UTS UNIVERSITY OF TECHNOLOGY SYDNEY

Deep Learning-Based Detection of Pulmonary Involvement in Malignant Disease Using CT and ^{18}F-FDG PET/CT Scans

by Maryam Fallahpoor

Thesis submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy

Under the supervision of
Distinguished Professor Biswajeet Pradhan
Dr. Subrata Chakraborty, Dr. Shilpa Bade-Gite

University of Technology Sydney
Faculty of Engineering and Information Technology

March 2025

Certificate of original authorship

I, Maryam Fallahpoor declare that this thesis, is submitted in fulfilment of the requirements for the award of Doctor of Philosophy, in the School of Civil and Environmental Engineering, Faculty of Engineering and Information Technology at the University of Technology Sydney. This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis. This document has not been submitted for qualifications at any other academic institution. This research is supported by the Australian Government Research Training Program.

Production Note:
Signature removed
Signature: prior to publication. Date: 11/03/2025

Name: Maryam Fallahpoor

Acknowledgments

I am deeply thankful to Prof. Biswajeet Pradhan and Dr. Subrata Chakraborty for giving me the opportunity to conduct this research under their expert guidance. Their unwavering support, insightful advice, and engaging scholarly discussions have profoundly shaped my research journey and enriched my understanding.

I would like to extend special thanks to the University of Technology Sydney for their generous support through the UTS President's Scholarship and the International Research Scholarship, which have enabled me to pursue my PhD at this esteemed institution. I am also grateful to UTS for their financial support through the VCCf fund, which facilitated my participation in the significant international conference, ICRR 2024 in Lyon, France. I sincerely appreciate the computational resources provided by the UTS eResearch High-Performance GPUs, which have significantly enhanced the efficiency of my research endeavors.

To my family, whose constant support has been my source of strength, I owe a profound debt of gratitude. I am especially grateful to my parents for their steadfast encouragement and to my husband for his unwavering support during the challenging days when my dedication to research made me less available. His emotional support and encouragement enabled me to give my best to my work. It is to them that I dedicate this thesis.

Maryam Fallahpoor

March 2025

Sydney, Australia

List of papers/publications

Published peer-reviewed journal papers

1. Fallahpoor M, Chakraborty S, Heshejin MT, Chegeni H, Horry MJ, Pradhan B. Generalizability assessment of COVID-19 3D CT data for deep learning-based disease detection. Computers in Biology and Medicine. 2022 Jun 1;145:105464. https://www.sciencedirect.com/science/article/pii/S0010482522002566
2. Fallahpoor M, Chakraborty S, Pradhan B, Faust O, Barua PD, Chegeni H, Acharya R. Deep learning techniques in PET/CT imaging: A comprehensive review from sinogram to image space. Computer Methods and Programs in Biomedicine. 2023 Oct 21:107880. https://www.sciencedirect.com/science/article/abs/pii/S0169260723005461

Other publications during the candidature

- Horry MJ, Chakraborty S, Pradhan B, Fallahpoor M, Chegeni H, Paul M. Factors determining generalization in deep learning models for scoring COVID-CT images. Mathematical Biosciences and Engineering. 2021 Oct 27;18(6):9264-93. https://www.aimspress.com/article/doi/10.3934/mbe.2021456

All the aforementioned papers have been published during my Ph.D. candidature.

Table of Contents

List of Figures

List of Tables

List of Acronyms

FCN	Fully connected network
CNN	Convolutional neural network
^{18}F FDG	2-deoxy-2-[fluorine-18]fluoro- D-glucose
PET	Positron emission tomography
CT	Computed tomography
DL	Deep learning
AI	Artificial intelligence
PPV	Positive predictive value
NPV	Negative predictive value
AUC	Area under the Curve
ROC	Receiver Operating Characteristic
GPU	Graphics processing unit
DICOM	Digital Imaging and Communications in Medicine
NIFTI	Neuroimaging Informatics Technology Initiative
PACS	Picture archiving and communication system
SHAP	SHapley Additive exPlanations
XAI	Explainable artificial intelligence
SUV_{max}	Maximum standardized uptake value

Abstract

Pulmonary involvements, including pulmonary nodules, pose a risk of lung complications and lung cancer, the leading cause of cancer-related deaths globally. Diagnosis and categorization of pulmonary nodules is critical for lung cancer detection and staging other malignancies due to the high incidence of metastasis in the lungs from various cancers, which is typically performed using common medical imaging such as ^{18}F FDG PET/CT imaging.

Diagnosis of pulmonary involvement require significant experience, are labor-intensive, and prone to error. Artificial intelligence (AI), particularly deep learning, has shown promising results in medical applications. This study aims to develop a reliable 3D deep learning-based approach to assist physicians in accurately diagnosing pulmonary nodules from ^{18}F FDG PET/CT imaging. We investigated lung nodules originating from malignancies other than lung cancer and excluded lung masses. Most studies focus only on lung nodules from lung cancer, with limited attention given to classifying nodules from other malignancies.

The study has three main objectives: First, classification of pulmonary involvements, including COVID-19, using optimized deep learning models, generalizability assessment, and identification of the optimal pre-processing steps. Second objective is classification of pulmonary nodules from a subset of a ^{18}F FDG PET/CT dataset using the state-of-the-art deep learning models to determine the best model and approach for multi-class classification of pulmonary nodules similar to real clinical scenarios. And the third objective is to develop an optimized deep learning-based approach for higher accuracy in pulmonary nodule detection and classification using a large ^{18}F FDG PET/CT dataset. For the first objective, classification of lung involvement from COVID-19 was performed using four datasets of CT images and seven deep learning models in 5-fold cross-validation. To achieve the second objective, different deep learning models were tested on a subset of PET/CT dataset in a multi-class classification approach to refine models and methodologies through a trial-and-error approach. Finally, adjustments were made to the top-performing model's layers and pre-processing steps for optimization of lung nodule classification from 1304 ^{18}F FDG PET/CT images.

The results of this study indicated that the combining 80% of one dataset with at least 40% from another yields comparable results to using the full combination. We have achieved overall accuracy of 89% for lung nodule classification. The multi-class simulation, classifying benign, malignant, and suspicious cases, closely mimics real-world conditions, demonstrating the capability of the model to handle complex scenarios and its potential to assist medical professionals in making more accurate diagnoses.

CHAPTER 1

1. INTRODUCTION

1.1 Background

There is a risk that pulmonary involvements, including pulmonary nodules, lead to lung complications and lung cancer. Lung cancer is the cause of most cancer-related deaths worldwide and the second most commonly diagnosed cancer (Barta et al., 2019; Schabath & Cote, 2019; Sung et al., 2021). Pulmonary nodule detection is considered as one of the most critical tasks for lung cancer detection and staging other malignancies (Hammer & Byrne, 2022; Loverdos et al., 2019; Vlahos et al., 2018). There is a high incidence of metastasis in the lungs from different types of cancers (Ferlito et al., 2001; Fidler, 1989; Herold et al., 1996; Van Schil, 2007; Zeidman, 1957). Pulmonary nodules that originate from cancers outside the lung are commonly referred as metastatic pulmonary nodules. These nodules represent the spread of malignancies from other primary sites in the body to the lungs. Metastasis to the lungs occurs because the lungs receive a significant portion of the blood flow of the body, making them a common site for metastatic deposits. The presence of metastatic pulmonary nodules generally indicates that the cancer is at an advanced stage. This has important implications for the prognosis and treatment options of the patient. Therefore, early diagnosis and characterization of them are of critical importance. Benign nodules are often caused by non-cancerous conditions such as infections, inflammation, or granulomas. Malignant nodules, on the other hand, are often associated with primary lung cancer or metastasis from other cancers. The size and growth rate of pulmonary nodules are important factors in determining the risk of malignancy. Nodules larger than 8 mm or those that exhibit rapid growth over time are more likely to be malignant. Regular monitoring and follow-up imaging are often recommended for smaller nodules to assess changes over time. Generally, pulmonary nodules are diagnosed using chest radiography, computed tomography (CT), and positron emission tomography (PET) (Dewan et al., 1993; Hadique et al., 2020; Teramoto et al., 2014; Townsend, 2008). PET combined with CT (PET/CT) offers superior diagnostic accuracy compared to conventional X-ray and standalone CT imaging for lung nodules. PET/CT provides both anatomical and metabolic information, allowing for a more comprehensive evaluation of nodules. While CT identifies the size, shape, and location of the nodule, PET detects metabolic activity, which is critical for differentiating between benign and malignant nodules. Studies have shown that PET/CT has higher sensitivity and specificity for lung nodule diagnosis, particularly for identifying malignancies, and is more effective in staging lung cancer (Hadique et al., 2020). Detection and classification of pulmonary involvement requires years of experience for physicians, it is a burden of work, and prone to error (Liu et al., 2020). Artificial intelligence (AI),

specifically machine/deep learning (DL), has been found to have promising results in various medical applications in recent years (Cheng et al., 2016; Cicero et al., 2017; Hwang et al., 2018; Lee, 2020; Schwyzer et al., 2018). AI-based detection could assist medical professionals in better and faster diagnosis (Johnson et al., 2021; Kumar et al., 2023; Shi et al., 2020; Umapathy et al., 2023; Vaishya et al., 2020). The AI models can take the form of machine learning (ML) or DL. ML algorithms can only make decisions on lowdimensional data. ML for PET/CT analysis necessitates a feature extraction step, transforming imaging data into a low-dimensional feature vector. This process relies on expert knowledge of feature extraction algorithms, but it is constrained by the risk of information loss. As a result, feature engineering is less effective when dealing with extensive and diverse datasets. In contrast, DL algorithms can directly analyze high-dimensional data, such as PET/CT images, eliminating the need for manual feature engineering and overcoming the limitations associated with information loss. This practical advantage translates into superior performance of DL methods for classifying high-volume PET/CT imaging studies compared to traditional machine learning approaches. One of the obstacles in machine learning approaches is manual feature extraction. Using DL, as an upper class of machine learning algorithms, automatically extracts features needed for the given task. DL has also achieved more reliable results than machine learning since more data is used as input. Currently, DL is the most advanced method for learning, selecting, and extracting features (LeCun et al., 2015). Convolutional neural networks (CNNs) are a class of DL methods used for object detection, classification, and analysis. CNN-based methods have proven to be quite effective for image-level diagnostics since CNNs have achieved comparable or better performance in object-classification tasks than classical methods (He et al., 2015; Russakovsky et al., 2015). Several image recognition and classification tasks have been accomplished by CNNs, challenging the accuracy of medical experts in some cases (Ghafoorian et al., 2017; Litjens et al., 2017; Shen et al., 2017). In fact, 'meaningful' image features for a given task can be implicitly learned by CNNs. In deeper layers, nonlinear mappings are used to learn higher-level features from images, and these features are then used for prediction. Building upon the success of CNNs in medical image diagnostics, the evolution of DL models has led to the development of more sophisticated architectures capable of handling diverse and complex tasks. Models such as Residual Networks (ResNet), DenseNets, and Vision Transformers (ViTs) have been introduced to mitigate issues like vanishing gradients and to enhance feature propagation, allowing for deeper, more accurate models (Dosovitskiy, 2020; He et al., 2016; Huang et al., 2017). These advancements have extended the application of DL beyond simple classification, enabling tasks like multi-class segmentation, object detection, and even anomaly detection with unprecedented accuracy. For instance, advanced models such as U-Net and its variations have been particularly successful in medical image segmentation, aiding in precise localization of pathological regions (Ronneberger

et al., 2015). These innovations have paved the way for DL algorithms to not only match but, in some cases, surpass human-level performance in specific diagnostic tasks, revolutionizing medical imaging and providing a powerful tool for aiding clinicians in disease detection and diagnosis.

1.2. PET/CT imaging and application of artificial intelligence

PET/CT is a medical imaging technique that fuses functional imaging from PET, which shows the spatial distribution of metabolic or biochemical processes within the human body, and anatomical imaging from CT (Torigian et al., 2007). Figure 1-1 depicts the typical slice of PET and CT fusion to form a PET/CT scan. A PET/CT scan is created by fusion (or co-registration of) PET and CT images. In recent years, PET/CT imaging has become more prevalent, resulting in a significant increase in data volume per scan (Meadows & Allie). This, in turn, has amplified the workload for experts tasked with interpreting these images. Furthermore, the growing volume of imaging data has led to both inter- and intra-operator variability (Motwani, 2022). Within a practical timeframe, it is economically infeasible to examine all the available evidence presented by high-resolution PET/CT images. Hence, the reading expert has to decide where to place the focus of attention; thus, there is a risk of missing another critical information available in the image. The aforementioned issues pose significant challenges as a direct consequence of improvements that are generally considered progress for PET/CT technology. In other words, the progress in PET/CT imaging leads to increased workload and increased intra- and inter-operator variability which can reduce the diagnosis quality. From a general standpoint, the root cause of the problem is the implicit assumption that a skilled practitioner can effectively interpret the images provided and that this entire process is efficient.

The process of interpreting medical images for diagnosis is a specialized and profound task (Razzak et al., 2018) when identifying all the anatomical structures shown in a PET/CT image. In a diagnosis context, it is more important to identify disease-specific characteristic deviations from the normal or, indeed, the characterization of disease-specific objects, such as tumors. These specialized and intricate analysis tasks are time-consuming for human experts, and they are susceptible to both inter- and intra-observer variability (Lindner et al., 2016). AI could potentially offer a solution to these challenges by providing diagnostic support (Tizhoosh et al., 2021).

The role of AI in this context can be compared to that of a human practitioner, as both require extensive training and validation before they can be trusted to assist in clinical decision-making. Just as a radiologist undergoes years of education, practical training, and rigorous testing to become proficient in interpreting medical images, AI algorithms also require a structured training process. However, unlike human radiologists who draw from a combination of theoretical knowledge, clinical experience, and interpretative skills, AI models are entirely trained based on

the data they are provided. This makes the quality, quantity, and diversity of training data pivotal to the performance of AI systems.

In the context of PET/CT analysis, the training data typically consists of imaging studies that have been interpreted and labeled by human experts. These expert annotations serve as the ground truth, guiding the AI in learning how to distinguish between normal and abnormal findings, identify specific disease patterns, and provide diagnostic insights. The ability of AI models to accurately analyze new PET/CT scans is directly dependent on the representativeness and thoroughness of the training data it was exposed to. If the training data set lacks diversity, for example, in terms of patient demographics or disease variations, the AI model might struggle to generalize its findings across different patient populations or rare conditions. The critical role of training data cannot be overstated. For AI models, training data is not merely a component of their development—it is the foundation upon which all their capabilities are built. If the data is biased, incomplete, or poorly annotated, the AI model's output will likely reflect those shortcomings, potentially leading to diagnostic inaccuracies or errors. Despite these challenges, AI models offer several advantages over traditional methods. They can process vast amounts of data rapidly, identify subtle patterns that might be overlooked by human eyes, and provide consistent, objective analyses. This is particularly valuable in PET/CT analysis, where the integration of metabolic and anatomical information can be complex and time-consuming. However, it is important to recognize the differences in the training of AI models compared to the education of human radiologists. While radiologists develop their expertise through years of exposure to a wide variety of cases, continuous learning, and the ability to synthesize information from multiple sources, AI models rely solely on the imaging data and associated labels provided during the training phase. They do not have the ability to understand or reason about the underlying biological processes; instead, they detect statistical patterns that correlate with certain diagnoses. This means that the performance of AI is confined within the boundaries of its training data. For example, an AI model trained predominantly on images of a certain type of cancer may excel at detecting that particular cancer but might underperform when confronted with a less common or unrepresented variant. The effectiveness of AI is directly tied to how comprehensively its training data reflects the diversity of real-world cases.

While AI holds tremendous promise in augmenting the diagnostic capabilities of radiologists, particularly in complex imaging modalities like PET/CT, the success of these AI systems is fundamentally linked to the quality and comprehensiveness of their training data. The careful selection, curation, and continuous updating of training datasets are crucial to developing AI models that can reliably contribute to medical practice. As AI technology continues to evolve, the collaboration between AI developers and clinical practitioners will be essential to ensure that these tools are not only powerful but also safe and effective in real-world healthcare settings.

The use of DL models as a subset of AI, focuses on neural networks with multiple layers, the application of this technology in PET/CT analysis has shown significant potential. DL models, particularly CNNs, are exceptionally well-suited for analyzing complex imaging data due to their ability to automatically learn hierarchical features from raw data. In PET/CT analysis, DL models can be trained to recognize intricate patterns in both metabolic and anatomical data, often surpassing traditional machine learning methods in terms of accuracy and efficiency. These models are capable of learning from large imaging dataset, identifying features that may be too subtle or complex for traditional algorithms or even human experts to discern. However, the performance of DL models is highly dependent on the quality of the training data, as they require large, diverse, and well-annotated datasets to avoid overfitting and to generalize well across different patient populations. As a result, the development of DL models in PET/CT analysis necessitates an accurate approach to data preparation and model training to ensure that these advanced tools can provide reliable and clinically meaningful insights in practice (Dayarathna et al., 2023).

1.3. PET/CT imaging and application of DL for different malignancies

The application of PET/CT imaging spans a wide array of malignancies, offering enhanced diagnostic accuracy and more effective treatment planning (Langer, 2010; Townsend et al., 2004). Its ability to reveal both primary tumors and distant metastases in a single scan makes it particularly valuable for staging cancer and monitoring treatment response. For instance, PET/CT can differentiate between viable tumor tissue and necrosis or fibrosis, which is crucial for determining the effectiveness of therapeutic interventions and making timely adjustments to treatment plans (Ambrosini & Fanti, 2011; Endo et al., 2006; Mahajan & Cook, 2017; Trotter et al., 2023). The DL technologies have further augmented the capabilities of PET/CT imaging. DL-based systems have been developed to assist in the interpretation of PET/CT scans, offering decision support that can lead to faster and more accurate diagnoses. Moreover, the integration of DL in PET/CT imaging is not only improving the detection of malignancies but also aiding in the prediction of patient outcomes and personalization of treatment strategies. By analyzing vast amounts of imaging data, DL models can identify prognostic markers that help in stratifying patients according to their risk levels and likely response to specific treatments. In this section, we introduce the different diseases for which DL-based PET/CT diagnosis support systems have been developed. Furthermore, we detail the DL technology that was used to establish the decision support.

1.3.1. *Body part/malignancy*

PET/CT is used for a variety of purposes, including identifying malignant tumors, staging lymph nodes, detecting metastasis locations, evaluating response to therapy, and indicating whether the malignancy has been recurrent (Jerusalem et al., 2003). PET/CT is used for various medical

conditions including carcinoma of unknown primary, adrenal cancer, blood malignancy (Leukemia), bone tumor, brain tumor, breast cancer, cardiac conditions, cervix cancer, CNS-related epilepsy, neurodegenerative diseases, neurovascular diseases, and other miscellaneous CNS disorders. It is also used for the imaging of colorectal (anal) cancer, esophageal cancer, gastric cancer, gastrointestinal stromal tumor (GIST), germ cell tumor, head and neck cancer, hepatobiliary cancer, Hodgkin lymphoma, non-Hodgkin lymphoma, lung cancer, melanoma, non-melanoma skin cancer, mesothelioma, multiple myeloma, neuroendocrine tumor, ovarian cancer, pancreas cancer, parathyroid cancer, pheochromocytoma, prostate cancer, retinoblastoma, salivary gland cancer, sarcoidosis, sarcoma (soft tissue), small intestine cancer, testicular cancer, thyroid cancer/thymic carcinoma, urinary system conditions (kidney, ureter, urethra, bladder, vaginal, vulvar, penile), uterine cancer, and other related conditions.

1.3.2. *Lung*

Tang et al. (Tang et al., 2019) studied the ^{18}F FDG PET/CT scan's ability to diagnose pulmonary nodules based on their size. They used one hundred ^{18}F FDG PET/CT scans of patients who were diagnosed as having solitary pulmonary nodules (SPNs) of different sizes. Their results showed that, compared to other methods, ^{18}F FDG PET/CT has an excellent performance in identifying different-sized SPNs, particularly those between 11 and 20 mm in diameter. In addition, ^{18}F-PET/CT was found to be more sensitive than CT scans in characterizing pulmonary nodules (Chin et al., 2006; Jeong et al., 2008; Kim et al., 2007). Moreover, PET-CT is useful in the management of patients with lung diseases such as non-small cell lung cancer (NSCLC) because it is able to detect disease sites even if the underlying structure is normal on CT (Ettinger et al., 2017; Kumar et al., 2019; Silvestri et al., 2013).

The main criteria for analysing pulmonary nodules are their size, density, and metabolic activity reported by maximum standardized uptake value (Garcia-Velloso et al., 2016; Groheux et al., 2016) (SUV_{max}). The size of pulmonary nodules usually ranges from 4 to 30 mm. Usually, pulmonary nodules bigger than 3 cm are classified as lung masses and malignant (Ost et al., 2003; Suo et al., 2016). Typically, pulmonary nodules in ^{18}F-PET/CT are stratified into three categories: (i) malignant nodules suspected to be metastatic, (ii) benign or not related to the patient's malignancy, and (iii) indeterminate nodules that need follow-up (Grisanti et al., 2021). Regarding location, it has been observed that most malignant nodules are located in the upper lobes and especially the right upper lobe (Khan et al., 2011; Larici et al., 2017; Sinsuat et al., 2011). The SUV_{max} in PET/CT scan is a measure that determines the maximum concentration of ^{18}F-FDG in the region of interest (ROI) (Chan et al., 2010; Fletcher & Kinahan, 2010; Nahmias & Wahl, 2008). As a result, SUV_{max} harmonizes the interpretation reports, and for pulmonary involvements, SUV_{max} of 2.5 or higher indicates malignancy (Divisi et al., 2017; Takeda et al., 2014).

1.3.3. *Brain*

Neurology continues to be one of the foremost areas in which functional imaging can provide unique information, both for clinical and research purposes. PET/CT or PET/MRI, which provide both functional and morphological information, may be useful for detecting brain tumors in the early stages (Muoio et al., 2018). While FDG continues to be the most common radiotracer for the evaluation of glucose metabolism, alternative radiotracers have been developed over the last few years (Farwell et al., 2014). Various radiotracers can be used to measure brain activity associated with different metabolic processes. For example, regional glucose metabolism can be assessed with FDG (Silverman, 2004), functional integrity of pre-synaptic dopaminergic function with [^{18}F]-DOPA (Storch et al., 2013), amino acid uptake can be assessed using [^{18}F]-fluoroethyltyrosine (FET) (Dunkl et al., 2015), and amyloid protein can be measured using various radiotracers. Along with its vast application in research studies, brain PET is also used clinically for detecting cancer, determining whether cancer has spread to the brain, diagnosing dementias, including Alzheimer's disease, determining Parkinson's disease from other conditions, and preparing for epilepsy surgery.

1.3.4. *Breast*

Despite mammography being the first-line imaging test used to detect and screen breast cancer, ultrasound, MRI, and PET/CT are also commonly used in secondary screenings. In breast cancer patient evaluations, PET/CT is considered an adjunct imaging modality. Clinically, FDG PET/CT was proven useful in staging recurrent or metastatic breast cancer and for evaluating the response to neoadjuvant and post-treatment of locally advanced breast cancer (Garcia & Singh, 2023).

1.3.5. *Prostate*

Detection and localizing prostate tumors in the early stages remains a challenging task for physicians. The most commonly effective approach for restaging biochemically recurrent prostate cancer is the use of PET/CT scans. PET/CT imaging has become a common practice for detecting and localizing prostate tumors in cases of biochemical recurrence after prostate cancer treatment (Piert et al., 2009; Umbehr et al., 2013), and for determining response to therapy. Different imaging techniques are used to determine the extent of prostate cancer following a diagnosis. An initial prostate cancer screening and treatment strategy can be based on PET/CT scans. PET imaging using prostate-specific membrane antigen (PSMA) has significantly improved prostate cancer diagnosis and treatment (Hope et al., 2018). One of the most common radioactive imaging agents for prostate cancer diagnosis is ^{68}Ga-PSMA which detects prostate cancer cells by binding to the PSMA protein. Another commonly used radiotracer in prostate cancer imaging is ^{11}C-choline. This tracer can also show tumors in a PET image due to its participation in the process of cell membrane synthesis (Kanda et al., 2008).

1.3.6. *Lymphoma/lymph node related*

PET scans are used to look for lymph node cancer, although CT scans are considered the most convenient method of detecting lymph nodes. The PET/CT imaging method is widely used to detect and assess metastatic lymph nodes in different malignancies such as lung, breast, bladder, cervical, and so forth (Borm et al., 2019; Schmidt-Hansen et al., 2014; Sironi et al., 2006; Swinnen et al., 2010). The PET/CT is very useful in determining the most appropriate initial treatment for lymphoma. It can be used to determine the baseline staging for the disease and provide important prognostic information. For evaluating treatment response, PET/CT can provide a more accurate image of residual viable malignant lesions (Cronin et al., 2010). In this way, the findings will influence treatments that need to be added or substituted.

1.3.7. *Head and neck*

In general, head and neck cancers refer to cancers that develop in the head and neck region. There are a variety of types of head and neck cancer based on anatomic distributions. Sinuses, nasal passages, mouths, salivary glands, and the throat can be affected. Despite being different, they are treated similarly and are considered a group. Skin cancers that begin in the head and neck area are also considered as head and neck cancer categories. Head and neck cancer staging has been significantly improved by PET/CT imaging in recent years. In head and neck cancer, PET/CT is recommended as the most effective imaging modality for detecting, assessing response to therapy, and separating benign inflammation from tumor persistence/recurrence (Eyassu & Young, 2021).

1.4. Challenges in applying DL models to PET/CT imaging

The applications of DL models using PET/CT images encounter several challenges. PET/CT imaging is a complex modality requiring specialized equipment and expertise for acquisition, resulting in a limited availability of training data. This scarcity can lead to the common issue of overfitting, where models become excessively tailored to training data and struggle to generalize to new data, resulting in errors and subpar real-world performance. An additional hurdle is the need for data annotation, requiring human experts to label images and pinpoint specific features—a time-consuming process reliant on specialized knowledge, impeding DL model development. Moreover, PET/CT images are high-dimensional, posing difficulties in extracting relevant information, while their complex structure, encompassing anatomical and functional aspects, adds complexity to model design. The medical imaging industry's stringent regulations pertaining to data protection and privacy present further obstacles to the access and utilization of necessary data for DL model training.

Figure 1.1. A typical slice of (a) CT scan, (b) PET scan and (c) PET/CT scan which is formed after fusion (co-registration) of PET and CT images.

1.5. The critical role of PET/CT imaging in distinguishing lung nodules in patients with non-lung malignancies

PET/CT imaging plays a pivotal role in the detection and characterization of lung nodules, particularly in patients whose primary malignancy originates outside the lung (Griffeth, 2005). This imaging modality is uniquely suited to address the complex diagnostic challenges associated with evaluating lung nodules in such patients. For individuals with cancers like breast cancer, colon cancer, or lymphoma, the appearance of lung nodules can signal a range of possibilities, from benign inflammatory or infectious conditions to potentially life-threatening metastatic lesions. The difficulty in distinguishing between benign and malignant nodules in these patients is a significant clinical challenge. Lung nodules that appear in patients with a known history of non-lung cancers always raise the concern of metastatic spread, necessitating a thorough and careful assessment to guide appropriate treatment decisions. In this context, PET/CT imaging, particularly when combined with advanced DL models, offers a sophisticated approach to improve diagnostic accuracy. The use of ^{18}F FDG PET/CT scans allows for the metabolic characterization of lung nodules, providing additional insights beyond the structural details visible on CT alone. By analyzing the metabolic activity within the nodules, PET/CT can help differentiate between benign, malignant, and suspicious lesions with greater precision. The study highlighted here leverages DL models trained on PET/CT data from a diverse array of malignancies, including breast cancer, colon cancer, and lymphoma. By refining our understanding of how lung nodules should be interpreted in patients with these non-lung cancers, the research aims to develop tools that can accurately distinguish between different nodule types. This, in turn, empowers clinicians to make more informed decisions, potentially improving patient outcomes by identifying metastatic disease earlier and more accurately.

1.6. Problem Statement

The detection and classification of pulmonary nodules present significant challenges in medical imaging, particularly in the context of malignancies originating outside of the lungs. Despite

advancements in DL approaches, a major limitation in the current research lies in the generalizability of these models across diverse datasets. Many studies demonstrate success with DL models on specific datasets but often fail to maintain their accuracy when applied to different datasets, imaging protocols, or patient populations. This inconsistency underscores the critical need for rigorous evaluation of model generalizability, especially in detecting pulmonary involvement across varied clinical scenarios. The challenge of generalization is particularly relevant in cases of COVID-19 detection from CT scans and nodule detection in ^{18}F FDG PET/CT imaging, where imaging variations are common.

Another key issue is the detection of pulmonary nodules in patients with non-lung cancers, such as breast, colon, or lymphoma. In these cases, the potential for metastasis complicates the diagnostic process, and the traditional binary classification of lung nodules into benign and malignant categories often falls short. Some nodules require ongoing monitoring due to their indeterminate nature, a nuance that is typically overlooked in most existing DL models. Additionally, there is a notable gap in research on the detection and classification of nodules originating from non-lung malignancies, an area critical to improving early intervention and treatment outcomes. Moreover, the scarcity of publicly available PET/CT datasets limits the development of robust and generalizable DL models. To address this limitation, researchers often use transfer learning, where models pre-trained on one dataset are adapted to another. However, the effectiveness of transfer learning depends heavily on the similarity between the source and target datasets. When substantial differences exist, model performance deteriorates, highlighting the need for further exploration of generalization techniques and model adaptation strategies.

This study aims to address these challenges by developing a novel DL-based approach for the detection and multi-class classification of pulmonary nodules using ^{18}F FDG PET/CT imaging. The research will emphasize improving model generalizability across different datasets and imaging conditions, and it will rigorously assess how DL models perform in varied clinical scenarios. Additionally, the study will focus on tackling the complexity of nodule classification, especially for patients with malignancies originating outside the lungs. By refining model architectures, optimizing transfer learning strategies, and utilizing larger, more diverse datasets, this study seeks to create a more accurate, generalizable, and clinically relevant solution for pulmonary nodule detection and classification, aiding in better-informed clinical decision-making.

1.7. Research gap

Regarding the detection and classification of lung involvement, the challenge of generalizability remains underexplored in the existing literature. Achieving good results with a DL model on one dataset does not guarantee similar success on another, making it essential to rigorously evaluate the generalizability of these models in studies focused on pulmonary involvement detection and

classification. Furthermore, early detection of malignant pulmonary nodules is crucial, especially considering the high prevalence of lung metastases, as this significantly improves the chances of successful treatment. Most studies that employ DL for pulmonary nodule detection using ^{18}F-FDG PET/CT imaging have concentrated on lung cancer, while other malignancies that can cause malignant lung metastases have been overlooked. There is a notable gap in research focusing on the detection of pulmonary nodules that originate from malignancies other than lung cancer. Additionally, most studies have simplified the characterization of pulmonary nodules into two categories: benign and malignant. In clinical practice, however, there are nodules that require ongoing monitoring before a definitive diagnosis can be made. This oversimplification does not reflect the complexities encountered in real-world medical settings. There is a notable gap in research focusing on the detection of pulmonary nodules that originate from malignancies other than lung cancer. Additionally, most studies have simplified the characterization of pulmonary nodules into two categories: benign and malignant. In clinical practice, however, there are nodules that require ongoing monitoring before a definitive diagnosis can be made. This oversimplification does not reflect the complexities encountered in real-world medical settings. Moreover, there are limitations due to the scarcity of publicly available PET/CT datasets and the difficulties associated with accessing private datasets. As a result, many researchers rely on transfer learning, where models are pre-trained on one dataset and then adapted to a new, related task. While transfer learning can be effective, its success depends heavily on the similarity between the initial and target problems. This effectiveness of the method diminishes when there are substantial differences between the datasets used for training and those used for the target application. Moreover, the amount of data available and the need for robust generalization studies are critical factors in improving lung involvement detection and classification. Despite this, many existing studies have been constrained by the use of small PET/CT image datasets, limiting their ability to produce widely applicable results. To advance the field, future research must focus on enhancing model generalizability, exploring the detection of pulmonary nodules from a variety of malignancies, and refining the characterization of nodules to better align with clinical realities.

This study aims at developing a DL model for pulmonary involvement detection and classification in ^{18}F FDG PET/CT imaging. A DL-based approach will be developed to train multiple models for pulmonary nodule detection and classification. This study is uniquely positioned to not only enhance the detection of lung nodules but also to specifically tackle the complex challenge of evaluating lung nodules in patients who have malignancies originating outside the lung. In such patients, the risk of metastatic disease is significant, complicating the diagnostic process. Lung nodules detected in these individuals could represent a range of possibilities—from benign conditions, such as inflammatory or infectious processes, to metastatic lesions from a primary malignancy located elsewhere in the body. The traditional challenge in clinical practice has been

the differentiation between benign and malignant lung nodules, particularly in patients with a known history of non-lung cancers. For instance, patients with breast cancer, colon cancer, or lymphoma often develop lung nodules that may be incidental findings, but the possibility of these nodules representing metastatic disease always necessitates careful assessment and sometimes it needs extra follow-up examinations. This study leverages the power of DL models applied to ^{18}F-FDG PET/CT scans to provide a more sophisticated approach to this differentiation. By focusing on a diverse array of malignancies—such as colon cancer, breast cancer, and lymphoma and so forth—this research tries to refine our understanding of how lung nodules should be interpreted in patients with non-lung malignancies. The goal is to develop DL-based approach that can accurately distinguish between benign, malignant, and suspicious nodules, thus providing clinicians with a powerful way to make more informed decisions.

1.8. Hypothesis

The DL application for the detection and classification of pulmonary involvement, including COVID-19, from CT scans can enhance the generalizability of DL models across different datasets. It is further hypothesized that applying these advancements to ^{18}F-FDG PET/CT scans will improve the classification of lung nodules, particularly in categorizing them into benign, malignant, and indeterminate nodules, including those originating from malignancies outside the lung.

1.9. Motivation

The accurate detection and classification of pulmonary conditions, such as those associated with COVID-19, are crucial for effective clinical management. With the surge in COVID-19 cases, the demand for reliable DL models that can generalize well across different populations and imaging conditions has become increasingly evident. However, a significant challenge remains: achieving consistent performance across diverse datasets. Many existing models perform well on specific datasets but fail to maintain accuracy when applied to different patient populations or imaging protocols. This inconsistency highlights the importance of rigorous evaluation of model generalizability in the context of pulmonary involvement detection and classification. This study addresses this gap by focusing on the development of DL models optimized for generalizability in detecting and classifying pulmonary involvement, including COVID-19, from CT scans. By refining pre-processing steps, such as lung segmentation and CT resampling, and exploring model architectures, the research aims to create DL models that perform reliably across varied datasets. This is essential for improving diagnostic accuracy in real-world clinical settings, where variability in imaging data is common. The research also explores the effectiveness of transfer learning, where models pre-trained on one dataset are adapted to a new, related task. This strategy is particularly relevant when dealing with limited data, but its success depends heavily on the similarity between the initial and target datasets. By optimizing this process and combining

datasets strategically, the study seeks to overcome the limitations of small sample sizes and improve the overall performance of DL models in lung involvement classification. Building on the advancements made in the generalizability of CT-based models, this study further narrows its focus to the detection and classification of lung nodules using ^{18}F-FDG PET/CT scans. The accurate classification of lung nodules is particularly challenging in patients with a history of non-lung malignancies, where nodules could represent benign conditions, metastatic lesions, or other suspicious abnormalities. The traditional binary classification into benign or malignant does not adequately reflect the complexities encountered in clinical practice, where some nodules require ongoing monitoring before a definitive diagnosis can be made. This research is uniquely positioned to address these challenges by leveraging large ^{18}F FDG PET/CT datasets and state-of-the-art DL models. The goal is to develop a DL-based approach that can not only differentiate between benign and malignant nodules but also identify those that require follow-up, thus aligning more closely with clinical realities. By focusing on nodules that may originate from malignancies outside the lung, such as breast cancer, colon cancer, or lymphoma, the study aims to refine the diagnostic process and provide clinicians with a powerful tool for more accurate and informed decision-making. This approach is critical for improving patient outcomes, particularly in cases where the risk of metastatic disease is significant. A key challenge in developing robust DL models for lung nodule detection is the scarcity of large, publicly available PET/CT datasets. This study addresses this limitation by utilizing a large dataset, which is essential for training models that can generalize well across different clinical scenarios.

