routinely available inputs, MPMT-Pneumo holds potential as a scalable diagnostic support tool in both academic and community settings.

4.7 Conclusion

MPMT-Pneumo provides an effective diagnostic solution for distinguishing BP from NBP by integrating multi-plane CT imaging and inflammatory biomarkers. Its consistent performance across institutions with varied imaging protocols supports its potential for widespread clinical adoption. Prospective validation in larger and more diverse populations will be essential to establish its role in real-time decision support and in guiding more appropriate antibiotic use.

CHAPTER 5 Development of a Multicenter Distribution Adaptation (MDA) Radiomics Framework for Differentiating Tuberculous and Fungal Pneumonia on CT

5.1 Introduction

TB and FP remain major causes of respiratory infections worldwide. The latest estimates from the World Health Organization indicate that in 2023, approximately 10.8 million people developed TB, corresponding to 134 cases per 100,000 population, with about 1.25 million deaths (WHO, 2025). Following the disruptions of the COVID-19 pandemic, TB has once again become the leading cause of death from a single infectious agent, posing a substantial global health threat [129]. FP also represents a significant burden, with global analyses reporting over 2 million cases annually, primarily affecting individuals with chronic lung diseases, immunosuppression, or critical illness, and associated mortality rates ranging from 40% to 80% depending on the underlying condition [130]. Both TB and FP present major clinical challenges due to their severity, complexity of treatment, and considerable public health impact.

Despite the critical need for accurate etiological diagnosis, distinguishing TB from FP remains challenging in routine clinical practice. Clinical symptoms, laboratory findings, and radiological features frequently overlap, leading to delays or errors in treatment initiation [57]. Studies have shown that a considerable proportion of patients initially treated for smear-negative TB were later found to have fungal infections [131]. Microbiological confirmation remains the gold standard but faces substantial challenges. TB culture, while highly specific, requires several weeks for growth and shows limited sensitivity [132]. Molecular methods such as PCR have improved diagnostic speed but still suffer from reduced sensitivity in low-burden samples and limited accessibility in many regions [133]. For FP, definitive diagnosis often requires invasive procedures such as bronchoalveolar lavage or biopsy, which carry procedural risks and can yield inconsistent results depending on the patient's immune status and the pathogen involved [134]. Moreover, serological and antigen-based tests, although valuable, lack universal sensitivity and specificity, especially in non-HIV populations or early-stage disease [135].

Chest CT plays a key role in evaluating pulmonary infections and can reveal imaging features suggestive of TB or FP. Certain CT signs, such as the "tree-in-bud" appearance for TB or the halo sign for FP, have been shown to aid differentiation [136, 137]. However, common imaging features, including cavitation, nodules, consolidation, and ground-glass opacities, are frequently observed in both diseases, limiting the specificity of visual CT interpretation [138, 139]. To date, there are no CT-based models specifically tailored to distinguish TB from FP, leaving a major gap in non-invasive diagnostic approaches.

Radiomics, a technique that extracts high-dimensional quantitative features from medical images, offers a potential solution by capturing subtle differences in lesion texture, shape, and intensity beyond human visual perception [46]. Previous studies have demonstrated that radiomics can assist in characterizing infectious lung diseases, including distinguishing TB or FP from other pneumonias [58, 140, 141]. However, most radiomics models have been developed using single-center datasets and often show decreased performance when applied to external multicenter cohorts [142]. A key barrier is domain shift, arising from variations in CT acquisition protocols, scanner characteristics, and reconstruction algorithms, which can distort radiomic feature distributions and impair model robustness [143]. To address this challenge, feature-level

DA techniques have been proposed to align feature distributions across centers, enhancing the reliability of radiomics models for multicenter applications.

In this multicenter study, we aim to develop and validate a CT-based radiomics model, adjusted using feature-level DA techniques, to non-invasively differentiate TB from FP across multiple centers, while providing interpretable results to facilitate clinical adoption.

5.2 Materials and Methods

This section describes the methods used in our study. We aimed to see if information from chest CT scans could help tell the difference between TB and FP. We also wanted to create a prediction model that works well across different hospitals. First, we will explain the study plan and how we selected our patients. Then, we will describe how CT scans were taken and prepared to make sure the data was consistent and of good quality. After that, we'll cover how we found specific features in the images, chose the most important ones, and built our model to handle differences in data from various centers. Finally, we will explain the statistical tests we used to check how well our model performed.

5.2.1 Study Design and Patient Population

This retrospective multicenter study was approved by the Institutional Review Boards of all four participating centers, with the requirement for written informed consent waived. Between June 2017 and March 2024, a total of 778 patients were initially screened from four medical centers. Eligible patients had pathologically confirmed TB or FP. TB was diagnosed by positive Mycobacterium tuberculosis culture, nucleic acid amplification test from respiratory specimens, or histopathological evidence. FP was confirmed by positive fungal culture, histopathology, or serological markers combined with clinical and radiological findings.

Inclusion criteria were: (1) age ≥18 years; and (2) availability of a chest CT scan as part of diagnostic evaluation. Exclusion criteria were: (1) incomplete or poor-quality CT images; (2) prior anti-TB or anti-fungal treatment before CT imaging; and (3) significant comorbid pulmonary diseases (e.g., lung cancer, advanced COPD, or interstitial lung disease) that could confound image interpretation. The detailed process of patient selection and the number of cases at each stage are illustrated in **Figure 5.1**.

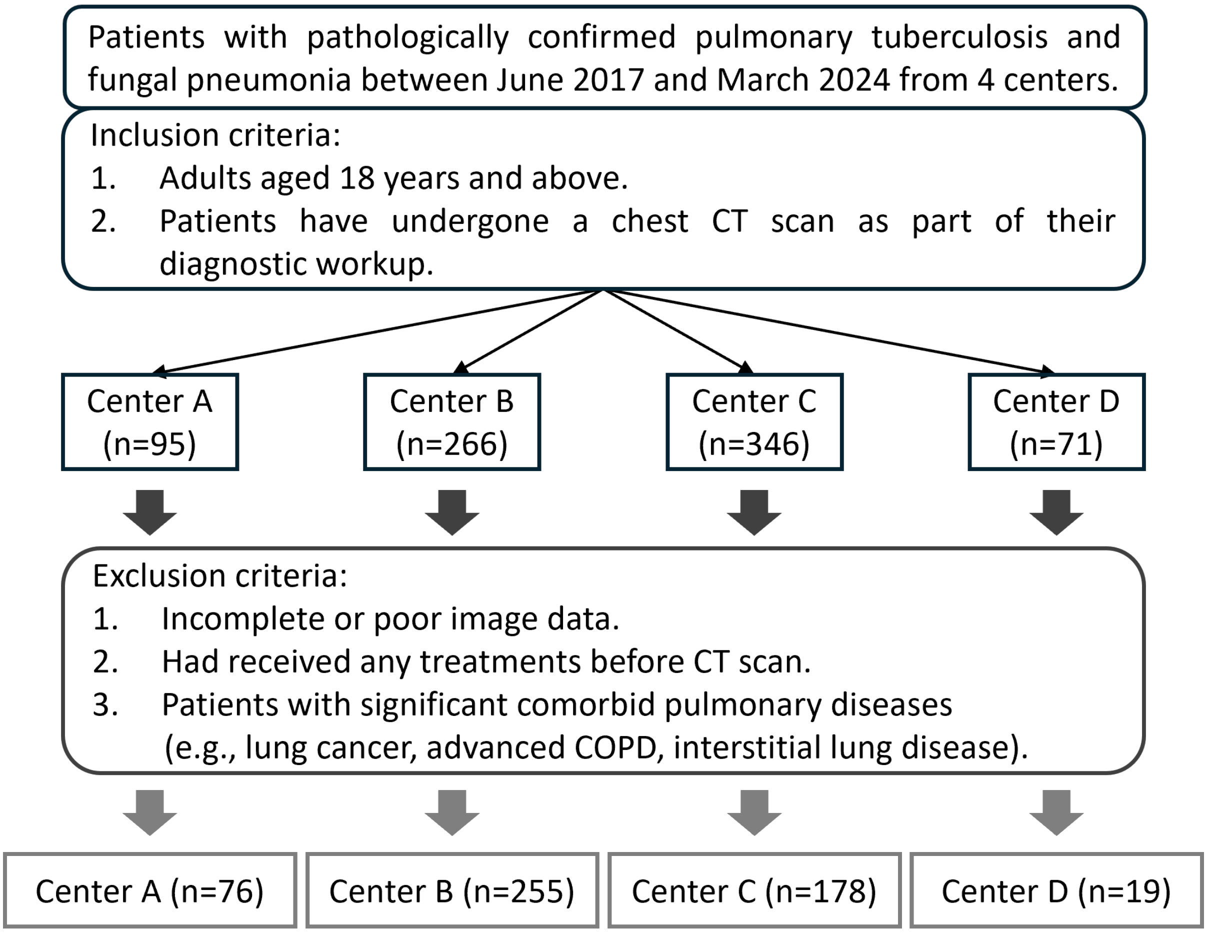


Figure 5.1. Flowchart illustrates the inclusion and exclusion criteria for patient selection in this multicenter study.

5.2.2 CT Image Acquisition and Preprocessing

Chest CT scans were acquired using multi-detector CT systems at all four centers, with varying acquisition protocols. Tube voltage ranged from 100 to 120 kV, and slice thickness varied between 1 mm and 5 mm. To standardize the data for radiomic analysis, all DICOM images were first converted to NIfTI format. Only lung window reconstructions (window level −600 HU, width 1500 HU) were included for further processing. Lung and lung lobe regions were automatically segmented using the open-source DL model TotalSegmentator [117]. Infection regions were then segmented using a customized nnU-Net [66] model trained on an internal dataset [118] of pneumonia cases. The resulting segmentation masks were independently reviewed by two board-certified thoracic radiologists (10 and 15 years of experience). Discrepancies were resolved by consensus. After expert review, all CT image volumes were resampled to 1 × 1 × 1 mm using B-spline interpolation, while all corresponding segmentation masks were resampled using nearest-neighbor interpolation. The overall imaging pre-processing and radiomics workflow are illustrated in **Figure 5.2**.

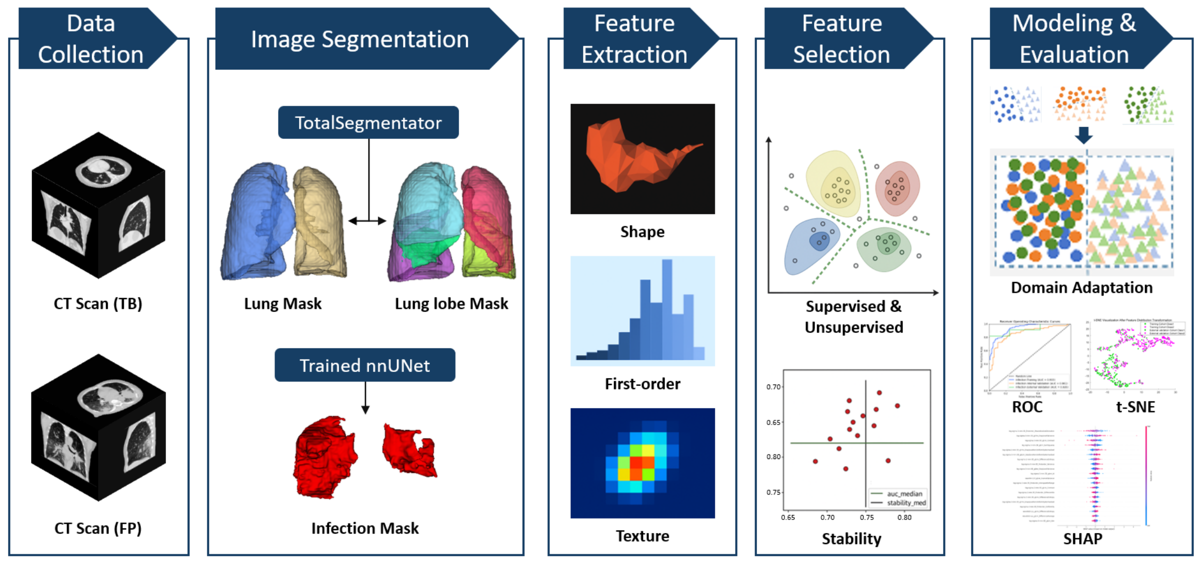


Figure 5.2. The radiomics workflow in the prediction of tuberculosis (TB) and fungal pneumonia (FP).

5.2.3 Radiomic Feature Extraction

Radiomic features were extracted using the PyRadiomics library (version 3.0.1), following Image Biomarker Standardization Initiative guidelines. Each resampled CT scan and its corresponding segmentation masks—including the infection region, lung, and lung lobes—served as input. The extracted features were grouped into three categories: shape features, first-order features, and texture features. Texture features were derived from five standard matrixes: gray-level co-occurrence matrix, run length matrix, size zone matrix, dependence matrix, and neighboring gray-tone difference matrix. Shape features were computed from the resampled CT image, while first-order and texture features were also extracted from filtered images generated by wavelet decomposition (coif1 wavelet, level 2) and Laplacian of Gaussian filtering with sigma values of 1, 2, and 3 mm. A fixed bin width of 25 Hounsfield Units was used for intensity discretization. Features that failed extraction due to segmentation or image artefacts were excluded. All outputs were saved in structured format for quality control and downstream analysis.

5.2.4 Feature Selection

Following feature extraction, a large number of radiomics features were obtained. Because the dimensionality of these features substantially exceeded the sample size, applying ML models directly to such high-dimensional data can lead to the “curse of dimensionality” [144], which manifests as data sparsity and may compromise model generalizability. To address this, a robust FS strategy was implemented to remove redundant and irrelevant features before model development. Additionally, we emphasized evaluating feature stability during the FS process, since instability may lead to variability in identified biomarkers and thus undermine their clinical reproducibility.

As delineated in [144], the research team developed a Python package encompassing 33 FS methodologies, where supervised approaches were strategically employed for irrelevancy mitigation while unsupervised techniques targeted redundancy elimination. In the current investigation, we systematically identified four supervised methods—S1 (F-score [145]), S2 (T-score [144]), S3 (ReliefF [146]), S4 (Fisher-score [144])—and six unsupervised methods—U1 (Laplacian-score [144]), U2 (SPEC [144]), U3 (MCFS [144]), U4 (NDFS [144]), U5 (UDFS [144]), U6 (Pearson-score [144])—for experimental validation. To optimize both irrelevancy and redundancy reduction, we exhaustively generated 24 distinct combinations of supervised and unsupervised (SU) approaches. Furthermore, ensuring the stability of selected features across multiple iterations emerged as a critical methodological consideration. Following [147], we implemented a frequency-based stability metric to quantitatively assess the reproducibility of each SU combination throughout the experimental workflow. Specifically, let $\boldsymbol{Z}$ be a binary matrix of size $M \times d$:

$$\boldsymbol{Z} = \begin{bmatrix} z_{11} & \cdots & z_{1d} \\ \cdots & \cdots & \cdots \\ z_{M1} & \cdots & z_{Md} \end{bmatrix} \tag{5.1}$$

where each row represents the FS result of the d features in response to a SU combination in one try. As an example, $z_{Md} = 1$ means that the d-th feature in the M-th try is selected. Then stability based on the frequency-based criterion can be defined as

$$Stability(\boldsymbol{Z}) = 1 - \frac{\frac{1}{d}\sum_{f=1}^{d}\left[\frac{M}{M-1}\left(\frac{1}{M}\sum_{i=1}^{M} z_{if}\right)\left(1-\frac{1}{M}\sum_{i=1}^{M} z_{if}\right)\right]}{\frac{\frac{1}{M}\sum_{i=1}^{M}\sum_{f=1}^{d} z_{if}}{d}\left(1-\frac{\frac{1}{M}\sum_{i=1}^{M}\sum_{f=1}^{d} z_{if}}{d}\right)} \tag{5.2}$$

This metric ranges from 0 to 1, with higher values indicating better stability. In this study, the SU combination with the highest product of AUC and stability (AUC × Stability) was selected as the final method for feature reduction.

5.2.5 Modelling

In multicenter investigations, inherent heterogeneity in data acquisition protocols and standardization criteria across participating institutions may induce significant discrepancies in both conditional probability distributions and marginal distributions across centers. Such distributional divergence poses substantial challenges to the cross-institutional generalizability of predictive models. To address this fundamental limitation, we propose a novel MDA framework that implements dynamic domain-invariant feature alignment through distribution adaptation mechanisms. This architecture specifically optimizes the joint distribution across different centers in the training cohort, with the explicit objective of enhancing external validation performance on unseen multicenter datasets.

Suppose we have a training cohort containing two centers for model training, the first center is denoted as $D^{s1} = \left\{x_i^{s1}, y_i^{s1}\right\}_{i=1}^{n}$ and the second center is denoted as $D^{s2} = \left\{x_i^{s2}, y_i^{s2}\right\}_{j=1}^{m}$, where x_i^{s1} (x_j^{s2}) represents the feature vector and y_i^{s1} (y_j^{s2}) represents the corresponding label. We assume that their marginal probability distribution and the conditional probability distribution, namely $P(x_i^{s1}) \neq P(x_j^{s2})$ and $P(y_i^{s1}|x_i^{s1}) \neq P(y_j^{s2}|x_j^{s2})$. The purpose of MDA is to attempt to minimize the marginal probability distribution and the conditional probability distribution across different centers. In this study, MDA can be formulated by minimizing the following objectives,

$$\begin{aligned} &\min\{diff(D^{s1}, D^{s2})\} \\ &= \min\{(1-\mu)diff\big(P(x^{s1}), P(x^{s2})\big) + \mu diff(P\big(y^{s1}|x^{s1}\big), P\big(y^{s2}|x^{s2}\big))\} \end{aligned} \tag{5.3}$$

where $diff(D^{s1}, D^{s2})$ is a function used to measure the distribution difference, $\mu \in [0, 1]$ is a balanced parameter. When the parameter μ approaches 0, it indicates that D^{s1} and D^{s2} are more dissimilar. In this case, the marginal distribution becomes more dominant.

Conversely, when the parameter μ approaches 1, it reveals that D^{s1} and D^{s2} are more similar. In this scenario, conditional distribution becomes more important to adapt. The balance parameter μ is determined empirically by grid search over [0, 1] with a step of 0.1, using a validation set split from the training data. The optimal value was selected based on the one that yielded the highest average AUC across validation folds, thus ensuring an effective trade-off between adapting to marginal (scanner/protocol-related) and conditional (disease-specific) distribution shifts. With MMD, Eq. (3) can be further objectivized as follows,

$$
\begin{aligned}
& min\{diff(D^{s1}, D^{s2})\} \\
&= min\left\{(1-\mu)\left\|\frac{1}{n}\sum_{i=1}^{n} f(x_i^{s1}) - \frac{1}{m}\sum_{j=1}^{m} f(x_j^{s2})\right\|_F^2 \right. \\
&\quad \left. + \mu \sum_{c=1}^{C}\left\|\frac{1}{n_c}\sum_{x_i^{s1}\in D_{(c)}^{s1}} f(x_i^{s1}) - \frac{1}{m_c}\sum_{x_i^{s2}\in D_{(c)}^{s2}} f(x_i^{s2})\right\|_F^2\right\}
\end{aligned}
\tag{5.4}
$$

where $f(\cdot)$ is a classifier, $c \in \{1,2,\ldots,C\}$ represents the distinct class label, $D_{(c)}^{s1}$ and $D_{(c)}^{s2}$ represent the samples belonging to the c^{th} class from D^{s1} and D^{s2}. n_c and m_c are the sizes of $D_{(c)}^{s1}$ and $D_{(c)}^{s2}$. In this study, $f(\cdot)$ is set to $f(\mathbf{x}) = \mathbf{A}^T\mathbf{x}$, therefore, Eq. (4) can be specified as follows,

$$
\begin{aligned}
& \min_{\boldsymbol{A}}\{diff(D^{s1}, D^{s2})\} \\
&= \min_{\boldsymbol{A}}\left\{\boldsymbol{A}^T\boldsymbol{X}[(1-\mu)\boldsymbol{M}_0 + \mu\sum_{c=1}^{C}\boldsymbol{M}_c]\boldsymbol{X}\boldsymbol{A}^T\right\}
\end{aligned}
\tag{5.5}
$$

where $\mathbf{X} = [x_{1,}x_{2,\ldots,}x_{n,}x_{n+1,}x_{n+2,\ldots,}x_{n+m,}]^T$ represents the training cohort, $\mathbf{A}$ is the transformation (mapping) matrix, $\mathbf{M}_0$ and $\mathbf{M}_c$ are defined as follows,

$$\mathbf{M}_0 = \begin{cases} \dfrac{1}{n^2} & x_i \in D^{s1}\ and\ x_j \in D^{s1} \\ \dfrac{1}{m^2} & x_i \in D^{s2}\ and\ x_j \in D^{s2} \\ -\dfrac{1}{mn} & otherwise \end{cases} \tag{5.6}$$

$$\mathbf{M}_c = \begin{cases} \dfrac{1}{n_c^2} & x_i \in D^{s1}_{(c)}\ and\ x_j \in D^{s1}_{(c)} \\ \dfrac{1}{m_c^2} & x_i \in D^{s2}_{(c)}\ and\ x_j \in D^{s2}_{(c)} \\ -\dfrac{\mathbf{1}}{m_c n_c} & \begin{matrix} x_i \in D^{s1}_{(c)}, x_j \in D^{s2}_{(c)} \\ x_i \in D^{s2}_{(c)}, x_j \in D^{s1}_{(c)} \end{matrix} \\ 0 & otherwise \end{cases} \tag{5.7}$$

Therefore, finally, the classifier embedded with balanced distribution adaptation can be formulated as follows,

$$\min_{\boldsymbol{A}} \left\{ \|(\boldsymbol{A}^T\boldsymbol{X} - \boldsymbol{Y})\|_F^2 + \alpha\|\boldsymbol{A}\|_F^2 + \beta(\boldsymbol{A}^T\boldsymbol{X}((1-\mu)\boldsymbol{M}_0 + \mu\sum_{c=1}^{C}\boldsymbol{M}_c)\boldsymbol{X}^T\boldsymbol{A}) \right\} \tag{5.8}$$

where $\mathbf{Y}= [y_{1,}y_{2,\dots,}y_{n,}y_{n+1,}y_{n+2,\dots,}y_{n+m,}]^T$. α and β are two regularized parameters. Eq. (8) can be solved by introducing Lagrange multipliers, the closed-form solution of **A** is

$$\boldsymbol{A} = \left[\boldsymbol{X}^T\left(\beta\left((1-\mu)\boldsymbol{M}_0 + \mu\sum_{c=1}^{C}\boldsymbol{M}_c\right) + (\alpha+1)\boldsymbol{I}\right)\boldsymbol{X}\right]^{-1}\boldsymbol{X}^T\boldsymbol{Y} \tag{5.9}$$

where **I** is a $(n+m) \times (n+m)$ identity matrix. With **A,** for an unlabeled sample $\tilde{\mathbf{x}}$ in the external validation cohort, its prediction can be obtained by $\mathbf{A}^T\tilde{\mathbf{x}}$**.** Please note that MDA can be easily modified to fit the scenarios that the training cohort contains more than two centers.

5.2.6 Statistical Analysis

Model performance was evaluated for each ROI using multiple ML algorithms. For each iteration, 70% of patients in the training cohort were randomly selected for model fitting, and the remaining 30% were used for internal validation. This process was repeated for all seven

ROIs and each leave-one-center-out cross-validation split. The area under the receiver operating characteristic (AUC-ROC) curve was used as the primary metric for performance evaluation.

5.3 Results

5.3.1 Baseline characteristics of patients

A total of 528 patients (317 with TB, 211 with FP) from four medical centers (Centers A–D) were included in the analysis. Clinical and demographic characteristics for each center and disease subgroup are summarized in **Table 5.1**. Across the centers, patients with FP tended to be older than those with TB, with significant age differences observed in Centers A and C. In Center B, there were more males in the TB group than in the FP group. Other clinical variables were generally similar between groups. For model development and evaluation, a leave-one-center-out approach was used, with each center serving as the external validation cohort in turn.

Table 5.1. Summary of clinical characteristics of enrolled patients.

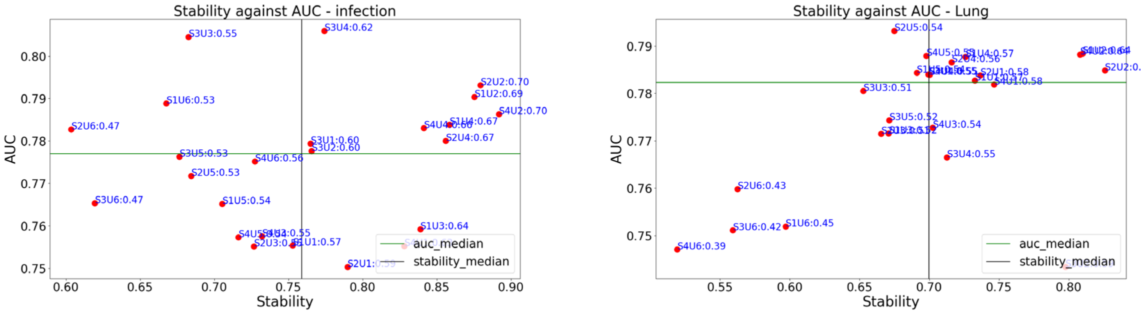
Dataset	**Age (years)**	**Gender (%)**	
		Male (%)	**Female (%)**
Center A (n=76)			
TB (n=34)	58.97±16.34	19 (55.88%)	15 (44.12%)
FP (n=42)	72.62±15.87	24 (57.14%)	18 (42.86%)
p-value	0.002	0.924	
Center B (n=255)			
TB (n=194)	51.19±15.73	161 (82.99%)	33 (17.01%)
FP (n=61)	52.25±14.46	36 (59.02%)	25 (40.98%)
p-value	0.667	0.001	
Center C (n=178)			
TB (n=59)	48.97±13.14	40 (67.80%)	19 (32.20%)
FP (n=119)	66.23±16.79	76 (63.87%)	43 (36.13%)
p-value	0.001	0.864	
Center D (n=19)			
TB (n=11)	44.4±16.92	6 (54.55%)	5 (45.45%)
FP (n=8)	61.75±25.28	4 (50.0%)	4 (50.0%)
p-value	0.102	0.997	

5.3.2 Results of Feature Selection

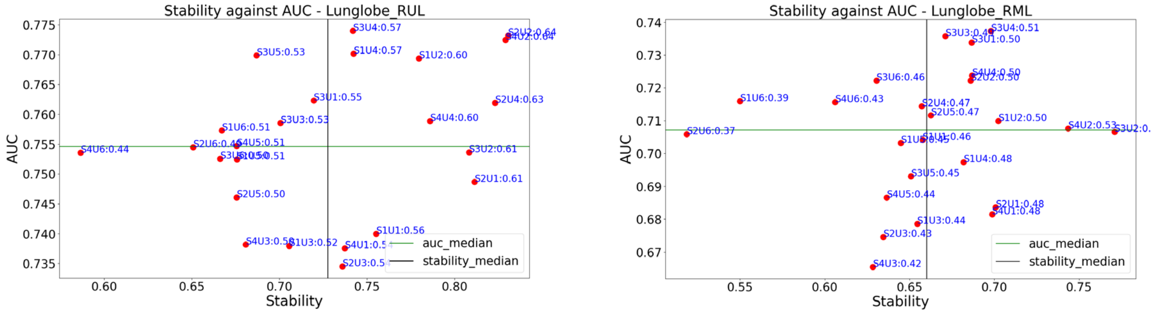
FS was performed to identify the most robust and informative radiomic features across different ROIs and external validation settings. As shown in **Figure 5.3**, we visualized the performance and stability of all SU FS combinations by plotting AUC versus stability for each of the ROIs, using the setting where Centers A, B, and D form the training cohort and Center C serves as the external validation cohort. For each ROI, the SU combination with the highest product of AUC and stability was selected. For example, on the infection ROI, the S2U2 combination (T-score + SPEC) achieved the optimal balance of model performance and feature stability and was thus chosen for downstream analysis.

To systematically evaluate the robustness of FS strategies under different cohort splits, we repeated this process by taking each center in turn as the external validation cohort. The optimal SU combinations determined for each scenario are summarized in **Table 5.2**. This highlights the variability in the best FS strategies across different ROIs and external validation settings.

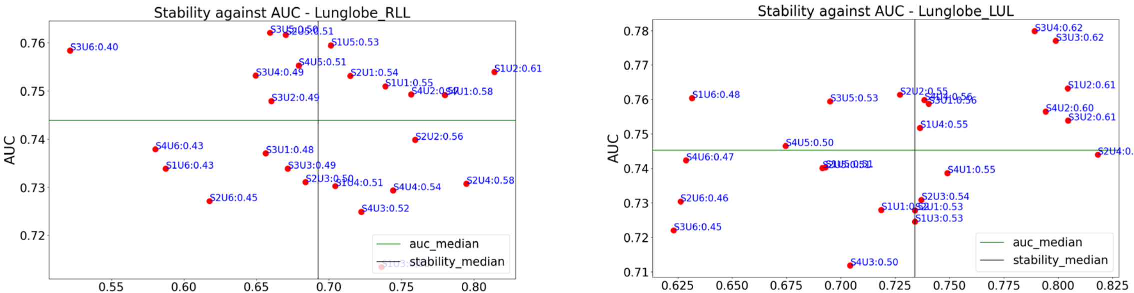
After determining the optimal SU combination for each ROI and cohort split, we applied the corresponding method to the training data to select features for model development. The importance of the selected features was further assessed using SHAP values, as depicted in **Figure 5.4** for the seven ROIs. These SHAP summary plots illustrate the contribution and impact of each radiomic feature on model prediction and reveal region-specific and cohort-specific differences in feature relevance.



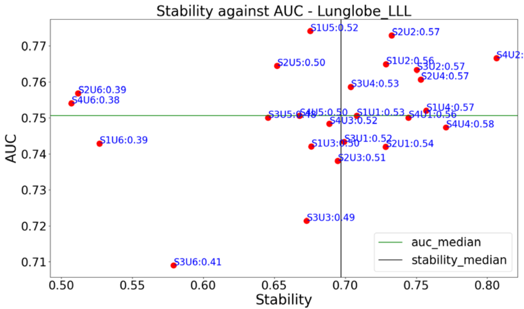
(a) Infection (b) Lung



(c) Right Upper Lobe (RUL) (d) Right Middle Lobe (RML)



(e) Right Lower Lobe (RLL) (f) Left Upper Lobe (LUL)



(g) Left Lower Lobe (LLL)

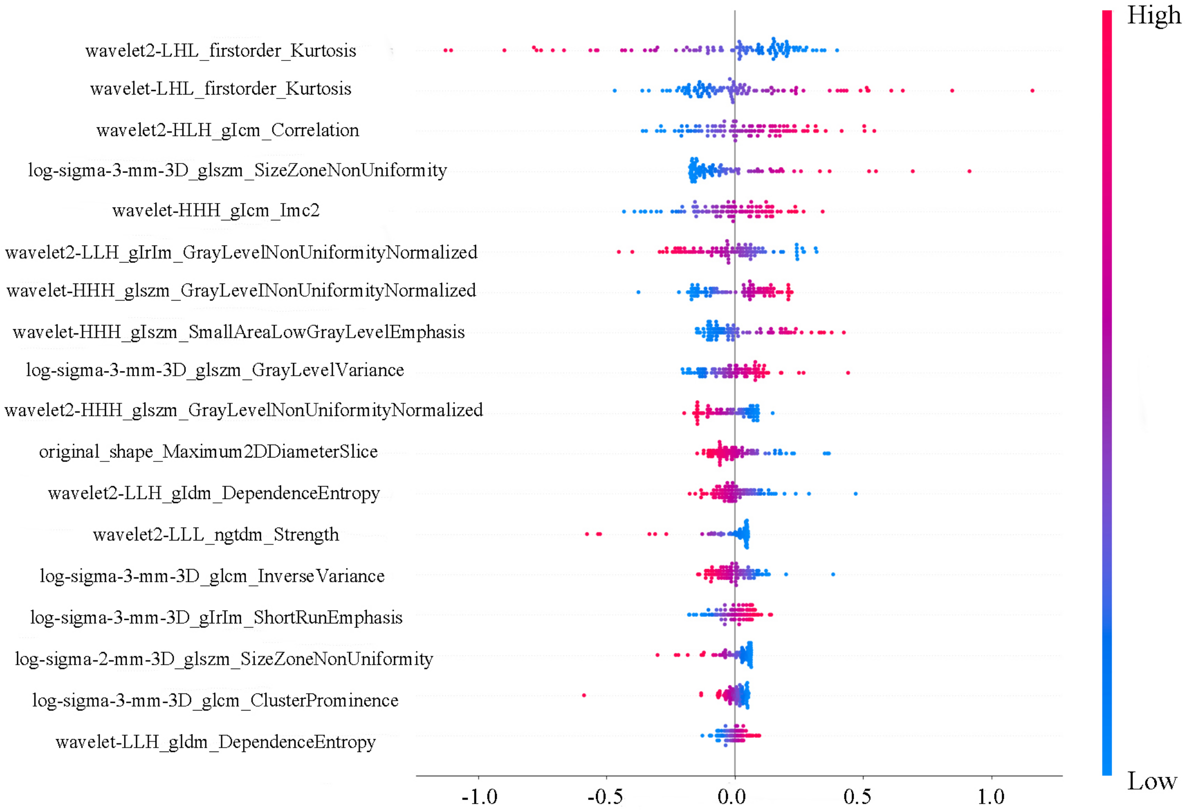
Figure 5.3. Scatter plots of AUC versus feature selection stability for ROIs. **(a)** Infection, **(b)** Full Lung, **(c)** Right Upper Lobe (RUL), **(d)** Right Middle Lobe (RML), **(e)** Right Lower Lobe (RLL), **(f)** Left Upper Lobe (LUL), and **(g)** Left Lower Lobe (LLL). Each point represents a different feature selection method. The green and black lines indicate the median AUC and median stability, respectively.

Table 5.2. SU combination determination in terms of AUC*Stability when Center A, B, C, and D are taken as the external validation cohort, respectively.