1.10. Research objectives

1.10.1. *Aim*

We are going to develop a model architecture derived from DL approach that will be reliable and that will assist physician for more accurate diagnosis of pulmonary nodules from ^{18}F FDG PET/CT imaging.

1.10.2. *Objective 1*

The objective of this research is to develop and apply DL models for detecting and classifying pulmonary involvement, including COVID-19, from CT images. The study aims to evaluate the performance and generalizability of these models across different datasets and clinical scenarios. Additionally, the research seeks to identify the most effective pre-processing steps, such as segmentation and data augmentation, to enhance the accuracy and efficiency of the DL models in accurately diagnosing pulmonary involvement from CT scans.

1.10.3. *Objective 2*

Building on the insights gained from Objective 1, this objective focuses on detecting and performing multi-class classification of pulmonary nodules from a subset of ^{18}F FDG PET/CT

data using state-of-the-art DL models. By leveraging the pre-processing techniques and generalizability assessments identified in Objective 1, this phase aims to determine the best-performing DL model and approach for training ^{18}F FDG PET/CT data. The goal is to refine the methodology for accurately classifying pulmonary nodules into multiple categories.

1.10.4. *Objective 3*

Building on the knowledge and methodologies developed in Objectives 1 and 2, this objective aims to create an optimized DL-based approach to enhance accuracy in pulmonary nodule detection and multi-class classification using a large ^{18}F FDG PET/CT dataset. By applying the best-performing models and pre-processing techniques identified earlier, the goal is to further improve the classification of pulmonary nodules into benign, malignant, and suspicious categories.

1.11. Research Questions

The following research questions will be addressed in my thesis.

Based on objective 1:

What are the pre-processing steps that improve pulmonary involvement classification?

How is the state-of-the-art 3D CNNs performance on large dataset (for COVID-19) for pulmonary involvement classification?

Can different combinations of datasets assist generalization?

Based on objective 2:

With hybrid imaging, such as ^{18}F FDG PET/CT, what are the steps required for feeding images to the DL models, considering the small size of pulmonary nodules?

Which available state-of-the-art 3D CNNs perform better on ^{18}F FDG PET/CT imaging for pulmonary nodule detection?

Can the DL models classify the pulmonary nodules seen in ^{18}F FDG PET/CT scans in more categories than malignant-benign?

Based on objective 3:

What is the effect of using large PET/CT dataset for pulmonary nodule detection?

What are the characteristics of the developed model for pulmonary nodule detection and classification using ^{18}F FDG PET/CT scans?

How much does the developed model improve the results of pulmonary nodule detection and classification compared to available DL models?

1.12. Novelty and contribution of the research

1.12.1. *Fundamental novelty and contribution*

This study makes a significant fundamental contribution by advancing the application of optimized DL models and techniques for the detection and classification of pulmonary involvements, including COVID-19 and lung nodules, from CT and ^{18}F-FDG PET/CT images. Through innovative approaches in dataset combination, model optimization, and pre-processing enhancements, the research demonstrates that high levels of accuracy and generalizability can be achieved even in scenarios where large datasets are not readily available. A key insight from the study is the effective strategy of combining 80% of one dataset with a 40% or greater contribution from another, which yields results comparable to using the entire combined datasets. This finding is particularly important in the context of clinical applications, where the availability of large, well-curated datasets is often limited. Another key contribution of this study is performing multi-class classification, rather than the typical binary classification, for lung nodule diagnosis using a large ^{18}F FDG PET/CT image dataset. The nodules are categorized into three classes: benign, malignant, and those requiring further follow-up, better reflecting real-world clinical scenarios. Furthermore, the study highlights the crucial role of pre-processing steps—such as lung segmentation, PET/CT resampling, fusion, and the exclusion of zero-valued voxels—in significantly enhancing the efficiency and performance of DL models. These advancements contribute to the broader understanding of how to optimize DL models for complex medical imaging tasks, laying the groundwork for future developments in DL-driven diagnostic tools.

1.12.2. *Application novelty and contribution*

A significant contribution of this study is the discovery that combining 80% of one dataset with at least 40% of another dataset and above, produces results comparable to using the full combined datasets. This approach is particularly valuable for clinical applications, where access to large, well-curated datasets is often limited. Another application contribution of this study is particularly noteworthy in the context of pulmonary nodule detection and classification using ^{18}F-FDG PET/CT data. By evaluating state-of-the-art DL models and incorporating advanced techniques like varying patch slice thicknesses, the research enhances the robustness and classification accuracy of these models. The study successfully meets its objective of detecting and categorizing pulmonary nodules, demonstrating high sensitivity and specificity for malignant nodules, which exhibit distinct metabolic activity. However, it also identifies the ongoing challenges in accurately classifying benign and suspicious nodules due to their overlapping imaging features. This underscores the need for continuous refinement of DL models to improve diagnostic accuracy across a broader spectrum of pulmonary conditions. Additionally, the research emphasizes the importance of using well-curated clinical datasets, which provide more reliable training data compared to publicly available datasets that may suffer from biases or inaccuracies. The

optimized DL model developed in this study, with its enhanced 3D convolutional layers and fine-tuned architecture, shows strong potential for clinical application, particularly in the accurate and timely diagnosis of various types of lung malignancies. The findings of this study have significant implications for improving diagnostic accuracy and patient outcomes in pulmonary imaging, offering a powerful tool for clinicians in the management of lung-related diseases.

1.13. Thesis outline

Chapter 1 introduces the significance of detecting pulmonary nodules, particularly in the context of malignancies originating outside the lungs, like metastasis from other cancers. It discusses the challenges of using ^{18}F FDG PET/CT imaging for such diagnoses and the role of DL models in improving diagnostic accuracy. The chapter also lays out the problem statement, research gap, hypothesis, motivation, aim and objectives of the study, which aim to enhance the generalizability of DL models in detecting pulmonary involvement across diverse datasets and clinical conditions. The comprehensive literature review of existing studies has been presented in chapter 2 which covers the application of DL techniques in medical imaging, with a focus on pulmonary nodules and lung involvement from COVID-19 and other malignancies. It compares various DL approaches, examines the challenges of generalizability, and highlights gaps in the current research, such as the need for robust multi-class classification methods in PET/CT imaging. Chapter 3 presents the datasets, pre-processing methods, and DL models used in the study. It includes detailed descriptions of the CT and PET/CT data, segmentation and augmentation processes, and the deep learning architectures implemented. It also outlines the classification tasks for COVID-19 lung involvement and pulmonary nodules, as well as the transfer learning experiments conducted to assess generalizability across different datasets. Chapter 4 discusses the results of the experiments, including the performance of DL models on COVID-19 lung involvement classification and multi-class lung nodule detection using PET/CT images. The chapter explores the accuracy, sensitivity, and specificity of models, evaluates their generalizability across datasets, and discusses the use of explainable AI techniques like SHAP to interpret model outcomes. Finally, chapter 5 provides a general conclusion based on the research findings, emphasizing the success of the proposed DL models in achieving high accuracy in lung involvement and nodule classification. It also discusses the limitations of the study and provides recommendations for future research, particularly regarding the need for larger and more diverse datasets.

CHAPTER 2

2. LITERATURE REVIEW

2.1. Introduction

The intersection of DL and medical imaging has ushered in a new era of diagnostic precision and efficiency, particularly in the field of pulmonary diseases. As the COVID-19 pandemic unfolded, the need for rapid, accurate, and automated diagnostic tools became paramount, driving significant advancements in the use of DL models for analyzing CT scans of affected lungs. These models have shown promise in detecting, classifying, and segmenting lung involvement in COVID-19 patients, providing critical support in managing the global health crisis. However, the potential of DL extends far beyond COVID-19, includes a wide range of pulmonary imaging applications, including the detection and classification of lung nodules using advanced imaging modalities such as ^{18}F-FDG PET/CT. First, we give a comprehensive literature review on the application of DL models in medical imaging focusing on CT and PET/CT images. Moreover, this review identifies key research gaps that must be addressed to further advance the field. These include the need for multi-class classification models that reflect the complexities of clinical decision-making, the incorporation of 3D data to enhance diagnostic accuracy, and the development of robust DL models that are trained from scratch on specialized datasets. Additionally, the review highlights the importance of comprehensive model comparisons and the expansion of dataset sizes to improve the real-world applicability of these technologies. Then, this comprehensive literature review explores the dual focus of DL in medical imaging: the automated diagnosis of lung involvement in COVID-19 from CT images and the detection and characterization of pulmonary nodules using ^{18}F-FDG PET/CT. In both areas, while significant progress has been made, challenges remain, particularly regarding the generalizability of DL models across diverse datasets and clinical scenarios. The review critically examines the methodologies employed in existing studies, the results achieved, and the limitations encountered, with particular attention to issues such as dataset bias, the role of lung segmentation, and the effectiveness of multi-dataset validation.

By synthesizing findings across a broad spectrum of studies, this review not only provides a detailed understanding of the current state of DL in pulmonary imaging but also offers insights into the future directions necessary to fully realize the potential of these technologies in clinical practice. Through this exploration, the review underscores the critical role of DL in transforming diagnostic radiology and enhancing patient outcomes across various pulmonary conditions.

2.2. Literature review on deep learning applications in medical imaging

DL has emerged as a powerful tool in various domains, particularly those requiring the analysis of large and complex datasets. It is a subset of ML that utilizes artificial neural networks to learn patterns from data, making it highly effective in tasks like image and speech recognition. Unlike traditional ML models, DL models automatically learn from data through multiple layers of abstraction, enabling them to recognize intricate patterns without the need for manual feature extraction. The ability to process vast amounts of data efficiently has positioned DL as a game-changer in fields like healthcare, especially in medical imaging.

The application of DL in medical imaging began to gain prominence after the breakthrough of AlexNet in 2012 (Alom et al., 2018), which showed that CNNs could outperform traditional methods in image classification. Medical imaging, which plays a critical role in disease diagnosis, treatment planning, and monitoring, was a natural candidate for DL due to the complexity of the images involved (Alzubaidi et al., 2021; Lee et al., 2017; Li et al., 2023; Sarvamangala & Kulkarni, 2022; Yadav & Jadhav, 2019). Figure 2.1 illustrates the typical CNN architecture.

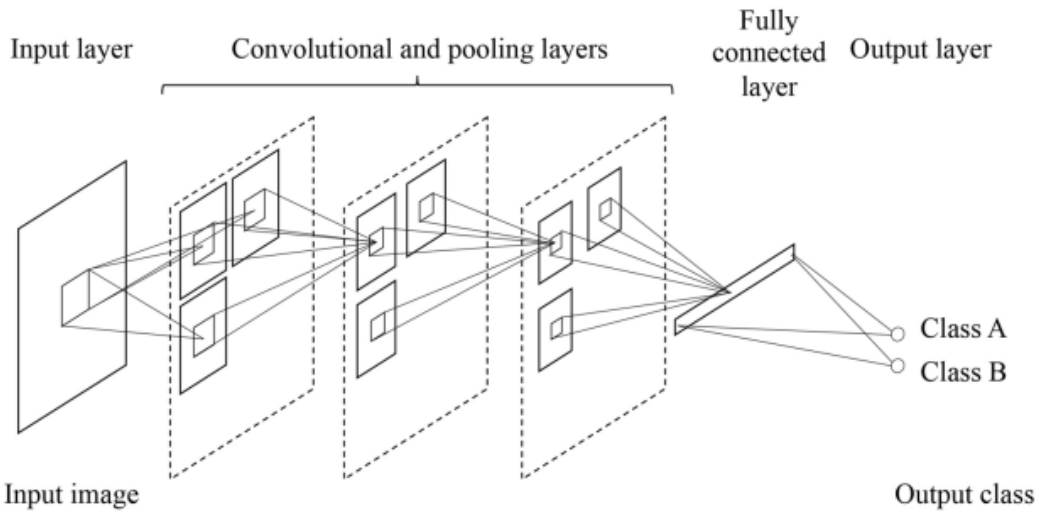


Figure 2.1. A typical CNN architecture (Suzuki, 2017)

Initially, the adoption of DL in medical imaging was slow, as researchers had to overcome several challenges, including limited availability of annotated medical datasets, computational limitations, and the necessity for highly accurate predictions. However, as larger annotated datasets became available and computing power improved, DL-based models started to outperform conventional image analysis methods, particularly in radiology (Artesani et al., 2024; Liu et al., 2024). CT imaging, a cornerstone in diagnostic radiology, provides high-resolution cross-sectional images of the body, enabling clinicians to detect and diagnose various conditions, particularly in oncology, neurology, and cardiology (Di Carli et al., 2016; Ghouri et al., 2019; LeVine III, 2010; Panayides et al., 2020; Rubin, 2014). DL has been applied to CT imaging in

several ways, enhancing the accuracy and efficiency of tasks that previously relied on manual interpretation by radiologists. Low-dose CT scans are critical in minimizing patient radiation exposure, but they often suffer from poor image quality due to noise. DL models, particularly CNNs, have been developed to denoise these images by mapping low-dose scans to their high-dose counterparts, allowing for improved visualization of anatomical structures without compromising diagnostic accuracy (Domingues et al., 2020). The residual encoder-decoder networks proposed by Chen et al. (Chen et al., 2017) have been particularly successful in this regard, significantly enhancing the quality of low-dose CT images. In oncology, early detection of tumors is essential for effective treatment (Jiang et al., 2023). DL-based segmentation algorithms, particularly U-Net architectures, have been employed for tumor detection in CT images. These models are capable of learning complex features from medical images, enabling them to accurately delineate tumor boundaries. Fully automated systems using U-Nets have been widely adopted in lung and liver cancer detection, providing a more precise and efficient alternative to manual segmentation (Rayed et al., 2024).

The outbreak of the COVID-19 pandemic saw an urgent need for rapid and accurate diagnostic tools. DL models were quickly adapted to detect COVID-19-related pneumonia on chest CT scans (Bhatele et al., 2024; Fusco et al., 2021; Gunraj et al., 2020; Harmon et al., 2020; Heidari et al., 2022; Zandehshahvar et al., 2021). These models were trained on datasets from patients with confirmed COVID-19 cases, learning to identify characteristic features such as ground-glass opacities and consolidation. The ability of these models to quickly and accurately identify COVID-19 in CT scans highlighted the adaptability of DL in responding to emerging medical crises (Harmon et al., 2020).

PET/CT imaging offers a powerful tool for the detection and monitoring of various diseases, particularly cancer (Ambrosini et al., 2012; Endo et al., 2006; Farwell et al., 2014; Glaudemans et al., 2013; Lawrence et al., 2010; Weber & Figlin, 2007). PET/CT imaging provides both metabolic and anatomical information, making it a crucial tool in oncology. However, PET imaging is often associated with high radiation exposure, and the images can be degraded by noise due to the randomness of the physical processes involved in PET signal acquisition. Reducing the radiation dose in PET scans is a major goal in the field of medical imaging. DL models have been employed to generate high-quality PET images from low-dose scans (Gong et al., 2019; Kaplan & Zhu, 2019). For example, CNN-based models can map low-dose PET images to full-dose equivalents, preserving the diagnostic quality while minimizing patient exposure to harmful radiation (Kaplan & Zhu, 2019; Xiang et al., 2017).

Accurate segmentation of lesions in PET/CT images is critical for the diagnosis and treatment of cancer. DL-based models, particularly 3D CNNs, have been developed for the segmentation of tumors and lesions in PET/CT scans (Kao & Yang, 2022). These models not only delineate tumor

boundaries but also provide volumetric data, which is crucial for treatment planning, particularly in radiation therapy. For example, Zhong et al. demonstrated the use of 3D FCNs for the segmentation of tumors in PET/CT images, showing improved accuracy over traditional methods (Zhong et al., 2018).

One of the advantages of PET/CT is the ability to combine metabolic information from PET with anatomical information from CT. Similarly, DL models have been developed to integrate data from both modalities, allowing for a more comprehensive analysis of disease. This has proven particularly useful in the detection of metastases, where cancer may spread to multiple organs, and a single modality may not provide sufficient information.

2.2.1. *Applications of Deep learning in lung studies*

In some studies, the combination of DL and radiomics is employed for lung cancer studies. Afshar et al. (Afshar et al., 2020), developed a DL-based radiomics model (CNN-based) to predict time-to-event outcome which predicts the risk of death or recurrence based on input images. They concluded that DL-based radiomics outperforms hand-crafted radiomics and has the capacity to be used in patient-specific management of lung cancer. Their proposed model consists of two channels to train on both CT and PET components of PET/CT. Han et al. (Han et al., 2021) conducted a study about the classification of histologic subtypes of non-small cell lung cancer and developed an optimal model from radiomics-based machine learning and the DL algorithm using FDG PET/CT images. They examined 10 different feature selection methods, 10 different machine learning models, and Vgg16 as the DL model, which outperformed all methods. Lin et al. (Lin et al., 2024), developed a combined model integrating DL, radiomics, and clinical data to classify lung nodules as benign or malignant, further refining the classification into specific pathological subtypes and Lung-RADS scores. The model was trained on the 3D CT images from Lung Nodule Analysis 2016 (LUNA16) public dataset and two private CT datasets (Lung Nodule Received Operation (LNOP) and Lung Nodule in Health Examination (LNHE). It utilized a stacked ensemble approach with the AutoGluon-Tabular classifier, combining 3D CNN features, radiomics, and clinical data. The model achieved high accuracy, with 92.8% for benign/malignant classification and strong performance in both subtype and Lung-RADS classification tasks. The datasets included 1004 nodules from the LUNA16, 1027 from LNOP, and 1525 from LNHE. The model's performance in distinguishing between benign and malignant nodules, as well as its accuracy in classifying pathological subtypes and Lung-RADS scores, demonstrates its potential for enhancing lung cancer diagnostics. This combined approach could significantly improve the precision of lung nodule evaluation in clinical settings, aiding in the early and accurate diagnosis of lung cancer.

Multimodality image fusion is another focus of researchers. DL models are increasingly being used to fuse data from multiple imaging modalities, such as combining PET and CT or MRI and

CT data. This allows for more accurate diagnosis and treatment planning, as it combines the strengths of each modality to provide a more detailed view of the patient's condition.

2.2.2. *Comparison between 2-dimentional and 3-dimentional CNNs*

Three-dimensional imaging has become increasingly important in medical diagnostics. Medical imaging in 3D is a technology that uses 2D slices in order to create 3D imaging to provide a depth of the internal body for analysis. It is believed that the reason for superiority of using 3D CNN over 2D CNN and 1D is that 3D CNN can extract both spectral and spatial features from 3D input image and in this way, it can reduce false-positive reduction. Whereas 2D CNN only can extract spatial features and 1D CNN models can extract spectral features from volumetric input data (Firat et al., 2023; Shen et al., 2024). DL models are computationally intensive when implementing 3D medical images. However, the recent advancements in high-speed GPUs allow researchers to cope with this issue. Also, patch-based methods can reduce the computational burden. There are studies demonstrating that DL application using 3D medical images outperform 2D. For instance, Yan et al. (Yu et al., 2020), showed better classification performance of 3D CNN for pulmonary nodule malignancy over 2D CNN. Xiao et al. (Xiao et al., 2020), found that a 3D-Res2Unet model outperformed a 2D network in the task of lung nodule segmentation, further supporting the superiority of 3D approaches in certain medical imaging applications. The passage argues that while 3D CNNs require more computational resources, their ability to capture more detailed and comprehensive features from 3D medical images makes them a more effective tool in certain medical imaging applications, particularly with the help of modern computational advancements.

DL models have been particularly effective in analyzing 3D medical images, such as those generated by CT, MRI, and PET/CT scans. The use of 3D CNNs has allowed for more accurate segmentation and classification of volumetric data. For example, in the case of lung cancer, 3D CNNs can be used to segment tumors across multiple slices of a CT or PET/CT scan, providing a more comprehensive view of the tumor's size, shape, and location (Kao & Yang, 2022). DL models have also been used to reconstruct 3D images from 2D slices, which is particularly useful in surgical planning and radiation therapy. By reconstructing a 3D model of a patient's anatomy, clinicians can better plan interventions and tailor treatments to the patient's specific needs.

While DL has made significant advancements in medical imaging, several challenges remain. One of the key challenges is the generalizability of DL models. Many models perform well on specific datasets but fail to generalize to new data from different patient populations or imaging devices. This highlights the need for larger, more diverse datasets and more robust training methods.

Another challenge is the interpretability of DL models. While these models can provide highly accurate predictions, they are often considered "black boxes," making it difficult for clinicians to

understand how a model arrived at a particular decision. Future research should focus on developing more interpretable DL models that can provide insights into their decision-making processes.

Finally, the integration of DL with emerging technologies, such as augmented reality and virtual reality, could further enhance the application of DL in medical imaging. These technologies could allow clinicians to interact with 3D reconstructions of patient anatomy in real-time, providing more accurate and personalized care.

Deep learning has revolutionized medical imaging by providing more accurate, efficient, and automated solutions to many of the challenges faced by radiologists. From improving image quality in low-dose CT scans to providing more accurate tumor segmentation in PET/CT images, DL models have become essential tools in modern healthcare. As the field continues to evolve, the focus will likely shift towards improving the generalizability and interpretability of DL models, as well as integrating them with emerging technologies to further enhance their clinical utility.

2.3. Literature review on lung involvement from COVID-19 CT images

There are many studies that used publicly available DL models for lung involvement detection, classification, and segmentation. Upon the global outbreak of the recent COVID-19 pandemic, the need for computer-aided diagnosis methods has significantly increased (Ardakani et al., 2021; Gudigar et al., 2021; Rubin, 2019; Shoeibi et al., 2024). Most studies conducted on automated COVID-19 diagnosis from CT images using a single, internal dataset for training, validation, and testing DL models, resulting in high classification metrics (He et al., 2021; Wang et al., 2020). It is not possible to assess whether these results are driven by classifier sensitivity to disease pathology or bias introduced by class imbalance, patient selection, or confounding bias. This is particularly a concern where disease-positive and negative disease patients have been sourced independently, potentially introducing systematic differences in image classes related to CT acquisition apparatus, operational parameters, and regional patient morphological differences. Such biases have been found to result in considerably lower classification metrics when these models are tested against external datasets (Bassi & Attux, 2022; Maguolo & Nanni, 2021). A small number of studies focused on investigating the generalization of AI-based COVID-19 diagnosis (Harmon et al., 2020; Horry et al., 2021; Nguyen et al., 2021). Harmon et al. (Harmon et al., 2020) combined four datasets into combinations of training, validation, and testing image corpora by excluding one dataset consisting of 147 patients as a holdout test set. They used DenseNet-121 as 3D CNN and implemented both segmented lung and full 3D image classification, considering one complete volume at a fixed size. They achieved 90.8% accuracy, 84% sensitivity, and 93% specificity. In this study, a total combination of datasets was performed, and the results were tested on a comparably small dataset. The authors considered one fixed

dataset as a test dataset and there is a lack of external validation on each four datasets to demonstrate their network capability to achieve similar results.

In the current study, we seek the results of one trained dataset when tested on the other datasets and the effect of data augmentation on generalization.

In separate work, Nguyen et al. (Nguyen et al., 2021) used four different datasets, including one internal dataset at UT Southwestern (UTSW) (337 patients) and three external datasets: 1) China Consortium of Chest CT Image Investigation (CC–CCII), 2) COVID-CT set and 3) MosMedData (Morozov et al., 2020). They implemented nine combinations of these datasets for two classes of COVID-19 positive and COVID-19 negative cases. They both trained the different combinations and tested on an external test dataset and trained the different combinations and tested on a holdout test set from one of the datasets used for training. They used different models for training on 3D CT images from which the best results were for the models trained on multiple datasets and evaluated on a test set from one of the datasets used for training (accuracy of 86–97%). Despite these promising internal classification results, classification metrics for these models were reduced to pure chance when evaluated against an external dataset, with an area under the curve (AUC) of 0.5 calculated for all models. Nguyen et al. (Nguyen et al., 2021) adjusted the disease positive probability threshold to maximize accuracy in their simulations, thereby tightly binding model performance to the test dataset. This study did not segment the lung field from the CT images to reduce signal noise from features including ribs/bone and surrounding areas. In the present study, we have used 0.5 as the disease positive probability threshold for all models to decouple results from datasets, and lung segmentation was performed in the pre-processing part of the current study.

Bhuyan et al. (Bhuyan et al., 2022), presented a system for mass segmentation and classification of COVID-19 from X-Rays or CT scans using YOLO-based DL model. In their 2D study, they used 2794 images from 150 COVID-19 patients for training and 1061 images from 50 COVID-19 patients for test. They stated that this system can be used for high accuracy and sensitivity. Ni et al. (Ni et al., 2020), conducted a study for automatic COVID-19 detection and compared the results with the specialists' assessments. In their 3D CNN approach, they used convolutional MVP-Net to detect infected area and 3D U-Net to segment COVID-19 involvements of 14435 patients with chest CT images. They obtained better results compared with radiological residents. Segmentation of COVID-19 involvements is also studied in a number of studies. Ghomi et al. (Ghomi et al., 2020), conducted a study to segment COVID-19 involvements in a 2D study using 2469 CT slices and DL. They obtained 0.954 accuracy for COVID-19 involvement segmentation from slices and stated that DL approached have acceptable results in COVID-19 involvement segmentation and can assist physicians to find and localize suspected abnormal areas in CT scans.

More recently, two comprehensive studies addressed generalization aspect of COVID-19 classification task. Li et al. (Li et al., 2021), proposed the contrastive multi-task convolutional neural network (CMT-CNN) as a multi-task framework to increase generalizability. The authors stated that there is no need for further annotation to improve generalization using CMT-CNN. They used 3D volumes of CT images from two datasets: one from CC-CCII2 (Zhang et al., 2020) with 4356 CT images and one from their hospital consisting of 402 CT images from 108 COVID-19 diagnosed patients confirmed by RT-PCR test. For X-ray, they used three datasets, including two public datasets from Cohen et al. (Cohen et al., 2020) and Kaggle and one from their hospital-based dataset with 231 COVID-19 cases in total. They used Mendeley Data website, containing 4007 pneumonia and 1583 normal cases as their normal control instances. Certain augmentation methods, including distortion, painting, and perspective transformations, improved representational learning capability. The results of their study indicate 5.49–6.45% generalization accuracy improvement for CT and 0.96–2.42% for X-ray images. In another study, Aversano et al. (Aversano et al., 2021) combined three pre-trained deep neural networks, including VGG-19 (Simonyan & Zisserman, 2014), Xception (Chollet, 2017), and ResNet-50 (He et al., 2016), evolved with a direct coding scheme based on genetic programming to develop an ensemble classifier for each lung lobe (superior, middle, and inferior). The main parts of their proposed ensemble architecture are multiple deep neural networks based on pre-trained models and a voting strategy. For the training phase, they used two volumetric CT datasets, Extensive COVID-19 X-Ray and CT Chest Images Dataset and Coronavirus (COVID-19) CC-19 dataset (Kumar et al., 2021), then clustered them into three sub-datasets comprising images of each lung lobe. To evaluate the results on external data, they used SARS-COV-2 Ct-Scan Dataset (Soares et al., 2020). The pre-trained transfer learning CNN models combined with VGG-19, ResNet-50, and Xception were re-trained for the binary classification of CT images of COVID-19 versus normal cases. The genetic algorithm in this study executes an evolutionary process to identify the best architecture adaptation of the pre-trained models. The evaluation results on the external test dataset showed F1 score of 0.903 while it was 0.94–0.95 for their integrated dataset. In Refs. (Aversano et al., 2021; Li et al., 2021) studies, whole datasets were considered training and test datasets to assess generalizability. There is a lack of true external validation for each dataset (i.e., considering each dataset as an external test dataset in different simulations), and the applicability of trained models to real-life clinical situations is unknown.

A few previous studies have assessed the generalizability of CNN models trained on 2D CT slices and X-ray images. In a study by Silva et al. (Silva et al., 2020), that was performed on 2D CT slices, EfficientCovidNet, was proposed along with a voting-based approach and a cross-dataset analysis for COVID-19 detection. They evaluated EfficientCovidNet on three setups and with the two largest public CT datasets, including a cross-dataset analysis. The results of this study

indicated the accuracy drops from 87.68 to 56.16% for the external COVID-19 test set. Ahmed et al. (Ahmed et al., 2021) demonstrated a significant gap between the model tested on before-seen data (same source) and the model tested on external data in COVID-19 detection from X-ray images. Their developed model reached the AUC of 1.00 when tested on seen data while it was only 0.38 on external data. Hence, they recommended further investigations into finding/focusing on features that can be generalized across datasets.

Bassi et al. (Bassi & Attux, 2022) tested the effect of segmentation on X-ray image classification of COVID-19, normal, and pneumonia cases. It was shown that segmenting lung has a positive effect on the model generalization capability, increasing the mean accuracy score on the external test dataset by 4.7% and the Bayesian estimation means by 4.4%. The results when tested on the external dataset, showed 85% sensitivity for COVID-19 detection in the case of the segmented lung being used while it was 81% for non-segmented lung. They stated that the improvement in accuracy might be due to the attention of DNN to the lung region. Lung segmentation can also reduce dataset bias and improve generalization.

Shah et al. (Shah et al., 2021) proposed an AI-driven framework for COVID-19 diagnosis using 3D CT images, integrating a multi-level feature extraction strategy. Utilizing both internal and external datasets, they emphasized the importance of dataset diversity. The framework achieved a classification accuracy of 92.4% on internal datasets but saw performance decline to 75.6% on external datasets, underscoring the generalization issue. They introduced an adaptive feature fusion technique to enhance model robustness across different datasets. They also emphasized the necessity of using multiple datasets for assessing generalization. While they incorporated adaptive feature fusion, their study primarily focused on internal validation with limited external dataset evaluation. This raises concerns about the practical applicability and robustness of their model in diverse clinical settings. Chen et al. focused on the significance of transfer learning for COVID-19 diagnosis using CT images. They fine-tuned a pre-trained ResNet-50 model on a combination of internal and external datasets, finding that transfer learning markedly improved performance, achieving an accuracy of 89.1% on external datasets. They also emphasized the necessity of lung field segmentation to enhance model generalization capabilities. Wang et al. (S. Wang et al., 2021) developed a DL algorithm using a modified Inception model to screen for COVID-19 from CT images. The dataset comprised 905 confirmed COVID-19 patients and 1,127 non-COVID-19 patients. The model was trained to differentiate between COVID-19 and other types of pneumonia. The study reported an AUC of 0.98, with a sensitivity of 94% and specificity of 88% on the internal test set. The study highlighted the model's potential in accurately identifying COVID-19 cases but did not provide extensive external validation. While they achieved high accuracy on their internal test set, the lack of external validation raises concerns about the model's generalizability. Our study addresses this limitation by validating the model on multiple external

datasets, ensuring robust performance across different clinical settings and reducing the risk of overfitting to a specific dataset. In a study conducted by Li et al. (Li et al., 2020) a DL model was developed by combining UNet++ for lung segmentation and a convolutional neural network for classification to distinguish COVID-19 from community-acquired pneumonia on chest CT scans. The study included CT scans from 3,322 patients across 10 hospitals in China. The model achieved an AUC of 0.96, with 90% sensitivity and 96% specificity on the internal test set. The lung segmentation step aimed to improve model accuracy by focusing on relevant regions of interest. Despite the high internal validation performance, the study did not extensively evaluate the model on external datasets. They demonstrated the effectiveness of lung segmentation in improving classification accuracy. However, their reliance on internal validation limits the assessment of model generalizability. Our study builds on this by incorporating external validation and testing models on separate datasets, providing a more comprehensive evaluation of generalization performance. Additionally, we emphasize the importance of lung segmentation in our pre-processing pipeline to enhance model robustness and accuracy. Bai et al. (Bai et al., 2020) explored the use of artificial intelligence to augment radiologist performance in distinguishing COVID-19 from other types of pneumonia on chest CT scans. The study included 1,136 CT scans from multiple institutions. The AI model, a CNN, was trained to classify the images and then used to assist radiologists in their diagnoses. The combined AI-radiologist approach achieved an AUC of 0.95, sensitivity of 91%, and specificity of 88%, showing significant improvement over radiologists working alone. However, the study's evaluation was based on an internal dataset, limiting the assessment of model generalizability. They explored the use of artificial intelligence to augment radiologist performance in distinguishing COVID-19 from other types of pneumonia on chest CT scans. The study included 1,136 CT scans from multiple institutions. The AI model, a CNN, was trained to classify the images and then used to assist radiologists in their diagnoses. The combined AI-radiologist approach achieved an AUC of 0.95, sensitivity of 91%, and specificity of 88%, showing significant improvement over radiologists working alone. However, the study's evaluation was based on an internal dataset, limiting the assessment of model generalizability.

2.3.1. *Research gap in COVID-19 classification*

The key focus of our study is to assess the generalization of computer vision models trained on 3D CT images for automated COVID-19 diagnosis. According to the literature above, it can be seen that most available studies have selected one fixed dataset as their external test dataset. Therefore, there is a need for the study to test external validation on each dataset involved in the study since the results may vary significantly on different external test sets. Hence, we dedicated a part of this study to investigating this issue. Another research gap identified in the above studies is that all available datasets were combined together for training and testing. Although a large

number of data leads to more accurate results, the results from combinations of different data portions have not been investigated yet. We addressed the analysis of the results acquired from different combinations of dataset portions in both fully trained and transfer learning approaches. In the case of satisfactory results, the need for large dataset combinations is alleviated.

Finally, we have found that lung field segmentation plays a pivotal role in promoting model generalizations and recommend that this procedure be a standard part of the 3D CT image pre-processing pipeline for CNN-based COVID-19 diagnosis from medical images.

2.4. Literature on deep learning in PET/CT imaging across different malignancies

To gain a comprehensive understanding, we reviewed studies that applied DL techniques to the analysis of PET/CT images. Our search was based on the keywords “deep learning”, [AND] “PET/CT”, [OR] “positron emission tomography/computed tomography” (Figure 2.2) in the title and abstract of the articles. We used five databases including Scopus, PubMed, Web of Science, IEEE, and Google Scholar to include the most recent papers that have not been found in other databases. This approach resulted in 99 distinct studies, that were conducted during the time frame from 2017 to 2022 to encompass the latest research papers and progress. We interpret this as an indication that the topic has become more important in recent years.

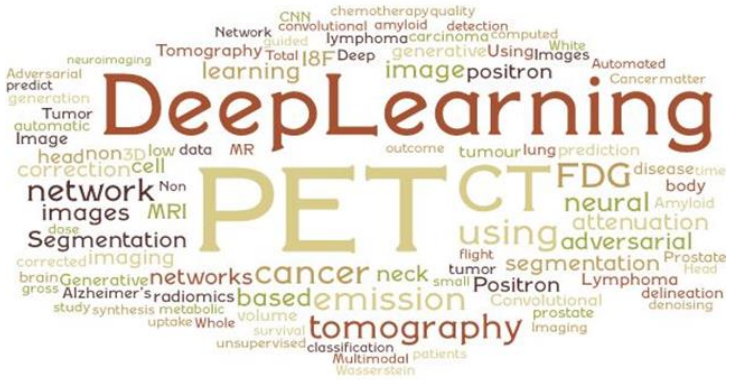


Figure 2.2. Cloud tag of the application of DL on PET/CT imaging review

Table 2.1 provides information on the organs/malignancy that have been studied, the number of studies conducted, the prevalent DL model used, the primary task performed, and the main types of data analyzed. According to the table, the most commonly studied organ/malignancy in the PET/CT imaging studies is the brain followed by lung/lung cancer, with 24 studies performed for the brain and 15 studies for the lung. Several studies (26 studies) selected the most available PET/CT images to study the particular task regardless of malignancy.

Table 2.1. The number of papers that studied different cancers and malignancies.