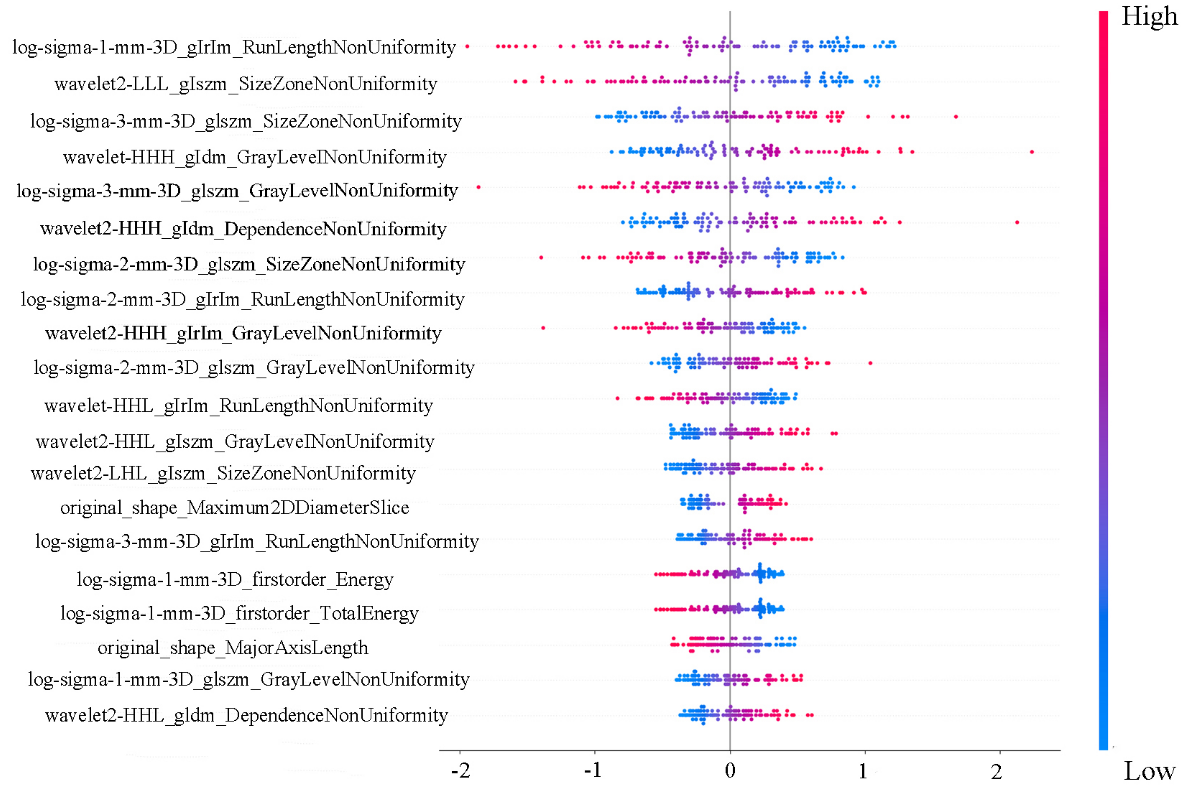
ROI	**BCD/A**	**ACD/B**	**ABD/C**	**ABC/D**
Infection	S3U2	S2U4	S2U2	S2U2
Full Lung	S1U2	S1U2	S2U2	S4U1
RUL	S1U2	S2U2	S2U2	S2U2
RML	S2U2	S1U2	S3U2	S1U2
RLL	S4U2	S2U2	S1U2	S1U4
LUL	S3U2	S1U2	S3U4	S3U2
LLL	S2U2	S2U2	S4U2	S2U2

BCD/A means Centers B, C, and D are taken as the training cohort, Center A is taken as the external validation cohort.

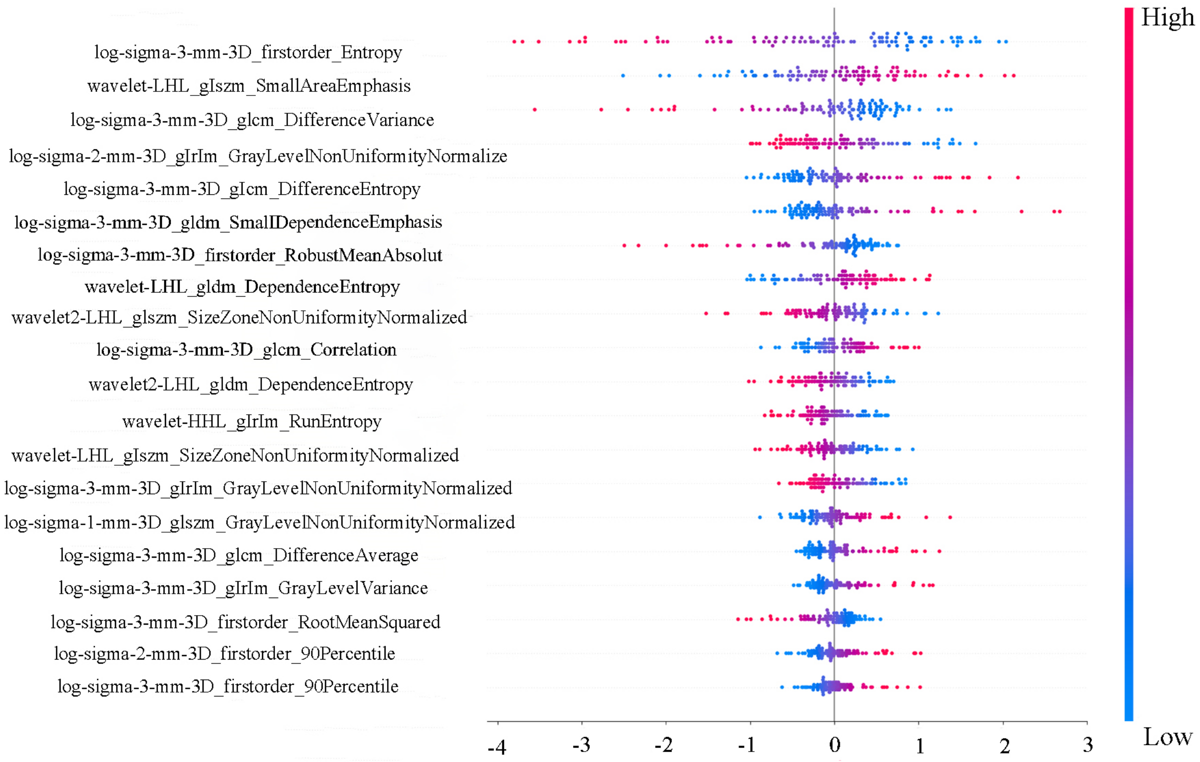
Abbreviation: RUL, right upper lobe; RML, right middle lobe; RLL, right lower lobe; LUL, left upper lobe; LLL, left lower lobe; S1, F-score; S2, T-score; S3, ReliefF; S4, Fisher-score, U1, Laplacian-score; U2, SPEC; U3, MCFS; U4, NDFS; U5, UDFS; U6, Person-score.



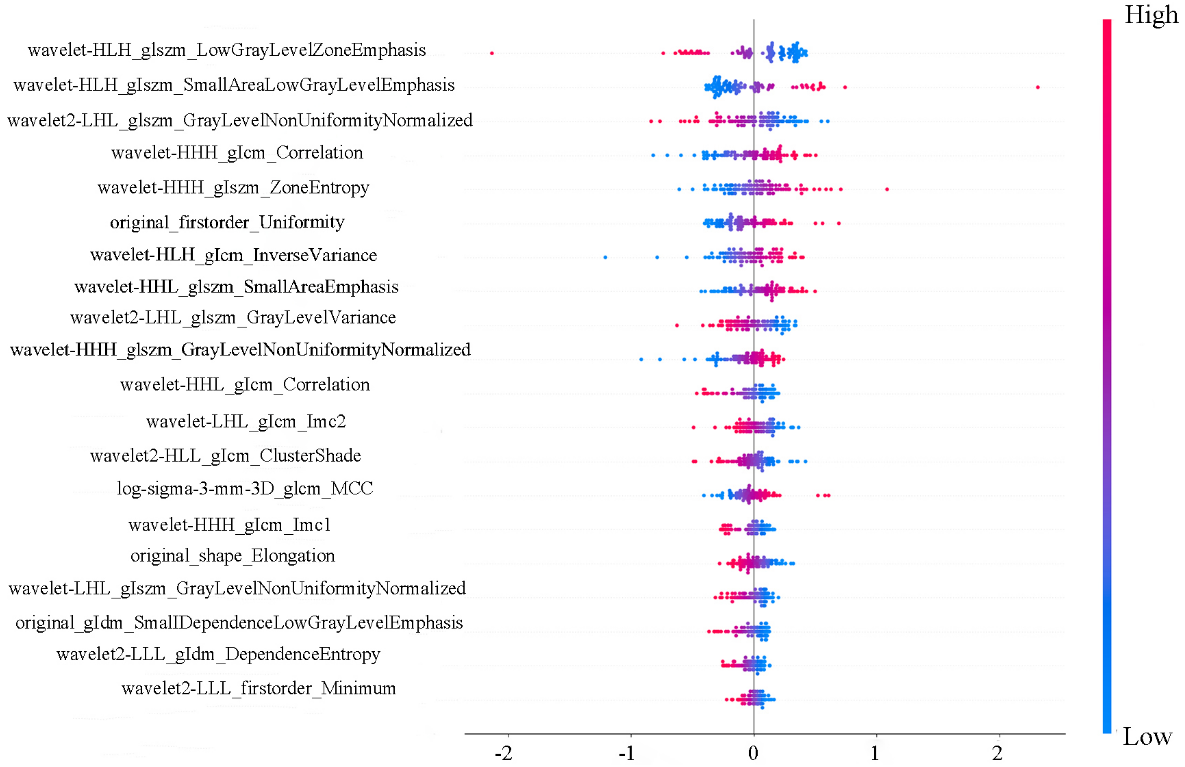
(a) Infection



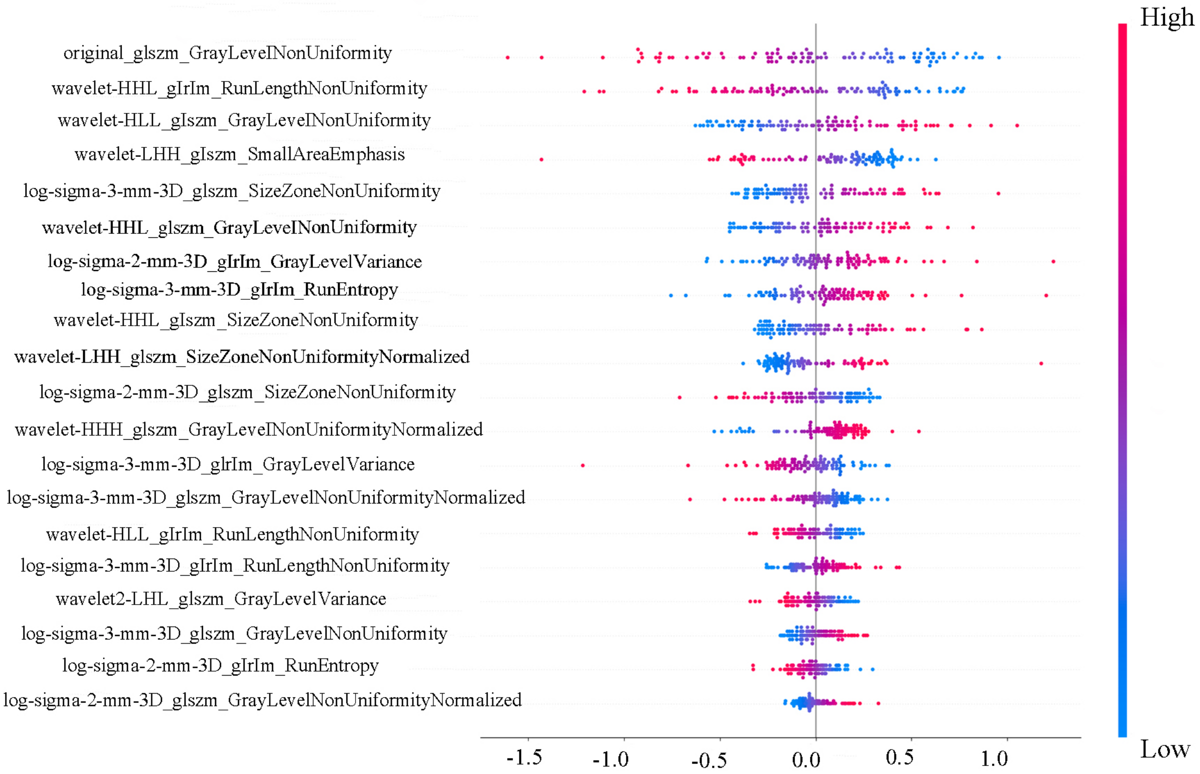
(b) Full Lung



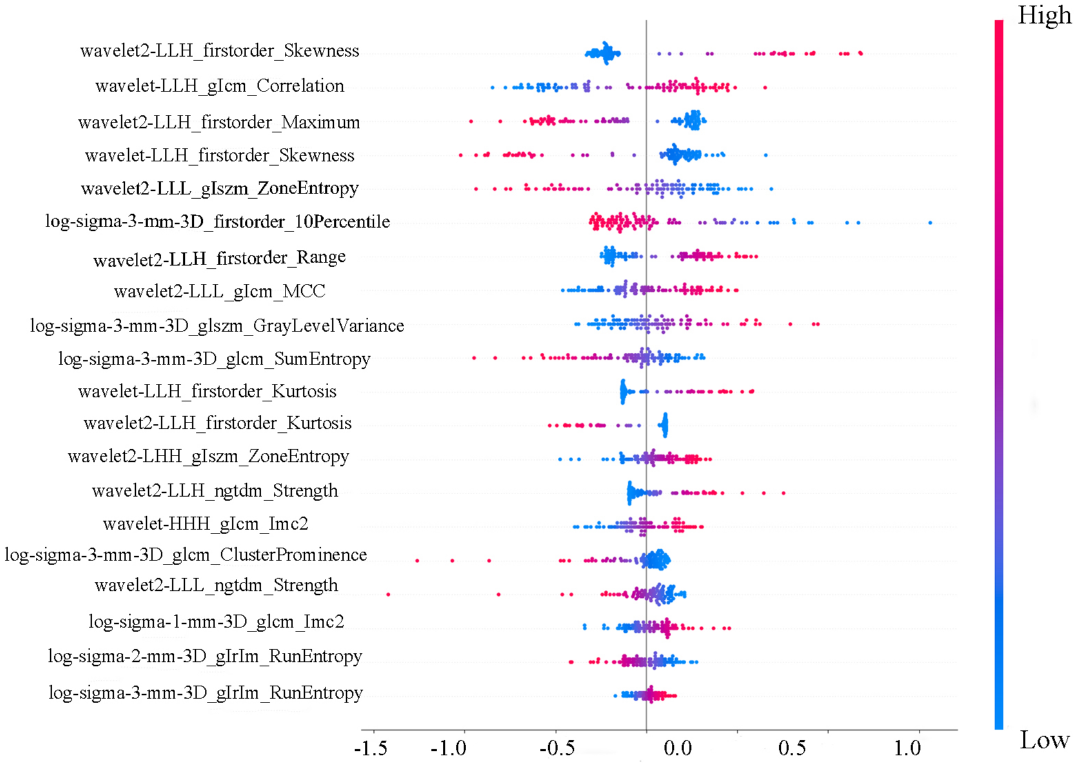
(c) Right Upper Lobe (RUL)



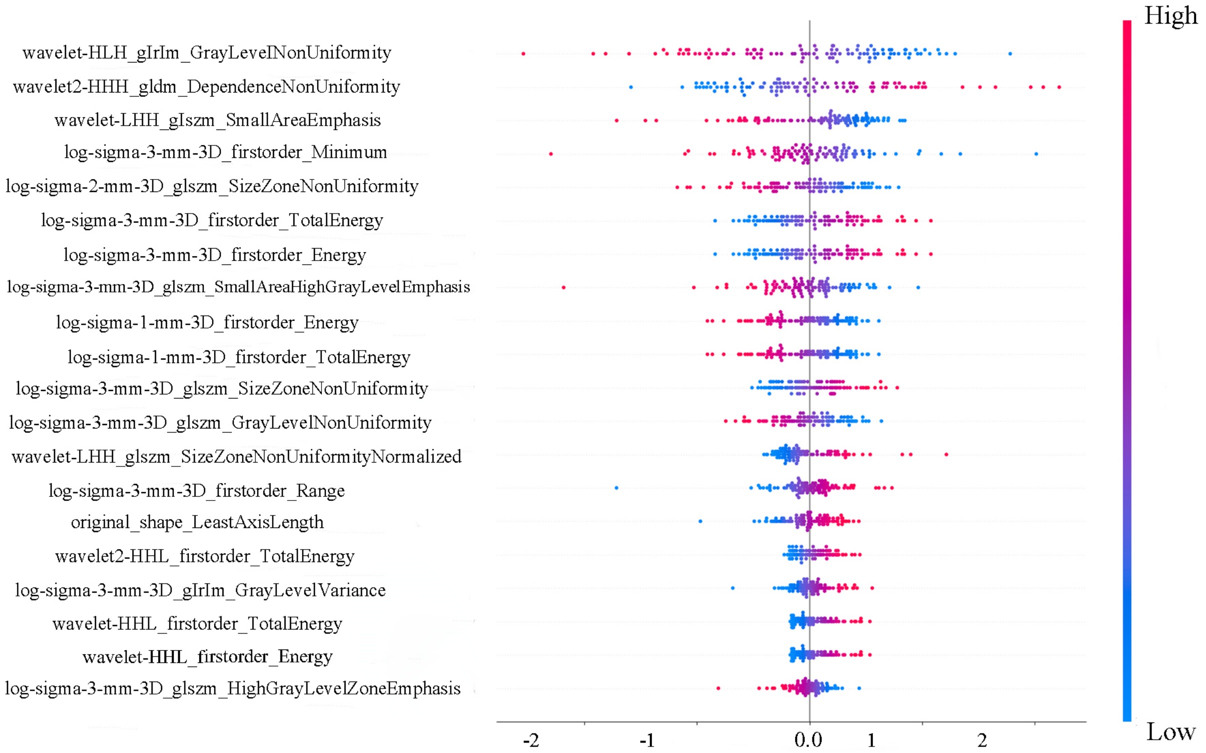
(d) Right Middle Lobe (RML)



(e) Right Lower Lobe (RLL)



(f) Left Upper Lobe (LUL)



(g) Left Lower Lobe (LLL)

Figure 5.4. SHAP summary plots illustrate the importance and impact of the top radiomic features on model output for ROIs.

(a) Infection, **(b)** Full Lung, **(c)** Right Upper Lobe (RUL), **(d)** Right Middle Lobe (RML), **(e)** Right Lower Lobe (RLL), **(f)** Left Upper Lobe (LUL), and **(g)** Left Lower Lobe (LLL)

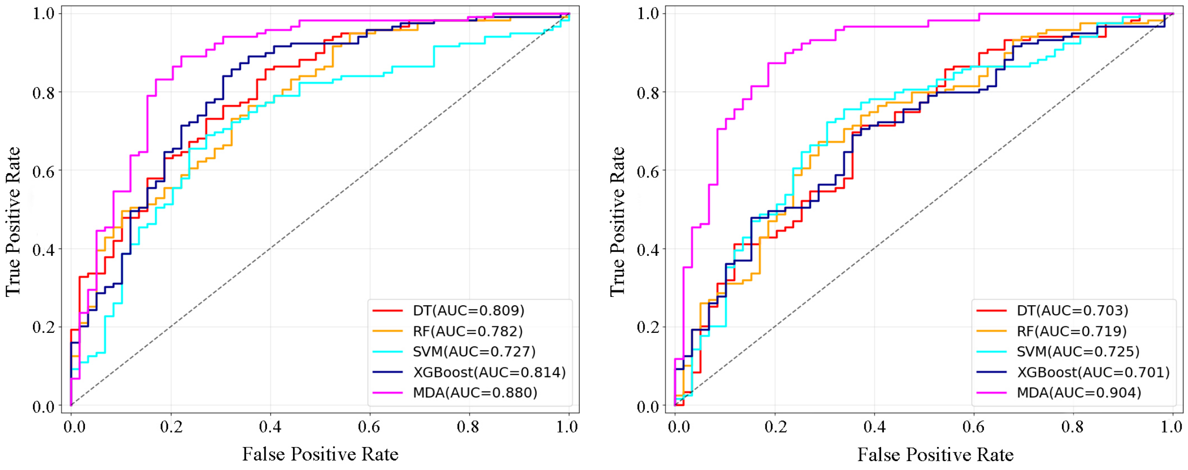
5.3.3 Performance of MDA

Table 5.3 summarizes the AUCs for each model and ROI in the training, internal validation, and external validation cohorts (with Center C as the external validation cohort). Among all models, MDA consistently achieved the highest AUCs in both internal and external validation across most ROIs. For example, in the infection ROI, our MDA reached an external validation AUC of 0.881, outperforming SVM, RF, decision tree (DT), and XGBoost. Similar trends were observed for the remaining ROIs, with MDA generally outperforming traditional pipelines. **Figure 5.5** visualizes the ROC curves for the seven ROIs in the external validation cohort, illustrating that MDA provided superior discrimination compared to other models.

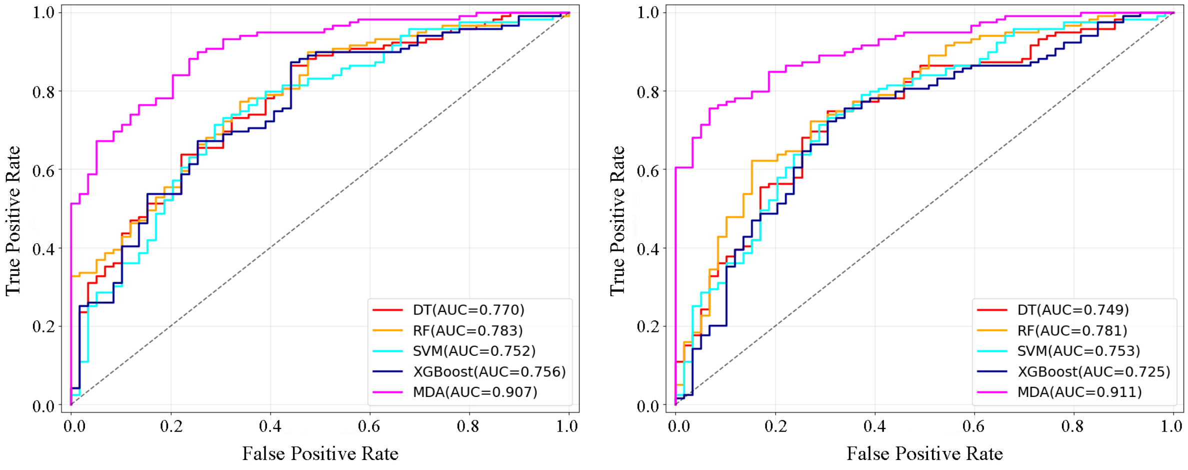
To further validate the generalizability of our models, each center was, in turn, used as the external validation cohort while the other three served as the training set. The results of these cross-center external validations are reported in **Tables 5.4–5.6**. Across all cross-validation scenarios, Our MDA remained the best-performing model, with consistently higher AUC values than the other algorithms.

Table 5.3. Classification performance in terms of AUC on the 7 ROIs (Center C is the external validation cohort).

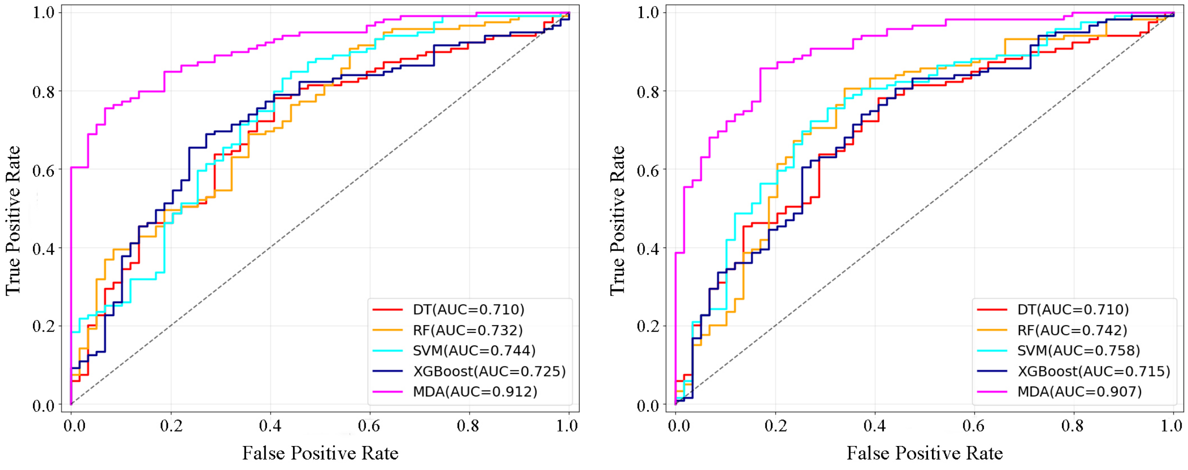
ROI	**Models**	**Training**			**Internal Validation**			**External Validation**		
		AUC (95% CI)	*p*-value	Corrected *p*-value	AUC (95% CI)	*p*-value	Corrected *p*-value	AUC (95% CI)	*p*-value	Corrected *p*-value
Infection	DT	0.856	<0.001	<0.001	0.765	<0.001	<0.001	0.809	<0.001	<0.001
	RF	0.907	0.021	0.035	0.827	<0.001	<0.001	0.779	<0.001	<0.001
	SVM	0.886	<0.001	<0.001	0.812	<0.001	<0.001	0.727	<0.001	<0.001
	XGBoost	0.875	<0.001	<0.001	0.751	<0.001	<0.001	0.811	<0.001	<0.001
	MDA	**0.927**	**-**	**-**	**0.915**	**-**	**-**	**0.881**	**-**	**-**
Full Lung	DT	0.889	<0.001	<0.001	0.709	<0.001	<0.001	0.703	<0.001	<0.001
	RF	0.865	<0.001	<0.001	0.744	<0.001	<0.001	0.719	<0.001	<0.001
	SVM	0.862	<0.001	<0.001	0.789	<0.001	<0.001	0.721	<0.001	<0.001
	XGBoost	0.851	<0.001	<0.001	0.731	<0.001	<0.001	0.701	<0.001	<0.001
	MDA	**0.939**	**-**	**-**	**0.928**	**-**	**-**	**0.901**	**-**	**-**
RUL	DT	0.900	<0.001	<0.001	0.741	<0.001	<0.001	0.770	<0.001	<0.001
	RF	0.859	<0.001	<0.001	0.769	<0.001	<0.001	0.783	<0.001	<0.001
	SVM	0.898	<0.001	<0.001	0.801	<0.001	<0.001	0.752	<0.001	<0.001
	XGBoost	0.865	<0.001	<0.001	0.763	<0.001	<0.001	0.752	<0.001	<0.001
	MDA	**0.950**	**-**	**-**	**0.927**	**-**	**-**	**0.906**	**-**	**-**
RML	DT	0.908	<0.001	<0.001	0.734	<0.001	<0.001	0.749	<0.001	<0.001
	RF	0.869	<0.001	<0.001	0.762	<0.001	<0.001	0.781	<0.001	<0.001
	SVM	0.881	<0.001	<0.001	0.809	<0.001	<0.001	0.756	<0.001	<0.001
	XGBoost	0.861	<0.001	<0.001	0.767	<0.001	<0.001	0.728	<0.001	<0.001
	MDA	**0.948**	**-**	**-**	**0.932**	**-**	**-**	**0.910**	**-**	**-**
RLL	DT	0.906	<0.001	<0.001	0.721	<0.001	<0.001	0.710	<0.001	<0.001
	RF	0.867	<0.001	<0.001	0.749	<0.001	<0.001	0.735	<0.001	<0.001
	SVM	0.876	<0.001	<0.001	0.792	<0.001	<0.001	0.741	<0.001	<0.001
	XGBoost	0.857	<0.001	<0.001	0.754	<0.001	<0.001	0.728	<0.001	<0.001
	MDA	**0.962**	**-**	**-**	**0.932**	**-**	**-**	**0.910**	**-**	**-**
LUL	DT	0.880	<0.001	<0.001	0.797	<0.001	<0.001	0.710	<0.001	<0.001
	RF	0.858	<0.001	<0.001	0.801	<0.001	<0.001	0.740	<0.001	<0.001
	SVM	0.875	<0.001	<0.001	0.812	<0.001	<0.001	0.759	<0.001	<0.001
	XGBoost	0.821	<0.001	<0.001	0.721	<0.001	<0.001	0.711	<0.001	<0.001
	MDA	**0.941**	**-**	**-**	**0.921**	**-**	**-**	**0.903**	**-**	**-**
LLL	DT	0.879	<0.001	<0.001	0.749	<0.001	<0.001	0.765	<0.001	<0.001
	RF	0.848	<0.001	<0.001	0.767	<0.001	<0.001	0.774	<0.001	<0.001
	SVM	0.890	<0.001	<0.001	0.798	<0.001	<0.001	0.762	<0.001	<0.001
	XGBoost	0.878	<0.001	<0.001	0.765	<0.001	<0.001	0.752	<0.001	<0.001
	MDA	**0.947**	**-**	**-**	**0.932**	**-**	**-**	**0.913**	**-**	**-**



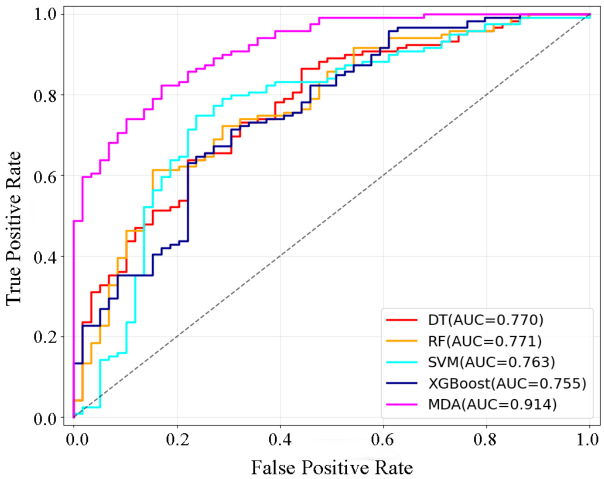
(a) Infection (b) Full Lung



(c) Right Upper Lobe (RUL) (d) Right Middle Lobe (RML)



(e) Right Lower Lobe (RLL) (f) Left Upper Lobe (LUL)



(g) Left Lower Lobe (LLL)

Figure 5.5. Receiver operating characteristic (ROC) curves for ROIs.

(a) Infection, **(b)** Full Lung, **(c)** Right Upper Lobe (RUL), **(d)** Right Middle Lobe (RML), **(e)** Right Lower Lobe (RLL), **(f)** Left Upper Lobe (LUL), and **(g)** Left Lower Lobe (LLL). Each plot compares the performance of different classification models, with the AUC for each model indicated in the legend. The dashed line represents random classification.

Table 5.4. Classification performance in terms of AUC on the 7 ROIs (Center A is the external validation cohort).

ROI	**Models**	**Training**			**Internal Validation**			**External Validation**		
		AUC (95% CI)	*p*-value	Corrected *p*-value	AUC (95% CI)	*p*-value	Corrected *p*-value	AUC (95% CI)	*p*-value	Corrected *p*-value
Infection	DT	0.898	<0.001	<0.001	0.868	<0.001	<0.001	0.804	<0.001	<0.001
	RF	0.903	0.015	0.018	0.887	<0.001	<0.001	0.828	<0.001	<0.001
	SVM	0.883	<0.001	<0.001	0.862	<0.001	<0.001	0.797	<0.001	<0.001
	XGBoost	0.906	0.014	0.017	0.860	<0.001	<0.001	0.828	<0.001	<0.001
	MDA	**0.933**	**-**	**-**	**0.928**	**-**	**-**	**0.913**	**-**	**-**
Full Lung	DT	0.874	<0.001	<0.001	0.857	<0.001	<0.001	0.796	<0.001	<0.001
	RF	0.881	<0.001	<0.001	0.850	<0.001	<0.001	0.780	<0.001	<0.001
	SVM	0.893	<0.001	<0.001	0.879	<0.001	<0.001	0.805	<0.001	<0.001
	XGBoost	0.889	<0.001	<0.001	0.824	<0.001	<0.001	0.784	<0.001	<0.001
	MDA	**0.952**	**-**	**-**	**0.928**	**-**	**-**	**0.916**	**-**	**-**
RUL	DT	0.909	<0.001	<0.001	0.871	<0.001	<0.001	0.832	<0.001	<0.001
	RF	0.921	<0.001	<0.001	0.900	0.014	0.034	0.867	0.023	0.035
	SVM	0.943	<0.001	<0.001	0.907	<0.001	<0.001	0.809	<0.001	<0.001
	XGBoost	0.902	<0.001	<0.001	0.876	<0.001	<0.001	0.832	<0.001	<0.001
	MDA	**0.971**	**-**	**-**	**0.920**	**-**	**-**	**0.885**	**-**	**-**
RML	DT	0.870	<0.001	<0.001	0.849	<0.001	<0.001	0.800	<0.001	<0.001
	RF	0.900	<0.001	<0.001	0.876	<0.001	<0.001	0.806	<0.001	<0.001
	SVM	0.906	<0.001	<0.001	0.863	<0.001	<0.001	0.806	<0.001	<0.001
	XGBoost	0.911	<0.001	<0.001	0.857	<0.001	<0.001	0.811	<0.001	<0.001
	MDA	**0.954**	**-**	**-**	**0.912**	**-**	**-**	**0.884**	**-**	**-**
RLL	DT	0.909	<0.001	<0.001	0.879	<0.001	<0.001	0.813	<0.001	<0.001
	RF	0.922	<0.001	<0.001	0.880	<0.001	<0.001	0.812	<0.001	<0.001
	SVM	0.905	<0.001	<0.001	0.880	<0.001	<0.001	0.831	<0.001	<0.001
	XGBoost	0.904	<0.001	<0.001	0.875	<0.001	<0.001	0.805	<0.001	<0.001
	MDA	**0.978**	**-**	**-**	**0.941**	**-**	**-**	**0.918**	**-**	**-**
LUL	DT	0.921	<0.001	<0.001	0.899	<0.001	<0.001	0.823	<0.001	<0.001
	RF	0.922	<0.001	<0.001	0.908	<0.001	<0.001	0.866	<0.001	<0.001
	SVM	0.945	<0.001	<0.001	0.921	<0.001	<0.001	0.893	<0.001	<0.001
	XGBoost	0.904	<0.001	<0.001	0.878	<0.001	<0.001	0.829	<0.001	<0.001
	MDA	**0.982**	**-**	**-**	**0.965**	**-**	**-**	**0.925**	**-**	**-**
LLL	DT	0.924	<0.001	<0.001	0.882	<0.001	<0.001	0.804	<0.001	<0.001
	RF	0.912	<0.001	<0.001	0.878	<0.001	<0.001	0.801	<0.001	<0.001
	SVM	0.886	<0.001	<0.001	0.841	<0.001	<0.001	0.789	<0.001	<0.001
	XGBoost	0.901	<0.001	<0.001	0.868	<0.001	<0.001	0.790	<0.001	<0.001
	MDA	**0.951**	**-**	**-**	**0.924**	**-**	**-**	**0.895**	**-**	**-**

Table 5.5. Classification performance in terms of AUC on the 7 ROIs (Center B is the external validation cohort).