Studied organ/malignancy	Number of studies	DL model	Performed task	Data types
Lung	15	6 ResNet-based, 4 U-net based, 3 CNN-based	9 classification (prediction), 6 segmentation	15 private, 1 public
Brain and related disease	24	9 GAN-based, 9 ResNet-based, 4 U-net based	10 image generation, 7 classification, 5 Segmentation, 4 attenuation correction	21 private, 4 public
Whole body/Total body	26	14 U-net based, 5 GAN based, 3 ResNet-based	7 attenuation correction, 7 image generation, 4 denoising	26 private, 2 public
Breast	2	1 CNN-based, 1 U-net based	1 classification, 1 segmentation	2 private, 0 public
Head and neck	12	6 U-net based, 2 GAN, 2 CNN based, 2 DenseNet	6 segmentation, 3 classification, 2 image generation	12 private, 1 public
Lymphoma	6	4 U-net based, 1 CNN based, 1 ResNet-based, 1 DenseNet	3 segmentation, 4 classification	6 private, 0 public
Prostate	5	2 CNN-based, 1 ResNet, 1 U-net based	4 classification, 1 segmentation	5 private, 0 public
Others	9	2 U-net based, 2 CNN-based	5 classification, 3 segmentation, 1 denoising	9 private, 0 public

2.5. Literature review on nodule detection in ^{18}F FDG PET/CT2.4.1.

2.5.1. Pulmonary involvement diagnosis and classification using ^{18}F FDG PET/CT imaging

FDG absorption in the FDG-avid area is used as a source of functional information in ^{18}F-PET, and low-dose CT provides anatomical information. The combination of PET and CT has collaborative benefits of acquiring either separately, minimizing their individual limitations

(Almuhaideb et al., 2011; Beyer et al., 2000; Schöder et al., 2003). PET/CT can be used to diagnose and stage many malignancies with high efficiency and can be used for monitoring response to therapy (Eubank & Mankoff, 2005). Hadique et al. (Hadique et al., 2020), focused on the role of 18F-FDG PET/CT in evaluating pulmonary nodules detected during low-dose computed tomography (LDCT) lung cancer screening and assessed 75 patients with identified lung nodules, using 18F-FDG PET/CT to determine the likelihood of malignancy. They found that ^{18}F FDG PET/CT had a sensitivity of 94% and a specificity of 82% for detecting malignancies. Notably, 86% of the patients with malignant or indeterminate nodules on PET/CT proceeded to biopsy, confirming its value in clinical decision-making. The study highlighted PET/CT's utility in detecting incidental findings, which influenced further diagnostic and therapeutic interventions in nearly half of the patients. A study run by Kukava et al. (Kukava & Baramia, 2022), compared ^{18}F-FDG PET/CT with FAPI PET/CT in the early detection of progressive pulmonary fibrosis in patients with interstitial lung disease. The findings revealed that FAPI PET/CT was more effective than FDG PET/CT in predicting disease progression, with higher sensitivity in identifying fibrotic changes. This study underscores the evolving role of PET/CT modalities in managing non-oncological lung diseases, suggesting that different radiotracers may offer distinct advantages depending on the clinical scenario. Tang et al. (Tang et al., 2019) studied the ^{18}F FDG PET/CT scan ability to diagnose pulmonary nodules based on their size. They used one hundred ^{18}F FDG PET/CT scans of patients who were diagnosed having solitary pulmonary nodules (SPNs) of different sizes. The results of this study showed that compared to other methods, ^{18}F FDG PET/CT has an excellent performance in identifying different-sized SPNs, particularly those between 11 and 20 mm in diameter. In addition, ^{18}F-PET/CT has been shown to be more sensitive than CT scans in characterizing pulmonary nodules (Chin et al., 2006; Jeong et al., 2008; Kim et al., 2007). Moreover, PET-CT is beneficial in the management of patients with lung diseases such as non-small cell lung cancer (NSCLC) because it is able to detect disease sites even if the underlying structure is normal on CT. In terms of detecting primary malignancies and evaluating lymph nodes in patients with these malignancies, ^{18}F-FDG PET/CT is reported to have 99% sensitivity, 94% accuracy, and 100% positive predictive value (Budak et al., 2018).

2.5.2. *Pulmonary nodule size, location, and SUV_{max}*

Pulmonary nodules are small, rounded growths in the lung tissue, and their evaluation is critical in determining potential malignancy. Numerous studies have been conducted to explore the relationship between the size and location of pulmonary nodules in the lungs, as well as the standardized uptake value (SUV_{max}) values, and how these factors correlate with the likelihood of malignancy. Based on studies, key factors for analyzing these nodules include size, density, and metabolic activity, typically assessed using the maximum SUVmax on an ^{18}F-FDG PET/CT scan

(Garcia-Velloso et al., 2016; Groheux et al., 2016). The size of these nodules generally ranges from 4 to 30 millimeters. Nodules larger than 30 mm are usually classified as lung masses, which have a higher likelihood of being malignant. (Ost et al., 2003; Suo et al., 2016). The size of a nodule plays a significant role in risk assessment. For instance, nodules larger than 3 cm are often considered malignant. However, the malignancy risk also depends on other factors such as the nodule's growth rate and characteristics observed on imaging. It's important to note that nodules between 4-6 mm have a malignancy risk of less than 1%, whereas those between 8-30 mm have a progressively increasing risk. Typically, pulmonary nodules in ^{18}F-PET/CT are stratified into three categories: malignant nodules suspected to be metastatic, benign or not related to the patient's malignancy, and indeterminate nodules that need follow-up (Garcia-Velloso et al., 2016; Grisanti et al., 2021). Malignant nodules are frequently located in the upper lobes of the lungs, with a notable prevalence in the right upper lobe. The metabolic activity, as indicated by SUVmax, can vary depending on the histological type of the tumor. For instance, squamous cell carcinoma and adenocarcinoma, the two most common types of NSCLC, often show high SUVmax values, which can correlate with more aggressive disease and worse prognosis (Khan et al., 2011; Larici et al., 2017; Sinsuat et al., 2011). The SUV_{max} in PET/CT scan is a measure that determines the maximum concentration of ^{18}F-FDG in the region of interest (ROI) (Chan et al., 2010; Fletcher & Kinahan, 2010; Nahmias & Wahl, 2008). The SUV_{max} is an important metric in PET/CT scans, indicating the maximum concentration of the radiotracer in the nodule, which correlates with its metabolic activity. A higher SUV_{max} generally suggests higher metabolic activity, which is often associated with malignancy. A commonly used threshold is an SUV_{max} of 2.5 or higher to indicate potential malignancy. However, some studies suggest that higher cut-off values, like 11.6, might be more specific in certain contexts, especially in advanced non-small cell lung cancer (NSCLC) (Zhao et al., 2021). Studies also demonstrate that SUV_{max} can vary with tumor size and histological type. Larger tumors generally show higher SUV_{max} values, and different histological types (like squamous cell carcinoma vs. adenocarcinoma) can exhibit different metabolic behaviors as measured by SUV_{max}.

2.5.3. *Deep learning for lung nodule detection, classification, and segmentation*

Due to its crucial importance, many studies focus on detection and classification of pulmonary nodules in lung cancer (Kao & Yang, 2022; Liu et al., 2021). Nasrullah et al. (Nasrullah et al., 2019), presented a system to precisely detect malignant nodules in lung cancer using physiological symptoms, CT scan analysis, and clinical biomarkers. For CT analysis and lung nodule detection and classification, they used two deep three-dimensional (3D) customized mixed link network (CMixNet) and the results showed sensitivity of 94% and specificity of 91% on LIDC-IDRI datasets. Tong et al. (Tong et al., 2018) performed pulmonary nodule segmentation using U-Net architecture in combination with a residual block. Using CT slices of LUNA16

dataset, they obtained a Dice coefficient of 0.736. Bianconi et al. (Bianconi et al., 2021) developed a semi-automatic method for pulmonary nodule segmentation using DL and compared it with conventional approaches. They used 383 axial ROIs and 12 CNN (a combination of four segmentation models) for training the network. The results of the presented study showed superiority of DL over conventional methods. Wang et al. (Wang et al., 2017), aimed to improve the detection of lung nodules using a Computer-Aided Diagnosis (CAD) system. They addressed the challenge of limited medical data by using DL through transfer learning, combined with hand-crafted features, to enhance feature extraction. The dataset used was the public JSRT database, consisting of 154 cases with confirmed lung nodules and 93 normal cases. Their method, which fused DL and hand-crafted features, achieved higher sensitivity (69.3%) and specificity (96.2%) at a lower false positive rate compared to using hand-crafted features alone. This approach promises more effective lung nodule detection in medical imaging.

2.5.4. *Deep learning for lung nodule detection, classification, and segmentation using ^{18}F FDG PET/CT scans*

The use of DL for the detection and segmentation of pulmonary nodules from ^{18}F FDG PET/CT scans has recently been brought to the attention of researchers. Alves et al. (Alves et al., 2024), developed and validated a 3D Convolutional Neural Network (CNN) model to classify pulmonary nodules as benign or malignant using 2-[^{18}F]FDG PET/CT images. They used a dataset of 113 participants, each with one nodule, retrospectively selected. They employed three types of 3D CNN architectures (Stacked 3D CNN, VGG-like, and Inception-v2-like) and also tested transfer learning with ResNet-50. The data was split into five sets, with four sets used for 4-fold cross-validation to train and evaluate the models, and the fifth set was held out for final model testing. They employed data augmentation techniques to increase the dataset size, creating approximately 4,900 images for model training. The final model, a Stacked 3D CNN, achieved an area under the ROC curve of 0.8385 in the test set, with a sensitivity of 80.00%, specificity of 69.23%, and accuracy of 73.91% using an optimized decision threshold. The results suggest that the 3D CNN model effectively distinguished between benign and malignant pulmonary nodules, with performance slightly better than the SUVmax value, though the difference was not statistically significant. Additionally, the study utilized Grad-CAM for model explainability, showing that the model focused on the nodule region during decision-making. Despite the promising results, the study notes limitations, such as the relatively small dataset and the retrospective nature of the data collection.

Kirienko et al. (Kirienko et al., 2018), presented a CNN-based algorithm to classify lung cancer as T1-T2 or T3-T4 (assessing T-parameter) on staging FDG PET/CT scans. They used 3D patches of 472 PET/CT images for training, validation and test and obtained 69% accuracy on test hold-out dataset. Their study concluded that CNNs were feasible and promising for assessing lung

cancer T-parameters. One popular approach in DL studies is using a pre-trained model on a new task which is called "transfer learning". Schwyzer et al. (Schwyzer et al., 2018) employed transfer learning to detect lung cancer from standard full-dose, tenfold reduced dose, and thirtyfold reduced dose PET images to study whether the similar information can be acquired from the low-dose PET images. They used a deep residual neural network for training and testing 3936 PET slices. According to their study, lung cancer detection from PET images is possible even at very low radiation doses of 0.11 mSv and in this way smaller radiation hazards will be imposed to the patients. Park et al. (Park et al., 2021) also leveraged transfer learning to classify and diagnose lung cancer mass of 359 ^{18}F FDG PET/CT images. They used ResNet-18 as a DL model and pretrained weights from ImageNet dataset. From PET/CT images, SUV_{max} and lesion size were derived as metadata. The results of this study indicated that these metadata can improve the DL performance. In a study conducted by Park et al. (Park et al., 2023), a two-stage U-Net architecture was employed to segment lung cancer regions using ^{18}F FDG PET/CT scans. They used volume-of-interests (VOIs) extracted from the 887 ^{18}F FDG PET/CT images of lung cancer patients as network input and obtained a Dice coefficient of more than 0.78. The authors concluded that the method presented can greatly assist in the accurate segmentation of lung cancer using ^{18}F FDG PET/CT.

In a separate study conducted by Park et al. (Park et al., 2021) metadata containing SUV_{max} and lesion size derived from PET/CT was used to aid in detection of lung nodules/masses from CT images. In their 2-dimentional study, they employed transfer learning with a pre-trained ResNet-18 model to facilitate the differential diagnosis of lung cancer.

Schwyzer et al., (Schwyzer et al., 2018) in their initial work, assessed the performance of the DL in detection of lung cancer patients (n = 50) from controls (n = 50). They also assessed the diagnostic efficacy of the transfer learning approach using standard clinical dose PET images (PET100%), along with doses reduced by tenfold (PET10%) and thirtyfold (PET3.3%) to approximately 0.11 mSv. They applied transfer learning for the classification of 2D PET slices. A total of 3936 PET slices were converted from DICOM to JPEG files. Using a pre-trained deep residual neural network, they split the images into 10 subsets for testing via tenfold cross-validation.

The findings indicated that the area under the curve for detecting lung cancer was 0.989, 0.983, and 0.970 for standard dose images (PET100%), PET images with a tenfold reduction in dose (PET10%), and those with a thirtyfold reduction (PET3.3%) in reconstruction, respectively. Also, the network achieved sensitivities of 95.9% and 91.5%, and specificities of 98.1% and 94.2% at standard dose and ultralow dose PET3.3%, respectively.

In their recent work, Schwyzer et al., (Schwyzer et al., 2020) assessed the use of artificial intelligence and the impact of image reconstructions on small FDG-positive pulmonary nodules

detection. In their 2-dimentional approach, they used Fifty-seven PET/CT images of patients with 92 ^{18}F-FDG-avid lung nodules (all ≤ 2 cm). They studied 8824 PET slices reconstructed using block sequential regularized expectation maximization (BSREM) and ordered subset expectation maximization (OSEM) image reconstruction methods. An analysis was conducted focusing on sensitivity using two methods: one based on the maximum standardized uptake value (SUV_{max}) and another based on size, with subgroups categorized by nodule size.

The annotation of PET slices was done by a specialist in radiology and hybrid imaging in two steps: in the first step, boundaries were set for both the upper and lower extents of each lung. Then, each slice within this volume was labelled either 0 or 1 to indicate the absence or presence of a nodule, respectively. Additionally, the specialist documented the SUV_{max} of each pulmonary nodule using a standardized volume of interest (VOI) tool in both OSEM and BSREM datasets. In this work, fast.ai library based on a pretrained Res-Net-34 was used for DL training which essentially employing transfer learning. The DL model in this study was pre-trained on the Image-Net dataset, therefore, they also converted 8824 DICOM of PET images into JPEG files. Both OSEM and BSREM reconstruction sets were split to 80% for training, 10% for validation, and 10% for test in a tenfold cross-validation mode of training.

The areas under the Receiver-operator characteristic (ROC) curve (AUC) of the DL algorithm in nodule detection using OSEM reconstruction was 0.796 (95% CI: 0.772–0.869), while it was 0.848 (95% CI: 0.828–0.869) with BSREM reconstruction. The AUC significantly favored BSREM over OSEM ($p = 0.001$). In a slice-wise examination, sensitivity and specificity were 66.7% and 79.0% for OSEM, respectively, and 69.2% and 84.5% for BSREM. In a nodule-wise assessment, the overall sensitivity of OSEM stood at 81.5%, compared to 87.0% for BSREM.

The limited size of the dataset in this study restricts the generalizability of their findings. Additionally, the use of 2-dimensional slices as input data lacks the richness of information present in 3D data. The absence of lung area segmentation may introduce ambiguity in nodule detection. Employing transfer learning on pre-trained data might reduce accuracy compared to training from scratch. Furthermore, converting DICOM images to JPEG format may lead to information loss during the conversion process.

In the study by Borrelli et al. (Borrelli et al., 2021) used an AI-based approach to segment theragnostic and textural information of lung cancer and to detect lung masses automatically and then calculates the total lesion glycolysis (TLG) on FDG PET-CT images. In this retrospective study, FDG PET-CT images of 112 patients with suspected/known lung cancer were involved. A specialist manually segmented lung lesions to create training data. The findings indicated that the AI tool achieved a 90% sensitivity in lesion detection. The positive predictive value was 88%, while the negative predictive value reached 100%. Additionally, there was a strong correlation ($R2 = 0.74$) between the manual and AI-derived TLG measurements.

Given that they examined patients with suspected or confirmed lung cancer, the detectability and complexity of lesions were heightened, making them more detectable compared to identifying lung nodules. However, with a limited dataset, the utilization of a more complex model has a greater risk of overfitting, while a simpler CNN architecture, when yielding favorable outcomes, may indicate potential issues with overfitting or underfitting. Additionally, the study raises concerns regarding the generalizability of its findings to larger datasets or real-world scenarios due to its small dataset size. Moreover, there is a lack of comparison between their CNN and other more complex DL models, which would provide insights into performance, model complexity, and computational demands.

2.5.5. *Explainable AI for pulmonary nodule detection*

As explainable AI has gained attention in recent years for interpreting deep learning models, particularly in medical research, some studies have specifically focused on this topic. Wang et al. (Wang et al., 2024) introduces ExPN-Net, a multi-task deep-learning model designed to improve both the accuracy and interpretability of pulmonary nodule diagnosis. Lung cancer remains the leading cause of cancer-related deaths, and early detection is crucial for effective treatment. However, diagnosing pulmonary nodules relies heavily on radiologists' expertise, which can be both subjective and labor-intensive. While computer-aided diagnosis systems have significantly enhanced diagnostic accuracy, the lack of model transparency and reliability continues to hinder their widespread clinical adoption. This research aims to bridge that gap by developing a model that not only classifies pulmonary nodules as benign or malignant but also identifies and localizes specific nodule characteristics, enhancing interpretability for radiologists.

The proposed ExPN-Net follows a multi-task learning approach that integrates nodule classification with characteristic identification. It employs 3D nnU-Net for automatic nodule segmentation, generating probability maps that serve as attention mechanisms for subsequent classification. The network further incorporates an Anatomical Attention Gate to introduce explicit spatial attention, ensuring that the model focuses on relevant regions, and a Soft Activation Map module to generate fine-grained visual activation maps. These visualizations enable radiologists to understand where the AI is focusing and why certain predictions are made, making the model more transparent. The model was evaluated on thoracic CT images from two datasets: the public LIDC/IDRI dataset, consisting of 624 participants with 1226 nodules, and an in-house dataset from a medical institution, containing 807 participants with 812 nodules. While the LIDC dataset was used for benign versus malignant classification, the in-house dataset aimed to differentiate between invasive lung cancer subtypes with micropapillary or solid patterns.

Experimental results demonstrated that ExPN-Net outperformed existing models, achieving an AUC of 0.992 on the LIDC dataset and 0.923 on the in-house dataset. The study further showed that incorporating nodule characteristic identification not only improved model interpretability

but also enhanced classification accuracy. The explicit attention mechanism provided by the AAG module proved to be more effective than self-attention methods, ensuring that the model accurately focused on clinically relevant regions. Additionally, the SAM-generated activation maps provided detailed visual explanations, helping radiologists assess the AI's decision-making process more effectively. Their study presents ExPN-Net as a reliable and explainable AI model for pulmonary nodule diagnosis, addressing a critical limitation of existing deep-learning-based CAD systems. By combining high diagnostic accuracy with enhanced interpretability, the model provides a solution that can be more effectively integrated into clinical practice. The ability to identify and visually highlight nodule characteristics ensures that radiologists can better interpret AI-generated diagnoses, fostering greater trust in AI-assisted diagnostic tools.

Fernandes et al. (Fernandes et al., 2024) conducted a study focusing on the application of explainable AI in cancer classification. The primary aim was to identify the most influential genes involved in classifying five recurrent cancer types in women: breast cancer, lung adenocarcinoma, thyroid cancer, ovarian cancer, and colon adenocarcinoma.

The researchers utilized machine learning models, including decision trees, random forests, and XGBoost, trained on RNA sequencing gene expression data sourced from The Cancer Genome Atlas. To enhance interpretability, they applied the SHAP method, which allowed them to determine the features that significantly influenced the decision-making processes of the models.

The study focused on gene expression data derived from RNA sequencing, which provided a comprehensive view of the genes' activity levels across the different cancer types.

The models achieved high accuracy rates, with random forests reaching up to 99.82%. Through the application of SHAP, the researchers identified a subset of genes that were most influential in the classification process, thereby reducing the feature set to a more manageable and interpretable size.

Mahua Pal (Pal, 2023) performed a research on the development of an explainable AI model for detecting malignancy in pulmonary nodules. The primary aim was to enhance trust in AI-based lung cancer detection by addressing the black-box nature of AI models and reducing AI risk. The study sought to interpret AI predictions using XAI tools, thereby making AI-based diagnostic decisions more transparent and acceptable to medical practitioners. The research employed an XGBoost classifier for binary classification of pulmonary nodules into benign and malignant categories. Three explainability techniques were integrated: two post-hoc methods, SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-Agnostic Explanations), along with an ante-hoc method leveraging XGBoost's inherent explainability. The study used the LIDC-IDRI dataset, which contains thoracic CT scans with radiologist-annotated nodules. A feature selection process was conducted based on XAI outputs to identify the most relevant biomarkers, which included attributes such as calcification, lobulation, sphericity, texture, and

subtlety. To address class imbalance in the dataset, the ADASYN oversampling technique was employed, improving the training process.

The results demonstrated that the model achieved high predictive performance, with an AUC of 0.952, accuracy of 0.929, and specificity of 0.861. After feature selection, a refined model was built using the most influential biomarkers, leading to improved performance metrics, including an accuracy of 0.939 and specificity of 0.888. SHAP and LIME provided visual explanations of feature contributions, reinforcing trust in AI predictions. The study concluded that integrating XAI tools enhances model interpretability, making AI-driven lung cancer diagnostics more transparent and clinically acceptable.

2.5.6. *Research Gap in pulmonary nodule detection*

While significant progress has been made in the use of ^{18}F-FDG PET/CT imaging and DL models for pulmonary nodule detection and classification, several notable gaps remain:

- Binary classification limitation: The majority of existing studies have focused on the binary classification of pulmonary nodules as either benign or malignant. However, this approach oversimplifies the complexity of clinical scenarios, where many nodules cannot be definitively categorized as benign or malignant. In reality, a significant number of cases fall into an indeterminate category that requires further follow-up and monitoring. There is a critical need for research that explores multi-class classification models that better reflect the nuanced decisions physicians make in clinical practice, including the categorization of nodules that are suspicious or require additional diagnostic evaluation.
- Limited focus on using PET/CT images: The majority of research on pulmonary nodule detection or segmentation focuses on CT images (Bianconi et al., 2021; Sourlos et al., 2022). Some studies used PET images primarily to enhance nodule detection on CT scans rather than incorporating them directly into the DL input (Apostolopoulos et al., 2021; Park et al., 2021), (Park et al., 2021).
- Inconsistent use of 3D data: Much of the current DL-based research on lung nodules, particularly in the context of PET/CT imaging, relies on 2D slices rather than 3D data, which could provide richer and more informative input for models. There is a need for studies that develop and validate models using 3D data to potentially improve diagnostic accuracy.
- Prevalence of transfer learning: Most studies on lung nodule detection using PET/CT images have employed transfer learning rather than training models from scratch. While transfer learning offers the advantage of leveraging pre-trained models and typically requires less computational power and data, it may not fully capture the unique features of PET/CT imaging for lung nodule detection. Future research could benefit from

developing models trained from scratch on PET/CT datasets, potentially leading to more accurate and specialized detection algorithms.

- Small dataset sizes: Many studies, including those using advanced techniques like transfer learning, have been constrained by small dataset sizes, which can result in overfitting and limit the generalizability of the models. Expanding datasets and including more diverse cases would help to train more robust models suitable for real-world application.
- Lack of comprehensive model comparison: There has been limited comparison between different DL architectures in terms of performance, model complexity, and computational demands. Future research should focus on benchmarking various DL models, including complex architectures, to identify the most effective approaches for specific clinical tasks in PET/CT imaging.
- Further research is needed on explainable AI to enhance the interpretation of deep learning models and improve their interoperability for reliable use. Additionally, existing studies have primarily focused on CT images, with no research conducted on the interpretability of PET/CT images for pulmonary nodule detection.

Addressing these gaps would significantly advance the development of DL models that align more closely with clinical realities, improving the accuracy and utility of pulmonary nodule detection and classification using ^{18}F-FDG PET/CT imaging.

2.6. Summary

This chapter presents a comprehensive literature review on the intersection of DL and medical imaging, particularly in the context of pulmonary diseases. The review focuses heavily on the advancements driven by the COVID-19 pandemic, where DL models were rapidly developed for analyzing lung CT scans to detect, classify, and segment COVID-19-related abnormalities. While these models have proven effective, the review identifies a major challenge in the generalizability of DL models across diverse datasets. Most models show high accuracy on internal datasets but struggle when applied to external datasets, highlighting issues such as dataset bias, the need for lung segmentation, and the importance of multi-dataset validation.

The chapter extends its analysis to DL applications beyond COVID-19, including lung nodule detection using PET/CT imaging. It reviews studies that apply DL to various pulmonary imaging tasks and identifies research gaps such as the need for multi-class classification models to handle complex clinical scenarios, better use of 3D data, and the development of models trained from scratch on specialized datasets. Additionally, the review highlights the need for larger datasets and comprehensive model comparisons to improve the real-world applicability of DL technologies.

This literature review offers a detailed examination of the current state of DL in pulmonary imaging and underscores the critical research gaps and challenges that must be addressed to fully realize the potential of DL models in clinical practice.

CHAPTER 3

3. MATERIALS AND METHODS

3.1. Introduction

This section presents a comprehensive approach for the classification of COVID-19 lung involvement using 3D DL models and CT images, and the multi-classification of lung nodules using ^{18}F-FDG PET/CT images. The methodology is structured around the development and evaluation of advanced DL models, with a particular focus on assessing their generalizability across diverse datasets.

The first part of the methodology involves the classification of COVID-19 lung involvement. Four independently sourced CT datasets are utilized, each subjected to rigorous pre-processing steps, including resampling, cropping, and intensity clipping, to standardize the data and improve model performance. Segmentation techniques are applied to isolate lung regions, enhancing the accuracy of subsequent classifications. The study employs various 3D CNNs, including ResNet and DenseNet architectures, and evaluates their performance using k-fold cross-validation. The generalizability of the models is thoroughly tested across different datasets, ensuring robustness and reliability in real-world applications.

The second part of the methodology focuses on the multi-classification of lung nodules using ^{18}F-FDG PET/CT images. The dataset, which includes 1304 PET/CT images from patients with a variety of malignancies, is carefully curated to exclude cases of primary lung cancer to reduce bias. The images undergo extensive pre-processing, including resizing, segmentation, and normalization, to prepare them for DL model training. The study explores various DL architectures, with a particular emphasis on Inception-ResNet-v2 models, to classify lung nodules into benign, malignant, and suspicious categories. The methodology also addresses the challenge of class imbalance through oversampling techniques and applies data augmentation to improve model robustness. Throughout the study, multiple evaluation metrics are employed to assess the performance of the DL models, including accuracy, sensitivity, specificity, and the AUC of the ROC curve. These metrics provide a comprehensive understanding of the capabilities of the models and their potential application in clinical settings.

Figure 3.1 presents a structured overview of the approach used in this study, divided into three main objectives. Each objective highlights a specific area of focus, utilizing advanced DL techniques and extensive datasets to address critical challenges in medical imaging and diagnosis.

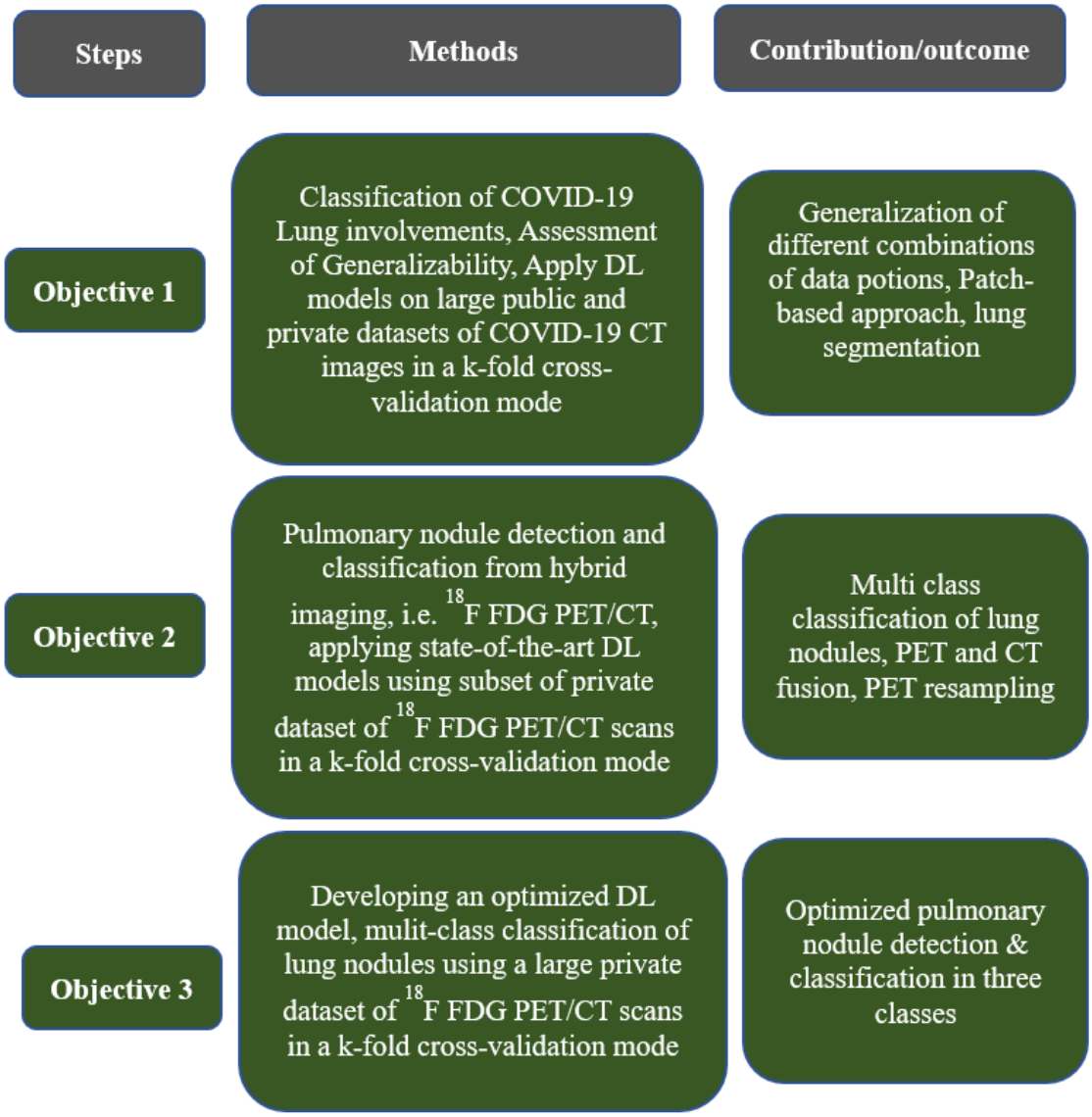


Figure 3.1. Overview of the study approach based on three objectives: (1) COVID-19 lung involvement classification, (2) Pulmonary nodule detection using ^{18}F FDG PET/CT, and (3) Optimizing DL models for multi-class nodule classification, with outcomes linked to advanced imaging and cross-validation techniques.

The main goal of our first objective was to investigate in the absence of several large datasets, how much of combination of another dataset gives the acceptable results. The flowchart in Figure 3.2 summarizes the procedures implemented to perform first objective, and each step is described below. To achieve this goal, we begin by outlining data characterization, various pre-processing steps and initial simulations to determine the optimal approach.

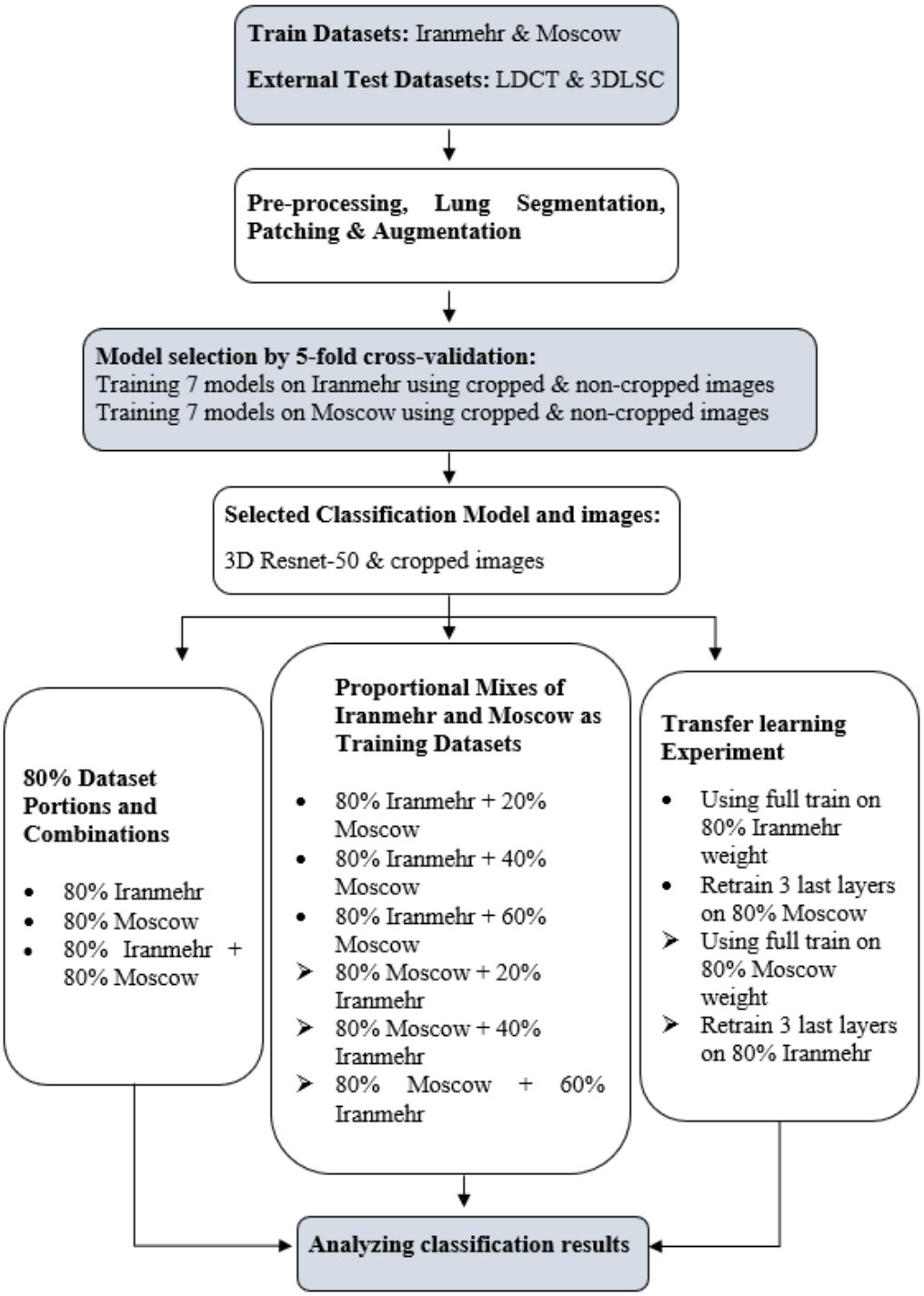


Figure 3.2. The overview of first objective

3.2. CT Dataset characterizations

Our study employs four independently sourced datasets. The first dataset was collected from Iranmehr hospital, located in Tehran, Iran, and we name this dataset as "Iranmehr". Digital Imaging and Communications in Medicine (DICOM) data of chest CT images of 1110 patients were collected from Iranmehr hospital picture archiving and communication system (PACS). This

dataset was collected from February 2020 to March 2020, when COVID-19 was at its peak. Imaging was done on GE Medical Brightspeed 16 detector multislice CT scan machine; low dose spiral high-resolution CT imaging technique was employed. CT images were collected as a screening protocol before hospitalization of the patients for COVID-19 infection detection. Pulmonary COVID-19 involvement score was based on the interpretation of two expert independent radiologists who had access to clinical data of the patients. Radiology specialists validated the gathered data, so only normal, and COVID-19 patients were included. Iranmehr Hospital specialists supervised the collection of all patient data. Data was collected under the policies of Iranmehr hospital, which allow anonymized data to be used for research purposes. The data collection, subsequent anonymization (done onsite under strict supervision), and usage for this study were undertaken with proper authorization and following international data privacy standards. The second dataset was sourced from hospitals located in Moscow and made available by Morozov et al. (Morozov et al., 2020). This dataset has been assessed and labeled by expert radiologists according to COVID-19 lung involvement and grouped into four classes at 25% intervals. The first class, named as CT-0 contains 254 images with no lung involvement representing a normal CT image. Classes CT-1 to CT-4 represent 25% to 100% lung involvement and contain 854 images. This dataset is referred to as "Moscow" in this paper. We used two additional external test datasets to assess the validity of our results. First, the low-dose and ultra-low-dose (LDCT) containing CT images of 104 COVID-19 positive cases, and 56 normal cases, collected in Babak Imaging Center, Tehran, Iran. The second dataset is the 3DLSC-COVID dataset (X. Wang et al., 2021) which is publicly available and contains 100 COVID-19 positive cases and 96 normal. The LDCT image format is DICOM and 3DLSC is NIFTI.