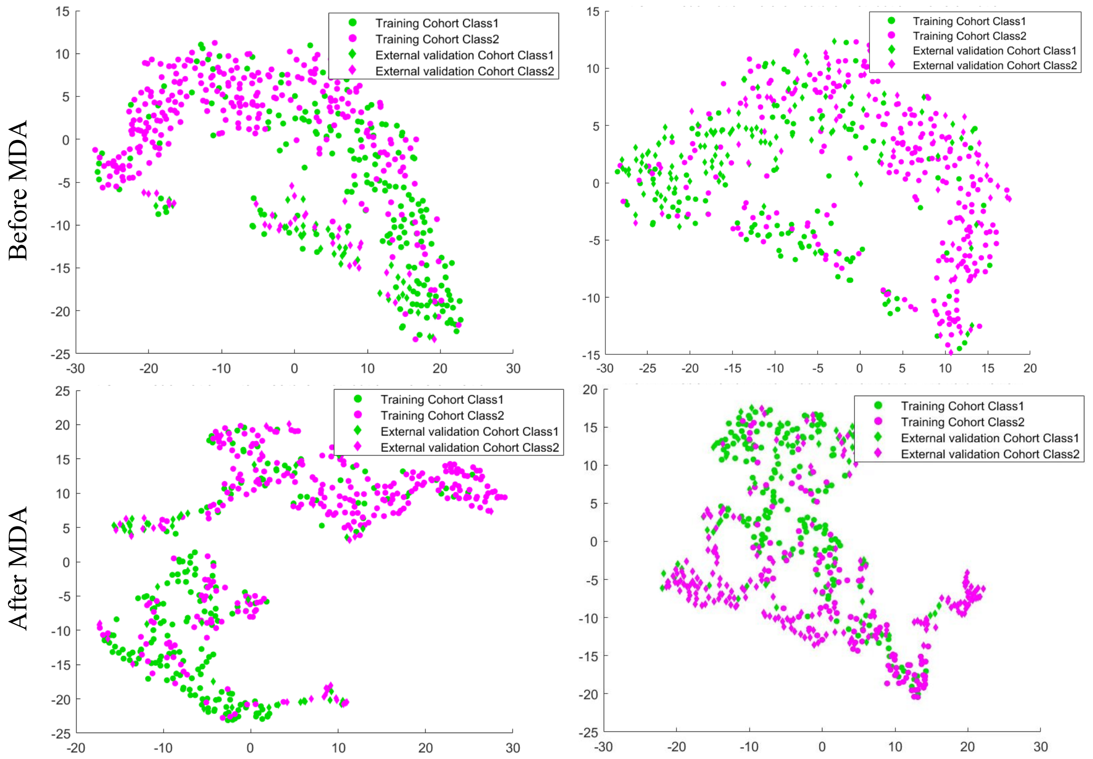
ROI	Models	**Training**			**Internal Validation**			**External Validation**		
		AUC (95% CI)	*p*-value	Corrected *p*-value	AUC (95% CI)	*p*-value	Corrected *p*-value	AUC (95% CI)	*p*-value	Corrected *p*-value
Infection	DT	0.849	<0.001	<0.001	0.818	<0.001	<0.001	0.751	<0.001	<0.001
	RF	0.916	<0.001	<0.001	0.879	<0.001	<0.001	0.823	<0.001	<0.001
	SVM	0.904	<0.001	<0.001	0.869	<0.001	<0.001	0.807	<0.001	<0.001
	XGBoost	0.866	<0.001	<0.001	0.833	<0.001	<0.001	0.811	<0.001	<0.001
	MDA	**0.958**	**-**	**-**	**0.932**	**-**	**-**	**0.905**	**-**	**-**
Full Lung	DT	0.856	<0.001	<0.001	0.809	<0.001	<0.001	0.728	<0.001	<0.001
	RF	0.892	<0.001	<0.001	0.821	<0.001	<0.001	0.764	<0.001	<0.001
	SVM	0.897	<0.001	<0.001	0.843	<0.001	<0.001	0.775	<0.001	<0.001
	XGBoost	0.908	<0.001	<0.001	0.825	<0.001	<0.001	0.756	<0.001	<0.001
	MDA	**0.930**	**-**	**-**	**0.920**	**-**	**-**	**0.894**	**-**	**-**
RUL	DT	0.890	<0.001	<0.001	0.854	<0.001	<0.001	0.808	<0.001	<0.001
	RF	0.908	<0.001	<0.001	0.860	<0.001	<0.001	0.810	<0.001	<0.001
	SVM	0.901	<0.001	<0.001	0.859	<0.001	<0.001	0.797	<0.001	<0.001
	XGBoost	0.890	<0.001	<0.001	0.841	<0.001	<0.001	0.798	<0.001	<0.001
	MDA	**0.936**	**-**	**-**	**0.909**	**-**	**-**	**0.879**	**-**	**-**
RML	DT	0.905	<0.001	<0.001	0.817	<0.001	<0.001	0.776	<0.001	<0.001
	RF	0.917	<0.001	<0.001	0.865	<0.001	<0.001	0.802	<0.001	<0.001
	SVM	0.902	<0.001	<0.001	0.843	<0.001	<0.001	0.803	<0.001	<0.001
	XGBoost	0.896	<0.001	<0.001	0.875	<0.001	<0.001	0.824	<0.001	<0.001
	MDA	**0.951**	**-**	**-**	**0.921**	**-**	**-**	**0.893**	**-**	**-**
RLL	DT	0.887	<0.001	<0.001	0.816	<0.001	<0.001	0.813	<0.001	<0.001
	RF	0.920	<0.001	<0.001	0.859	<0.001	<0.001	0.842	<0.001	<0.001
	SVM	0.901	<0.001	<0.001	0.872	<0.001	<0.001	0.824	<0.001	<0.001
	XGBoost	0.879	<0.001	<0.001	0.923	<0.001	<0.001	0.805	<0.001	<0.001
	MDA	**0.948**	**-**	**-**	**0.924**	**-**	**-**	**0.886**	**-**	**-**
LUL	DT	0.902	<0.001	<0.001	0.853	<0.001	<0.001	0.806	<0.001	<0.001
	RF	0.898	<0.001	<0.001	0.826	<0.001	<0.001	0.769	<0.001	<0.001
	SVM	0.890	<0.001	<0.001	0.818	<0.001	<0.001	0.754	<0.001	<0.001
	XGBoost	0.885	<0.001	<0.001	0.826	<0.001	<0.001	0.760	<0.001	<0.001
	MDA	**0.941**	**-**	**-**	**0.913**	**-**	**-**	**0.885**	**-**	**-**
LLL	DT	0.884	<0.001	<0.001	0.817	<0.001	<0.001	0.765	<0.001	<0.001
	RF	0.900	<0.001	<0.001	0.864	<0.001	<0.001	0.810	<0.001	<0.001
	SVM	0.904	<0.001	<0.001	0.858	<0.001	<0.001	0.780	<0.001	<0.001
	XGBoost	0.917	<0.001	<0.001	0.868	<0.001	<0.001	0.790	<0.001	<0.001
	MDA	**0.946**	**-**	**-**	**0.897**	**-**	**-**	**0.871**	**-**	**-**

Table 5.6. Classification performance in terms of AUC on the 7 ROIs (Center D is the external validation cohort).

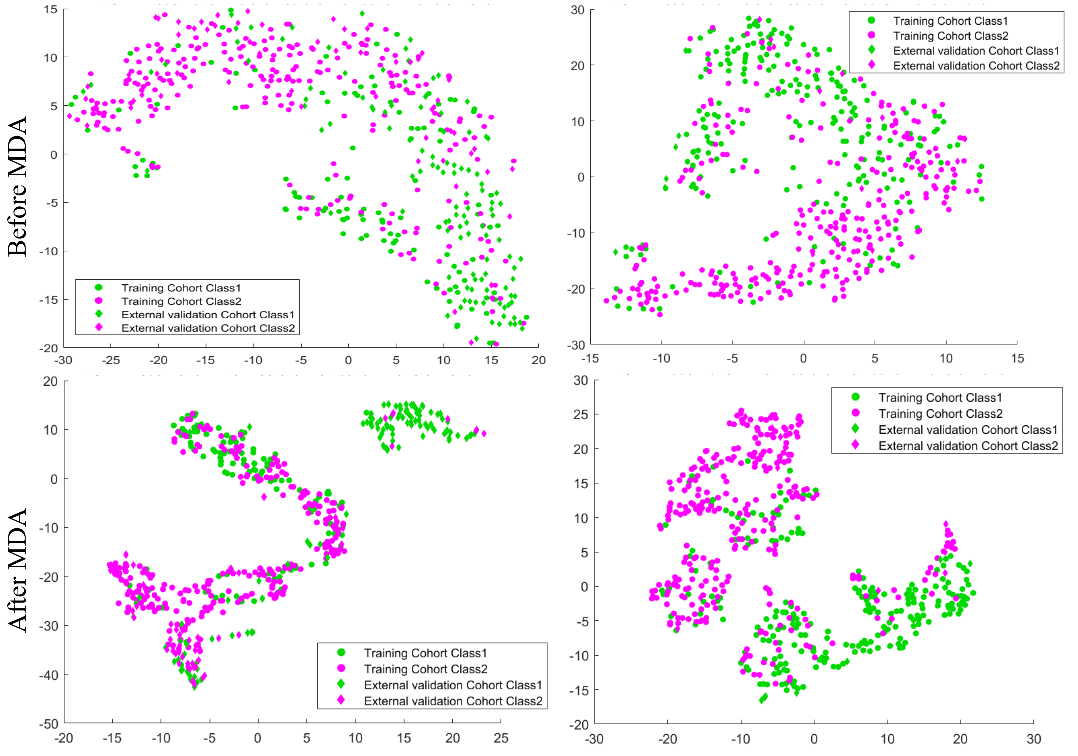
ROI	**Models**	**Training**			**Internal Validation**			**External Validation**		
		AUC (95% CI)	*p*-value	Corrected *p*-value	AUC (95% CI)	*p*-value	Corrected *p*-value	AUC (95% CI)	*p*-value	Corrected *p*-value
Infection	DT	0.887	<0.001	<0.001	0.786	<0.001	<0.001	0.830	<0.001	<0.001
	RF	0.931	0.045	0.049	0.848	<0.001	<0.001	0.761	<0.001	<0.001
	SVM	0.907	<0.001	<0.001	0.864	<0.001	<0.001	0.750	<0.001	<0.001
	XGBoost	0.907	0.004	0.007	0.793	<0.001	<0.001	0.807	<0.001	<0.001
	MDA	**0.933**	**-**	**-**	**0.929**	**-**	**-**	**0.920**	**-**	**-**
Full Lung	DT	0.906	<0.001	<0.001	0.720	<0.001	<0.001	0.693	<0.001	<0.001
	RF	0.888	<0.001	<0.001	0.767	<0.001	<0.001	0.705	<0.001	<0.001
	SVM	0.895	<0.001	<0.001	0.735	<0.001	<0.001	0.784	<0.001	<0.001
	XGBoost	0.879	<0.001	<0.001	0.768	<0.001	<0.001	0.716	<0.001	<0.001
	MDA	**0.922**	**-**	**-**	**0.918**	**-**	**-**	**0.909**	**-**	**-**
RLL	DT	0.906	<0.001	<0.001	0.704	<0.001	<0.001	0.784	<0.001	<0.001
	RF	0.944	<0.001	<0.001	0.819	<0.001	<0.001	0.864	<0.001	<0.001
	SVM	0.956	0.016	0.034	0.827	<0.001	<0.001	0.864	<0.001	<0.001
	XGBoost	0.884	<0.001	<0.001	0.717	<0.001	<0.001	0.875	<0.001	<0.001
	MDA	**0.987**	**-**	**-**	**0.981**	**-**	**-**	**0.977**	**-**	**-**
RUL	DT	0.872	<0.001	<0.001	0.749	<0.001	<0.001	0.852	<0.001	<0.001
	RF	0.931	<0.001	<0.001	0.909	<0.001	<0.001	0.909	<0.001	<0.001
	SVM	0.953	<0.001	<0.001	0.827	<0.001	<0.001	0.886	<0.001	<0.001
	XGBoost	0.865	<0.001	<0.001	0.865	<0.001	<0.001	0.841	<0.001	<0.001
	MDA	**0.985**	**-**	**-**	**0.980**	**-**	**-**	**0.977**	**-**	**-**
RML	DT	0.886	<0.001	<0.001	0.648	<0.001	<0.001	0.852	<0.001	<0.001
	RF	0.952	<0.001	<0.001	0.802	<0.001	<0.001	0.875	<0.001	<0.001
	SVM	0.951	<0.001	<0.001	0.862	<0.001	<0.001	0.830	<0.001	<0.001
	XGBoost	0.845	<0.001	<0.001	0.678	<0.001	<0.001	0.830	<0.001	<0.001
	MDA	**0.996**	**-**	**-**	**0.972**	**-**	**-**	**0.955**	**-**	**-**
LUL	DT	0.871	<0.001	<0.001	0.775	<0.001	<0.001	0.761	<0.001	<0.001
	RF	0.923	0.046	0.054	0.841	<0.001	<0.001	0.750	<0.001	<0.001
	SVM	0.836	<0.001	<0.001	0.746	<0.001	<0.001	0.716	<0.001	<0.001
	XGBoost	0.906	<0.001	<0.001	0.723	<0.001	<0.001	0.784	<0.001	<0.001
	MDA	**0.932**	**-**	**-**	**0.925**	**-**	**-**	**0.920**	**-**	**-**
LLL	DT	0.874	<0.001	<0.001	0.706	<0.001	<0.001	0.818	<0.001	<0.001
	RF	0.926	<0.001	<0.001	0.807	<0.001	<0.001	0.852	<0.001	<0.001
	SVM	0.951	0.032	0.046	0.786	<0.001	<0.001	0.761	<0.001	<0.001
	XGBoost	0.924	<0.001	<0.001	0.786	<0.001	<0.001	0.841	<0.001	<0.001
	MDA	**0.961**	**-**	**-**	**0.944**	**-**	**-**	**0.932**	**-**	**-**

5.3.4 Visualization of Feature Distribution

To further assess the effect of DA on feature distribution, we employed t-SNE to visualize the distribution of radiomic features before and after DA across different centers. **Figure 5.6** presents the t-SNE maps, with each column corresponding to a scenario in which one center is used as the external validation cohort and the other three centers as the training cohort. The first row displays the feature distributions before DA. In these plots, green and magenta dots represent class 1 and class 2 samples, respectively; circles denote samples from the training cohort, while diamonds denote samples from the external validation cohort. As shown, there are clear distribution discrepancies between the training and external validation cohorts, indicating domain shift. The second row presents the feature distribution after DA. Following adaptation, the feature distributions of the external validation cohort are better aligned with those of the training cohort, as evidenced by greater overlap and reduced separation in the t-SNE space. This improvement demonstrates the effectiveness of the DA strategy in reducing the impact of inter-center variability.



(a) Center A (b) Center B



(c) Center C (d) Center D

Figure 5.6. t-SNE visualization of feature distributions for each center used as the external validation cohort.

5.4 Discussion

Addressing the critical challenge of domain shift in multicenter radiomics, this study successfully develops and validates a CT-based model, significantly enhanced by a novel feature-level MDA framework, to differentiate TB from FP. Our principal finding demonstrates the MDA-enhanced model's robust diagnostic performance, achieving high AUCs across both internal and external validation cohorts and notably outperforming traditional ML classifiers like SVM and Random Forest across most ROIs and leave-one-center-out cross-validation scenarios. Crucially, this approach effectively mitigated the impact of inter-center data variability, as visualized by t-SNE, and offered interpretable feature importance through SHAP analysis.

Furthermore, the SHAP analysis offers insights into the potential clinical relevance of key features. For instance, the consistent importance of texture features may reflect the underlying tissue heterogeneity in infectious lesions. In tuberculosis, this could correspond to caseous necrosis, whereas in fungal pneumonia, it might represent a mix of hyphal growth, hemorrhage, and inflammatory cell infiltration. Similarly, shape features such as sphericity and compactness, while less prominent, could relate to the distinct growth patterns of these pathogens. Although these interpretations are correlational, they provide a plausible link between the quantitative radiomic signatures and the known pathophysiology, enhancing the model's clinical plausibility.

This robust performance is particularly significant given the well-documented challenge of accurately and non-invasively differentiating TB from FP. Overlapping clinical and radiological presentations often hinder timely diagnosis, and while microbiological confirmation is the gold standard, it has limitations in sensitivity and speed. Radiomics extracts quantitative imaging biomarkers, with prior studies indicating its potential for infectious lung diseases. However, poor generalizability due to domain shift from varied CT protocols has been a critical barrier to clinical translation. Our study directly confronts this by explicitly accounting for inter-center heterogeneity. Indeed, while DA techniques have been explored in medical imaging, their specific application to differentiate TB and FP using multicenter radiomics, coupled with a rigorous feature-level adaptation and stability-focused FS, constitutes a key advancement presented in this work. The performance of our MDA model is highly encouraging, especially considering the multicenter validation.

Several strengths underpin this study. Firstly, the multicenter design involving four distinct centers provides a rigorous test of model generalizability. Secondly, the introduction of the MDA framework is a core methodological innovation, offering a robust solution to domain shift by successfully aligning feature distributions. Thirdly, our comprehensive FS, optimizing for both predictive performance and stability, enhances the reliability of the identified radiomic signatures. Furthermore, SHAP values provide valuable interpretability, highlighting key differentiating radiomic features, which is vital for clinical acceptance. The initial standardized ROI definition via automated segmentation also minimized operator variability.

Despite these strengths, several limitations warrant consideration. This retrospective analysis may be subject to inherent selection biases. While our cohort is multicenter, future studies with larger sample sizes would further enhance robustness and enable detailed subgroup analyses. The variability in CT acquisition parameters, though a realistic test for MDA, might leave residual effects. Accuracy of initial segmentations, despite review, can influence extracted features. Moreover, while MDA is effective, mathematical transformations may not always have direct biological correlations, a common challenge in complex ML, although feature importance methods like SHAP analysis provide some insight into the model's decision-making based on the adapted features. Finally, excluding patients with significant comorbidities, while necessary here, might limit applicability to more complex clinical presentations.

The clinical implications of these findings are substantial. An accurate, non-invasive MDA radiomics tool could significantly expedite the differential diagnosis of TB and FP, facilitating earlier, targeted treatment, particularly in resource-limited settings or when microbiological results are delayed. Future research should prioritize prospective validation of this MDA radiomics model in diverse populations and settings. Integrating these robust radiomic features with clinical data and other biomarkers could further elevate diagnostic accuracy. Exploring the biological underpinnings of SHAP-identified features may also offer deeper insights into imaging manifestations of TB and FP.

In conclusion, this multicenter study demonstrates that a CT-based radiomics model, when enhanced by a novel feature-level MDA framework, can effectively and robustly differentiate TB from FP. By directly addressing domain shift, our approach offers a valuable non-invasive tool to support clinical decision making and exemplifies a promising strategy for improving the generalizability of radiomics in complex diagnostic challenges.

CHAPTER 6 Discussion

This thesis presented two complementary AI approaches to improve pneumonia diagnosis using chest CT scans. The first part introduced the MPMT-Pneumo for classifying pneumonia as bacterial or non-bacterial. The second part developed the MDA radiomics framework for distinguishing TB from FP. Across both approaches, our key finding is that integrating diverse data sources and addressing data heterogeneity can substantially enhance diagnostic accuracy and generalizability in pneumonia classification.

6.1 Summary of Current Findings

In the MPMT-Pneumo model development, we combined information from multiple CT imaging planes with routine inflammatory biomarkers to differentiate BP from NBP. This multi-modal strategy mirrors how a clinician considers both imaging and laboratory findings together. Our results demonstrate that this integrated transformer-based model achieved excellent discrimination between BP and NBP. The model's area under the ROC curve (AUC) was approximately 0.87–0.90, indicating high overall accuracy in identifying the pneumonia etiology. Importantly, the sensitivity was high (around 88% for detecting bacterial cases) while maintaining comparably high specificity, meaning the model could catch most bacterial infections without over-calling non-bacterial cases as bacterial. This performance was consistent across the internal test data and an independent external cohort from a second hospital, showing balanced sensitivity and specificity across two distinct clinical sites. Notably, MPMT-Pneumo outperformed SOTA methods [52, 53], which had considerably lower AUC. The substantial improvement underscores the benefit of our approach: by fusing multi-view CT features with laboratory biomarkers, the model captured a more complete picture of the infection. Another practical advantage is that MPMT-Pneumo achieved this high performance without requiring contrast-enhanced scans or handcrafted image features. It leveraged standard non-contrast CT images and automatically learned relevant features, which suggests the model can be deployed with routine imaging data and minimal manual intervention. These findings confirm that a multi-plane transformer architecture, enriched with clinical data, can significantly improve differentiation of bacterial versus non-bacterial pneumonia. Clinically, this is important because it could help guide early therapy – for instance, identifying likely bacterial infections that need antibiotics versus viral or other pneumonias where antibiotics would be ineffective.

The second major contribution of this work is the development of a feature-level DA radiomics framework for classifying TB versus FP on CT. Differentiating TB from fungal infections is a challenging task, especially across different hospitals, because imaging appearances can overlap and scanner settings or patient demographics can vary by site. In this study, we extracted quantitative radiomic features from CT images of patients with TB or FP and introduced a novel DA technique to improve the model's robustness across multiple centers. Our approach MDA framework adjusted the radiomics feature distributions to be more consistent between institutions, tackling the problem of domain shift – the changes in data characteristics from one hospital to another. The model, enhanced with DA, showed strong performance in distinguishing TB from FP. We observed high AUC values in the range of roughly 0.80–0.85 on external validation sets, approaching the performance seen in internal validation (which was around 0.85–0.89). For example, when training on three hospitals and testing on a fourth, the domain-adapted model's AUC remained in the mid-0.80s, whereas without adaptation the performance would have dropped more sharply (in some cases into the 0.70s). This indicates that our MDA strategy successfully narrowed the gap between training and test domain performance. Sensitivity for detecting, say, TB cases remained high even when the model was applied to scans from an unseen center, illustrating improved generalizability. By addressing inter-center variability, the radiomics framework maintained reliable accuracy across four distinct clinical centers included in our study. Methodologically, this is a significant result – it demonstrates that feature-level DA can make a traditionally brittle radiomics model much more generalizable. Moreover, the multi-center design of this experiment itself is a strength: involving four separate hospitals provided a rigorous test of how well the AI model generalizes. Our principal finding in this part is that incorporating DA is essential for achieving robust cross-center performance in CT-based pneumonia classification. This is a direct response to a critical challenge in medical AI deployment: models often perform well on data from one hospital but fail when deployed elsewhere. Our framework shows a viable solution to this problem in the context of pulmonary infections. Clinically, a reliable radiomics model that works across centers could assist in early differentiation of TB vs. FP, which is crucial because the treatments for these two infections are very different. By improving diagnostic confidence and consistency across hospitals, this approach could help ensure patients get the correct treatment faster, even in centers that were not part of the original training data.

The contributions of both the MPMT-Pneumo model and the MDA radiomics framework advance the field of CT-based pneumonia differentiation. They illustrate two important strategies for building robust medical AI: (1) integrating multiple data modalities to boost diagnostic performance, and (2) explicitly handling domain differences to ensure generalization in diverse settings. The emphasis on multi-center evaluation throughout this thesis, whether testing the DL model on an independent hospital or evaluating the radiomics model across four centers, is a distinguishing aspect of our work. By assessing model performance across diverse clinical settings and patient populations, we have strengthened the evidence that these AI tools could maintain their accuracy in the real world, outside of a single controlled research environment. In summary, this thesis demonstrated how novel multi-plane

DL and domain-adapted radiomics techniques can complement each other to address complex pneumonia diagnoses. It provided new insights and methods that bridge the gap between algorithm development and practical clinical implementation, bringing us closer to AI solutions that are both high-performing and generalizable for respiratory infection diagnosis.

6.2 Limitations and Future Directions

Despite the encouraging results, several important limitations must be addressed. First, while the multi-center design is a key strength, the participating institutions were all urban tertiary hospitals with relatively similar patient populations and CT protocols. The generalizability of our models to more diverse clinical settings, such as community hospitals, rural clinics, or centers using different scanner manufacturers and protocols (e.g., low-dose CT), remains unverified. Future work should prioritize validation on a more heterogeneous collection of datasets to test the true robustness of the MPMT-Pneumo and MDA frameworks.

Second, the current work was conducted in a retrospective manner, which has inherent limitations. All CT scans and clinical data were collected from past records, and outcomes were known, which can introduce selection biases. For example, only patients who got certain tests were included, which might bias the sample toward more severe or atypical cases that underwent extensive workups. Also, retrospective models might inadvertently learn spurious correlations specific to the collected. To truly gauge real-world performance, prospective studies will be needed. A prospective trial could measure not only the model’s accuracy on new patients, but also how it influences clinician decision-making and patient outcomes. Such studies would provide stronger evidence of clinical utility and could uncover unanticipated issues when the system is used live.

Another limitation is the dependence on annotations and specific inputs in our modeling approach. The MPMT-Pneumo model relies on having certain lab biomarkers available alongside the CT; however, in practice, those lab tests might not always be taken prior to diagnosis or might vary in timing and quality. If a patient's biomarker data is missing or delayed, the model's multi-modal design could be less useful. Future versions of the model should be flexible. For example, able to produce a result from imaging alone when lab data aren't yet available, perhaps with a confidence measure or incorporating other readily available clinical information. On the radiomics side, our framework required that the ROI be identified for feature extraction. In this study, we had the benefit of radiologists' expertise and clinical records to confirm which areas of the lung corresponded to TB or fungal infection. This manual or semi-manual annotation of lesions can be time-consuming and subject to inter-observer variability. It also limits scalability, since not every new case will have a carefully delineated infection region beforehand. To mitigate this, future work could integrate automated lung and lesion segmentation using DL, so that radiomics features are extracted consistently without manual intervention. By combining an automated segmentation step with our DA pipeline, we could streamline the radiomics workflow and reduce the dependency on expert annotations.

Interpretability of the models is another area for improvement. The DL transformer in MPMT-Pneumo, while highly accurate, operates as a complex "black box," making it challenging to explain why it predicts a case as bacterial or not. Likewise, the radiomics model yields a score based on many features, and although radiomics features have semantic meaning, it's not immediately clear to clinicians how the model arrives at a TB vs fungal decision. This lack of transparent reasoning can hinder clinical trust and adoption. Future research should pursue deeper interpretability for both models.

A further set of limitations involves the practical barriers to deploying these models in a real-world clinical setting. A primary challenge is technical integration with existing hospital infrastructure. Our models were developed using curated, pre-processed datasets, but a production-grade system must seamlessly interface with a hospital's Picture Archiving and Communication System (PACS). This requires building a robust pipeline capable of ingesting raw DICOM images from a variety of scanner manufacturers, handling inconsistencies in metadata, and returning results in a format that integrates smoothly into the clinical workflow. Furthermore, any AI tool intended for diagnostic support is considered a medical device and requires stringent regulatory approval, such as FDA clearance in the United States or a CE mark in Europe. This process involves extensive validation, documentation of quality control, and risk management protocols that represent a substantial undertaking beyond the scope of this academic research but are a critical prerequisite for clinical use. Finally, the question of model maintenance must be addressed. While the MDA framework showed strong generalizability, deploying it at a new institution with unique patient demographics or imaging protocols may necessitate a site-specific recalibration or localized fine-tuning to ensure optimal performance. A strategy for ongoing performance monitoring and periodic model updates would be essential for long-term clinical viability and safety.

Building on these findings, the path forward involves several concrete steps toward clinical translation. The immediate priority is the design of a prospective, multi-center trial. This trial would enroll patients with suspected pneumonia across a diverse range of healthcare settings to validate the models' performance in a real-time, real-world workflow. Key endpoints would include not only diagnostic accuracy but also the impact on clinical outcomes, such as time to appropriate antibiotic initiation and length of hospital stay. A second critical direction is the integration of the AI models with clinical decision support systems. Instead of being a standalone tool, the model's output could be linked to institutional antibiotic stewardship guidelines. For instance, a high probability of bacterial pneumonia could trigger a recommendation for a specific first-line antibiotic, while a low probability could support a "watchful waiting" approach, thereby directly contributing to more rational antimicrobial use.

Finally, the scope of the models must be expanded. Future work will focus on extending the classification framework to differentiate among other critical pneumonia subtypes, such as specific viral etiologies or atypical pathogens. This would transform the current binary classifiers into a comprehensive diagnostic panel, significantly increasing their clinical value. This could be achieved by developing a hybrid model that leverages a DL backbone for automated lesion segmentation and feature extraction, followed by the domain-adapted radiomics framework to ensure robust, generalizable performance across multiple etiologies and clinical sites.

Our multi-plane DL model and domain-adapted radiomics framework have made meaningful advances in pneumonia differentiation, there remain clear pathways for refinement before they can be fully realized in clinical practice. The work done here provides a strong proof-of-concept that advanced AI methods can distinguish between different pneumonia etiologies with high accuracy and robustness. With continued research and collaboration between clinicians, radiologists, and researchers, these tools can be further improved and ultimately translated into routine healthcare settings, potentially leading to faster diagnoses, more targeted treatments, and better outcomes for patients with pneumonia around the world.

CHAPTER 7 Conclusion

This study offers new and important ways to help solve difficult and ongoing problems in using CT scans to tell different types of pneumonia apart.

The research successfully culminated in the development of the MPMT-Pneumo model. This model has demonstrated its effectiveness in distinguishing bacterial pneumonia from non-bacterial pneumonia through an innovative approach that synergistically combines multi-plane CT image analysis with readily available clinical inflammatory biomarkers. The strong diagnostic performance of MPMT-Pneumo, its consistency when tested across different medical centers, and its clear potential to support more judicious antibiotic stewardship collectively underscore its significant clinical promise. Furthermore, this thesis introduced and validated the MDA-Radiomics framework, a novel feature-level DA technique. This framework has proven to demonstrably enhance the robustness and generalizability of CT-based radiomics models for the complex task of differentiating tuberculous pneumonia from FP across multiple, diverse clinical sites. By effectively tackling the issue of domain shift, the MDA-Radiomics framework offers a valuable non-invasive tool to aid clinicians in making challenging diagnostic decisions.

The rigorous multi-center validation methodologies employed for these AI solutions constitute a significant aspect of this research. Such validation is paramount for developing AI models that demonstrate not only accuracy in controlled studies but also reliability, transferability, and trustworthiness in diverse real-world healthcare environments. Consequently, this thesis extends beyond the development of two specific models to offer a methodological framework for creating clinically relevant and generalizable AI in medical imaging. It advocates for a holistic approach encompassing advanced model architecture, explicit strategies for managing data heterogeneity (such as DA), comprehensive multi-source validation, and a dedicated focus on achieving model interpretability and practical utility. This research contributes to the growing body of evidence supporting the transformative potential of AI in medicine. By furnishing tools that facilitate more accurate, timely, and objective pneumonia diagnoses, this work aims to enhance clinical decision-making, promote more rational and personalized treatment strategies, and ultimately improve health outcomes for patients afflicted with respiratory infections. The AI tools presented are conceptualized as decision-support systems, intended to augment clinical expertise by providing rapid, objective, and data-driven insights, particularly in complex or ambiguous cases, rather than replacing the crucial role of clinicians in patient care.

The integration of AI into routine healthcare practice remains an evolving process, but the advancements presented herein lay a clear foundation for the next phase of development. Future work will be centered on clinical translation through prospective validation, integration with decision support tools like antibiotic stewardship programs, and expansion to a broader range of pneumonia etiologies. By rigorously addressing these next steps, the frameworks developed in this thesis can move from proof-of-concept to becoming an effective and reliable partner for clinicians in the global initiative to combat pneumonia and enhance respiratory health.

REFERENCE

1. (UNICEF) TUNCsF. Pneumonia 2024 [updated November 2024. Available from: https://data.unicef.org/topic/child-health/pneumonia/.