3.3. Preprocessing and hyper parameter tuning

The matrix size of all CT images was 512 × 512 pixels, but they had different slice numbers. So, after loading DICOM CT images, they were initially resampled and interpolated to have the same slice number. We prepared two forms of datasets, including cropped and non-cropped images, to assess the effect of cropping in the training phase. For cropping sets, all images were cropped to remove surrounding areas that are not significant and then resampled to have the same size as 128 × 128 × 60. All CT images were resampled to a resolution of 1 mm × 1 mm × 1 mm and intensity clipped to (−1000, 400) Hounsfield Unit (HU) range which is considered as HU window for lung. We are not interested in HU values above 400, which are bony structures. The values below -1000 are also out of the range of the lung's HU. For non-cropping sets, we applied the same approach for resampling and kept the whole field of view (FOV) of the image. The normalization of the images was also performed. We assigned 1 for positive pulmonary COVID-19 involvement and 0 for normal images. We saved the preprocessed images as NumPy arrays to be fed as a network's input. The input of the 3D networks must have the same slice number. So, for 3D

CNNs, resampling is of great importance. Additionally, cropping and intensity clipping remove the less useful parts of the image, resulting in more efficient training. Different l1 and l2 regularization at variable values were tested, and the l2 regularization at 0.001 was optimal and used. Furthermore, dropout at 0.2 was employed for further regularization and "Adam" as model optimizer. To overcome the imbalanced number of COVID-19 and normal cases, we set a number of 6 patches per image for the majority class and 16 patches per image for the minority class.

3.3.1. *Segmentation*

The segmentation results showed that lung field from the CT image improved the results of classification and generalizability. We tested three different algorithms on our four datasets to assess whether we could have one single lung segmentation approach. The segmentation methods include DSB Lung Segmentation Algorithm from Kaggle, an algorithm developed by Zuidhof (Zuidhof, 2017), and a U-net based lung segmentation developed by Hofmanninger (Hofmanninger et al., 2020). Nevertheless, as can be seen in Figure 3.3, for each data format, one type of segmentation method performs better. This is probably due to the Neuroimaging Informatics Technology Initiative (NIFTI) format of images compared to DICOM format. The reason for this might be the loss of some information during the conversion of original DICOM images to NIFTI format. The DSB algorithm failed to segment peripheral parts of the lung which have COVID-19 involvement. Therefore, we applied a U-net based lung segmentation module on Iranmehr and LDCT datasets to have 3D segmented lung area. For segmentation of Moscow and 3DLSC datasets, we used the Zuidhof method. Figure 3.4 shows the result of the segmented lung used in the present study. On the other hand, the Zuidhof is not accurate as Hofmanninger approach, and there were 17 out of 1110 images from the Moscow dataset that were not segmented properly. We used original CT images for these 17 non-segmented cases. Next, normalization, zero-centering, and shuffling were done in a preprocessing part of the task.

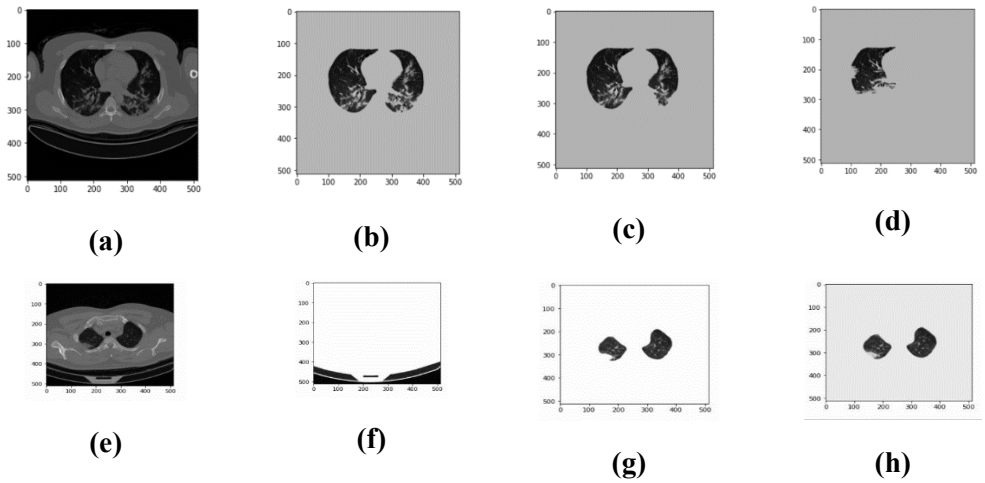
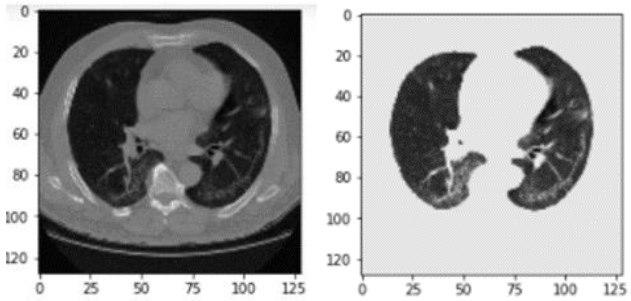


Figure 3.3. Results of different segmentation methonds on DICOM and NIFTI image format. (a) original NIFTI image segmented by (b) Zuidhof method, (c) DSB algorithm, and (d) Hofmanninger method. (e) Original DICOM image segmented by (f) Zuidhof method, (g) DSB algorithm, and (h) Hofmanninger method



(a) **(b)**

Figure 3.4. A typical slice of a) chest CT image and b) segmented lung. This slice is for a COVID-19 positive case.

3.3.2. *Patching and Augmentation*

During our trial-and-error experimentations, we found that simulations with data augmentation outperformed simulations without data by 2 to 5 percent. Random data augmentation prevents early overfitting and improves model performance. Furthermore, data augmentation produces different shapes and orientations of the images while still being recognizable, allowing the model to learn more features. Data augmentation steps were employed using random noise (mean: 0, standard deviation: 0.08), translation (shift with the size of random integer number between (-0.1, 0.1) × patch size in the x direction), random rotation (random rotation between 0° and 360°), distort elastic (alpha: 100, sigma: 10), flip (in the direction of x and z axis), 90-degree rotation (which provides random rotation of 90°, 180°, 270°), and scaling (zoom with the random size between 0.6 and 1.2). All the augmentations were applied in “on the fly” mode in the generator to prevent the network from overfitting.

We applied patching to train input images of a size that both covers the most part of the lung and is a reasonable size for patches. Logically, since the test is performed on the full image of the patient, not just a patch, the patch-size should be large enough to cover most of the lungs. Testing on the full image provides the most accurate results. When a patch is normal, it means the patient's decision was normal, however there may be one patch that is detected as COVID-19, so the full image is COVID-19. From different trials for hyper-parameters’ testing, the patch-size of 115 × 115 × 55 was applied for images since it had up to 2 percent better performance than other patch

sizes. For each dimension of x, y, and z, a random number was selected from the difference between the original size of the image and the patch size. The patch produced was from that random number to the patch size, plus that random number in each dimension. For big patch sizes, overlap would be very high. However, in the generator part of our network, we first patched the data and then implemented augmentation on each patch to have augmented data as much as possible. From our trials, we found that this method improved the model performance compared to when patching performed after data augmentation. The reason is that each patch undergoes a set of random augmentation and would be unique. The adopted approach for patching is shown is pseudocode below:

```
def generate_train_data(x_inp, y_inp, patchsize, is_Moscow=False):
    while True:
        selected_pat_x, selected_pat_y = select_image_randomly()
        patch_number = 6 if is_Moscow and selected_pat_y == 1 else 16
        for i in range(1, patch_number + 1):
            row = np.random.choice(range(IMAGE_X_SIZE - patchsize[0]))
            column = np.random.choice(range(IMAGE_Y_SIZE - patchsize[1]))
            slice_ = np.random.choice(range(IMAGE_Z_SIZE - patchsize[2]))
            xr = selected_pat_x[row:row + patchsize[0], column:column + patchsize[1], slice_:slice_
+ patchsize[2]]
            for augmentation in augmentation_methods:
                xr = augment_patch_randomly(xr, augmentation)
            yield (xr, selected_pat_y)
```

3.4. Classification

We used Tensorflow (Abadi et al., 2016) library as the platform. Different l1 and l2 regularization at variable values were tested, and the l2 regularization at 0.001 was optimal and used. Furthermore, dropout at 0.2 was employed for further regularization and “Adam” as model optimizer. To overcome the imbalanced number of COVID-19 and normal cases, we set a number of 6 patches per image for the majority class and 16 patches per image for the minority class. Also, we applied this approach to overcome the overfitting of the training. We saved the best weights of the trained model, so that overfitting didn’t affect the simulations, even if we had overfitting in certain scenarios.

The network was trained on advanced GPUs provided by the UTS Interactive High-Performance Computing (iHPC).

3.5. Deep learning model selection for COVID-19 involvement classification

For all the experiments in the present work, we employed 5-fold running. Using the k-fold technique, the dataset is randomly partitioned into 5 groups or folds of roughly equal size. In order to test the model performance, the first fold is kept, and the model is trained using k-folds. Validation is repeated k times, and each time a different fold or a different set of data points is used. Seven common models, including ResNet-50, ResNet-152, DenseNet-169, DenseNet-201, Resnext-50, Seresnext-50, and Seresnet-152 were assessed for generalizability. we selected state-of-the-art deep learning models through a trial-and-error approach, applying them images to determine which yielded the best results, while also considering their prior usage in the literature. In the present study, binary classification was performed using common models that have been used in previous studies (Mohandass et al., 2024; Xue & Abhayaratne, 2023; Yang et al., 2023). Furthermore, for our 3-dimensional data, more complex models, such as ResNets and DenseNets, provided better results than simpler models, such as VGG. We used 1110 3D CT images of Iranmehr and Moscow dataset in this stage and fed 100 percent of the data for training. Training was carried out on one dataset and was tested against another whole dataset, i.e., training on 1110 Iranmehr dataset and testing against 1110 Moscow dataset, and training on 1110 Moscow dataset and testing on 1110 Iranmehr dataset. Several hyper-parameters were tested, including learning rate, different patch sizes, and the number of training iterations. After hyper parameter tuning, training was performed using seven mentioned models for an initial learning rate of 10^{-4} and 250 epochs for 5-fold as the best selected hyper-parameters. Figure 3.5 illustrates 3D ResNet-50 structure, and Figure 3.6 shows the resultant learning curves.

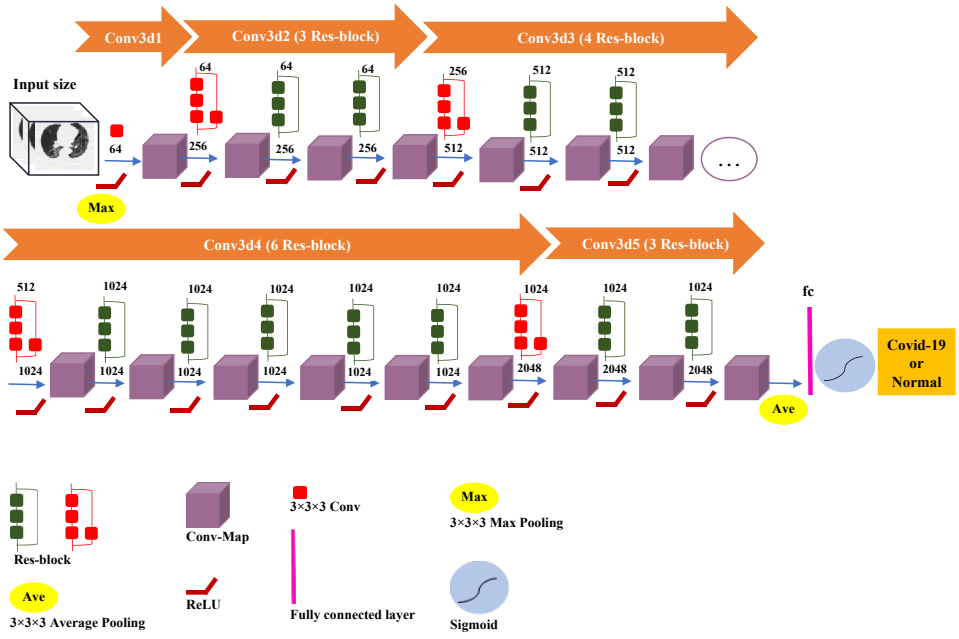
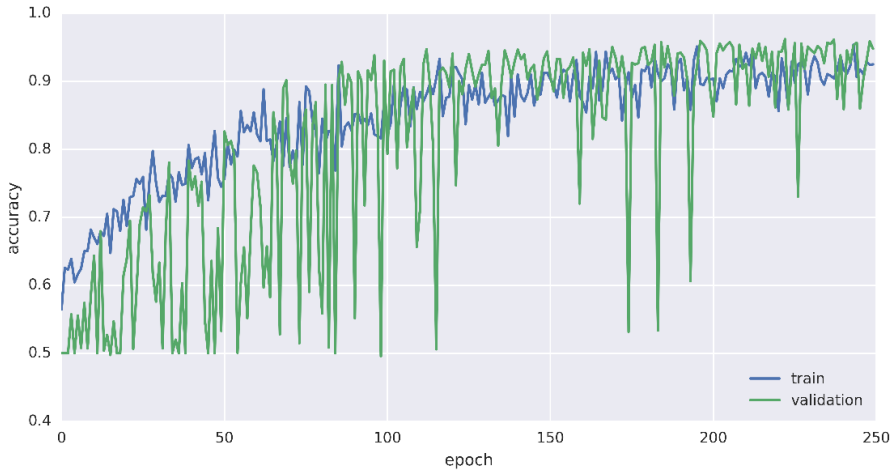
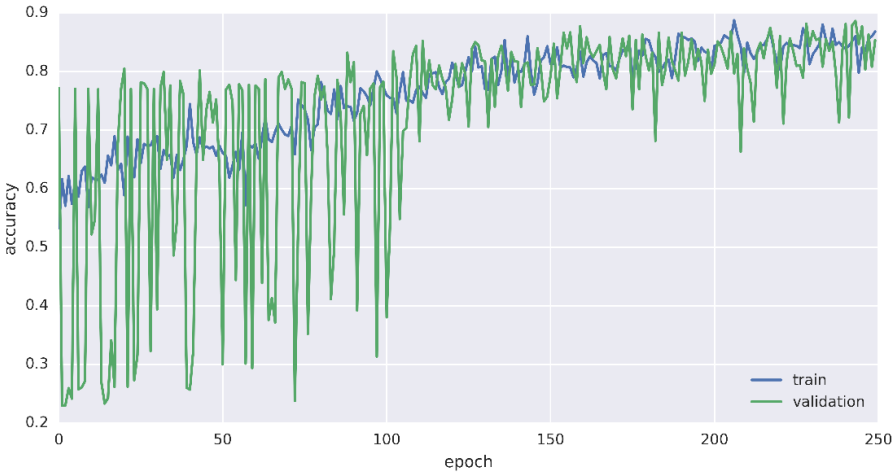


Figure 3.5. Architecture of 3D ResNet-50. The segmented lung images are fed to the model, and the model output would be the predicted probability of COVID-19 positive or normal.



(a)



(b)

Figure 3.6. Learning curve for training on (a) Iranmehr data using ResNet-50, and (b) Moscow data using ResNet-50.

3.6. Generalization assessment

To assess the generalizability, we performed a series of experiments. Based on 1110 3D CT images of Moscow dataset, we used the same number from Iranmehr dataset randomly. The

allocation to training, validation, and test groups and splitting was done randomly. We split each dataset into five segments of 20 percent each, with equal distribution of normal and COVID-19 in each segment. We separated one 20 percent segment from each dataset as holdout test set and kept it same for all experiments and all training parts were performed with remaining 80 percent from each dataset. The generalization evaluation was carried out in three experiments, as follows.

3.6.1. *80% Dataset Portions and Combinations*

Models were fully trained in 3 experiments using 80 percent Iranmehr, 80 percent Moscow, and 80 percent Iranmehr + 80 percent Moscow data, respectively. These models were tested with 20 percent holdout sets separately from Iranmehr and Moscow sets. The details shown in Table 1.2 serve as the base experiment results for generalization tests.

Table 3.1. Training and Testing data percentage for the 80% dataset portions and combinations.

Test No.	**Training data**	**Testing data**
1	80% Iranmehr	20% Iranmehr holdout set 20% Moscow holdout set LDCT 3DLSC
2	80% Moscow	20% Iranmehr holdout set 20% Moscow holdout set LDCT 3DLSC
3	80% Iranmehr + 80% Moscow	20% Iranmehr holdout set 20% Moscow holdout set LDCT 3DLSC

3.6.2. *Proportional Mixes of Iranmehr and Moscow as Training Datasets*

Models trained with 80 percent of Iranmehr dataset with the addition of an increasing portion of Moscow dataset (20%, 40%, 60%). Similarly, Models were trained with 80 percent Moscow dataset and an increasing portion of Iranmehr dataset (20%, 40%, 60%). All models were then tested on one 20 percent of holdout set from each dataset. Additionally, to evaluate the effect of

adding different combinations to the dataset, they were tested on two external datasets. Details of these six tests are given in Table 3.2.

Table 3.2. Training and Testing data percentage for the proportional mixes of Iranmehr and Moscow as training datasets.

Test No.	Training data	Testing data
1	Iranmehr 80% + Moscow 20%	20% Iranmehr holdout set 20% Moscow holdout set LDCT 3DLSC
2	Iranmehr 80% + Moscow 40%	20% Iranmehr holdout set 20% Moscow holdout set LDCT 3DLSC
3	Iranmehr 80% + Moscow 60%	20% Iranmehr holdout set 20% Moscow holdout set LDCT 3DLSC
4	Moscow 80% + Iranmehr 20%	20% Iranmehr holdout set 20% Moscow holdout set LDCT 3DLSC
5	Moscow 80% + Iranmehr 40%	20% Iranmehr holdout set 20% Moscow holdout set LDCT 3DLSC
6	Moscow 80% + Iranmehr 60%	20% Iranmehr holdout set 20% Moscow holdout set LDCT

		3DLSC

3.6.3. *Transfer learning Experiment*

Transfer learning is a machine learning technique where a pre-trained model, developed on one task, is reused or fine-tuned for another task (Abbas et al., 2020; Dutta et al., 2020; Kora et al., 2022; Yu et al., 2022). In deep learning, this approach is particularly useful for tasks with limited data, where leveraging the knowledge gained from a different, larger dataset can help improve performance. Transfer learning enables the model to generalize better across different but related datasets, which reduces the training time and the risk of overfitting, particularly when using complex architectures like CNNs for image-based tasks. In medical imaging, transfer learning has gained prominence due to the scarcity of labeled data and the high cost of manually annotating medical images. In particular, 3D medical imaging tasks, such as diagnosing COVID-19 from CT scans, benefit from transfer learning as the datasets used can vary in distribution, imaging conditions, and noise levels. Transfer learning allows models trained on one large medical imaging dataset to transfer their learned features—such as recognizing patterns, edges, and textures relevant to medical diagnosis—to another dataset.

One specific technique in transfer learning is to reuse the majority of the deep learning model's architecture while retraining only the last few layers for the target task. This method is particularly useful when working with datasets that share common low- and mid-level features, such as 3D CT scans of different patient populations. The early layers in a CNN tend to capture basic image features like textures and edges, which are often transferable across different datasets, while the later layers capture task-specific features that need retraining for new datasets.

Therefore, in this experiment, transfer learning was employed to check how it helps with generalization. This task included three steps for each dataset of Moscow and Iranmehr datasets.

1. Using weights of the full trained model with 80 percent of one dataset.
2. Retraining three last layers with 80 percent of the other dataset.
3. Testing on Iranmehr and Moscow holdout test data, and two external datasets: LDCT 3DLSC. (See Table 3.3)

In this experiment, we used the last Conv3D layer with 1048576 trainable parameters, the last batch normalization with 8192 trainable parameters, and the fully connected layer with 2049 trainable parameters. For instance, when we trained 80 percent of Moscow dataset using transfer learning, we loaded the weight obtained from 80 percent full training of Iranmehr dataset. A similar approach is done for transfer learning using 80 percent of Iranmehr dataset. This method of transfer learning leverages the shared characteristics between the two datasets while allowing the model to adapt its higher-level features to the specific target dataset. By using a trial-and-error

approach, the selection of the last three layers for retraining provided the best possible accuracy, ensuring that the model could generalize across both datasets and external ones.

Table 3.3. Training and Testing data for the Transfer learning Experiment.

Test No.	Training Data	Testing Data
1	80% Iranmehr	20% Iranmehr holdout set 20% Moscow holdout set LDCT 3DLSC
2	80% Moscow	20% Iranmehr holdout set 20% Moscow holdout set LDCT 3DLSC

3.7. Evaluation metrics

The performance of all trained models was assessed using a comprehensive set of statistical metrics. These metrics include accuracy (Eq. 1), sensitivity (Eq. 2), specificity (Eq. 3), F1-score (Eq. 4), positive predictive value (PPV, Eq. 5), negative predictive value (NPV, Eq. 6), and the AUC of the ROC curve (Hicks et al., 2022). Each of these metrics provides unique insights into different aspects of the model's performance, allowing for a thorough evaluation. Below are the detailed descriptions and mathematical formulations for each statistical measure:

3.7.1. *Accuracy*

Accuracy measures the proportion of all correctly classified instances (both true positives and true negatives) out of the total number of instances. It provides a general indication of how well the model performs across all classes. However, in the case of imbalanced datasets where one class dominates, accuracy alone can be misleading, as it may not adequately reflect the model's ability to correctly classify the minority class.

$$(Eq.1) \qquad Accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{\text{Total number of samples}}$$

3.7.2. *Sensitivity*

Sensitivity, also known as recall or the true positive rate, measures the proportion of actual positive cases that are correctly identified by the model. High sensitivity is crucial in situations

where it is important to minimize the number of false negatives, such as in medical diagnostics, where failing to identify a condition could have serious consequences.

$$(Eq.2) \qquad Sensitivity = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

3.7.3. *Specificity*

Specificity, or the true negative rate, measures the proportion of actual negative cases that are correctly identified by the model. High specificity is important in scenarios where the cost of false positives is high, such as when a positive result could lead to unnecessary treatment or intervention. Specificity complements sensitivity by focusing on the model's ability to correctly identify negative cases.

$$(Eq.3) \qquad Specificity = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}$$

3.7.4. *Positive predictive value*

PPV, also known as precision, measures the proportion of positive predictions that are actually correct. High PPV is important in contexts where false positives can have significant negative consequences, such as in medical testing where a false positive might lead to unnecessary anxiety or treatment. Precision complements recall by focusing on the accuracy of positive predictions.

$$(Eq.4) \qquad PPV = \frac{TP}{\mathrm{TP} + \mathrm{FP}}$$

3.7.5. *Negative predictive value*

The NPV measures the proportion of negative predictions that are actually correct. A high NPV is crucial in situations where missing a true positive (i.e., a false negative) could be particularly harmful, ensuring that negative predictions can be trusted. NPV complements specificity by focusing on the accuracy of negative predictions.

$$(Eq.5) \qquad NPV = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FN}}$$

3.7.6. *F1-score*

The F1-score is the harmonic mean of precision (positive predictive value) and recall (sensitivity). It balances the trade-off between precision and recall, making it a useful metric when you need to account for both false positives and false negatives. In this context, precision (PPV) is the proportion of positive predictions that are truly positive. Recall (Sensitivity)is the proportion of actual positives that are correctly identified. The F1-score is particularly valuable in cases where the class distribution is imbalanced, as it provides a more balanced view of the model's performance by considering both precision and recall.

$$(Eq.6) \qquad F1 - Score = \frac{2 \times Precision \times Recall}{\text{Precision} + \text{Recall}}$$

where, TP is the number of true positives, TN is true negative, FP stands for the number of false positives, and FN indicates the number of false negatives.

3.7.7. *Receiver operating characteristic curve*

The receiver operating characteristic (ROC) curve is a graphical representation that illustrates the diagnostic ability of a binary classifier as its discrimination threshold is varied. It plots the true positive rate (TPR or sensitivity) against the false positive rate (FPR or 1 - specificity) at various threshold levels. The ROC curve shows the trade-off between sensitivity and specificity, demonstrating how improving one metric may affect the other.

3.7.8. *Area under the curve*

The AUC is a single scalar value that summarizes the overall performance of the ROC curve. It represents the probability that a randomly chosen positive instance is ranked higher than a randomly chosen negative instance by the model. An AUC of 1 indicates perfect performance, while an AUC of 0.5 suggests that the model performs no better than random guessing. A higher AUC value indicates better overall performance of the model, reflecting a superior trade-off between sensitivity and specificity.

3.8. Lung nodule detection from ^{18}F FDG PET/CT images

The overview of the process for lung nodule classification from ^{18}F FDG PET/CT images using DL model is shown in Figure 3.7.

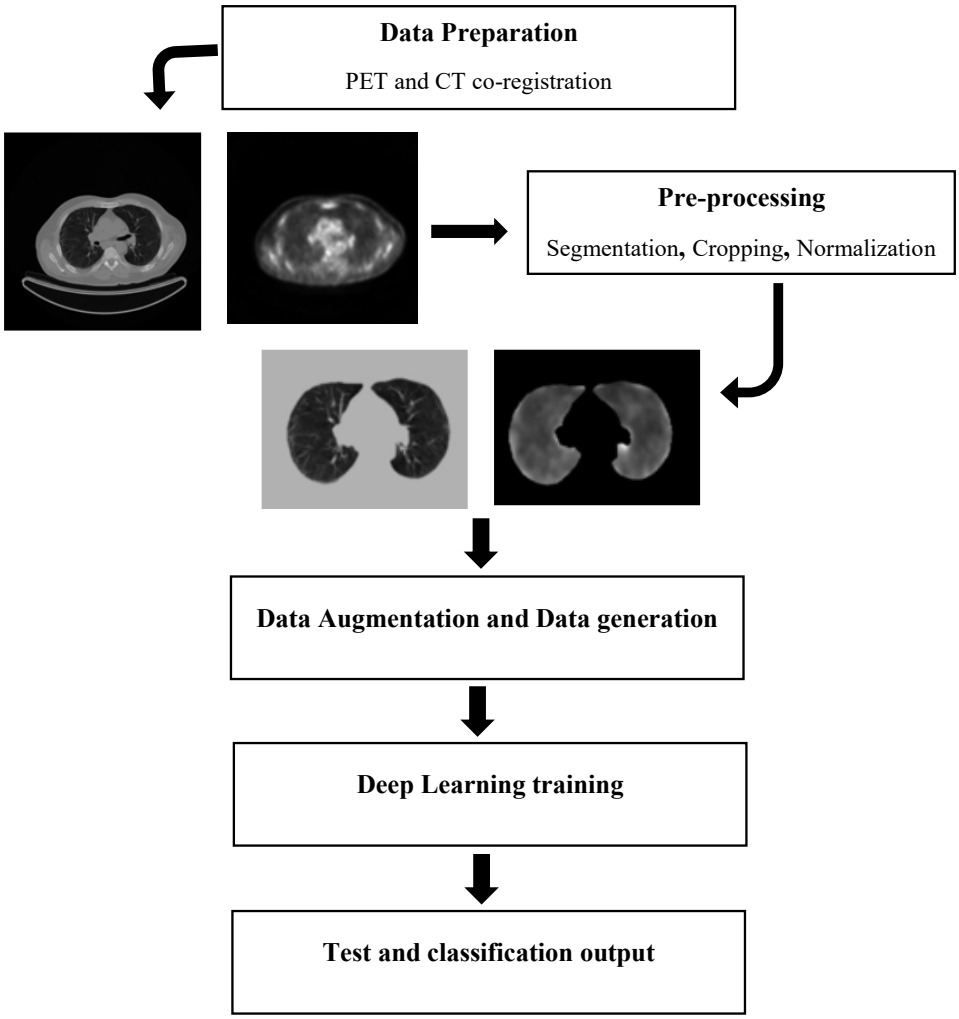


Figure 3.7. The workflow of the lung nodule detection from PET/CT images.

3.9. ^{18}F FDG PET/CT dataset characterization and imaging

The dataset comprised 1304 PET/CT images collected retrospectively from Shariati Hospital in Iran. The DICOM data were retrieved from the Shariati Hospital PACS for patients who underwent their routine ^{18}F-FDG PET/CT scan evaluation at Shariati Hospital between 2016 and 2021. At least two physicians involved in inspecting patients' scans and reporting the patients' status. The dataset included whole-body and total-body PET/CT images from patients with various malignancies. However, cases involving lung cancer and lung masses were excluded due to the extensive lung involvement typically observed in these conditions, which could introduce substantial bias into the training dataset. This exclusion was crucial to preserve the integrity of the data analysis and to enhance the accuracy of the subsequent modelling outcomes. These images were acquired using a Siemens Biograph 6 TruePoint PET-CT scanner. The CT scans were performed with parameters set at 80 kV for children and 110 kV for adults, with 30 mAs

and a pitch index of 1.5. This was followed by PET imaging in 3D mode. The direction of scanning (either craniocaudal or caudocranial) depended on whether the patients were positioned head-first or feet-first, aiming to minimize bed movement. The injected activity of ^{18}F-FDG was adjusted based on the patient's weight, with 10 mCi administered for those weighing 68 kg or less and a minimum of 0.7 mCi for those weighing 3 kg. All patients fasted before the FDG imaging, avoiding carbohydrates for 24 hours prior and refraining from eating, except drinking water, for 6-8 hours before the imaging procedure.

We randomly divided the dataset into two parts: 70% for training and validation, and 30% for a holdout test set. Two main directories were created: one for training and validation (referred to as the train folder) and another for testing (the test folder). Within each of these folders, sub-folders were organized by category—benign, malignant, and suspicious. The train folder contained 373 benign cases, 164 malignant cases, and 375 suspicious cases, while the test folder included 160 benign cases, 70 malignant cases, and 162 suspicious cases.

3.10. Demographic information of subjects

The dataset comprises 1304 PET/CT images (mean age: 52), categorized into benign lung nodules/without lung nodules, malignant lung nodules, and suspicious lung nodules requiring follow-up. The benign category includes 533 PET/CT images, the malignant category includes 234 PET/CT images, and the suspicious category contains 537 PET/CT images. Male and female patients were nearly equally represented across the groups. For the benign group, the mean age was 48 years, with a minimum age of 4 years and a maximum age of 93 years. Prominent malignancy types in the benign group were Hodgkin's Lymphoma (77 cases), Breast Cancer (45 cases), Colon Cancer (39 cases), and Non-Hodgkin's Lymphoma (28 cases). In the malignant group, the mean age was 54 years, with ages ranging from 5 to 87 years. The most common malignancies were Colon Cancer (30 cases), Breast Cancer (29 cases), Rectal Cancer (12 cases), Melanoma (10 cases), and Hodgkin's Lymphoma (10 cases). For the suspicious group, the mean age was 54 years, with a minimum age of 4 years and a maximum age of 86 years. The predominant malignancy types were Colon Cancer (57 cases), Breast Cancer (49 cases), Hodgkin's Lymphoma (31 cases), Rectal Cancer (23 cases), and Gastric Cancer (21 cases). Detailed information of the demographic information is presented in the Appendix.

This study, therefore, not only aims to improve lung nodule detection but also specifically addresses the challenge of assessing lung nodules in patients with malignancies outside the lung, where the risk of metastatic disease complicates the diagnostic process. By focusing on a diverse array of malignancies, this research seeks to enhance our understanding of how lung nodules should be interpreted in the context of various cancers, ultimately contributing to more accurate and early detection of malignancy in these high-risk groups.

3.11. Pre-processing and hyper parameter tuning

First, we aligned the size of the PET images to match the matrix size of the CT images, i.e. 512×512. We increased the PET image size to correspond with the CT size to maintain image quality for segmentation and preserve nodule information, which might be lost if the images were reduced in size. According our observations from previous objective, we utilized the most recent version of the lung segmentation code to segment the lungs from the CT images. We then multiplied the segmented lung image derived from CT by the corresponding resized PET image to obtain the segmented lung from the aligned PET image as well (Figure 3.8).

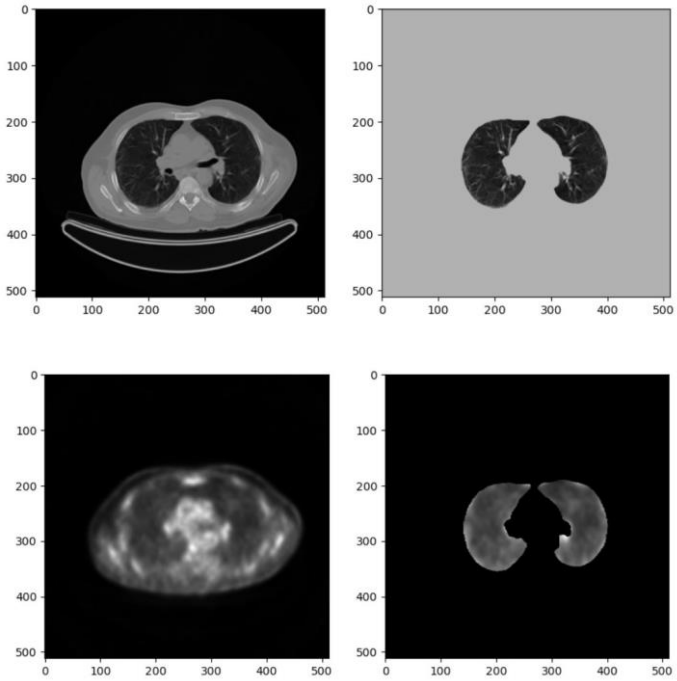


Figure 3.8. Segmentation of CT (top) and PET (bottom) images using python script.

Besides cropping unnecessary borders from the images, we removed all zero voxels, as they do not contain any helpful information. This step helps streamline the simulation process and reduces computational costs. To accommodate the DL model, we used different slice sizes without deforming the original shape of the images, preserving their structural integrity.

To address memory constraints, instead of creating a single numpy input data array, we saved each segmented and cropped image and saved it as a numpy array containing the patient's aligned and pre-processed PET and CT scans, and corresponding label based on the patient's annotation. The labels were as follows: Benign (labeled as "0"), Malignant (labeled as "1"), and Suspicious (labeled as "2"). Then, all images were split into the respective train and test folders. The training folder contained 70% of the dataset for training and validation, while the test folder contained the remaining 30% as holdout test data.

We also performed normalization and zero-centering on the CT and PET images to standardize the input data. For reliability, the simulations were conducted using a 5-fold cross-validation approach, ensuring more robust and reliable results.

3.12. Data generator

We adopted a patch-based approach for this study, building on our previous work. By dividing the images into smaller, manageable patches, we can more effectively analyse detailed regions of the images without overwhelming computational resources. This method allows for finer granularity in the analysis, which is essential for accurate diagnosis in medical imaging. We chose patch sizes that were sufficiently large, constrained only by the system's memory capacity, to ensure meaningful results when evaluating the entire image. This ensures that each patch contains sufficient contextual information to be useful for diagnostic purposes, thus improving the model's ability to make accurate predictions when the patches are combined to assess the entire image. This was essential for the objective of lung nodule diagnosis. In medical imaging datasets, class imbalance is a common issue that can lead to biased model performance. Given the data imbalance among the three classes (with malignant cases being about half as frequent as benign and suspicious cases), To counteract this, we employed an oversampling technique by generating additional patches for the malignant class. We balanced the training dataset by adjusting the number of patches created for malignant cases. This ensures that the model receives adequate representation from all classes, which is crucial for effectively distinguishing between different types of nodules.

3.13. Augmentation

Given the complexity of PET/CT datasets, we implemented several augmentation techniques to enhance our network's performance. These augmentations include flipping images in the x direction, rotating, scaling, rotating by three degrees, and shifting in the z direction. Flipping in the X Direction horizontally flips the images and ensures the model learns features regardless of their orientation, promoting invariance to the left-right positioning of the images. Rotating the images helps the model to recognize structures from different angles, improving its ability to handle varied patient positioning. Scaling adjusts the size of the images can help the model become invariant to different sizes of anatomical structures and lesions. Three-degree rotations are specific small angle rotations to ensure minor angular variations are also captured in the training process, further enhancing model robustness. Shifting in the Z Direction moves the images slightly along the z-axis (depth) and helps the model to be more adaptable to different slices and depths in volumetric scans.

We carefully selected augmentation techniques to ensure the integrity and diagnostic value of the medical images remained intact. Some augmentations, while potentially beneficial for general image processing, can distort the medical images in ways that render them less useful or even

unrecognizable. For instance, certain augmentations can alter the shape or structure of anatomical features, leading to a loss of crucial diagnostic information.

To illustrate, we avoided using noise augmentation for images containing pulmonary nodules. Pulmonary nodules are inherently small and challenging to detect; introducing additional noise could obscure these nodules further, making accurate detection even more difficult. Similarly, elastic distortion was excluded because it can significantly warp the shapes of organs or tissues. Such alterations might result in images that no longer represent the true anatomical structures, thus losing their diagnostic relevance.

These decisions are grounded in maintaining the clinical applicability and reliability of the augmented images. Figure 3.9 provides examples of images that were excluded due to inappropriate augmentations.

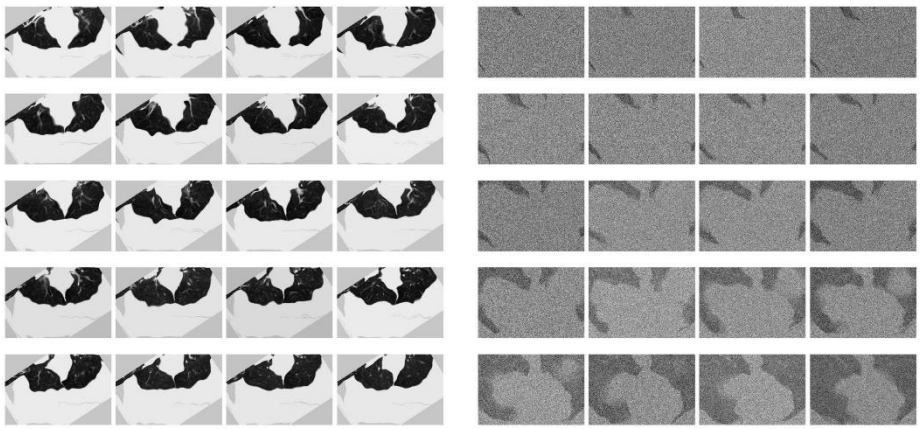


Figure 3.9. examples of excluded augmentations techniques that don't keep the original image information.

3.14. Evaluated deep learning models for lung nodule classification

State-of the art DL models including Resnet18, Resnet34, Resnet50, Resnet101, Resnext50, Resnext101, Seresnet18, Seresnet34, Seresnet50, Seresnet101, Seresnext50, Seresnext101 and Inception-Resnet-v2 have been employed for simulations to test which model outperforms other models.

We also experimented with various DL models, including EfficientNet, GraphNet, DenseNet, VGG, as well as deeper architectures such as ResNet152 and SE-ResNet152. However, due to challenges related to memory allocation and other technical complexities, we decided to exclude these models from our study. Addressing these issues was beyond the scope of our current research. Moreover, based on the outcomes of our initial experiments and the findings from

objective 1, we observed that the performance differences among most DL models were minimal. This observation is supported by research indicating that, beyond a certain point, increasing model complexity often yields diminishing returns in performance improvements, particularly when applied to similar tasks (Goodfellow et al., 2016).

3.14.1. *Selected DL model for nodule detection using ^{18}F FDG PET/CT images*

The Inception-ResNet model integrates Inception modules with residual connections. Each Inception module is modified to include a residual connection that bypasses the module and adds its input directly to the output. This helps train deep networks by mitigating the vanishing gradient problem and facilitating more effortless gradient flow. Like the original Inception model, inception blocks capture multi-scale features. Residual connections are added to each Inception block, allowing the model to benefit from the identity mappings that make optimization easier and improve convergence. Inception-ResNet incorporates the Inception modules, which can capture features at multiple scales within the same network layer. This leads to a richer representation than ResNet, which typically uses fixed-size filters. The residual connections in Inception-ResNet help mitigate the vanishing gradient problem, similar to ResNet, but with the added benefit of multi-scale feature extraction from the Inception modules. The Inception modules can be designed to be computationally efficient by including 1x1 convolutions that reduce dimensionality before applying more expensive operations like 3x3 and 5x5 convolutions. This efficiency is maintained in Inception-ResNet, allowing it to be deeper without an exponential increase in computational cost. The architecture of Inception-ResNet is more flexible in terms of handling different image sizes and complexities due to the varied convolutional filter sizes within the Inception modules. This can lead to better performance on diverse datasets.

3.15. Explainability in AI

Explainability in AI is imperative and cannot be overlooked. It is crucial for fostering trust and accountability in AI systems and making sure they are employed in an ethical and accountable fashion (Abedin, 2022). Explainability in AI is essential for fostering trust and accountability, especially in medical imaging applications such as PET/CT and CT scans. In healthcare, decisions made by AI systems can have significant implications for patient outcomes. Understanding how AI models arrive at their conclusions is crucial for clinicians and patients to trust these systems (Gunning & Aha, 2019). Without transparency, there may be reluctance to adopt AI technologies in clinical practice due to concerns about their reliability and fairness (Danks & London, 2017). Ethical considerations require that AI systems used in medical imaging be transparent to ensure responsible use. Explainable AI (XAI) can help identify and mitigate biases, errors, and unfair treatment embedded in algorithms. Providing clear insights into AI decision-making processes ensures that these systems align with ethical standards and legal requirements (Floridi & Cowls, 2022).

One of the significant challenges in ensuring the XAI models is their complexity, especially in DL models that can make highly accurate predictions and decisions. However, DL models are frequently perceived as a "black box" due to the lack of transparency in their internal workings and decision-making process, which can result in mistrust, bias, and errors in the AI systems (Bjerring & Busch, 2021).

Another challenge in explainability is the trade-off between accuracy and interpretability. Complex models can attain high accuracy by learning from vast amounts of data but can be difficult to understand and interpret. On the other hand, simple models with fewer parameters are easier to comprehend but may not achieve the same level of accuracy as complex models. As AI grows increasingly important in our daily lives, we must make explainability a priority, to make sure AI systems serve the benefit of all.

3.15.1. *Current Explainable AI (XAI) Methods in Medical Imaging*

Several XAI methods have been developed to address these challenges in medical imaging, each with distinct advantages and limitations:

3.15.1.1. *Model-Specific Methods*

Feature Visualization Feature visualization techniques help understand what features a neural network is learning at different layers by generating images that maximize neuron activation (Olah et al., 2017). This approach provides insights into how a model perceives and processes medical images, such as PET/CT and CT scans. This method Offers a visual understanding of internal representations and learned features. However, it Can be difficult to interpret for complex networks and may not provide complete explanations.

- Saliency Maps and Grad-CAM

  Saliency maps highlight the regions of an input image contributing most to the model's prediction. Grad-CAM (Gradient-weighted Class Activation Mapping) is a popular technique for generating visual explanations by producing heat maps that indicate important regions in the input image for model prediction (Selvaraju et al., 2020). It provides intuitive visual explanations of model predictions, highlighting crucial features in medical images. Grad-CAM is useful in identifying areas in scans that influence the AI's decisions, thus aiding radiologists in understanding the model’s focus. However, Saliency maps and Grad-CAM can produce noisy or ambiguous maps, and their effectiveness can be sensitive to model parameters and the specific layers chosen for visualization.

3.15.1.2. *Model-Agnostic Methods*

- LIME (Local Interpretable Model-agnostic Explanations)

LIME explains individual predictions by approximating the complex model locally with a simpler interpretable model (Ribeiro et al., 2016). LIME works by perturbing the input data around the prediction and observing how the predictions change. LIME then fits an interpretable model, such as a linear model, to these perturbed samples to explain the complex model's behavior in the local vicinity of the prediction. LIME offers local explanations for individual predictions, making it applicable across various models, regardless of their internal structure. However, local explanations may not generalize well to the entire model, potentially leading to misinterpretations if the local surrogate model does not accurately reflect the global behavior of the original model.

SHAP (SHapley Additive exPlanations)

SHAP values provide a unified measure of feature importance by assessing each feature's contribution to the model's prediction based on game theory (Lundberg & Lee, 2017). It offers a consistent and theoretically grounded approach to feature importance. However, it is computationally expensive, especially for large datasets and complex models.

3.15.1.3. *Rule-Based Methods*

- Decision Trees and Rule Extraction

These methods aim to extract decision rules from complex models like neural networks by approximating their decision boundaries with interpretable structures such as trees. This technique produces human-readable rules, offering insight into model decisions in medical imaging. However, it may oversimplify complex models, leading to a loss of accuracy.

3.16. Summary

This chapter presents a comprehensive methodology for evaluating the generalizability and interpretability of DL models across two major medical imaging tasks: (1) classifying COVID-19 lung involvement using 3D CT images, and (2) multi-classifying lung nodules using ^{18}FFDG PET/CT images. The study emphasizes the robustness of these models across diverse datasets and applies XAI techniques to ensure transparency in decision-making.

To classify COVID-19 lung involvement, the study utilized four independently sourced datasets, including the Iranmehr and Moscow datasets. Preprocessing steps included resampling to standardize slice numbers, intensity clipping, and lung segmentation using methods such as U-Net and Zuidhof. The models, including 3D CNNs like ResNet and DenseNet, were trained using k-fold cross-validation to evaluate their performance across different datasets.

Generalization experiments included three approaches:

80% Dataset Combinations: Models were trained on 80% of Iranmehr, Moscow, and combined datasets, and tested on the remaining 20% holdout sets, as well as external datasets (LDCT and 3DLSC).

Proportional Mixes: Increasing portions of one dataset were added to another to assess the impact of mixed data on model generalization.

Transfer Learning: Pretrained models were fine-tuned on the final layers using another dataset. The fine-tuned models were evaluated on internal and external holdout datasets.

The study also focused on classifying lung nodules using ^{18}F FDG PET/CT images from 1304 patients, excluding primary lung cancer cases to avoid bias. Preprocessing steps involved resizing, segmentation, and normalization. To handle class imbalance, oversampling and extensive data augmentation were applied. Various DL models were tested, with a focus on Inception-ResNet-v2, which was selected for its multi-scale feature extraction capabilities.

To ensure transparency and trustworthiness in the model's predictions, the study employed various XAI methods to visually explain the features driving the model's decisions. These methods generated heatmaps to highlight the regions of CT and PET/CT images that influenced the model's predictions the most. This layer of interpretability was critical for enhancing clinician trust and ensuring that the DL models could be adopted in real-world clinical settings.

The performance of the models was assessed using standard evaluation metrics such as accuracy, sensitivity, specificity, F1-score, and AUC of the ROC curve. These metrics provided a comprehensive view of the models' generalizability across diverse datasets. The use of XAI ensured that the models were not only accurate but also interpretable, making them suitable for real-world medical applications.

CHAPTER 4

4. RESULTS AND DISCUSSION

4.1. Introduction

This section presents the results and discussions of the current study. The results are systematically analyzed to evaluate the performance, generalizability, and robustness of the employed DL models across different datasets and tasks. The first part of the results focuses on the classification of COVID-19 lung involvement. The ResNet-50 model, selected based on its superior performance in preliminary tests, was extensively evaluated using multiple datasets, including Iranmehr, Moscow, LDCT, and 3DLSC. The results are presented in terms of accuracy, sensitivity, specificity, F1-score, and AUC for different dataset combinations and pre-processing methods. A significant emphasis is placed on assessing the model's generalizability, especially in how well it performs when trained on one dataset and tested on another. The results demonstrate that while ResNet-50 generally performs well across all datasets, its performance varies depending on the specific dataset and pre-processing techniques applied, such as image cropping.

The second part of the results section addresses the multi-classification of lung nodules into benign, malignant, and suspicious categories using ^{18}F-FDG PET/CT images. The InceptionResNet-v2 model is highlighted for its high accuracy and discriminative power across the different nodule classes, with performance metrics such as sensitivity, specificity, PPV, and NPV providing detailed insights into the model's effectiveness. Additionally, the study explores the use of feature visualization and SHAP representation to interpret the model's decision-making process, revealing the critical regions and features the model focuses on during classification.

A significant portion of the results discusses the generalizability of the models across different datasets. The study investigates various combinations of datasets, testing the capability of the models to generalize beyond the data they were trained on. This includes experiments with proportional mixes of datasets and transfer learning, where models trained on one dataset are fine-tuned on another. The findings suggest that combining datasets and using transfer learning can enhance model performance, although the results also highlight challenges, such as the variability in specificity when testing on different external datasets. The results demonstrate the importance of dataset diversity, pre-processing techniques, and model selection in achieving robust and generalizable DL models for medical image classification.

4.2. Results for COVID-19 involvement classification from CT images and generalizability

We selected ResNet-50 as the 3D model for the COVID-19 classification due to the better results (Table 4.1) and lower runtime, which is consistent with previous studies into COVID-19

classification from 3D CT images. An overview of Table 4.1 shows that ResNet-50 outperforms other models. In particular, the results of the ResNet-50 are better than the DenseNet-169 on the Moscow dataset in terms of accuracy and standard deviation. Regarding the comparison of cropped and non-cropped images, all models had better results on cropped images, except for ResNet-152 and DenseNet-201.

In the generalizability assessment approach, all models were evaluated by the mean of k-fold running parameters on external tests. Since our study is a binary simulation, we used sigmoid activation function. The "jitter" that is noticeable on the earlier training epochs is caused by fluctuations in training loss which is the consequence of training a very large network (ResNet50 with approximately 50 million trainable parameters) using datasets of 1110 3D CT scans. Also, COVID-19 lung involvement is not apparent or it may be very subtle in some patient CT slices. As a result, when there are a number of such slices in a training batch, the validation loss for that batch will be very small because gradient descent is minimal. As training proceeds, the neural network becomes more tolerant of these adversarial images resulting in a smoother training curve in the later training epochs.

Table 4.1. Accuracy (mean ± std) for 5-fold cross-validation on cropped and non-cropped images of Iranmehr and Moscow datasets.

	Iranmehr cropped	**Iranmehr non-cropped**	**Moscow cropped**	**Moscow non-cropped**
Densenet-169	**0.942 ± 0.012**	0.938 ± 0.004	0.857 ± 0.027	0.843 ± 0.019
ResNet-50	0.939 ± 0.014	0.938 ± 0.006	**0.864 ± 0.028**	**0.855 ± 0.012**
Resnext-50	0.940 ± 0.011	0.932 ± 0.006	0.856 ± 0.029	0.837 ± 0.007
Densenet-201	0.937 ± 0.012	**0.942 ± 0.006**	0.862 ± 0.032	0.834 ± 0.015
Resnet-152	0.935 ± 0.012	0.937 ± 0.006	0.859 ± 0.035	0.846 ± 0.013
Seresnet-152	0.936 ± 0.013	0.925 ± 0.009	0.857 ± 0.024	0.838 ± 0.015
Seresnext-50	0.934 ± 0.013	0.931 ± 0.006	0.856 ± 0.027	0.845 ± 0.013

The statistic and AUC results for model selection (external-validation evaluation) for training on Iranmehr dataset and tested on Moscow dataset, LDCT, and 3DLSC are presented in tables 4.2 to 4.4, respectively. The reverse process, i.e., training on Moscow dataset and testing on Iranmehr dataset, LDCT, and 3DLSC are presented in tables 4.5 to 4.7, respectively. According to the

tables, the results of training on Iranmehr and testing on LDCT are higher than other test datasets. However, it should be noted that compared to LDCT and 3DLSC, Moscow and Iranmehr datasets contain 1110 images, which increase the testing validity.

The accuracy of a different combination of datasets of experiment phases is given in table 4.8. According to table 4.8, the combination of 80 percent of one dataset with the addition of different of the other has close accuracy to the accuracy of the total combination. Table 4.9 presents the AUC results of different combination of datasets of experiment phases. According to table 4.8, all AUC results are near to the AUC of total combination. In table 4.10, the statistical transfer learning results are presented. By general overview of table 4.10, it can be found that for all metrics except for specificity, the results of retraining on 80 percent Moscow dataset using the weights of a full run of Iranmehr dataset are higher compared to the retraining on 80 percent Iranmehr dataset using the weights of Moscow dataset full run.

Table 4.2. Different model results when trained with Iranmehr dataset and tested against Moscow dataset.

3D Models	Accuracy	Sensitivity	Specificity	F1-Score	AUC
DenseNet-169	**0.814**	0.777	0.888	0.858	87±0.01
DenseNet-201	0.802	0.762	0.876	0.846	86±0.01
ResNet-152	0.796	0.775	**0.896**	0.858	88±0.01
ResNet-50	0.800	0.799	0.818	0.861	**89±0.01**
ResnNext-50	0.788	**0.813**	0.765	**0.862**	89±0.02
Seresnet-152	0.800	0.763	0.889	0.849	88±0.00
Seresnext50	0.792	0.758	0.9	0.847	88±0.01

Table 4.3. Different model results when trained with Iranmehr dataset and tested against LDCT dataset.

3D Models	Accuracy	Sensitivity	Specificity	F1-Score	AUC
DenseNet-169	**0.935**	0.901	0.982	**0.943**	0.96±0.01
DenseNet-201	0.917	0.875	**0.978**	0.927	95±0.02
ResNet-152	0.916	0.892	0.960	0.932	96±0.01
ResNet-50	0.920	0.892	0.964	0.933	96±0.01
ResnNext-50	0.910	**0.907**	0.971	**0.943**	**96±0.00**
Seresnet-152	0.897	0.878	0.957	0.924	95±0.01
Seresnext50	0.912	0.886	0.964	0.930	94±0.00

Table 4.4. Different model results when trained with Iranmehr dataset and tested against 3DLSC dataset.

3D Models	Accuracy	Sensitivity	Specificity	F1-Score	AUC
DenseNet-169	**0.889**	0.794	**0.970**	0.870	93±0.02
DenseNet-201	0.869	0.78	0.910	0.832	93±0.03
ResNet-152	0.871	0.808	0.958	0.873	**94±0.01**
ResNet-50	0.857	0.798	0.922	0.850	94±0.03
ResnNext-50	0.848	0.842	0.887	0.861	94±0.03
Seresnet-152	0.866	**0.844**	0.922	**0.879**	94±0.02
Seresnext50	0.863	0.786	0.943	0.853	94±0.02

Table 4.5. Different model results when trained with Moscow dataset and tested against Iranmehr dataset.

3D Models	Accuracy	Sensitivity	Specificity	F1-Score	AUC
DenseNet-169	**0.765**	0.943	0.507	0.765	92±0.01
DenseNet-201	0.716	0.921	0.636	0.809	92±0.01
ResNet-152	0.677	0.952	0.552	0.795	92±0.02
ResNet-50	0.741	0.946	0.595	0.805	92±0.02
ResnNext-50	0.699	**0.96**	0.492	0.781	92±0.02
Seresnet-152	0.716	0.953	0.548	0.797	92±0.01
Seresnext50	0.745	0.918	**0.681**	**0.822**	**93±0.01**

Table 4.6. Different model results when trained with Moscow dataset and tested against LDCT dataset.

3D Models	Accuracy	Sensitivity	Specificity	F1-Score	AUC
DenseNet-169	0.880	0.915	0.764	0.898	94±0.01
DenseNet-201	0.827	0.905	0.835	0.909	94±0.02
ResNet-152	0.847	**0.930**	0.782	0.910	**95±0.01**
ResNet-50	**0.889**	0.913	**0.857**	**0.918**	**95±0.01**
ResnNext-50	0.850	0.923	0.757	0.900	94±0.01
Seresnet-152	0.877	0.909	0.775	0.897	94±0.01
Seresnext50	0.846	0.894	0.814	0.898	94±0.01

Table 4.7. Different model results when trained with Moscow dataset and tested against 3DLSC dataset.

3D Models	Accuracy	Sensitivity	Specificity	F1-Score	AUC
DenseNet-169	0.728	0.886	0.672	**0.805**	90±0.03
DenseNet-201	0.715	0.882	**0.693**	0.811	90±0.02
ResNet-152	0.710	**0.954**	0.472	0.777	91±0.02
ResNet-50	0.771	0.91	0.662	0.817	**92±0.02**
ResnNext-50	0.737	0.9	0.591	0.786	89±0.03
Seresnet-152	**0.765**	0.888	0.662	**0.805**	90±0.03
Seresnext50	0.727	0.928	0.587	0.800	**92±0.02**

Table 4.8. Accuracy (%) of trained ResNet-50 for 80% dataset portions, combinations and proportional mixes of Iranmehr and Moscow as training datasets experiments.

Training setting	Test sets			
	Iranmehr test data (20%)	**Moscow test data (20%)**	**LDCT**	**3DLSC**
80% Iranmehr	90.5	79.5	90.7	84.28
80% Moscow	71.8	82.1	86.3	76.5
80% Iranmehr + 20% Moscow	91.2	81.0	92.2	88.0
80% Iranmehr + 40% Moscow	91.7	81.0	91.4	83.4

80% Iranmehr + 60% Moscow	91.3	82.6	91.3	84.0
80% Moscow + 20% Iranmehr	87.2	80.8	89.8	73.6
80% Moscow + 40% Iranmehr	89.1	82.1	90.7	77.9
80% Moscow + 60% Iranmehr	89.7	80.9	90.3	86.5
80% Iranmehr + 80% Moscow	91.5	83.6	91.9	73.9

Table 4.9. AUC ± std of trained ResNet-50 for 80% dataset portions, combinations and proportional mixes of Iranmehr and Moscow as training datasets experiments.

Training setting	**Test sets**			
80% Iranmehr	**Iranmehr test data (20%)**	**Moscow test data (20%)**	**LDCT**	**3DLSC**
80% Moscow	0.096±0.00	0.87±0.03	0.96±0.01	0.95±0.02
80% Iranmehr + 20% Moscow	0.92±0.01	0.91±0.01	0.94±0.01	0.90±0.02
80% Iranmehr + 40% Moscow	0.95±0.01	0.86±0.02	0.95±0.01	0.96±0.01
80% Iranmehr + 60% Moscow	0.95±0.01	0.87±0.03	0.95±0.01	0.96±0.01
80% Moscow + 20% Iranmehr	0.94±0.01	0.89±0.01	0.96±0.01	0.95±0.02

80% Moscow + 40% Iranmehr	0.92±0.01	0.88±0.01	0.94±0.00	0.94±0.01
80% Moscow + 60% Iranmehr	0.94±0.00	0.89±0.01	0.95±0.01	0.95±0.02
80% Iranmehr + 80% Moscow	0.94±0.00	0.88±0.02	0.95±0.00	0.95±0.03
	0.95±0.01	0.90±0.02	0.95±0.01	0.95±0.01

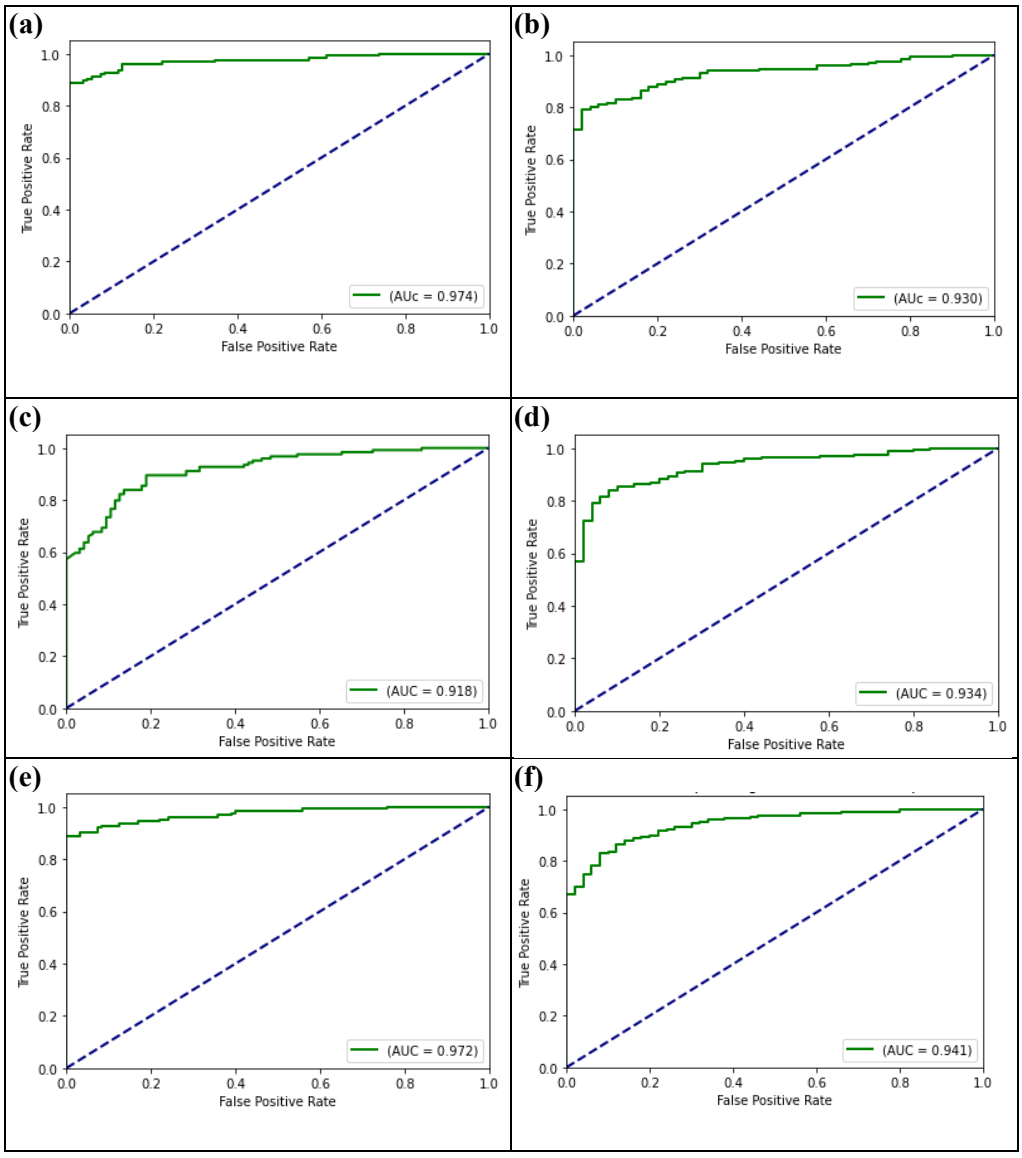
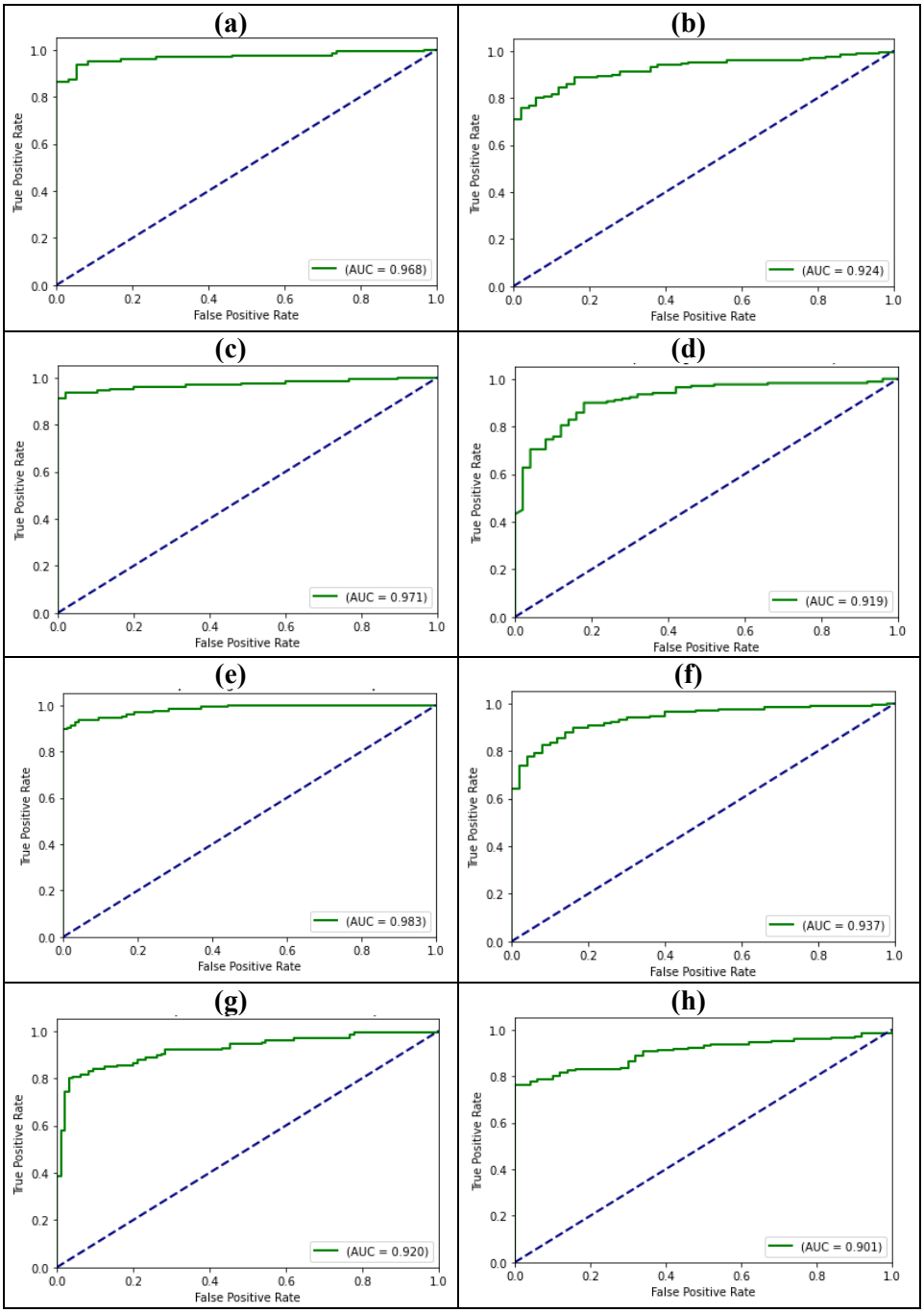


Figure 4.1. AUC of base experiment for a) training with 80% IranMehr, testing on 20% IranMehr, b) training with 80% IranMehr, testing on 20% MOSCOW, c) training with 80% MOSCOW and testing on 20% IranMehr, d) training with 80% MOSCOW and testing on 20% MOSCOW, e) training with 80% IranMehr + 80% MOSCOW and testing on 20% IranMehr and, f) training with 80% IranMehr + 80% MOSCOW and testing on 20% MOSCOW.



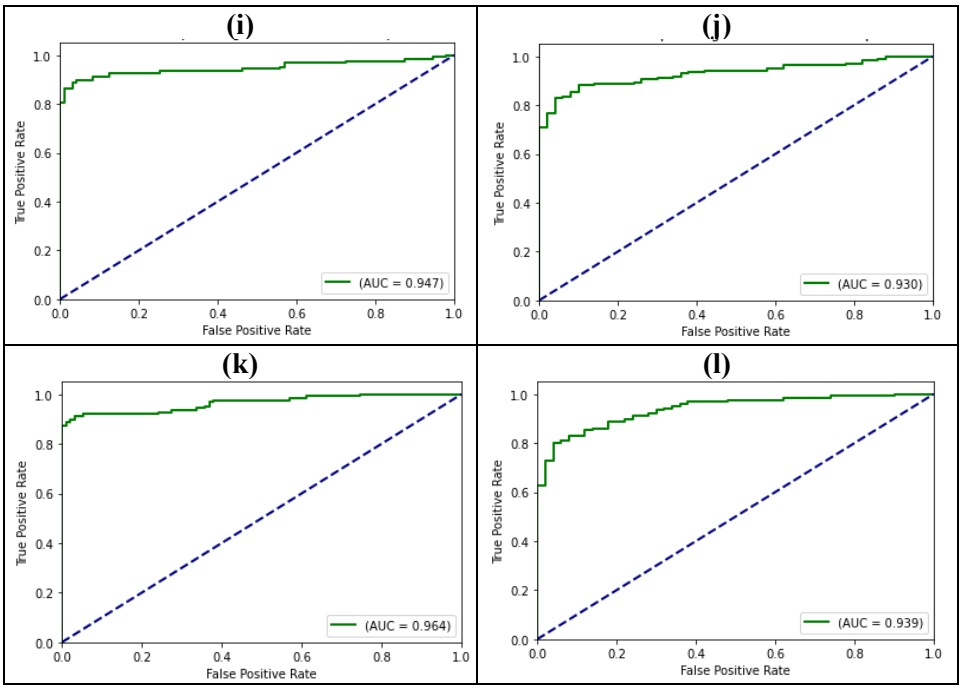


Figure 4.2. AUC of first experiment for (a) training with 80% IranMehr + 20% MOSCOW and testing on 20% IranMehr, (b) training with 80% IranMehr + 20% MOSCOW and testing on 20% MOSCOW, (c) training with 80% IranMehr + 40% MOSCOW and testing on 20% IranMehr, (d) training with 80% IranMehr + 40% MOSCOW and testing on 20% MOSCOW, (e) training with 80% IranMehr + 60% MOSCOW and testing on 20% IranMehr, (f) training with 80% IranMehr + 60% MOSCOW and testing on 20% MOSCOW, (g) training with 80% MOSCOW + 20% IranMehr and testing on 20% IranMehr (h) training with 80% MOSCOW + 20% IranMehr and testing on 20% MOSCOW, (i) training with 80% MOSCOW + 40% IranMehr and testing on 20% IranMehr, (j) training with 80% MOSCOW + 40% IranMehr and testing on 20% MOSCOW, (k) training with 80% MOSCOW + 60% IranMehr and testing on 20% IranMehr, (l) training with 80% MOSCOW + 60% IranMehr and testing on 20% MOSCOW

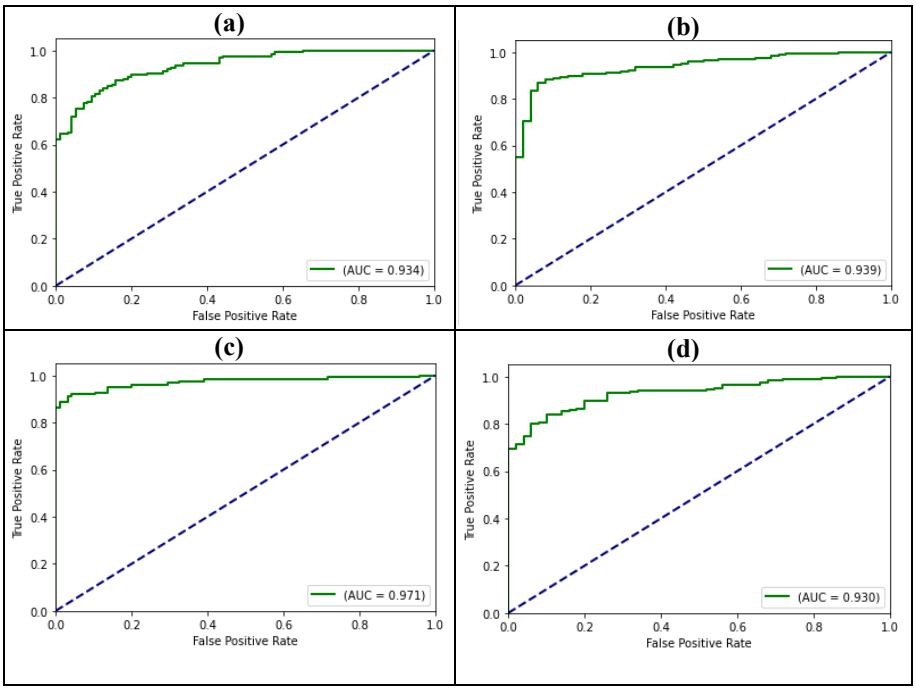


Figure 4.3. AUC of second experiment for (a) trained weights with 80% MOSCOW, retrain 3 layers on 80% IranMehr data, testing on 20% IranMehr, (b) trained weights with 80% MOSCOW, retrain 3 layers on 80% IranMehr data, testing on 20% MOSCOW, (c) trained weights with 80% IranMehr, retrain 3 layers on 80% MOSCOW data, testing on 20% IranMehr and, (d) trained weights with 80% IranMehr, retrain 3 layers on 80% MOSCOW data, testing on 20% MOSCOW.