2. Infections GBDLR, Antimicrobial Resistance C. Global, regional, and national incidence and mortality burden of non-COVID-19 lower respiratory infections and aetiologies, 1990-2021: a systematic analysis from the Global Burden of Disease Study 2021. Lancet Infect Dis. 2024;24(9):974-1002. https://doi.org/10.1016/S1473-3099(24)00176-2

3. Olson G, Davis AM. Diagnosis and Treatment of Adults With Community-Acquired Pneumonia. JAMA. 2020;323(9):885-886. https://doi.org/10.1001/jama.2019.21118

4. Henig O, Yahav D, Leibovici L, et al. Guidelines for the treatment of pneumonia and urinary tract infections: evaluation of methodological quality using the Appraisal of Guidelines, Research and Evaluation II instrument. Clin Microbiol Infect. 2013;19(12):1106-1114. https://doi.org/10.1111/1469-0691.12348

5. Vaughn VM, Dickson RP, Horowitz JK, et al. Community-Acquired Pneumonia: A Review. JAMA. 2024;332(15):1282-1295. https://doi.org/10.1001/jama.2024.14796

6. Miron M, Blaj M, Ristescu AI, et al. Hospital-Acquired Pneumonia and Ventilator-Associated Pneumonia: A Literature Review. Microorganisms. 2024;12(1). https://doi.org/10.3390/microorganisms12010213

7. Waite S, Jeudy J, White CS. Acute lung infections in normal and immunocompromised hosts. Radiol Clin North Am. 2006;44(2):295-315, ix. https://doi.org/10.1016/j.rcl.2005.10.009

8. Gramegna A, Sotgiu G, Di Pasquale M, et al. Atypical pathogens in hospitalized patients with community-acquired pneumonia: a worldwide perspective. BMC Infect Dis. 2018;18(1):677. https://doi.org/10.1186/s12879-018-3565-z

9. Brown L, Rautemaa-Richardson R, Mengoli C, et al. Polymerase Chain Reaction on Respiratory Tract Specimens of Immunocompromised Patients to Diagnose Pneumocystis Pneumonia: A Systematic Review and Meta-analysis. Clin Infect Dis. 2024;79(1):161-168. https://doi.org/10.1093/cid/ciae239

10. Cilloniz C, Dela Cruz CS, Dy-Agra G, et al. World Pneumonia Day 2024: Fighting Pneumonia and Antimicrobial Resistance. Am J Respir Crit Care Med. 2024;210(11):1283-1285. https://doi.org/10.1164/rccm.202408-1540ED

11. Collaborators GBDAR. Global burden of bacterial antimicrobial resistance 1990-2021: a systematic analysis with forecasts to 2050. Lancet. 2024;404(10459):1199-1226. https://doi.org/10.1016/S0140-6736(24)01867-1

12. Regunath H, Oba Y. Community-Acquired Pneumonia. StatPearls. Treasure Island (FL)2025. Available from: https://www.ncbi.nlm.nih.gov/pubmed/28613500

13. Pahal P, Rajasurya V, Nguyen AD. Typical Bacterial Pneumonia. StatPearls. Treasure Island (FL)2025. Available from: https://www.ncbi.nlm.nih.gov/pubmed/30485000

14. Shen F, Zubair M, Sergi C. Sputum Analysis. StatPearls. Treasure Island (FL)2025. Available from: https://www.ncbi.nlm.nih.gov/pubmed/33085342

15. Murdoch DR, O'Brien KL, Driscoll AJ, et al. Laboratory methods for determining pneumonia etiology in children. Clin Infect Dis. 2012;54 Suppl 2:S146-152. https://doi.org/10.1093/cid/cir1073

16. Stralin K, Tornqvist E, Kaltoft MS, et al. Etiologic diagnosis of adult bacterial pneumonia by culture and PCR applied to respiratory tract samples. J Clin Microbiol. 2006;44(2):643-645. https://doi.org/10.1128/JCM.44.2.643-645.2006

17. Gadsby NJ, Russell CD, McHugh MP, et al. Comprehensive Molecular Testing for Respiratory Pathogens in Community-Acquired Pneumonia. Clin Infect Dis. 2016;62(7):817-823. https://doi.org/10.1093/cid/civ1214

18. Yang S, Lin S, Khalil A, et al. Quantitative PCR assay using sputum samples for rapid diagnosis of pneumococcal pneumonia in adult emergency department patients. J Clin Microbiol. 2005;43(7):3221-3226. https://doi.org/10.1128/JCM.43.7.3221-3226.2005

19. Azar MM. A Diagnostic Approach to Fungal Pneumonia: An Infectious Diseases Perspective. Chest. 2024;165(3):559-572. https://doi.org/10.1016/j.chest.2023.10.005

20. Jaramillo Cartagena A, Asowata OE, Ng D, et al. An overview of the laboratory diagnosis of Pneumocystis jirovecii pneumonia. J Clin Microbiol. 2025;63(3):e0036124. https://doi.org/10.1128/jcm.00361-24

21. Kepka S, Zarca K, Ohana M, et al. Cost effectiveness of imaging strategies in the emergency department for the diagnostic workup of community-acquired pneumonia: a real-life retrospective study. Health Econ Rev. 2025;15(1):41. https://doi.org/10.1186/s13561-025-00625-8

22. Al Nufaiei ZF, Alshamrani KM. Comparing Ultrasound, Chest X-Ray, and CT Scan for Pneumonia Detection. Med Devices (Auckl). 2025;18:149-159. https://doi.org/10.2147/MDER.S501714

23. Pochepnia S, Grabczak EM, Johnson E, et al. Imaging in pulmonary infections of immunocompetent adult patients. Breathe (Sheff). 2024;20(1):230186. https://doi.org/10.1183/20734735.0186-2023

24. Claessens YE, Debray MP, Tubach F, et al. Early Chest Computed Tomography Scan to Assist Diagnosis and Guide Treatment Decision for Suspected Community-acquired Pneumonia. Am J Respir Crit Care Med. 2015;192(8):974-982. https://doi.org/10.1164/rccm.201501-0017OC

25. Klug M, Sobeh T, Green M, et al. Denoised Ultra-Low-Dose Chest CT to Assess Pneumonia in Individuals Who Are Immunocompromised. Radiol Cardiothorac Imaging. 2025;7(2):e240189. https://doi.org/10.1148/ryct.240189

26. Altmayer S, Zanon M, Pacini GS, et al. Comparison of the computed tomography findings in COVID-19 and other viral pneumonia in immunocompetent adults: a systematic review and meta-analysis. Eur Radiol. 2020;30(12):6485-6496. https://doi.org/10.1007/s00330-020-07018-x

27. Ito I, Ishida T, Togashi K, et al. Differentiation of bacterial and non-bacterial community-acquired pneumonia by thin-section computed tomography. Eur J Radiol. 2009;72(3):388-395. https://doi.org/10.1016/j.ejrad.2008.08.004

28. Franquet T. Imaging of pneumonia: trends and algorithms. Eur Respir J. 2001;18(1):196-208. https://doi.org/10.1183/09031936.01.00213501

29. Garg M, Prabhakar N, Gulati A, et al. Spectrum of imaging findings in pulmonary infections. Part 1: Bacterial and viral. Pol J Radiol. 2019;84:e205-e213. https://doi.org/10.5114/pjr.2019.85812

30. Oikonomou A, Muller NL, Nantel S. Radiographic and high-resolution CT findings of influenza virus pneumonia in patients with hematologic malignancies. AJR Am J Roentgenol. 2003;181(2):507-511. https://doi.org/10.2214/ajr.181.2.1810507

31. Kim SH, Kim MY, Hong SI, et al. Invasive Pulmonary Aspergillosis-mimicking Tuberculosis. Clin Infect Dis. 2015;61(1):9-17. https://doi.org/10.1093/cid/civ216

32. Lee KS, Im JG. CT in adults with tuberculosis of the chest: characteristic findings and role in management. AJR Am J Roentgenol. 1995;164(6):1361-1367. https://doi.org/10.2214/ajr.164.6.7754873

33. Koo HJ, Lim S, Choe J, et al. Radiographic and CT Features of Viral Pneumonia. Radiographics. 2018;38(3):719-739. https://doi.org/10.1148/rg.2018170048

34. Durhan G, Ardali Duzgun S, Baytar Y, et al. Two in one: Overlapping CT findings of COVID-19 and underlying lung diseases. Clin Imaging. 2023;93:60-69. https://doi.org/10.1016/j.clinimag.2022.11.005

35. Rizzetto F, Berta L, Zorzi G, et al. Diagnostic Performance in Differentiating COVID-19 from Other Viral Pneumonias on CT Imaging: Multi-Reader Analysis Compared with an Artificial Intelligence-Based Model. Tomography. 2022;8(6):2815-2827. https://doi.org/10.3390/tomography8060235

36. Felder FN, Walsh SLF. Exploring computer-based imaging analysis in interstitial lung disease: opportunities and challenges. ERJ Open Res. 2023;9(4). https://doi.org/10.1183/23120541.00145-2023

37. Delaney L, Alabed S, Maiter A, et al. Meta-Analysis of Interobserver Agreement in Assessment of Interstitial Lung Disease Using High-Resolution CT. Radiology. 2024;313(1):e240016. https://doi.org/10.1148/radiol.240016

38. Hopstaken RM, Witbraad T, van Engelshoven JM, et al. Inter-observer variation in the interpretation of chest radiographs for pneumonia in community-acquired lower respiratory tract infections. Clin Radiol. 2004;59(8):743-752. https://doi.org/10.1016/j.crad.2004.01.011

39. Aziz ZA, Wells AU, Hansell DM, et al. HRCT diagnosis of diffuse parenchymal lung disease: inter-observer variation. Thorax. 2004;59(6):506-511. https://doi.org/10.1136/thx.2003.020396

40. Avanzo M, Stancanello J, Pirrone G, et al. The Evolution of Artificial Intelligence in Medical Imaging: From Computer Science to Machine and Deep Learning. Cancers (Basel). 2024;16(21). https://doi.org/10.3390/cancers16213702

41. Yamashita R, Nishio M, Do RKG, et al. Convolutional neural networks: an overview and application in radiology. Insights Imaging. 2018;9(4):611-629. https://doi.org/10.1007/s13244-018-0639-9

42. Mienye ID, Swart TG, Obaido G, et al. Deep Convolutional Neural Networks in Medical Image Analysis: A Review. Information. 2025;16(3). https://doi.org/10.3390/info16030195

43. Dosovitskiy A, Beyer L, Kolesnikov A, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale2020 October 01, 2020:[arXiv:2010.11929 p.]. Available from: https://ui.adsabs.harvard.edu/abs/2020arXiv201011929D http://arxiv.org/pdf/2010.11929. https://doi.org/10.48550/arXiv.2010.11929

44. Azad R, Kazerouni A, Heidari M, et al. Advances in medical image analysis with vision Transformers: A comprehensive review. Med Image Anal. 2024;91:103000. https://doi.org/10.1016/j.media.2023.103000

45. Lambin P, Leijenaar RTH, Deist TM, et al. Radiomics: the bridge between medical imaging and personalized medicine. Nat Rev Clin Oncol. 2017;14(12):749-762. https://doi.org/10.1038/nrclinonc.2017.141

46. Mayerhoefer ME, Materka A, Langs G, et al. Introduction to Radiomics. J Nucl Med. 2020;61(4):488-495. https://doi.org/10.2967/jnumed.118.222893

47. Majumder S, Katz S, Kontos D, et al. State of the art: radiomics and radiomics-related artificial intelligence on the road to clinical translation. BJR Open. 2024;6(1):tzad004. https://doi.org/10.1093/bjro/tzad004

48. Lee CH, Pan CT, Lee MC, et al. RDAG U-Net: An Advanced AI Model for Efficient and Accurate CT Scan Analysis of SARS-CoV-2 Pneumonia Lesions. Diagnostics (Basel). 2024;14(18). https://doi.org/10.3390/diagnostics14182099

49. Zhang K, Liu X, Shen J, et al. Clinically Applicable AI System for Accurate Diagnosis, Quantitative Measurements, and Prognosis of COVID-19 Pneumonia Using Computed Tomography. Cell. 2020;181(6):1423-1433 e1411. https://doi.org/10.1016/j.cell.2020.04.045

50. Tew J, Calenoff L, Berlin BS. Bacterial or nonbacterial pneumonia: accuracy of radiographic diagnosis. Radiology. 1977;124(3):607-612. https://doi.org/10.1148/124.3.607

51. Thomas J, Pociute A, Kevalas R, et al. Blood biomarkers differentiating viral versus bacterial pneumonia aetiology: a literature review. Ital J Pediatr. 2020;46(1):4. https://doi.org/10.1186/s13052-020-0770-3

52. Wang F, Li X, Wen R, et al. Pneumonia-Plus: a deep learning model for the classification of bacterial, fungal, and viral pneumonia based on CT tomography. Eur Radiol. 2023;33(12):8869-8878. https://doi.org/10.1007/s00330-023-09833-4

53. Shao J, Ma J, Yu Y, et al. A multimodal integration pipeline for accurate diagnosis, pathogen identification, and prognosis prediction of pulmonary infections. Innovation (Camb). 2024;5(4):100648. https://doi.org/10.1016/j.xinn.2024.100648

54. Sun H, Ren G, Teng X, et al. Artificial intelligence-assisted multistrategy image enhancement of chest X-rays for COVID-19 classification. Quant Imaging Med Surg. 2023;13(1):394-416. https://doi.org/10.21037/qims-22-610

55. Lam NFD, Sun H, Song L, et al. Development and validation of bone-suppressed deep learning classification of COVID-19 presentation in chest radiographs. Quant Imaging Med Surg. 2022;12(7):3917-3931. https://doi.org/10.21037/qims-21-791

56. Garg M, Prabhakar N, Gulati A, et al. Spectrum of imaging findings in pulmonary infections. Part 2: Fungal, mycobacterial, and parasitic. Pol J Radiol. 2019;84:e214-e223. https://doi.org/10.5114/pjr.2019.85813

57. Bernal-Martinez L, Herrera L, Valero C, et al. Differential Diagnosis of Fungal Pneumonias vs. Tuberculosis in AIDS Patients by Using Two New Molecular Methods. J Fungi (Basel). 2021;7(5). https://doi.org/10.3390/jof7050336

58. Li P, Wang J, Tang M, et al. A CT-based radiomics predictive nomogram to identify pulmonary tuberculosis from community-acquired pneumonia: a multicenter cohort study. Front Cell Infect Microbiol. 2024;14:1388991. https://doi.org/10.3389/fcimb.2024.1388991

59. Eche T, Schwartz LH, Mokrane FZ, et al. Toward Generalizability in the Deployment of Artificial Intelligence in Radiology: Role of Computation Stress Testing to Overcome Underspecification. Radiol Artif Intell. 2021;3(6):e210097. https://doi.org/10.1148/ryai.2021210097

60. Guan H, Liu M. Domain Adaptation for Medical Image Analysis: A Survey. IEEE Trans Biomed Eng. 2022;69(3):1173-1185. https://doi.org/10.1109/TBME.2021.3117407

61. Ahmad IS, Dai JJ, Xie YQ, et al. Deep learning models for CT image classification: a comprehensive literature review. Quantitative Imaging in Medicine and Surgery. 2025;15(1):962-1011. https://doi.org/10.21037/qims-24-1400

62. Takahashi S, Sakaguchi Y, Kouno N, et al. Comparison of Vision Transformers and Convolutional Neural Networks in Medical Image Analysis: A Systematic Review. J Med Syst. 2024;48(1):84. https://doi.org/10.1007/s10916-024-02105-8

63. Zaremba W, Sutskever I, Vinyals O. Recurrent Neural Network Regularization2014 September 01, 2014:[arXiv:1409.2329 p.]. Available from: https://ui.adsabs.harvard.edu/abs/2014arXiv1409.2329Z http://arxiv.org/pdf/1409.2329. https://doi.org/10.48550/arXiv.1409.2329

64. Goodfellow IJ, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets. Advances in neural information processing systems, 2014. https://doi.org/10.48550/arXiv.1406.2661

65. de Margerie-Mellon C, Chassagnon G. Artificial intelligence: A critical review of applications for lung nodule and lung cancer. Diagn Interv Imaging. 2023;104(1):11-17. https://doi.org/10.1016/j.diii.2022.11.007

66. Isensee F, Jaeger PF, Kohl SAA, et al. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat Methods. 2021;18(2):203-211. https://doi.org/10.1038/s41592-020-01008-z

67. Kang M, Hong KS, Chikontwe P, et al. Quantitative Assessment of Chest CT Patterns in COVID-19 and Bacterial Pneumonia Patients: a Deep Learning Perspective. J Korean Med Sci. 2021;36(5):e46. https://doi.org/10.3346/jkms.2021.36.e46

68. Li L, Qin L, Xu Z, et al. Using Artificial Intelligence to Detect COVID-19 and Community-acquired Pneumonia Based on Pulmonary CT: Evaluation of the Diagnostic Accuracy. Radiology. 2020;296(2):E65-E71. https://doi.org/10.1148/radiol.2020200905

69. Wang Y, Liu C, Fan Y, et al. A multi-modal deep learning solution for precise pneumonia diagnosis: the PneumoFusion-Net model. Front Physiol. 2025;16:1512835. https://doi.org/10.3389/fphys.2025.1512835

70. Shi C, Shao Y, Shan F, et al. Development and validation of a deep learning model for multicategory pneumonia classification on chest computed tomography: a multicenter and multireader study. Quant Imaging Med Surg. 2023;13(12):8641-8656. https://doi.org/10.21037/qims-23-1097

71. Furtado A, da Purificacao CAC, Badaro R, et al. A Light Deep Learning Algorithm for CT Diagnosis of COVID-19 Pneumonia. Diagnostics (Basel). 2022;12(7). https://doi.org/10.3390/diagnostics12071527

72. Wu X, Hui H, Niu M, et al. Deep learning-based multi-view fusion model for screening 2019 novel coronavirus pneumonia: A multicentre study. Eur J Radiol. 2020;128:109041. https://doi.org/10.1016/j.ejrad.2020.109041

73. Xie S, Wu M, Shang Y, et al. Development and validation of an early diagnosis model for severe mycoplasma pneumonia in children based on interpretable machine learning. Respir Res. 2025;26(1):182. https://doi.org/10.1186/s12931-025-03262-1

74. Sourlos N, Vliegenthart R, Santinha J, et al. Recommendations for the creation of benchmark datasets for reproducible artificial intelligence in radiology. Insights Imaging. 2024;15(1):248. https://doi.org/10.1186/s13244-024-01833-2

75. Zhang WC, Guo Y, Jin QY. Radiomics and Its Feature Selection: A Review. Symmetry-Basel. 2023;15(10). https://doi.org/10.3390/sym15101834

76. Liu H, Ren H, Wu Z, et al. CT radiomics facilitates more accurate diagnosis of COVID-19 pneumonia: compared with CO-RADS. J Transl Med. 2021;19(1):29. https://doi.org/10.1186/s12967-020-02692-3

77. Feng Z, Li H, Liu Q, et al. CT Radiomics to Predict Macrotrabecular-Massive Subtype and Immune Status in Hepatocellular Carcinoma. Radiology. 2023;307(1):e221291. https://doi.org/10.1148/radiol.221291

78. Song F, Ma M, Zeng S, et al. CT enterography-based radiomics combined with body composition to predict infliximab treatment failure in Crohn's disease. Radiol Med. 2024;129(2):175-187. https://doi.org/10.1007/s11547-023-01748-w

79. Frix AN, Cousin F, Refaee T, et al. Radiomics in Lung Diseases Imaging: State-of-the-Art for Clinicians. J Pers Med. 2021;11(7). https://doi.org/10.3390/jpm11070602

80. Schniering J, Maciukiewicz M, Gabrys HS, et al. Computed tomography-based radiomics decodes prognostic and molecular differences in interstitial lung disease related to systemic sclerosis. Eur Respir J. 2022;59(5). https://doi.org/10.1183/13993003.04503-2020

81. Jiang J, Chen S, Zhang S, et al. A radiomics model utilizing CT for the early detection and diagnosis of severe community-acquired pneumonia. Bmc Med Imaging. 2024;24(1):202. https://doi.org/10.1186/s12880-024-01370-w

82. Zhao W, Xiong Z, Tian D, et al. The adding value of contrast-enhanced CT radiomics: Differentiating tuberculosis from non-tuberculous infectious lesions presenting as solid pulmonary nodules or masses. Front Public Health. 2022;10:1018527. https://doi.org/10.3389/fpubh.2022.1018527

83. Luo T, Guo J, Xi J, et al. Development and validation of a CT-based radiomics machine learning model for differentiating immune-related interstitial pneumonia. Int Immunopharmacol. 2025;156:114681. https://doi.org/10.1016/j.intimp.2025.114681

84. Li J, Si J, Yang Y, et al. The value of CT-based radiomics for differentiation of pleural effusions in bacterial pneumonia and Mycoplasma pneumoniae pneumonia in children. Transl Pediatr. 2025;14(1):70-79. https://doi.org/10.21037/tp-24-364

85. Gillies RJ, Kinahan PE, Hricak H. Radiomics: Images Are More than Pictures, They Are Data. Radiology. 2016;278(2):563-577. https://doi.org/10.1148/radiol.2015151169

86. Magnin CY, Lauer D, Ammeter M, et al. From images to clinical insights: an educational review on radiomics in lung diseases. Breathe (Sheff). 2025;21(1):230225. https://doi.org/10.1183/20734735.0225-2023

87. Yan Q, Zhao W, Kong H, et al. CT-based radiomics analysis of consolidation characteristics in differentiating pulmonary disease of non-tuberculous mycobacterium from pulmonary tuberculosis. Exp Ther Med. 2024;27(3):112. https://doi.org/10.3892/etm.2024.12400

88. Guo B, Lu D, Szumel G, et al. The Impact of Scanner Domain Shift on Deep Learning Performance in Medical Imaging: an Experimental Study2024 September 01, 2024:[arXiv:2409.04368 p.]. Available from: http://arxiv.org/pdf/2409.04368. https://doi.org/10.48550/arXiv.2409.04368

89. Madadi Y, Seydi V, Nasrollahi K, et al. Deep visual unsupervised domain adaptation for classification tasks: a survey. Iet Image Process. 2020;14(14):3283-3299. https://doi.org/10.1049/iet-ipr.2020.0087

90. Zhu JY, Park T, Isola P, et al. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. Ieee I Conf Comp Vis. 2017:2242-2251. https://doi.org/10.1109/Iccv.2017.244

91. Yoon JS, Oh K, Shin Y, et al. Domain Generalization for Medical Image Analysis: A Review. P Ieee. 2024;112(10):1583-1609. https://doi.org/10.1109/Jproc.2024.3507831

92. Huang H, Wu R, Li Y, et al. Self-Supervised Transfer Learning Based on Domain Adaptation for Benign-Malignant Lung Nodule Classification on Thoracic CT. IEEE J Biomed Health Inform. 2022;26(8):3860-3871. https://doi.org/10.1109/JBHI.2022.3171851

93. Feng Y, Luo Y, Yang J. Cross-platform privacy-preserving CT image COVID-19 diagnosis based on source-free domain adaptation. Knowl Based Syst. 2023;264:110324. https://doi.org/10.1016/j.knosys.2023.110324

94. Chen H, Jiang Y, Loew M, et al. Unsupervised domain adaptation based COVID-19 CT infection segmentation network. Appl Intell (Dordr). 2022;52(6):6340-6353. https://doi.org/10.1007/s10489-021-02691-x

95. Sanchez K, Hinojosa C, Arguello H, et al. CX-DaGAN: Domain Adaptation for Pneumonia Diagnosis on a Small Chest X-Ray Dataset. IEEE Trans Med Imaging. 2022;41(11):3278-3288. https://doi.org/10.1109/TMI.2022.3182168

96. He Y, Carass A, Zuo L, et al. Autoencoder based self-supervised test-time adaptation for medical image analysis. Med Image Anal. 2021;72:102136. https://doi.org/10.1016/j.media.2021.102136

97. Li H, Lin Z, Qiu Z, et al. Enhancing and Adapting in the Clinic: Source-Free Unsupervised Domain Adaptation for Medical Image Enhancement. IEEE Trans Med Imaging. 2024;43(4):1323-1336. https://doi.org/10.1109/TMI.2023.3335651

98. Kollias D, Arsenos A, Kollias S. Domain adaptation, Explainability & Fairness in AI for Medical Image Analysis: Diagnosis of COVID-19 based on 3-D Chest CT-scans. 2024 Ieee/Cvf Conference on Computer Vision and Pattern Recognition Workshops, Cvprw, 2024: 4907-4914. https://doi.org/10.1109/Cvprw63382.2024.00495

99. Feng Y, Wang Z, Xu X, et al. Contrastive domain adaptation with consistency match for automated pneumonia diagnosis. Med Image Anal. 2023;83:102664. https://doi.org/10.1016/j.media.2022.102664

100. Fan G, Zhou Y, Zhou F, et al. The mortality and years of life lost for community-acquired pneumonia before and during COVID-19 pandemic in China. Lancet Reg Health West Pac. 2024;42:100968. https://doi.org/10.1016/j.lanwpc.2023.100968

101. Sattar SBA, Nguyen AD, Sharma S. Bacterial pneumonia. StatPearls [Internet]: StatPearls Publishing; 2024. Available from: https://www.ncbi.nlm.nih.gov/books/NBK513321/

102. Torres A, Cilloniz C, Niederman MS, et al. Pneumonia. Nat Rev Dis Primers. 2021;7(1):25. https://doi.org/10.1038/s41572-021-00259-0

103. Womack J, Kropa J. Community-Acquired Pneumonia in Adults: Rapid Evidence Review. Am Fam Physician. 1052022. p. 625-630. Available from: https://www.ncbi.nlm.nih.gov/pubmed/35704808

104. Garin N, Marti C, Scheffler M, et al. Computed tomography scan contribution to the diagnosis of community-acquired pneumonia. Curr Opin Pulm Med. 2019;25(3):242-248. https://doi.org/10.1097/MCP.0000000000000567

105. Parekh M, Donuru A, Balasubramanya R, et al. Review of the Chest CT Differential Diagnosis of Ground-Glass Opacities in the COVID Era. Radiology. 2020;297(3):E289-E302. https://doi.org/10.1148/radiol.2020202504

106. Singh S, Pinsky P, Fineberg NS, et al. Evaluation of reader variability in the interpretation of follow-up CT scans at lung cancer screening. Radiology. 2011;259(1):263-270. https://doi.org/10.1148/radiol.10101254

107. Rello J, Bodi M, Mariscal D, et al. Microbiological testing and outcome of patients with severe community-acquired pneumonia. Chest. 2003;123(1):174-180. https://doi.org/10.1378/chest.123.1.174

108. Kareem A, Liu H, Sant P. Review on Pneumonia Image Detection: A Machine Learning Approach. Human-Centric Intelligent Systems. 2022;2(1):31-43. https://doi.org/10.1007/s44230-022-00002-2

109. Sharma S, Guleria K. A systematic literature review on deep learning approaches for pneumonia detection using chest X-ray images. Multimedia Tools and Applications. 2024;83(8):24101-24151. https://doi.org/10.1007/s11042-023-16419-1

110. Chen W, Han X, Wang J, et al. Deep diagnostic agent forest (DDAF): A deep learning pathogen recognition system for pneumonia based on CT. Comput Biol Med. 2022;141:105143. https://doi.org/10.1016/j.compbiomed.2021.105143

111. Li F, Kan J, Jin L, et al. A New Pneumonia Detection Model Based on Transformer with Improved Self-Attention Mechanism. International Conference on Cognitive based Information Processing and Applications: Springer, 2022: 505-514. https://doi.org/10.1007/978-981-19-9376-3_57

112. Jiang Z, Chen L. Multisemantic Level Patch Merger Vision Transformer for Diagnosis of Pneumonia. Comput Math Methods Med. 2022;2022:7852958. https://doi.org/10.1155/2022/7852958

113. Wang Y, Huang X, Li F, et al. Serum-integrated omics reveal the host response landscape for severe pediatric community-acquired pneumonia. Crit Care. 2023;27(1):79. https://doi.org/10.1186/s13054-023-04378-w

114. Wong PK, Yan T, Wang H, et al. Automatic detection of multiple types of pneumonia: Open dataset and a multi-scale attention network. Biomed Signal Process Control. 2022;73:103415. https://doi.org/10.1016/j.bspc.2021.103415

115. Mehrizi MHR, Gerritsen SH, de Klerk WM, et al. How do providers of artificial intelligence (AI) solutions propose and legitimize the values of their solutions for supporting diagnostic radiology workflow? A technography study in 2021. European Radiology. 2023;33(2):915-924. https://doi.org/10.1007/s00330-022-09090-x

116. Keys RG. Cubic Convolution Interpolation for Digital Image-Processing. Ieee T Acoust Speech. 1981;29(6):1153-1160. https://doi.org/10.1109/Tassp.1981.1163711

117. Wasserthal J, Breit HC, Meyer MT, et al. TotalSegmentator: Robust Segmentation of 104 Anatomic Structures in CT Images. Radiol Artif Intell. 2023;5(5):e230024. https://doi.org/10.1148/ryai.230024

118. Tsai E, Simpson S, Lungren MP, et al. Data from medical imaging data resource center (MIDRC)-RSNA international covid radiology database (RICORD) release 1C-Chest X-ray, covid+(MIDRC-RICORD-1C). The Cancer Imaging Archive2021. p. 4. https://doi.org/10.7937/91ah-v663.

119. Jang J, Hwang D. M3T: three-dimensional Medical image classifier using Multi-plane and Multi-slice Transformer. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022: 20718-20729. https://doi.org/10.1109/CVPR52688.2022.02006.

120. He KM, Zhang XY, Ren SQ, et al. Deep Residual Learning for Image Recognition. 2016 Ieee Conference on Computer Vision and Pattern Recognition (Cvpr). 2016:770-778. https://doi.org/10.1109/Cvpr.2016.90

121. Chen S, Ma K, Zheng Y. Med3d: Transfer learning for 3d medical image analysis. arXiv preprint arXiv:190400625. 2019. https://doi.org/10.48550/arXiv.1904.00625

122. Leng Z, Tan M, Liu C, et al. Polyloss: A polynomial expansion perspective of classification loss functions. arXiv preprint arXiv:220412511. 2022. https://doi.org/10.48550/arXiv.2204.12511

123. Lin T-Y, Goyal P, Girshick R, et al. Focal loss for dense object detection. Ieee T Pattern Anal. 2020;42:318-327. https://doi.org/10.1109/TPAMI.2018.2858826

124. Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics. 2020;21(1):6. https://doi.org/10.1186/s12864-019-6413-7

125. Ranschaert E, Topff L, Pianykh O. Optimization of radiology workflow with artificial intelligence. Radiologic Clinics. 2021;59(6):955-966. https://doi.org/10.1016/j.rcl.2021.06.006

126. Delevaux I, Andre M, Colombier M, et al. Can procalcitonin measurement help in differentiating between bacterial infection and other kinds of inflammatory processes? Annals of the rheumatic diseases. 2003;62(4):337-340. https://doi.org/10.1136/ard.62.4.337

127. Arnaout R, Curran L, Zhao Y, et al. An ensemble of neural networks provides expert-level prenatal detection of complex congenital heart disease. Nat Med. 2021;27(5):882-891. https://doi.org/10.1038/s41591-021-01342-5

128. Yu P, Zhang H, Wang D, et al. Spatial resolution enhancement using deep learning improves chest disease diagnosis based on thick slice CT. NPJ Digit Med. 2024;7(1):335. https://doi.org/10.1038/s41746-024-01338-8

129. Organization WH. Global Tuberculosis Report 2024. Geneva: World Health Organization; 2024. Available from: https://www.who.int/teams/global-tuberculosis-programme/tb-reports/global-tuberculosis-report-2024.

130. Denning DW. Global incidence and mortality of severe fungal disease. Lancet Infect Dis. 2024;24(7):e428-e438. https://doi.org/10.1016/S1473-3099(23)00692-8

131. Ekeng BE, Davies AA, Osaigbovo, II, et al. Pulmonary and Extrapulmonary Manifestations of Fungal Infections Misdiagnosed as Tuberculosis: The Need for Prompt Diagnosis and Management. J Fungi (Basel). 2022;8(5). https://doi.org/10.3390/jof8050460

132. Risso K, Michelangeli C, Gaudart A, et al. Time-to-detection in culture of Mycobacterium tuberculosis: performance for assessing index cases contact-positivity. Int J Infect Dis. 2023;134:280-284. https://doi.org/10.1016/j.ijid.2023.07.002

133. Babafemi EO, Cherian BP, Banting L, et al. Effectiveness of real-time polymerase chain reaction assay for the detection of Mycobacterium tuberculosis in pathological samples: a systematic review and meta-analysis. Syst Rev. 2017;6(1):215. https://doi.org/10.1186/s13643-017-0608-2

134. Hage CA, Knox KS, Wheat LJ. Endemic mycoses: overlooked causes of community acquired pneumonia. Respir Med. 2012;106(6):769-776. https://doi.org/10.1016/j.rmed.2012.02.004

135. Wheat LJ, Azar MM, Bahr NC, et al. Histoplasmosis. Infect Dis Clin North Am. 2016;30(1):207-227. https://doi.org/10.1016/j.idc.2015.10.009

136. Marchiori E, Marom EM, Zanetti G, et al. Reversed halo sign in invasive fungal infections: criteria for differentiation from organizing pneumonia. Chest. 2012;142(6):1469-1473. https://doi.org/10.1378/chest.12-0114

137. Marchiori E, Zanetti G, Hochhegger B, et al. Reversed halo sign on computed tomography: state-of-the-art review. Lung. 2012;190(4):389-394. https://doi.org/10.1007/s00408-012-9392-x

138. Yao Q, Zhou QH, Shen QL, et al. Imaging findings of pulmonary manifestations of chronic granulomatous disease in a large single center from Shanghai, China (1999-2018). Sci Rep. 2020;10(1):19349. https://doi.org/10.1038/s41598-020-76408-4

139. Kim SJ, Lee KS, Ryu YH, et al. Reversed halo sign on high-resolution CT of cryptogenic organizing pneumonia: diagnostic implications. AJR Am J Roentgenol. 2003;180(5):1251-1254. https://doi.org/10.2214/ajr.180.5.1801251

140. Thattaamuriyil Padmakumari L, Guido G, Caruso D, et al. The Role of Chest CT Radiomics in Diagnosis of Lung Cancer or Tuberculosis: A Pilot Study. Diagnostics (Basel). 2022;12(3). https://doi.org/10.3390/diagnostics12030739

141. Hu Y, Zhong L, Liu H, et al. Lung CT-based multi-lesion radiomic model to differentiate between nontuberculous mycobacteria and Mycobacterium tuberculosis. Med Phys. 2025;52(2):1086-1095. https://doi.org/10.1002/mp.17537

142. Song J, Hwang EJ, Yoon SH, et al. Emerging Trends and Innovations in Radiologic Diagnosis of Thoracic Diseases. Invest Radiol. 2025. https://doi.org/10.1097/RLI.0000000000001179

143. Pezzino S, Luca T, Castorina M, et al. Current trends and emerging themes in utilizing artificial intelligence to enhance anatomical diagnostic accuracy and efficiency in radiotherapy. Prog Biomed Eng (Bristol). 2025. https://doi.org/10.1088/2516-1091/adc85e

144. Li JD, Cheng KW, Wang SH, et al. Feature Selection: A Data Perspective. Acm Comput Surv. 2018;50(6). https://doi.org/10.1145/3136625

145. Tao P, Yi H, Wei C, et al. A method based on weighted F-score and SVM for feature selection. Chin Cont Decis Conf, 2013: 4287-4290. https://doi.org/10.1109/CCDC.2013.6561705

146. Xiao L, Xiaoli W, Qiang S. Feature selection of medical data sets based on RS-RELIEFF. I C Serv Syst Serv M, 2015: 1-5. https://doi.org/10.1109/ICSSSM.2015.7170275

147. Nogueira S, Sechidis K, Brown G. On the stability of feature selection algorithms. J Mach Learn Res. 182018. p. 1-54. Available from: https://www.jmlr.org/papers/v18/17-514.html