Table 4.10. Statistics of trained ResNet-50 for transfer learning experiment.

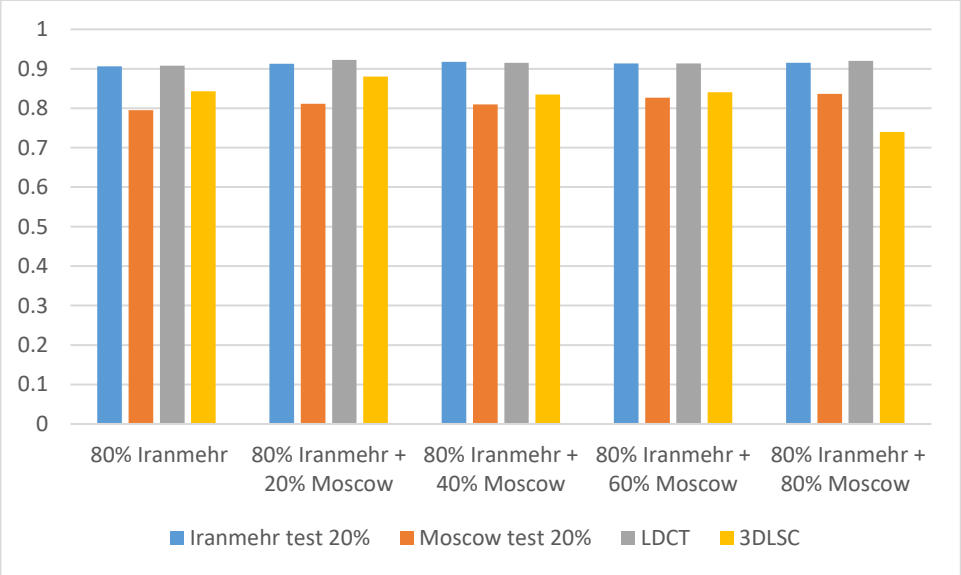
Test sets		**Training setting**	
		Retrain on 80% Iranmehr	Retrain on 80% Moscow
Accuracy (%)	Iranmehr test data (20%)	83.2	76.8
	Moscow test data (20%)	64.2	79.1
	LDCT	**85**	**82.3**
	3DLSC	77.6	71.2
Sensitivity (%)	Iranmehr test data (20%)	**79.8**	96.9

	Moscow test data (20%)	54.2	91.9
	LDCT	78.4	92.6
	3DLSC	63	**98.4**
Specificity (%)	Iranmehr test data (20%)	87.3	56.1
	Moscow test data (20%)	**100**	34
	LDCT	97.5	**64.2**
	3DLSC	93.1	41.4
F1-Score	Iranmehr test data (20%)	82.9	80.8
	Moscow test data (20%)	70.2	**86.9**
	LDCT	**87.2**	87.5
	3DLSC	74.2	77.4
AUC	Iranmehr test data (20%)	89±0.00	**0.95±0.00**
	Moscow test data (20%)	0.90±0.00	0.86±0.01
	LDCT	**0.93±0.00**	0.94±0.01
	3DLSC	86±0.00	0.94±0.01

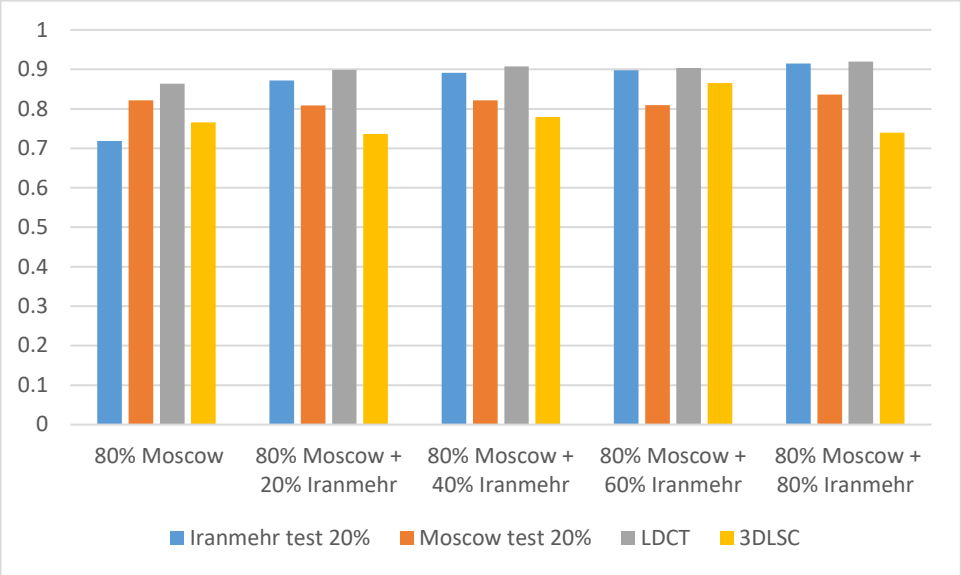
The comparison diagrams of accuracy, sensitivity, specificity, and F1 score are presented in figures 4.1 to 4.4. Additionally, figure 4.5 presents the results of the confusion matrix for all three experiments. According to figure 4.1, the accuracy of different combinations of datasets almost smoothly grows. We can see that the combination of 80 percent from one dataset with the addition of a different portion of the other datasets performs similarly on test sets. Figure 4.2 illustrates the sensitivity of different combinations. Based on the results shown in figure 4.2, for all the combinations, the results of testing on holdout test sets are considerably different from each other. Regarding to the specificity, it can be seen from figure 4.3 that the combination above 80 percent from one dataset and 40 percent of the other have similar results to the total combination. For F1

score, as shown in figure 10, the combination of 80 percent of one dataset added to 40 percent of the other reaches the results near to the total combination.

In Figure 4.5 (a-c), the results of the confusion matrix are presented when different combinations, as well as transfer learning results are tested against the unseen holdout dataset. The highest number of TPs belongs to the total combination, and other combinations have close results when tested on holdout test set of Moscow and Iranmehr. However, when testing on the external datasets, 3DLSC, we can see that the numbers of FPs are high in combination of 80 percent Moscow and 20 percent Iranmehr. Following the total combination, the combination of 80 percent Moscow and 20 percent Iranmehr and 80 percent Moscow added to 20 percent Iranmehr have the lowest number of FNs when tested on Iranmehr holdout test set.

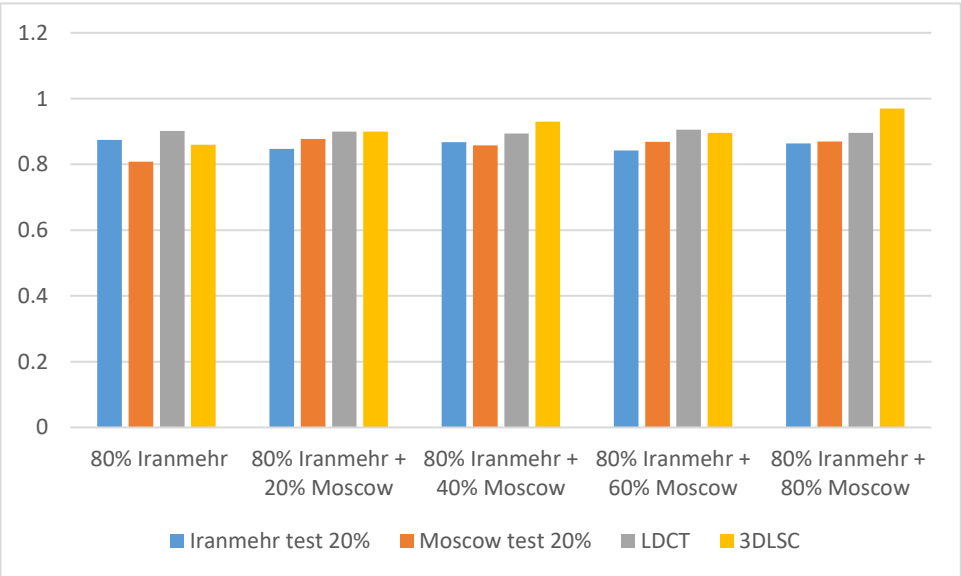


(a)

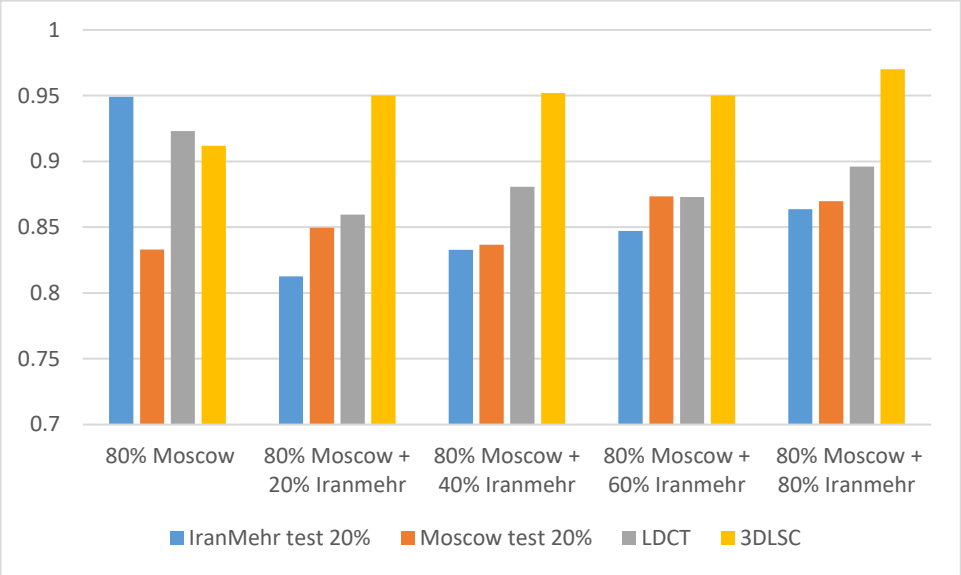


(b)

Figure 4.4. Accuracy results for a different portion of datasets when (a) 80% of the dataset is Iranmehr and splits of Moscow are added, and (b) 80% of the dataset is Moscow and splits of Iranmehr are added.

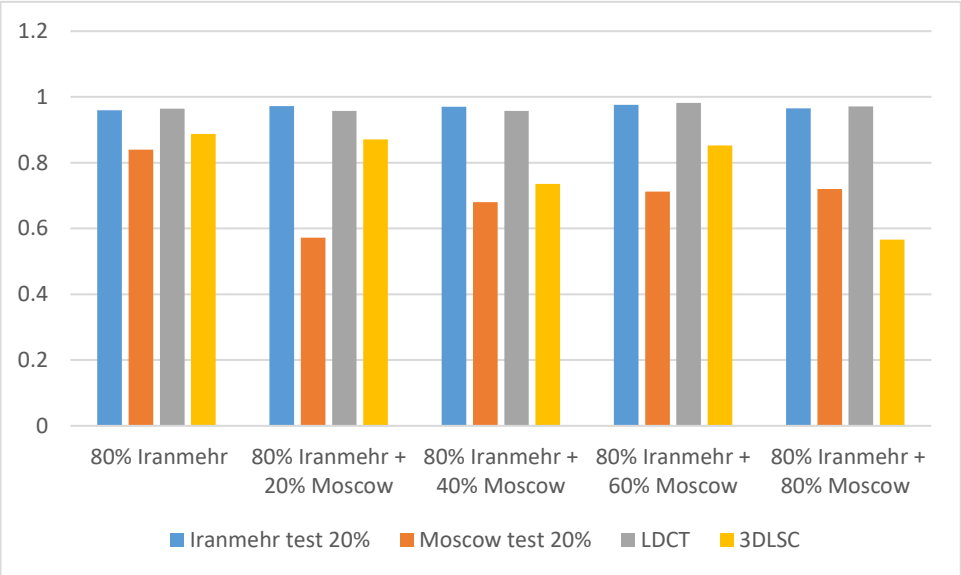


(a)

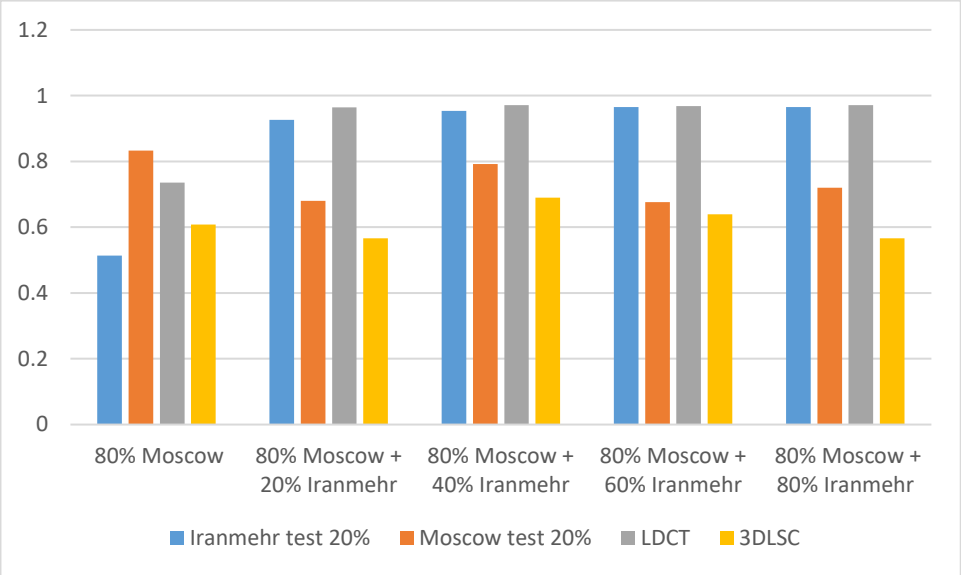


(b)

Figure 4.5. Sensitivity results for a different portion of datasets when (a) 80% of the dataset is Iranmehr and splits of Moscow are added, and (b) 80% of the dataset is Moscow and splits of Iranmehr are added.

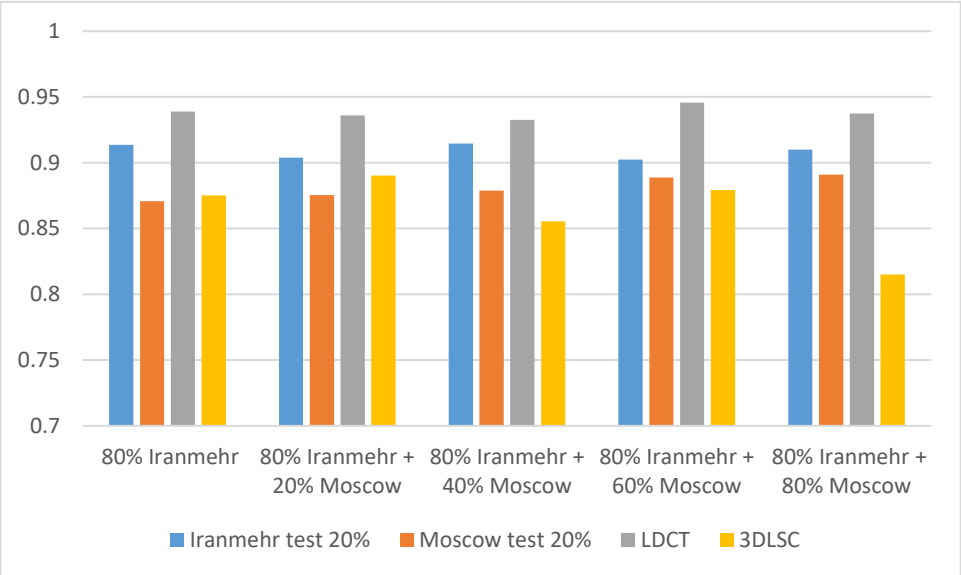


(a)

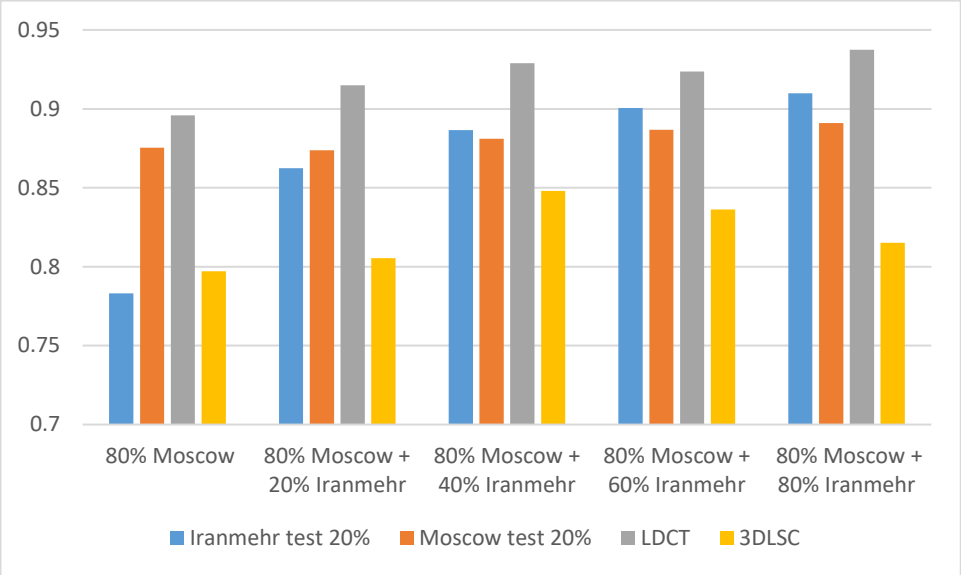


(b)

Figure 4.6. Specificity results for a different portion of datasets when (a) 80% of the dataset is Iranmehr and splits of Moscow are added, and (b) 80% of the dataset is Moscow and splits of Iranmehr are added.



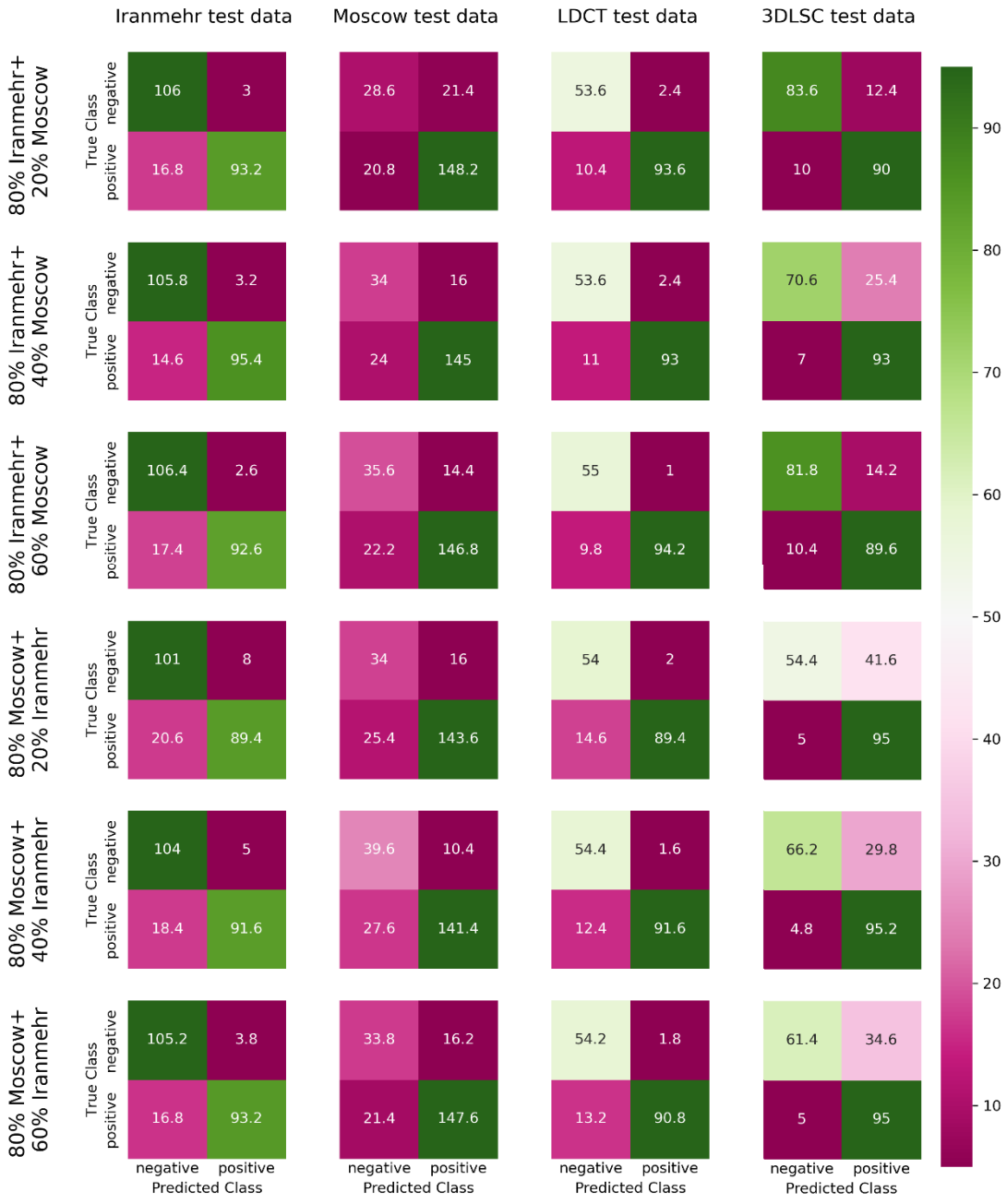
(a)



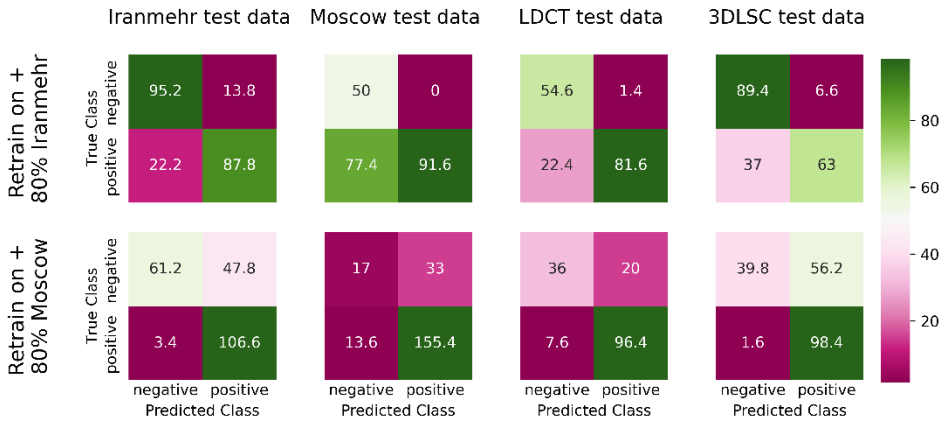
(b)

Figure 4.7. F1-score results for a different portion of datasets when (a) 80% of the dataset is Iranmehr and splits of Moscow are added, and (b) 80% of the dataset is Moscow and splits of Iranmehr are added.

(a)


Iranmehr test data
Moscow test data
LDCT test data
3DLSC test data
80% Iranmehr+ 20% Moscow
80% Iranmehr+ 40% Moscow
80% Iranmehr+ 60% Moscow
80% Moscow+ 20% Iranmehr
80% Moscow+ 40% Iranmehr
80% Moscow+ 60% Iranmehr
True Class
negative
positive
106 3 16.8 93.2
28.6 21.4 20.8 148.2
53.6 2.4 10.4 93.6
83.6 12.4 10 90
105.8 3.2 14.6 95.4
34 16 24 145
53.6 2.4 11 93
70.6 25.4 7 93
106.4 2.6 17.4 92.6
35.6 14.4 22.2 146.8
55 1 9.8 94.2
81.8 14.2 10.4 89.6
101 8 20.6 89.4
34 16 25.4 143.6
54 2 14.6 89.4
54.4 41.6 5 95
104 5 18.4 91.6
39.6 10.4 27.6 141.4
54.4 1.6 12.4 91.6
66.2 29.8 4.8 95.2
105.2 3.8 16.8 93.2
33.8 16.2 21.4 147.6
54.2 1.8 13.2 90.8
61.4 34.6 5 95
negative positive
Predicted Class
90
80
70
60
50
40
30
20
10

(b)



(c)

Figure 4.8. The results of confusion matrix for (a) 80% dataset portions and combinations experiment; (b) proportional mixes of Iranmehr and Moscow as training datasets experiment; and (c) Transfer learning experiment.

Through our trial and error experiments, we found that segmenting the data and excluding 0 values improved the model's overall performance by approximately 5%. Additionally, Inceptionresnet outperformed other employed DL models in terms of results and simulation speed. During the simulations, we saved the best and final weights corresponding to the highest validation accuracy and lowest loss, and then tested these weights on lung patches as well as full images. The findings indicated that the weights yielding the best accuracy on the full images outperformed all other weights.

4.3. Results for lung nodule classification from ^{18}F FDG PET/CT scans

4.3.1. *DL Model selection*

The performance of various DL models was evaluated on our dataset, with overall accuracy serving as the primary metric for comparison. Among the tested architectures, the InceptionResNet model achieved the highest accuracy at 89%, indicating its superior capability in handling our data. ResNet architectures showed varying levels of performance: ResNet50 attained an accuracy of 86%, which was higher than ResNet101 (84%), ResNet18 (81%), and ResNet34 (79%). The ResNeXt models exhibited competitive performance, with ResNeXt101 reaching an accuracy of 85%, closely followed by ResNeXt50 at 83%. In the SE-ResNet variants, SE-ResNet50 achieved an accuracy of 85%, similar to that of ResNeXt101. Other SE-ResNet models, such as SE-ResNet101 and SE-ResNeXt101, both scored 84%, while SE-ResNet18 and SE-ResNet34 attained accuracies of 78% and 81%, respectively. The results indicate that the InceptionResNet model outperforms the other tested architectures in this specific task (89%), with several models, particularly ResNet50 and SE-ResNet50, also demonstrating robust performance.

These findings highlight the variability in accuracy across different DL architectures, emphasizing the importance of model selection in optimizing performance for specific datasets.

4.3.2. *Classification report for lung nodule classification*

Table 4.10 presents the classification report for the three classes: benign, malignant, and suspicious. The report includes key metrics such as sensitivity, specificity, PPV, and NPV, which offer a detailed overview of the model's performance across each category.

Table 4.11. Classification report for benign, malignant, and suspicious classes.

	Sensitivity (%)	**Specificity (%)**	**PPV (%)**	**NPV (%)**
Benign	88	93.3	90.3	90.9
Malignant	90	85.7	86.3	89.5
Suspicious	87.6	90.9	87.7	91

The performance of the classification model is depicted through two key visualizations: the ROC curves and the training versus validation accuracy plot.

4.3.3. *ROC Curves and AUC*

The ROC curve plot presented in figure 4.9 shows the capability of the DL model to distinguish between classes (Søreide, 2009). The class-specific ROC curves for the InceptionResNet-v2 DL model indicate high discriminative power with AUC values of 0.92, 0.95, and 0.94 for class 0, class 1, and class 2, respectively (Figure 4.9). The highest AUC value was observed for class 1, indicating strong discriminative power in this category. These AUC values reflect the model's overall effectiveness in classifying the different classes.

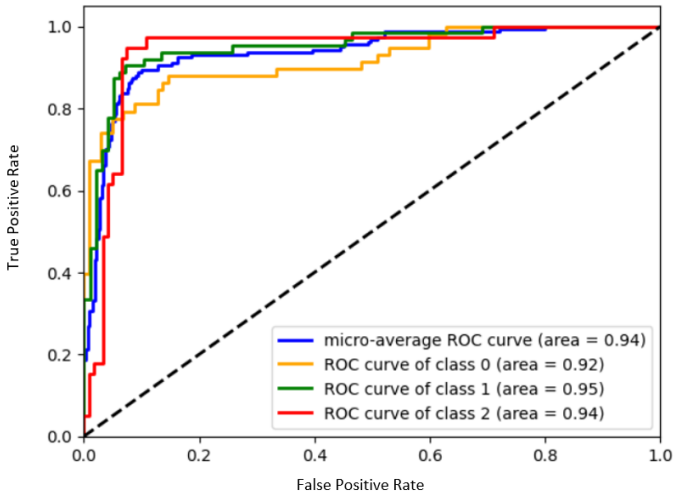


Figure 4.9. The ROC plot of the best accuracy weight.

4.3.4. *Training vs. validation performance*

Figure 4.10 shows the training and validation accuracy over 400 epochs. The validation accuracy, after some initial fluctuations, converges to a value close to the training accuracy. This pattern indicates effective learning by the model, with the validation accuracy closely tracking the training accuracy, suggesting minimal overfitting to the training data.

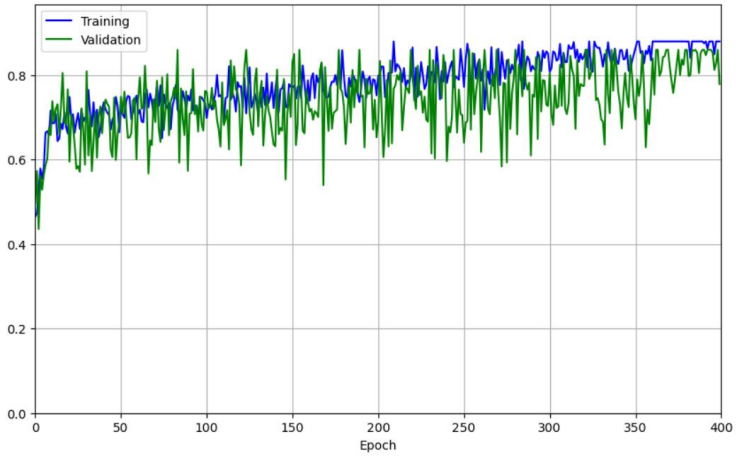


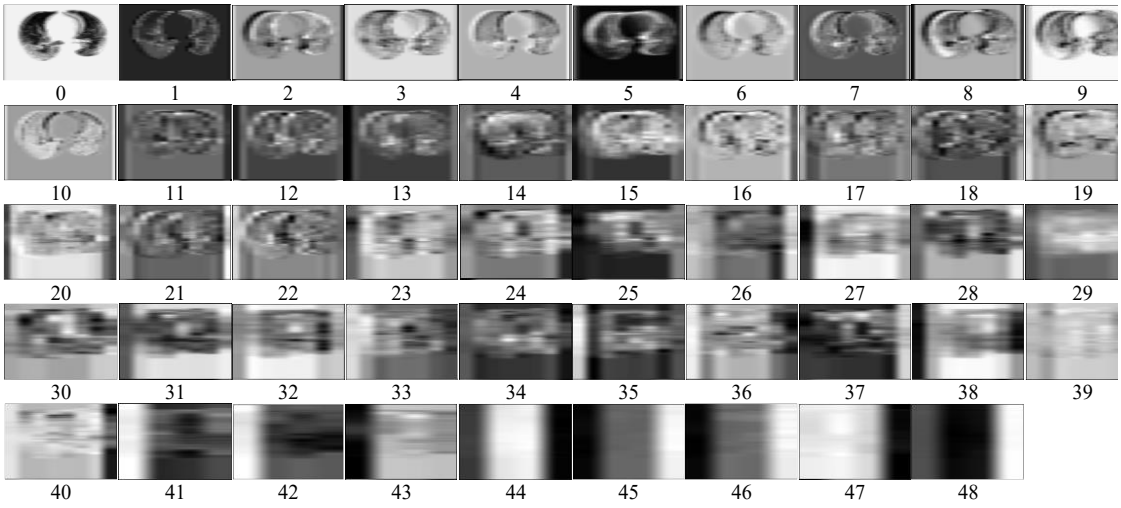
Figure 4.10. Training and validation performance across epochs.

4.4. Explainability assessments

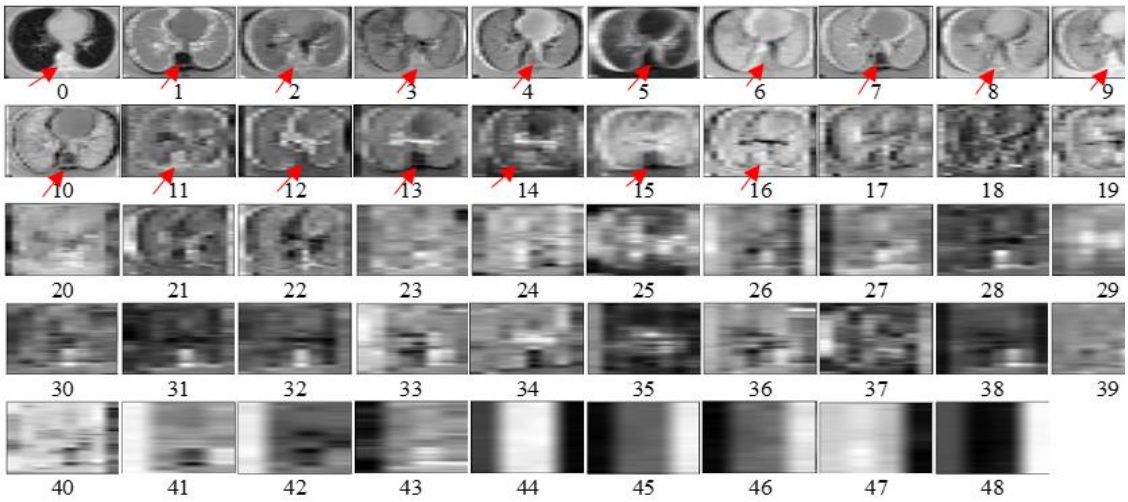
4.4.1. *Feature visualization*

In our study, we utilized feature visualization techniques, a key component of Explainable AI (XAI), to better understand the internal workings of our DL model when classifying lung CT images. Our simulations revealed that models trained on segmented lung CT images achieved

approximately 5 percent higher accuracy compared to those trained on non-segmented images. To further investigate the reasons behind this performance improvement, we employed feature visualization to examine the feature maps generated by the model for both segmented and non-segmented lung images. Figure 4.11(a) and Figure 4.11(b) display the feature map outputs for segmented and non-segmented lungs, respectively. Feature visualization allowed us to visually interpret the areas of the image that the model focuses on when making its predictions. In the case of non-segmented lung images (Figure 3.2(b)), the feature maps showed that the model was inadvertently paying attention to irrelevant regions, such as the ribs, spine, and surrounding tissues. These areas introduced noise and distractions, leading to less accurate classifications. In contrast, the feature maps for segmented lung images (Figure 3.2(a)) demonstrated that the model's focus was more aligned with the actual regions of interest—the lung tissues themselves. By removing the irrelevant areas through segmentation, the model could concentrate on the critical features necessary for accurate classification, leading to improved performance. This demonstrates the power of feature visualization within the XAI framework. By providing a visual representation of what the model "sees," we gain valuable insights into how preprocessing steps like segmentation can significantly influence the model's decision-making process. This not only helps us understand the model's behavior but also guides us in optimizing our data preprocessing techniques to enhance model performance.



(a)



(b)

Figure 4.11. (a) A typical segmented CT slice of feature map output from each convolutional layer of 3D ResNet-50 for a COVID-19 positive case; (b) A typical non-segmented CT slice of feature map output from each convolutional layer of 3D ResNet-50 for a COVID-19 positive case. Red arrows show the spine as an example surrounding area.

4.4.2. *SHAP Representation*

Figure 4.12 illustrates the SHAP value heatmaps overlaid on both PET and CT image patches, revealing the regions of the images that significantly influence the DL model's predictions. SHAP values provide an interpretable measure of the contribution of each pixel or region in the image to the output of the model. In these visualizations, the areas highlighted in red correspond to regions with high positive SHAP values, indicating their critical role in the model's decision-making process for predicting malignancies. These 'hot' regions are where the model detects features strongly associated with the presence of malignant tissue. The blue regions, representing negative or lower SHAP values, suggest areas that either contribute less to the prediction or potentially even detract from a malignant classification.

In both the PET and CT image patches, the model effectively identifies the critical regions that are known to correlate with malignancy. The PET images, characterized by metabolic activity, show higher SHAP values in areas typically associated with increased uptake, which is often indicative of cancerous growths. Similarly, in the CT images, the model focuses on structural anomalies—such as irregular masses or lesions—which are key indicators of malignancy. The visualization not only highlights the areas of the images that are most important for the model's prediction but also aligns with clinical understanding, where these 'hot' regions are indeed the focus of radiologists when identifying potential malignancies. By providing this interpretability,

the SHAP analysis offers reassurance that the model's predictions are based on medically relevant features, thereby enhancing the trustworthiness and clinical applicability of the DL model.

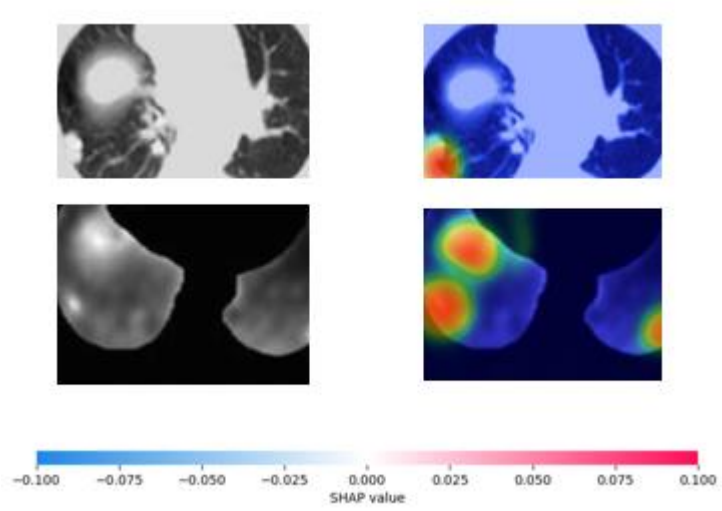


Figure 4.12. SHAP representation for the typical patches of malignant CT and PET. Heatmaps in the right side shows most of the malignant part has been detected by the DL model.

4.5. Discussion on COVID-19 classification and generalizability

The results of AI-based models seem to be more reliable when they use 3D CT images, and they are tested for generalizability (Serte & Demirel, 2021). The reason is that more features can be extracted in whole 3D slices compared to 2D implementations. As many COVID-19 CT images show, not all slices of a patient's image contain involvement. Therefore, considering the slice-base classification of COVID-19 and normal cases may not be as realistic as considering whole CT slices for each patient. This is especially true when the involvement is very small and its detection is possible only when the slice is compared with neighbor slices.

According to previous studies, and in our many experimental trials, lung segmentation improves the results of classification of the COVID-19 and normal cases and should be considered in preprocessing. This is probably due to the fact that it prevents the model from focusing on unwanted targets like bone and soft tissue. Segmentation results are also affected by image type. We used two different segmentation approaches to segment lung from NIFTI and DICOM CT images since there is no single segmentation method that works for all image formats. Also, a patch-based approach both for the compensation of imbalanced classes and to overcome overfitting showed the capacity to be considered for 3D medical datasets which may suffer from

a low number of images and imbalanced classes. Through trial and error in the current study, it has been shown that patch methods improved results by up to 2 percent over non-patches methods.

Based on our model selection simulations, which can be considered as external-validation evaluation, the generalizability of the 3D ResNet-50 along with procedures undertaken in this study indicates that the accuracy when trained with Iranmehr and tested on external datasets, is above 78 percent with the AUC of around 0.90. According to the statistics presented in Tables 5 to 10, Iranmehr dataset produces a generalizable model. This may be due to the precise data categorization in Iranmehr dataset as COVID-19 and normal patients for the training phase. Moreover, a general overview of the results reveals that Iranmehr and LDCT datasets have better results compared to Moscow and 3DLSC. This also may be related to the NIFTI format of these datasets, which seems to affect classification results compared to the DICOM format. Since two out of four external test datasets have a large number of images (1110 images), and experiments were carried out in a 5-fold approach, the test results are reproducible.

The main purpose of this study is to evaluate the effect of different portion combinations of datasets on generalizability. The results of this study confirm that, although the total combination produced the best results with less overfitting (as shown in Figure 4-13), different combinations of datasets provide close results. Moreover, in many studies, especially for the tasks related to medical images, accessibility, preparation, preprocessing may impose difficulties and sometimes be computationally expensive, especially in training on 3D images. With this aim, we divided two available 3D datasets, i.e., Iranmehr and Moscow, into the five 20 percent portions, and we evaluated the different combination results.

Regarding the accuracy, the combination of 80 percent of one dataset and 40 percent or 60 percent of the other reaches to acceptable results close to the results obtained from the total. It seems that when only 20 percent of each dataset is added to the other; the model encounters new features trying to learn them. However, since the number of images in the 20 percent portion is much lower than that of the 80 percent portion, the bias occurs, and the results on holdout test sets have a higher difference compared to other combinations. According to the obtained results, the sensitivity of different combinations succeeded in learning most of the features, and the results are near the total combination when the training set is Iranmehr. However, when the training set is Moscow, while the behavior of different combinations is similar, the results above 80 percent and 40 percent resemble those of the total combination. In terms of specificity, the results demonstrate that for different combinations, the test on the Iranmehr holdout test set and LDCT is more successful than that on the Moscow holdout test set and 3DLSC.

It was observed that specificity is dramatically low. According to our radiologists, and as it can be seen from Figure 4-14, there are some cases in CT0 (normal dataset) of Moscow dataset that are not normal lungs, so they can affect the classification results. We didn't remove any case from

the Moscow dataset to avoid data manipulation. Also, we see that for almost all metrics, Moscow dataset has an adverse effect. This indicates that public datasets still cannot be treated as ideal as real clinical data. Specifically, the results of specificity are much lower when tested on holdout 20 percent of Moscow test set and 3DLSC, even for total combination.

The high accuracy results in each combination or high F1-score shows the capability of the network either for COVID-19 detection or screening of similar disease type and the capability for screening. Another reason to demonstrate this capability is the much higher TP for each combination.

Several studies have combined smaller COVID-19 CT datasets into "supersets" to maximize the number of training samples for DL models. Previous studies have not investigated the effect of combining CT corpuses in this manner. In this study, we proved that, for the CT imaging mode, combining datasets is an effective approach to training DL models for COVID-19 detection. For the datasets investigated, we found a “saddle point” at the 80:40 percent mix of datasets. According to our interpretation, 80 percent of a primary dataset is adequate for fully training a model, and the additional fine-tuning using 40 percent of a secondary dataset helps the model generalize to a third, unseen dataset.

In our second experiment, we used transfer learning as an alternative DL approach for training models. It is clear that when we use the results of the full run for a similar image type, the results are better. It means that pretraining on medical images are more suitable for retraining medical images than using pretrained JPEG images such as ImageNet weights of generic images. Therefore, we used the weights from the full run of 80 percent of each dataset in retraining the last three layers using the other 80 percent dataset. According to table 16, the results show that when the weights come from the more accurate full run (here, full run using 80 percent Iranmehr), the result of retraining is better. However, the result of transfer learning is still lower than that of total combination full run, i.e., the full run of 80 percent of one dataset with the addition of 80 percent of the other dataset. Nevertheless, given to the results presented in Table 13, transfer learning technique allows for fewer data and faster training while providing close AUC and accuracy to those of total combination full run.

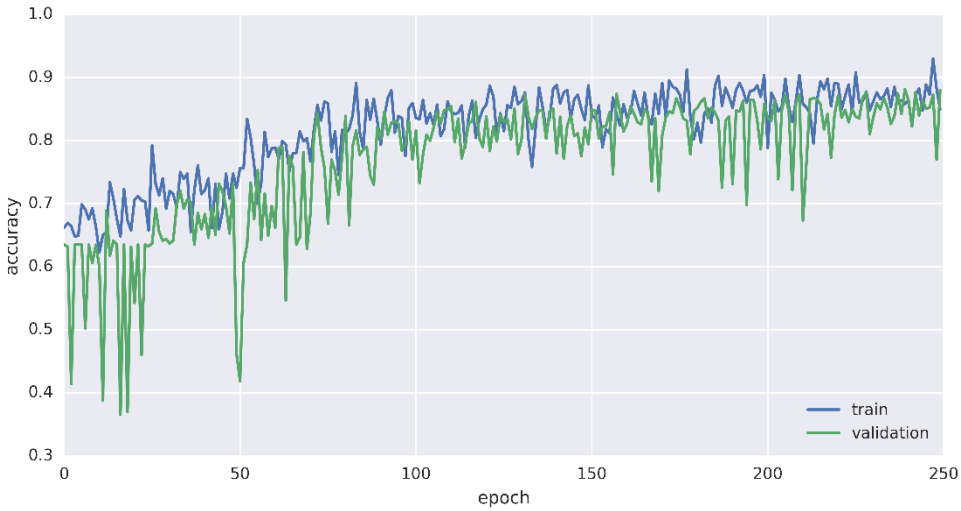
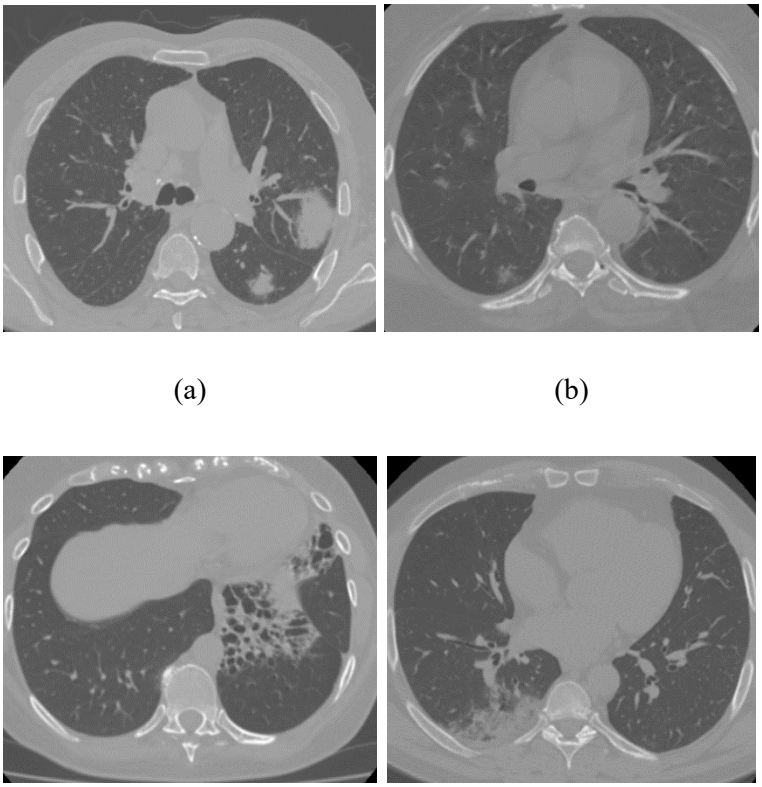


Figure 4.13. Learning curve for training on 80% Iranmehr 80% Moscow data using ResNet-50.



(a) (b)

(c) (d)

Figure 4.14. Some suspected cases involved in Moscow normal dataset folder. (a) case 7 diagnosed with patchy consolidation, which can be pneumonia including COVID-19 or

tumoral lesions, (b) case 34 is more probably a COVID-19 case with small involvement (c) case 68 diagnosed with honeycombing fibrosis at left lung lower lobe, and (d) case 77, diagnosed with wedge consolidation at base of the right lung which may arise due to the pulmonary thromboemboli or segmental pneumonia.

The simulations were run on our own medical data without using the pre-trained weights of models. Given the limited access to medical data, we attempted to achieve good results on our new datasets by cropping, segmentation, patching, etc., to improve training performance. This study would be useful in practice where datasets are limited and may be collected in different locations and settings.

4.6. Discussion on lung nodule classification from PET/CT

The model demonstrates a relatively high sensitivity for benign (88%) and suspicious (87.6%) classes, indicating it is effective at correctly identifying these cases. However, the sensitivity values also reveal that the model has a notable number of false negatives for these classes, meaning that some benign and suspicious cases are misclassified, incompletely capturing true cases. The specific challenge in distinguishing between benign and suspicious lung nodules is evident from these metrics. The PET/CT images for benign and suspicious nodules often present overlapping features, making it difficult for the model to differentiate between them accurately. This overlap can be attributed to the subtle and less distinct features that characterize these classes compared to malignant nodules.

The model performs better in identifying malignant cases, as indicated by the high sensitivity of 90%. This suggests the model is somewhat more accurate in correctly identifying true malignant cases than benign and suspicious ones. The specificity for benign and suspicious nodules is also high, at 93.3% and 90.9% respectively, further reinforcing the model's effectiveness in identifying non-malignant cases as not malignant. The improved performance for malignant nodules can be attributed to the distinct features in PET/CT images, such as the presence of "hot voxels." Hot voxels represent areas of high metabolic activity typically associated with malignant pulmonary nodules. These specific and distinct features make malignant nodules more easily recognizable for the DL model, thus improving sensitivity and specificity. High specificity means the model is less likely to incorrectly label non-diseased (benign or suspicious) cases as malignant, which is critical in reducing unnecessary anxiety and further invasive diagnostic procedures, such as biopsies, which can be risky and costly.

The PPV helps clinicians understand the likelihood that a positive test result (e.g., a nodule classified as malignant) truly represents the disease. A PPV of 90.3% for benign, 86.3% for

malignant, and 87.7% for suspicious, indicates that the test is fairly reliable in confirming the disease. The NPV helps clinicians understand the likelihood that a negative test result indicates the disease's absence. High NPVs, such as 90.9% for benign, 89.5% for malignant, and 91% for suspicious, mean the test is reliable in ruling out the disease. The model is quite effective at identifying suspicious nodules and ruling out non-suspicious cases. The high PPV indicates confidence in positive predictions for suspicious nodules, while the high NPV ensures reliable exclusion of non-suspicious cases. The ROC and AUC results demonstrate the model's strong performance in distinguishing between the different classes. The high AUC values for each class (0.92 for class 0, 0.95 for class 1, and 0.94 for class 2) indicate a high level of discriminative ability, with the micro-average AUC of 0.94 further reinforcing this performance across all classes.

It is noteworthy that the model sometimes struggled to differentiate between benign and suspicious classes. This difficulty is likely due to the similar features these classes share, such as small size and low SUV_{max} values, making them harder to distinguish. On the other hand, the SHAP representation for malignant patches highlights the model's ability to successfully identify hot regions associated with malignancies. The red areas in the heatmap indicate high SHAP values, showing the critical regions the model focuses on for its predictions. This interpretability aspect confirms that the model effectively detects malignant areas due to the presence of hot voxels, which are more distinct and easier for the DL model to recognize compared to benign and suspicious classes.

Using a large dataset of PET/CT scans in this study has led to more reliable results. Specifically, we employed 1304 ^{18}F FDG PET/CT scans using a 3D approach. Our methodology included a multi-classification task, reflecting the real-world practices of physicians. Notably, the model exhibited similar challenges to those faced by medical professionals, particularly in distinguishing between benign and suspicious cases. This mirrors the reality where some physicians may diagnose non-malignant nodules as benign or due to inflammation, while others might recommend follow-up.

One notable advantage of our study lies in the use of varying patch slice thicknesses, a technique that significantly enhances the robustness of our model. By employing different thicknesses, we can capture a wider range of information from the images, thereby reducing the necessity for modifications that might lead to information loss. This method also offers a practical solution for managing computational costs, as it allows us to process the data more efficiently. Furthermore, it plays a crucial role in preventing overfitting during the training phase, ensuring that our model generalizes well to new, unseen data. In addition to the advantages provided by varying patch slice thicknesses, we also addressed the challenge of data imbalance through a patch-based data generation approach. This strategy involves generating additional data patches to ensure a more

balanced representation of different classes within the dataset. By doing so, we further enhance the robustness and reliability of our model, making it better equipped to handle real-world scenarios where data may be inherently imbalanced. This comprehensive approach not only strengthens the model performance but also underscores the effectiveness of our methodology in overcoming common challenges in image processing and machine learning.

The multi-class simulation conducted in this study, which involves classifying cases into benign, malignant, and suspicious categories, presents a challenging yet highly realistic scenario. This approach mirrors the complexity of actual clinical environments where medical professionals must often differentiate between multiple categories of conditions, each with its own set of characteristics and implications.

By incorporating these three distinct classifications, our model is tested against a broader spectrum of possible diagnoses, thereby demonstrating its capability to manage complex and nuanced medical cases. This multi-faceted classification task is inherently more difficult than binary classification (such as simply distinguishing between benign and malignant cases) because it requires the model to discern more subtle differences between classes and make more refined decisions. The success of our model in this challenging multi-class simulation underscores its robustness and versatility. It highlights the model's potential to assist medical professionals by providing a reliable tool that can support more accurate and nuanced diagnoses. This is particularly important in medical practice, where the ability to accurately classify cases as benign, malignant, or suspicious can significantly impact treatment decisions and patient outcomes. Moreover, the model's performance in handling these complex scenarios suggests its potential for integration into clinical workflows, where it can serve as a valuable decision-support system. By helping to identify and classify suspicious cases that may require further investigation, the model can aid in early detection and timely intervention, ultimately contributing to improved patient care.

4.7. Summary

The results chapter presents a detailed evaluation of DL models applied to the classification of COVID-19 lung involvement and lung nodule classification using ^{18}FFDG PET/CT images. These results emphasize the models' generalizability across multiple datasets, the effectiveness of dataset combinations, and the success of transfer learning strategies.

The classification of COVID-19 lung involvement was performed using ResNet-50, chosen due to its superior performance in preliminary tests and relatively lower runtime. The model's results were presented using key metrics such as accuracy, sensitivity, specificity, F1-score, and AUC, with a particular focus on generalizability. ResNet-50 exhibited strong performance across multiple datasets, including Iranmehr, Moscow, LDCT, and 3DLSC, though the results varied depending on the specific dataset and preprocessing techniques applied. Cropped images

generally yielded better results compared to non-cropped images. When comparing datasets, the ResNet-50 outperformed DenseNet-169 in terms of accuracy and standard deviation on the Moscow dataset.

One of the most significant findings in this part was the generalizability of the models when training on one dataset and testing on others. Notably, the combination of 80% Iranmehr and 40% Moscow datasets showed strong results. This combination delivered results close to those obtained from training with the full dataset mix, indicating that such proportional combinations can yield robust models without requiring the full dataset. Another key result came from the 80% Moscow + 40% Iranmehr combination, further demonstrating the potential of dataset combinations to generalize effectively.

However, challenges remained, particularly with specificity. Specificity was inconsistent, particularly when tested on the Moscow dataset and the 3DLSC external dataset, where some data ambiguities affected classification performance. This issue of low specificity was partly attributed to mislabeled or abnormal cases within the Moscow dataset. For instance, cases marked as "normal" in the Moscow dataset were found to include pathologies that negatively impacted classification. Despite these challenges, the ResNet-50 model demonstrated overall strong performance, especially in terms of accuracy and AUC, making it a reliable tool for COVID-19 lung involvement classification across different datasets.

Transfer learning experiments provided additional insights into the model's ability to generalize across datasets. When retraining the final layers of ResNet-50 using weights trained on Iranmehr and testing on Moscow, the model achieved better results in accuracy, sensitivity, and F1-score compared to the reverse scenario.

Using feature visualization CT images as XAI method, the model successfully identified irregular masses and structural anomalies that aligned with known indicators of malignancy. This interpretability not only reinforced the model's accuracy but also aligned with clinical understanding of how these medical images are analyzed.

The second major focus of the chapter is on the classification of lung nodules into benign, malignant, and suspicious categories using ^{18}F FDG PET/CT images. InceptionResNet-v2 was highlighted as the most effective model for this task, achieving the highest accuracy of 89%, outperforming ResNet-50 and other architectures such as ResNeXt and SE-ResNet. The model demonstrated strong discriminative power across all three nodule categories, as evidenced by the high AUC values for each class. For instance, the model achieved AUCs of 0.92 for benign, 0.95 for malignant, and 0.94 for suspicious nodules, showcasing its ability to accurately distinguish between these categories.

The use of feature visualization and SHAP representation further enhanced the interpretability of the model's decisions. These techniques revealed the critical regions in the images that the model

focused on during classification, such as “hot voxels” in PET images, which correspond to regions of high metabolic activity typically associated with malignancy. Despite the overall high performance, the model faced challenges in distinguishing between benign and suspicious nodules. The overlap in characteristics between these classes, particularly in cases where the nodules were small or exhibited low SUV_{max} values, made classification difficult. Nonetheless, the model performed well in identifying malignant nodules, with a sensitivity of 90% and a specificity of 85.7%.

In terms of generalizability, the lung nodule classification model showed significant robustness across different FDG PET/CT images, benefiting from the use of a large dataset consisting of 1304 scans. The model’s success in handling multi-class classification tasks, which reflect real-world clinical practices, underscores its potential to support diagnostic decision-making. Moreover, the multi-class simulation approach, which involved differentiating between benign, malignant, and suspicious nodules, presented a more complex and realistic scenario than binary classification. The model’s ability to perform well in this complex task highlights its robustness and versatility.

The chapter concludes with a discussion on the broader implications of these findings for medical image classification. The combination of datasets proved to be an effective strategy for improving model performance and generalizability, particularly with the 80% + 40% dataset combinations. Furthermore, the application of transfer learning showed that models trained on medical images are better suited for fine-tuning on similar medical datasets, rather than using generic pretrained models such as those based on ImageNet. Despite the challenges in specificity, the results highlight the potential for DL models to generalize effectively across different datasets and imaging modalities, providing valuable tools for diagnostic support in medical practice.

CHAPTER 5

5. CONCLUSION

5.1. General Conclusion

This study represents a comprehensive exploration into the application of optimized DL models and DL techniques for the detection and classification of pulmonary involvements, including COVID-19 from CT, and lung nodules from ^{18}F FDG PET/CT imaging. Through a meticulous process of dataset combination, model optimization, and pre-processing enhancement, we have demonstrated that it is possible to achieve a high degree of accuracy and generalizability in DL models, even in scenarios where large datasets are not available.

One of the key findings of this research is the effectiveness of combining 80% of one dataset with a 40% or greater contribution from another, which yields results comparable to using the entire combined datasets. This is particularly relevant for clinical applications where the availability of large, well-curated datasets is often limited. The study also emphasizes the importance of optimizing pre-processing steps—such as lung segmentation, PET/CT resampling, fusion, and the exclusion of zero-valued voxels—which significantly contribute to the efficiency and performance of DL models.

A critical aspect of this study is the focus on the classification of different types of pulmonary nodules, including benign, malignant, and suspicious cases. This research highlights the challenges in distinguishing between these categories due to the overlapping features often present in PET/CT images. The ability of the DL models to accurately identify malignant nodules, which typically exhibit distinct metabolic activity, is particularly noteworthy. However, the study also underscores the inherent difficulties in accurately classifying benign and suspicious nodules, which often lack the pronounced metabolic features seen in malignancies. This finding points to the need for ongoing refinement of AI models to improve their diagnostic accuracy across a broader spectrum of pulmonary conditions.

Furthermore, the research underscores the importance of assessing different types of malignancies that can lead to lung involvement. The accurate detection and classification of various malignancies are crucial for ensuring timely and appropriate treatment decisions, which can significantly impact patient outcomes. The study’s approach, which includes varying patch slice thicknesses to enhance the model's robustness and minimize information loss, has proven effective in capturing a wider range of image details, thus improving the overall performance and reliability of the models.

5.2. Specific conclusion aligned with objectives

- Detection and classification of pulmonary involvement including COVID-19: The study successfully achieved the objective of detecting and classifying pulmonary involvement, including COVID-19, using optimized AI models. The integration of 80% of one dataset with 40% or more from another dataset proved to be an effective strategy, particularly in environments where large datasets are not accessible.
- Detection and categorization of pulmonary nodules using ^{18}F FDG PET/CT Data: The research met the objective of detecting and categorizing pulmonary nodules from a large ^{18}F FDG PET/CT dataset. By assessing various state-of-the-art DL models, we determined that the combination of different patch slice thicknesses and well-curated datasets improved the models' robustness and classification accuracy. The study found that while the models exhibited high sensitivity and specificity for malignant nodules, the classification of benign and suspicious nodules remained challenging due to overlapping imaging features. Moreover, the research highlights the importance of using clinical datasets, as they provide more reliable training data compared to some publicly available datasets, which may suffer from biases or incorrect diagnoses.
- Development of an optimized DL model for pulmonary nodule detection: In alignment with the objective to develop an optimized DL model, we introduced additional 3D convolutional layers and fine-tuned the model architecture and hyperparameters to enhance accuracy in pulmonary nodule detection and classification. The exclusion of zero-valued voxels during pre-processing was identified as a critical step that not only improved model performance but also addressed memory allocation issues, making the training process more robust and resource-efficient. The final model demonstrated strong performance in detecting malignant nodules, reinforcing its potential value in clinical applications, particularly for the accurate and timely diagnosis of various types of lung malignancies.

This research provides valuable insights into the optimization of AI models for pulmonary imaging, with significant impact on the diagnostic accuracy improvement and patient outcomes across a range of pulmonary conditions.

5.3. Limitations

This study faced several notable limitations, which highlight the challenges inherent in developing and DL models for medical imaging, particularly in the context of pulmonary nodule detection and classification using ^{18}F FDG PET/CT scans.

One of the primary limitations was the scarcity of large 3D datasets in clinical formats such as DICOM or NIFTI. These formats are essential for ensuring that the data used in research closely reflects real-world clinical images. Unfortunately, other 3D image formats, like JPEG, suffer from lower resolution and are not typically considered suitable for clinical applications. As a result, this study was constrained to using only two large datasets available in DICOM/NIFTI formats for training and two smaller datasets for testing. This limitation underscores the need for more extensive 3D CT datasets in clinical formats to be developed and made accessible for future research, especially if these models are to be integrated into clinical workflows. Additionally, a significant challenge encountered was the necessity to employ different lung segmentation approaches for the two image formats used in this study. The lack of a universal, accurate lung segmentation method that works consistently across various image formats added complexity to the pre-processing phase. This variation necessitated different pre-processing pipelines for each image format, which could potentially impact the consistency and reproducibility of the results. Future research should focus on developing a standardized segmentation approach that can be applied universally across different image formats to streamline the pre-processing workflow. Moreover, the study was subject to substantial computational and memory demands due to the large size of the PET and CT images and the requirement to process both types simultaneously. Even with the use of advanced GPUs, the computational intensity significantly extended the simulation times, with each run taking approximately five days to complete. This prolonged duration illustrates the trade-off between obtaining high-quality results and maintaining computational feasibility. The high resource demands also limited the number of simulations that could be feasibly conducted within the scope of this study. A further limitation was the unavailability of publicly annotated FDG PET/CT datasets, which restricted our ability to comprehensively test and validate the model across different data sources. The lack of external datasets hindered our ability to assess the model's robustness and generalizability in diverse scenarios, which is critical for ensuring its reliability in real-world clinical applications. Testing the model on various external FDG PET/CT datasets would have provided a more thorough evaluation of its performance and improved its potential applicability across different clinical environments. Additionally, it is important to note that some DL models could not be successfully implemented with our network and dataset due to technical limitations. Since resolving these technical issues was beyond the scope of this study, we opted to focus only on those models that were compatible and could be effectively trained and tested within the constraints of our computational resources. This decision may have limited the exploration of potentially more effective models, highlighting an area for future research to address these technical challenges and broaden the scope of model selection.

5.4. Future Recommendations

Building on the insights and limitations identified in this study, several key areas for future research and development are recommended to further advance the application of DL models in pulmonary imaging and enhance their clinical utility.There is a critical need for the creation and sharing of larger, high-quality 3D datasets in clinically relevant formats such as DICOM and NIFTI. Expanding the availability of these datasets will allow for more comprehensive training and validation of DL models, ensuring they can generalize effectively to a wider range of clinical scenarios. Collaborations between healthcare institutions, research centers, and data-sharing initiatives could play a pivotal role in overcoming the current data scarcity and improving model robustness.

To streamline the pre-processing workflow and ensure consistency across different studies, future research should focus on developing a lung segmentation algorithm that can accurately process images across various formats. This approach would reduce the complexity of pre-processing steps, making it easier to apply DL models to different types of imaging data and improving the reproducibility of results across different studies.

Given the substantial computational demands encountered in this study, future efforts should aim at optimizing the efficiency of DL models without compromising accuracy. This could involve the development of more efficient algorithms, the use of distributed computing, or the exploration of hardware accelerations such as tensor processing units (TPUs). By reducing the computational load, it would be possible to conduct more extensive simulations and improve the scalability of DL applications in clinical settings.

The absence of publicly annotated PET/CT datasets is a significant barrier to the thorough evaluation and validation of DL models. Future initiatives should prioritize the creation of such datasets, complete with detailed annotations and metadata, to facilitate the development of more robust and generalizable models. These datasets would allow for more rigorous testing across different institutions and patient populations, ultimately leading to AI tools that are more reliable in diverse clinical environments.

Future research should also focus on overcoming the technical challenges that prevented some DL models from being successfully implemented in this study. This could involve refining model architectures to be more compatible with the specific characteristics of medical imaging data or developing new methods to handle the unique demands of 3D image processing. By broadening the range of DL models that can be effectively applied, the potential for finding even more accurate and efficient models will be increased.

As medical imaging continues to evolve, there is an increasing potential for integrating data from multiple imaging modalities (e.g., PET, CT, MRI) into a single DL framework. Future research should explore how multimodal data can be effectively combined to enhance the accuracy and

reliability of DL models in detecting and classifying pulmonary conditions. This approach could provide a more comprehensive view of patient health and improve diagnostic precision.

To ensure the generalizability of DL models, it is essential that future research includes validation across diverse patient populations and clinical environments. This would involve testing models on datasets from different geographical regions, healthcare systems, and patient demographics to ensure that the models perform consistently well across varied contexts. Such validation is crucial for building trust in AI-driven diagnostic tools and ensuring their widespread adoption in clinical practice.

In this study, we utilized heatmap visualization techniques and SHAP to interpret the model's decisions. While these methods provided valuable insights, it is important to acknowledge that each XAI method carries inherent uncertainties and potential limitations that could impact the reliability and interpretability of the explanations. For instance, heatmap visualizations, such as those generated by Grad-CAM, may sometimes produce ambiguous or noisy outputs, while SHAP, though theoretically robust, can be computationally intensive and may not scale well with complex models or large datasets. Given these challenges, one key recommendation for future research is to perform a comprehensive comparison of all available XAI methods within the context of medical imaging. This comparative analysis should focus on evaluating the consistency and reliability of the explanations produced by different XAI techniques across a variety of model architectures and medical imaging modalities. By systematically assessing the degree of variation and potential biases introduced by each method, researchers can better understand the strengths and limitations of these approaches and make more informed choices when selecting XAI methods for clinical applications. Moreover, such a comparative study could help identify scenarios where certain XAI methods may outperform others, thereby guiding the development of more robust and interpretable AI models in medical imaging. This approach would contribute to the broader goal of ensuring that AI-driven insights in healthcare are both transparent and trustworthy, ultimately enhancing the interpretability and clinical utility of these technologies.

Ethical standard

This study was carried out under the ethics approval from the University of Technology Sydney, Australia, under "UTS HREC REF NO. ETH21-6536".

6. REFERENCES

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., & Isard, M. (2016). {TensorFlow}: a system for {Large-Scale} machine learning. 12th USENIX symposium on operating systems design and implementation (OSDI 16),

Abbas, A., Abdelsamea, M. M., & Gaber, M. M. (2020). Detrac: Transfer learning of class decomposed medical images in convolutional neural networks. *IEEE Access*, *8*, 74901-74913.

Abedin, B. (2022). Managing the tension between opposing effects of explainability of artificial intelligence: a contingency theory perspective. *Internet Research*, *32*(2), 425-453.

Afshar, P., Mohammadi, A., Tyrrell, P. N., Cheung, P., Sigiuk, A., Plataniotis, K. N., Nguyen, E. T., & Oikonomou, A. (2020). DRTOP: deep learning-based radiomics for the time-to-event outcome prediction in lung cancer. *Scientific reports*, *10*(1), 12366.

Ahmed, K. B., Goldgof, G. M., Paul, R., Goldgof, D. B., & Hall, L. O. (2021). Discovery of a generalization gap of convolutional neural networks on COVID-19 X-rays classification. *Ieee Access*, *9*, 72970-72979.

Almuhaideb, A., Papathanasiou, N., & Bomanji, J. (2011). 18F-FDG PET/CT imaging in oncology. *Annals of Saudi medicine*, *31*(1), 3-13.

Alom, M. Z., Taha, T. M., Yakopcic, C., Westberg, S., Sidike, P., Nasrin, M. S., Van Esesn, B. C., Awwal, A. A. S., & Asari, V. K. (2018). The history began from alexnet: A comprehensive survey on deep learning approaches. *arXiv preprint arXiv:1803.01164*.

Alves, V. M., dos Santos Cardoso, J., & Gama, J. (2024). Classification of Pulmonary Nodules in 2-[18F] FDG PET/CT Images with a 3D Convolutional Neural Network. *Nuclear Medicine and Molecular Imaging*, *58*(1), 9-24.

Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M. A., Al-Amidie, M., & Farhan, L. (2021). Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *Journal of Big data*, *8*, 1-74.

Ambrosini, V., & Fanti, S. (2011). Clinical applications of PET/CT in oncology: Neuroendocrine tumour. In *Principlesand Practice ofPET/CT Part 2 A Technologist's Guide* (pp. 102-107). European Association of Nuclear Medicine.

Ambrosini, V., Nicolini, S., Caroli, P., Nanni, C., Massaro, A., Marzola, M. C., Rubello, D., & Fanti, S. (2012). PET/CT imaging in different types of lung cancer: an overview. *European journal of radiology*, *81*(5), 988-1001.

Apostolopoulos, I. D., Pintelas, E. G., Livieris, I. E., Apostolopoulos, D. J., Papathanasiou, N. D., Pintelas, P. E., & Panayiotakis, G. S. (2021). Automatic classification of solitary pulmonary nodules in PET/CT imaging employing transfer learning techniques. *Medical & Biological Engineering & Computing*, *59*(6), 1299-1310.

Ardakani, A. A., Kwee, R. M., Mirza-Aghazadeh-Attari, M., Castro, H. M., Kuzan, T. Y., Altintoprak, K. M., Besutti, G., Monelli, F., Faeghi, F., & Acharya, U. R. (2021). A practical artificial intelligence system to diagnose COVID-19 using computed tomography: A multinational external validation study. *Pattern recognition letters*, *152*, 42-49.

Artesani, A., Bruno, A., Gelardi, F., & Chiti, A. (2024). Empowering PET: harnessing deep learning for improved clinical insight. *European Radiology Experimental*, *8*(1), 17.

Aversano, L., Bernardi, M. L., Cimitile, M., & Pecori, R. (2021). Deep neural networks ensemble to detect COVID-19 from CT scans. *Pattern recognition*, *120*, 108135.

[Record #154 is using a reference type undefined in this output style.]

Barta, J. A., Powell, C. A., & Wisnivesky, J. P. (2019). Global epidemiology of lung cancer. *Annals of global health*, *85*(1).

Bassi, P. R., & Attux, R. (2022). COVID-19 detection using chest X-rays: Is lung segmentation important for generalization? *Research on Biomedical Engineering*, *38*(4), 1121-1139.

Beyer, T., Townsend, D. W., Brun, T., Kinahan, P. E., Charron, M., Roddy, R., Jerin, J., Young, J., Byars, L., & Nutt, R. (2000). A combined PET/CT scanner for clinical oncology. *Journal of Nuclear Medicine*, *41*(8), 1369-1379.

Bhatele, K. R., Jha, A., Tiwari, D., Bhatele, M., Sharma, S., Mithora, M. R., & Singhal, S. (2024). Covid-19 detection: A systematic review of machine and deep learning-based approaches utilizing chest x-rays and ct scans. *Cognitive Computation*, *16*(4), 1889-1926.

Bhuyan, H. K., Chakraborty, C., Shelke, Y., & Pani, S. K. (2022). COVID-19 diagnosis system by deep learning approaches. *Expert Systems*, *39*(3), e12776.

Bianconi, F., Fravolini, M. L., Pizzoli, S., Palumbo, I., Minestrini, M., Rondini, M., Nuvoli, S., Spanu, A., & Palumbo, B. (2021). Comparative evaluation of conventional and deep learning methods for semi-automated segmentation of pulmonary nodules on CT. *Quantitative imaging in medicine and surgery*, *11*(7), 3286.

Bjerring, J. C., & Busch, J. (2021). Artificial intelligence and patient-centered decision-making. *Philosophy & technology*, *34*, 349-371.

Borm, K. J., Voppichler, J., Düsberg, M., Oechsner, M., Vag, T., Weber, W., Combs, S. E., & Duma, M. N. (2019). FDG/PET-CT–based lymph node atlas in breast cancer patients. *International Journal of Radiation Oncology* Biology* Physics*, *103*(3), 574-582.

[Record #117 is using a reference type undefined in this output style.]

Budak, E., Çok, G., & Akgün, A. (2018). The contribution of fluorine 18F-FDG PET/CT to lung cancer diagnosis, staging and treatment planning. *Molecular imaging and radionuclide therapy*, *27*(2), 73.

Chan, W. K., Mak, H. K., Huang, B., Yeung, D. W., Kwong, D. L.-W., & Khong, P.-L. (2010). Nasopharyngeal carcinoma: relationship between 18F-FDG PET-CT maximum standardized uptake value, metabolic tumour volume and total lesion glycolysis and TNM classification. *Nuclear medicine communications*, *31*(3), 206-210.

Chen, H., Zhang, Y., Kalra, M. K., Lin, F., Chen, Y., Liao, P., Zhou, J., & Wang, G. (2017). Low-dose CT with a residual encoder-decoder convolutional neural network. *IEEE Transactions on Medical Imaging*, *36*(12), 2524-2535.

Cheng, J.-Z., Ni, D., Chou, Y.-H., Qin, J., Tiu, C.-M., Chang, Y.-C., Huang, C.-S., Shen, D., & Chen, C.-M. (2016). Computer-aided diagnosis with deep learning architecture: applications to breast lesions in US images and pulmonary nodules in CT scans. *Scientific reports*, *6*(1), 24454.

Chin, A. Y., Lee, K. S., Kim, B.-T., Choi, J. Y., Kwon, O. J., Kim, H., Shim, Y. M., & Chung, M. J. (2006). Tissue characterization of solitary pulmonary nodule: comparative study between helical dynamic CT and integrated PET/CT. *Journal of Nuclear Medicine*, *47*(3), 443-450.

Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. Proceedings of the IEEE conference on computer vision and pattern recognition,

Cicero, M., Bilbily, A., Colak, E., Dowdell, T., Gray, B., Perampaladas, K., & Barfett, J. (2017). Training and validating a deep convolutional neural network for computer-aided detection and classification of abnormalities on frontal chest radiographs. *Investigative radiology*, *52*(5), 281-287.

Cohen, J. P., Morrison, P., Dao, L., Roth, K., Duong, T. Q., & Ghassemi, M. (2020). Covid-19 image data collection: Prospective predictions are the future. *arXiv preprint arXiv:2006.11988*.

Cronin, C. G., Swords, R., Truong, M. T., Viswanathan, C., Rohren, E., Giles, F. J., O'Dwyer, M., & Bruzzi, J. F. (2010). Clinical utility of PET/CT in lymphoma. *American Journal of Roentgenology*, *194*(1), W91-W103.

Danks, D., & London, A. J. (2017). Algorithmic Bias in Autonomous Systems. Ijcai,

Dayarathna, S., Islam, K. T., Uribe, S., Yang, G., Hayat, M., & Chen, Z. (2023). Deep learning based synthesis of MRI, CT and PET: Review and analysis. *Medical image analysis*, 103046.

Dewan, N. A., Gupta, N. C., Redepenning, L. S., Phalen, J. J., & Frick, M. P. (1993). Diagnostic efficacy of PET-FDG imaging in solitary pulmonary nodules: potential role in evaluation and management. *Chest*, *104*(4), 997-1002.

Di Carli, M. F., Geva, T., & Davidoff, R. (2016). The future of cardiovascular imaging. *Circulation*, *133*(25), 2640-2661.

Divisi, D., Barone, M., Bertolaccini, L., Rocco, G., Solli, P., Crisci, R., & Group, I. V. (2017). Standardized uptake value and radiological density attenuation as predictive and prognostic factors in patients with solitary pulmonary nodules: our experience on 1,592 patients. *Journal of Thoracic Disease*, *9*(8), 2551.

Domingues, I., Pereira, G., Martins, P., Duarte, H., Santos, J., & Abreu, P. H. (2020). Using deep learning techniques in medical imaging: a systematic review of applications on CT and PET. *Artificial Intelligence Review*, *53*, 4093-4160.

Dosovitskiy, A. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Dunkl, V., Cleff, C., Stoffels, G., Judov, N., Sarikaya-Seiwert, S., Law, I., Bøgeskov, L., Nysom, K., Andersen, S. B., & Steiger, H.-J. (2015). The usefulness of dynamic O-(2-18F-fluoroethyl)-L-tyrosine PET in the clinical evaluation of brain tumors in children and adolescents. *Journal of Nuclear Medicine*, *56*(1), 88-92.

Dutta, P., Upadhyay, P., De, M., & Khalkar, R. (2020). Medical image analysis using deep convolutional neural networks: CNN architectures and transfer learning. 2020 International Conference on Inventive Computation Technologies (ICICT),

Endo, K., Oriuchi, N., Higuchi, T., Iida, Y., Hanaoka, H., Miyakubo, M., Ishikita, T., & Koyama, K. (2006). PET and PET/CT using 18 F-FDG in the diagnosis and management of cancer patients. *International journal of clinical oncology*, *11*, 286-296.

Ettinger, D. S., Wood, D. E., Aisner, D. L., Akerley, W., Bauman, J., Chirieac, L. R., D'Amico, T. A., DeCamp, M. M., Dilling, T. J., & Dobelbower, M. (2017). Non–small cell lung cancer, version 5.2017, NCCN clinical practice guidelines in oncology. *Journal of the National Comprehensive Cancer Network*, *15*(4), 504-535.

Eubank, W. B., & Mankoff, D. A. (2005). Evolving role of positron emission tomography in breast cancer imaging. Seminars in nuclear medicine,

Eyassu, E., & Young, M. (2021). Nuclear medicine PET/CT head and neck cancer assessment, protocols, and interpretation.

Farwell, M. D., Pryma, D. A., & Mankoff, D. A. (2014). PET/CT imaging in cancer: current applications and future directions. *Cancer*, *120*(22), 3433-3445.

Ferlito, A., Shaha, A. R., Silver, C. E., Rinaldo, A., & Mondin, V. (2001). Incidence and sites of distant metastases from head and neck cancer. *ORL*, *63*(4), 202-207.

Fernandes, L., Pereira, T., & Oliveira, H. P. (2024). Exploring the differences between Multi-task and Single-task with the use of Explainable AI for lung nodule classification. 2024 IEEE 37th International Symposium on Computer-Based Medical Systems (CBMS),

Fidler, I. J. (1989). Origin and biology of cancer metastasis. *Cytometry: The Journal of the International Society for Analytical Cytology*, *10*(6), 673-680.

Firat, H., Asker, M. E., Bayindir, M. I., & Hanbay, D. (2023). 3D residual spatial–spectral convolution network for hyperspectral remote sensing image classification. *Neural Computing and Applications*, *35*(6), 4479-4497.

Fletcher, J., & Kinahan, P. (2010). PET/CT standardized uptake values (SUVs) in clinical practice and assessing response to therapy. *NIH Public Access*, *31*(6), 496-505.

Floridi, L., & Cowls, J. (2022). A unified framework of five principles for AI in society. *Machine learning and the city: Applications in architecture and urban design*, 535-545.

Fusco, R., Grassi, R., Granata, V., Setola, S. V., Grassi, F., Cozzi, D., Pecori, B., Izzo, F., & Petrillo, A. (2021). Artificial intelligence and COVID-19 using chest CT scan and chest X-ray images: machine learning and deep learning approaches for diagnosis and treatment. *Journal of Personalized Medicine*, *11*(10), 993.

Garcia-Velloso, M. J., Bastarrika, G., de-Torres, J. P., Lozano, M. D., Sanchez-Salcedo, P., Sancho, L., Nuñez-Cordoba, J. M., Campo, A., Alcaide, A. B., & Torre, W. (2016). Assessment of indeterminate pulmonary nodules detected in lung cancer screening: Diagnostic accuracy of FDG PET/CT. *Lung Cancer*, *97*, 81-86.

Garcia, D., & Singh, V. (2023). Nuclear medicine PET/CT thyroid cancer assessment, protocols, and interpretation. In *StatPearls [Internet]*. StatPearls Publishing.

Ghafoorian, M., Karssemeijer, N., Heskes, T., van Uden, I. W., Sanchez, C. I., Litjens, G., de Leeuw, F.-E., van Ginneken, B., Marchiori, E., & Platel, B. (2017). Location sensitive deep convolutional neural networks for segmentation of white matter hyperintensities. *Scientific reports*, *7*(1), 5110.

Ghomi, Z., Mirshahi, R., Bagheri, A. K., Fattahpour, A., Mohammadiun, S., Gharahbagh, A. A., Djavadifar, A., Arabalibeik, H., Sadiq, R., & Hewage, K. (2020). Segmentation of COVID-19 pneumonia lesions: A deep learning approach. *Medical Journal of the Islamic Republic of Iran*, *34*, 174.

Ghouri, M. A., Gupta, N., Bhat, A. P., Thimmappa, N. D., Saboo, S. S., Khandelwal, A., & Nagpal, P. (2019). CT and MR imaging of the upper extremity vasculature: pearls, pitfalls, and challenges. *Cardiovascular Diagnosis and Therapy*, *9*(Suppl 1), S152.

Glaudemans, A. W., de Vries, E. F., Galli, F., Dierckx, R. A., Slart, R. H., & Signore, A. (2013). The Use of 18F-FDG-PET/CT for diagnosis and treatment monitoring of inflammatory and infectious diseases. *Journal of Immunology Research*, *2013*(1), 623036.

Gong, K., Berg, E., Cherry, S. R., & Qi, J. (2019). Machine learning in PET: from photon detection to quantitative image reconstruction. *Proceedings of the IEEE*, *108*(1), 51-68.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep feedforward networks. *Deep learning*(1).

Griffeth, L. K. (2005). Use of PET/CT scanning in cancer patients: technical and practical considerations. Baylor University Medical Center Proceedings,

Grisanti, F., Zulueta, J., Rosales, J. J., Morales, M. I., Sancho, L., Lozano, M. D., Mesa-Guzman, M., & Garcia-Velloso, M. J. (2021). Diagnostic accuracy of visual analysis versus dual time-point imaging with 18F-FDG PET/CT for the characterization of indeterminate pulmonary nodules with low uptake. *Revista Española de Medicina Nuclear e Imagen Molecular (English Edition)*, *40*(3), 155-160.

Groheux, D., Quere, G., Blanc, E., Lemarignier, C., Vercellino, L., de Margerie-Mellon, C., Merlet, P., & Querellou, S. (2016). FDG PET-CT for solitary pulmonary nodule and lung cancer: literature review. *Diagnostic and Interventional Imaging*, *97*(10), 1003-1017.

Gudigar, A., Raghavendra, U., Nayak, S., Ooi, C. P., Chan, W. Y., Gangavarapu, M. R., Dharmik, C., Samanth, J., Kadri, N. A., & Hasikin, K. (2021). Role of artificial intelligence in COVID-19 detection. *Sensors*, *21*(23), 8045.

Gunning, D., & Aha, D. (2019). DARPA's explainable artificial intelligence (XAI) program. *AI magazine*, *40*(2), 44-58.

Gunraj, H., Wang, L., & Wong, A. (2020). Covidnet-ct: A tailored deep convolutional neural network design for detection of covid-19 cases from chest ct images. *Frontiers in medicine*, *7*, 608525.

Hadique, S., Jain, P., Hadi, Y., Baig, A., & Parker, J. E. (2020). Utility of FDG PET/CT for assessment of lung nodules identified during low dose computed tomography screening. *BMC Medical Imaging*, *20*, 1-6.

Hammer, M. M., & Byrne, S. C. (2022). Cancer risk in nodules detected at follow-up lung cancer screening CT. *American Journal of Roentgenology*, *218*(4), 634-641.

Han, Y., Ma, Y., Wu, Z., Zhang, F., Zheng, D., Liu, X., Tao, L., Liang, Z., Yang, Z., & Li, X. (2021). Histologic subtype classification of non-small cell lung cancer using PET/CT images. *European journal of nuclear medicine and molecular imaging*, *48*, 350-360.

Harmon, S. A., Sanford, T. H., Xu, S., Turkbey, E. B., Roth, H., Xu, Z., Yang, D., Myronenko, A., Anderson, V., & Amalou, A. (2020). Artificial intelligence for the detection of COVID-19 pneumonia on chest CT using multinational datasets. *Nature communications*, *11*(1), 4080.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. Proceedings of the IEEE international conference on computer vision,

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition,

He, X., Wang, S., Chu, X., Shi, S., Tang, J., Liu, X., Yan, C., Zhang, J., & Ding, G. (2021). Automated model design and benchmarking of deep learning models for covid-19 detection with chest ct scans. Proceedings of the AAAI conference on artificial intelligence,

Heidari, A., Navimipour, N. J., Unal, M., & Toumaj, S. (2022). The COVID-19 epidemic analysis and diagnosis using deep learning: A systematic literature review and future directions. *Computers in biology and medicine*, *141*, 105141.

Herold, C., Bankier, A., & Fleischmann, D. (1996). Lung metastases. *European radiology*, *6*, 596-606.

Hicks, S. A., Strümke, I., Thambawita, V., Hammou, M., Riegler, M. A., Halvorsen, P., & Parasa, S. (2022). On evaluation metrics for medical applications of artificial intelligence. *Scientific reports*, *12*(1), 5979.

Hofmanninger, J., Prayer, F., Pan, J., Röhrich, S., Prosch, H., & Langs, G. (2020). Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology problem. *European Radiology Experimental*, *4*, 1-13.

Hope, T. A., Afshar-Oromieh, A., Eiber, M., Emmett, L., Fendler, W. P., Lawhn-Heath, C., & Rowe, S. P. (2018). Imaging prostate cancer with prostate-specific membrane antigen PET/CT and PET/MRI: current and future applications. *American Journal of Roentgenology*, *211*(2), 286-294.

Horry, M. J., Chakraborty, S., Pradhan, B., Fallahpoor, M., Chegeni, H., & Paul, M. (2021). Factors determining generalization in deep learning models for scoring COVID-CT images. *Mathematical Biosciences and Engineering*.

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. Proceedings of the IEEE conference on computer vision and pattern recognition,

Hwang, D., Kim, K. Y., Kang, S. K., Seo, S., Paeng, J. C., Lee, D. S., & Lee, J. S. (2018). Improving the accuracy of simultaneously reconstructed activity and attenuation maps using deep learning. *Journal of Nuclear Medicine*, *59*(10), 1624-1629.

Jeong, S. Y., Lee, K. S., Shin, K. M., Bae, Y. A., Kim, B.-T., Choe, B. K., Kim, T. S., & Chung, M. J. (2008). Efficacy of PET/CT in the characterization of solid or partly solid solitary pulmonary nodules. *Lung Cancer*, *61*(2), 186-194.

Jerusalem, G., Hustinx, R., Beguin, Y., & Fillet, G. (2003). PET scan imaging in oncology. *European journal of cancer*, *39*(11), 1525-1534.

Jiang, X., Hu, Z., Wang, S., & Zhang, Y. (2023). Deep learning for medical image-based cancer diagnosis. *Cancers*, *15*(14), 3608.

Johnson, K. B., Wei, W. Q., Weeraratne, D., Frisse, M. E., Misulis, K., Rhee, K., Zhao, J., & Snowdon, J. L. (2021). Precision medicine, AI, and the future of personalized health care. *Clinical and translational science*, *14*(1), 86-93.

Kanda, T., Nakagomi, K., Goto, S., & Torizuka, T. (2008). Visualization of prostate cancer with 11C-choline positron emission tomography (PET): localization of primary and recurrent prostate cancer. *Hinyokika kiyo. Acta Urologica Japonica*, *54*(5), 325-332.

Kao, Y.-S., & Yang, J. (2022). Deep learning-based auto-segmentation of lung tumor PET/CT scans: a systematic review. *Clinical and Translational Imaging*, *10*(2), 217-223.

Kaplan, S., & Zhu, Y.-M. (2019). Full-dose PET image estimation from low-dose PET image using deep learning: a pilot study. *Journal of digital imaging*, *32*(5), 773-778.

Khan, A. N., Al-Jahdali, H. H., Irion, K. L., Arabi, M., & Koteyar, S. S. (2011). Solitary pulmonary nodule: A diagnostic algorithm in the light of current imaging technique. *Avicenna journal of medicine*, *1*(02), 39-51.

Kim, S. K., Allen-Auerbach, M., Goldin, J., Fueger, B. J., Dahlbom, M., Brown, M., Czernin, J., & Schiepers, C. (2007). Accuracy of PET/CT in characterization of solitary pulmonary lesions. *Journal of Nuclear Medicine*, *48*(2), 214-220.

Kirienko, M., Sollini, M., Silvestri, G., Mognetti, S., Voulaz, E., Antunovic, L., Rossi, A., Antiga, L., & Chiti, A. (2018). Convolutional neural networks promising in lung cancer T-parameter assessment on baseline FDG-PET/CT. *Contrast Media & Molecular Imaging*, *2018*(1), 1382309.

Kora, P., Ooi, C. P., Faust, O., Raghavendra, U., Gudigar, A., Chan, W. Y., Meenakshi, K., Swaraja, K., Plawiak, P., & Acharya, U. R. (2022). Transfer learning techniques for medical image analysis: A review. *Biocybernetics and Biomedical Engineering*, *42*(1), 79-107.

Kukava, S., & Baramia, M. (2022). Place and Role of PET/CT in the Diagnosis and Staging of Lung Cancer. In *Advances in Radiation Oncology in Lung Cancer* (pp. 85-111). Springer.

Kumar, A., Fulham, M., Feng, D., & Kim, J. (2019). Co-learning feature fusion maps from PET-CT images of lung cancer. *IEEE Transactions on Medical Imaging*, *39*(1), 204-217.

Kumar, R., Khan, A. A., Kumar, J., Golilarz, N. A., Zhang, S., Ting, Y., Zheng, C., & Wang, W. (2021). Blockchain-federated-learning and deep learning models for covid-19 detection using ct imaging. *IEEE Sensors Journal*, *21*(14), 16301-16314.

Kumar, Y., Koul, A., Singla, R., & Ijaz, M. F. (2023). Artificial intelligence in disease diagnosis: a systematic literature review, synthesizing framework and future research agenda. *Journal of ambient intelligence and humanized computing*, *14*(7), 8459-8486.

Langer, A. (2010). A systematic review of PET and PET/CT in oncology: a way to personalize cancer treatment in a cost-effective manner? *BMC health services research*, *10*, 1-16.

Larici, A. R., Farchione, A., Franchi, P., Ciliberto, M., Cicchetti, G., Calandriello, L., Del Ciello, A., & Bonomo, L. (2017). Lung nodules: size still matters. *European respiratory review*, *26*(146).

Lawrence, J., Rohren, E., & Provenzale, J. (2010). PET/CT today and tomorrow in veterinary cancer diagnosis and monitoring: fundamentals, early results and future perspectives. *Veterinary and Comparative Oncology*, *8*(3), 163-187.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436-444.

Lee, J.-G., Jun, S., Cho, Y.-W., Lee, H., Kim, G. B., Seo, J. B., & Kim, N. (2017). Deep learning in medical imaging: general overview. *Korean journal of radiology*, *18*(4), 570-584.

Lee, J. S. (2020). A review of deep-learning-based approaches for attenuation correction in positron emission tomography. *IEEE Transactions on Radiation and Plasma Medical Sciences*, *5*(2), 160-184.

LeVine III, H. (2010). *Medical imaging*. Bloomsbury Publishing USA.

Li, J., Zhao, G., Tao, Y., Zhai, P., Chen, H., He, H., & Cai, T. (2021). Multi-task contrastive learning for automatic CT and X-ray diagnosis of COVID-19. *Pattern recognition*, *114*, 107848.

Li, L., Qin, L., Xu, Z., Yin, Y., Wang, X., Kong, B., Bai, J., Lu, Y., Fang, Z., & Song, Q. (2020). Artificial intelligence distinguishes COVID-19 from community acquired pneumonia on chest CT. *Radiology*.

Li, M., Jiang, Y., Zhang, Y., & Zhu, H. (2023). Medical image analysis using deep learning algorithms. *Frontiers in Public Health*, *11*, 1273253.

Lin, C.-Y., Guo, S.-M., Lien, J.-J. J., Lin, W.-T., Liu, Y.-S., Lai, C.-H., Hsu, I.-L., Chang, C.-C., & Tseng, Y.-L. (2024). Combined model integrating deep learning, radiomics, and clinical data to classify lung nodules at chest CT. *La radiologia medica*, *129*(1), 56-69.

Lindner, C., Wang, C.-W., Huang, C.-T., Li, C.-H., Chang, S.-W., & Cootes, T. F. (2016). Fully automatic system for accurate localisation and analysis of cephalometric landmarks in lateral cephalograms. *Scientific reports*, *6*(1), 33581.

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., Van Der Laak, J. A., Van Ginneken, B., & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical image analysis*, *42*, 60-88.

Liu, B., Chi, W., Li, X., Li, P., Liang, W., Liu, H., Wang, W., & He, J. (2020). Evolving the pulmonary nodules diagnosis from classical approaches to deep learning-aided decision support: three decades' development course and future prospect. *Journal of cancer research and clinical oncology*, *146*, 153-185.

Liu, J., Xue, Q., Feng, Y., Xu, T., Shen, K., Shen, C., & Shi, Y. (2024). Enhancing Lesion Segmentation in PET/CT Imaging with Deep Learning and Advanced Data Preprocessing Techniques. *arXiv preprint arXiv:2409.09784*.

Liu, X., Li, K.-W., Yang, R., & Geng, L.-S. (2021). Review of deep learning based automatic segmentation for lung cancer radiotherapy. *Frontiers in oncology*, *11*, 717039.

Loverdos, K., Fotiadis, A., Kontogianni, C., Iliopoulou, M., & Gaga, M. (2019). Lung nodules: a comprehensive review on current approach and management. *Annals of thoracic medicine*, *14*(4), 226-238.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, *30*.

Maguolo, G., & Nanni, L. (2021). A critic evaluation of methods for COVID-19 automatic detection from X-ray images. *Information fusion*, *76*, 1-7.

Mahajan, A., & Cook, G. (2017). Clinical applications of PET/CT in oncology. *Basic science of PET imaging*, 429-450.

Meadows, A., & Allie, R. The growth and progression of PET-CT.

Mohandass, G., Krishnan, G. H., Selvaraj, D., & Sridhathan, C. (2024). Lung Cancer Classification using Optimized Attention-based Convolutional Neural Network with DenseNet-201 Transfer Learning Model on CT image. *Biomedical Signal Processing and Control*, *95*, 106330.

Morozov, S. P., Andreychenko, A. E., Blokhin, I. A., Gelezhe, P. B., Gonchar, A. P., Nikolaev, A. E., Pavlov, N. A., Chernina, V. Y., & Gombolevskiy, V. A. (2020). MosMedData: data set of 1110 chest CT scans performed during the COVID-19 epidemic. *Digital Diagnostics*, *1*(1), 49-59.

Motwani, M. (2022). Artificial intelligence primer for the nuclear cardiologist. *J Nucl Cardiol*, *12*, 120.

Muoio, B., Giovanella, L., & Treglia, G. (2018). Recent Developments of 18F-FET PET in Neuro-oncology. *Current Medicinal Chemistry*, *25*(26), 3061-3073.

Nahmias, C., & Wahl, L. M. (2008). Reproducibility of standardized uptake value measurements determined by 18F-FDG PET in malignant tumors. *Journal of Nuclear Medicine*, *49*(11), 1804-1808.

Nasrullah, N., Sang, J., Alam, M. S., Mateen, M., Cai, B., & Hu, H. (2019). Automated lung nodule detection and classification using deep learning combined with multiple strategies. *Sensors*, *19*(17), 3722.

Nguyen, D., Kay, F., Tan, J., Yan, Y., Ng, Y. S., Iyengar, P., Peshock, R., & Jiang, S. (2021). Deep learning–based COVID-19 pneumonia classification using chest CT images: model generalizability. *Frontiers in Artificial Intelligence*, *4*, 694875.

Ni, Q., Sun, Z. Y., Qi, L., Chen, W., Yang, Y., Wang, L., Zhang, X., Yang, L., Fang, Y., & Xing, Z. (2020). A deep learning approach to characterize 2019 coronavirus disease (COVID-19) pneumonia in chest CT images. *European radiology*, *30*, 6517-6527.

Olah, C., Mordvintsev, A., & Schubert, L. (2017). Feature visualization. *Distill*, *2*(11), e7.

Ost, D., Fein, A. M., & Feinsilver, S. H. (2003). The solitary pulmonary nodule. *New England Journal of Medicine*, *348*(25), 2535-2542.

Pal, M. (2023). An XAI Model for Malignancy Detection of the Pulmonary Nodules: Building Trust by Reducing AI Risk.

Panayides, A. S., Amini, A., Filipovic, N. D., Sharma, A., Tsaftaris, S. A., Young, A., Foran, D., Do, N., Golemati, S., & Kurc, T. (2020). AI in medical imaging informatics: current challenges and future directions. *IEEE journal of biomedical and health informatics*, *24*(7), 1837-1857.

Park, J., Kang, S. K., Hwang, D., Choi, H., Ha, S., Seo, J. M., Eo, J. S., & Lee, J. S. (2023). Automatic lung cancer segmentation in [18F] FDG PET/CT using a two-stage deep learning approach. *Nuclear Medicine and Molecular Imaging*, *57*(2), 86-93.

Park, Y.-J., Choi, D., Choi, J. Y., & Hyun, S. H. (2021). Performance evaluation of a deep learning system for differential diagnosis of lung cancer with conventional CT and FDG PET/CT using transfer learning and metadata. *Clinical nuclear medicine*, *46*(8), 635-640.

Piert, M., Park, H., Khan, A., Siddiqui, J., Hussain, H., Chenevert, T., Wood, D., Johnson, T., Shah, R. B., & Meyer, C. (2009). Detection of aggressive primary prostate cancer with 11C-choline PET/CT using multimodality fusion techniques. *Journal of Nuclear Medicine*, *50*(10), 1585-1593.

Rayed, M. E., Islam, S. S., Niha, S. I., Jim, J. R., Kabir, M. M., & Mridha, M. (2024). Deep learning for medical image segmentation: State-of-the-art advancements and challenges. *Informatics in medicine unlocked*, 101504.

Razzak, M. I., Naz, S., & Zaib, A. (2018). Deep learning for medical image processing: Overview, challenges and the future. *Classification in BioApps: Automation of decision making*, 323-350.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). " Why should i trust you?" Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining,

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18,

Rubin, D. L. (2019). Artificial intelligence in imaging: the radiologist's role. *Journal of the American College of Radiology*, *16*(9), 1309-1317.

Rubin, G. D. (2014). Computed tomography: revolutionizing the practice of medicine for 40 years. *Radiology*, *273*(2S), S45-S74.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., & Bernstein, M. (2015). Imagenet large scale visual recognition challenge. *International journal of computer vision*, *115*, 211-252.

Sarvamangala, D., & Kulkarni, R. V. (2022). Convolutional neural networks in medical image understanding: a survey. *Evolutionary intelligence*, *15*(1), 1-22.

Schabath, M. B., & Cote, M. L. (2019). Cancer progress and priorities: lung cancer. *Cancer epidemiology, biomarkers & prevention*, *28*(10), 1563-1579.

Schmidt-Hansen, M., Baldwin, D. R., Hasler, E., Zamora, J., Abraira, V., & i Figuls, M. R. (2014). PET-CT for assessing mediastinal lymph node involvement in patients with suspected resectable non-small cell lung cancer. *Cochrane Database of Systematic Reviews*(11).

Schöder, H., Erdi, Y. E., Larson, S. M., & Yeung, H. W. (2003). PET/CT: a new imaging technology in nuclear medicine. *European journal of nuclear medicine and molecular imaging*, *30*, 1419-1437.

Schwyzer, M., Ferraro, D. A., Muehlematter, U. J., Curioni-Fontecedro, A., Huellner, M. W., Von Schulthess, G. K., Kaufmann, P. A., Burger, I. A., & Messerli, M. (2018). Automated detection of lung cancer at ultralow dose PET/CT by deep neural networks–initial results. *Lung Cancer*, *126*, 170-173.

Schwyzer, M., Martini, K., Benz, D. C., Burger, I. A., Ferraro, D. A., Kudura, K., Treyer, V., von Schulthess, G. K., Kaufmann, P. A., & Huellner, M. W. (2020). Artificial intelligence for detecting small FDG-positive lung nodules in digital PET/CT: impact of image reconstructions on diagnostic performance. *European radiology*, *30*, 2031-2040.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2020). Grad-CAM: visual explanations from deep networks via gradient-based localization. *International journal of computer vision*, *128*, 336-359.

Serte, S., & Demirel, H. (2021). Deep learning for diagnosis of COVID-19 using 3D CT scans. *Computers in biology and medicine*, *132*, 104306.

Shah, V., Keniya, R., Shridharani, A., Punjabi, M., Shah, J., & Mehendale, N. (2021). Diagnosis of COVID-19 using CT scan images and deep learning techniques. *Emergency radiology*, *28*, 497-505.

Shen, D., Wu, G., & Suk, H.-I. (2017). Deep learning in medical image analysis. *Annual review of biomedical engineering*, *19*(1), 221-248.

Shen, J., Zhang, D., Dong, G., Sun, D., Liang, X., & Su, M. (2024). Classification of hyperspectral images based on fused 3D inception and 3D-2D hybrid convolution. *Signal, Image and Video Processing*, *18*(4), 3031-3041.

Shi, F., Wang, J., Shi, J., Wu, Z., Wang, Q., Tang, Z., He, K., Shi, Y., & Shen, D. (2020). Review of artificial intelligence techniques in imaging data acquisition, segmentation, and diagnosis for COVID-19. *IEEE reviews in biomedical engineering*, *14*, 4-15.

Shoeibi, A., Khodatars, M., Jafari, M., Ghassemi, N., Sadeghi, D., Moridian, P., Khadem, A., Alizadehsani, R., Hussain, S., & Zare, A. (2024). Automated detection and forecasting of covid-19 using deep learning techniques: A review. *Neurocomputing*, 127317.

Silva, P., Luz, E., Silva, G., Moreira, G., Silva, R., Lucio, D., & Menotti, D. (2020). COVID-19 detection in CT images with deep learning: A voting-based scheme and cross-datasets analysis. *Informatics in medicine unlocked*, *20*, 100427.

Silverman, D. H. (2004). Brain 18F-FDG PET in the diagnosis of neurodegenerative dementias: comparison with perfusion SPECT and with clinical evaluations lacking nuclear imaging. *Journal of Nuclear Medicine*, *45*(4), 594-607.

Silvestri, G. A., Gonzalez, A. V., Jantz, M. A., Margolis, M. L., Gould, M. K., Tanoue, L. T., Harris, L. J., & Detterbeck, F. C. (2013). Methods for staging non-small cell lung cancer: Diagnosis and management of lung cancer: American College of Chest Physicians evidence-based clinical practice guidelines. *Chest*, *143*(5), e211S-e250S.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Sinsuat, M., Saita, S., Kawata, Y., Niki, N., Ohmatsu, H., Tsuchida, T., Kakinuma, R., Kusumoto, M., Eguchi, K., & Kaneko, M. (2011). Influence of slice thickness on diagnoses of pulmonary nodules using low-dose CT: potential dependence of detection and diagnostic agreement on features and location of nodule. *Academic radiology*, *18*(5), 594-604.

Sironi, S., Buda, A., Picchio, M., Perego, P., Moreni, R., Pellegrino, A., Colombo, M., Mangioni, C., Messa, C., & Fazio, F. (2006). Lymph node metastasis in patients with clinical early-

stage cervical cancer: detection with integrated FDG PET/CT. *Radiology*, *238*(1), 272-279.

Soares, E., Angelov, P., Biaso, S., Froes, M. H., & Abe, D. K. (2020). SARS-CoV-2 CT-scan dataset: A large dataset of real patients CT scans for SARS-CoV-2 identification. *MedRxiv*, 2020.2004. 2024.20078584.

Søreide, K. (2009). Receiver-operating characteristic curve analysis in diagnostic, prognostic and predictive biomarker research. *Journal of clinical pathology*, *62*(1), 1-5.

Sourlos, N., Wang, J., Nagaraj, Y., van Ooijen, P., & Vliegenthart, R. (2022). Possible bias in supervised deep learning algorithms for CT lung nodule detection and classification. *Cancers*, *14*(16), 3867.

Storch, A., Wolz, M., Beuthien-Baumann, B., Loehle, M., Herting, B., Schwanebeck, U., Oehme, L., van den Hoff, J., Perick, M., & Graehlert, X. (2013). Effects of dopaminergic treatment on striatal dopamine turnover in de novo Parkinson disease. *Neurology*, *80*(19), 1754-1761.

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, *71*(3), 209-249.

Suo, S., Cheng, J., Cao, M., Lu, Q., Yin, Y., Xu, J., & Wu, H. (2016). Assessment of heterogeneity difference between edge and core by using texture analysis: differentiation of malignant from inflammatory pulmonary nodules and masses. *Academic radiology*, *23*(9), 1115-1122.

Suzuki, K. (2017). Overview of deep learning in medical imaging. *Radiological physics and technology*, *10*(3), 257-273.

Swinnen, G., Maes, A., Pottel, H., Vanneste, A., Billiet, I., Lesage, K., & Werbrouck, P. (2010). FDG-PET/CT for the preoperative lymph node staging of invasive bladder cancer. *European urology*, *57*(4), 641-647.

Takeda, A., Sanuki, N., Fujii, H., Yokosuka, N., Nishimura, S., Aoki, Y., Oku, Y., Ozawa, Y., & Kunieda, E. (2014). Maximum standardized uptake value on FDG-PET is a strong predictor of overall and disease-free survival for non–small-cell lung cancer patients after stereotactic body radiotherapy. *Journal of Thoracic Oncology*, *9*(1), 65-73.

Tang, K., Wang, L., Lin, J., Zheng, X., & Wu, Y. (2019). The value of 18F-FDG PET/CT in the diagnosis of different size of solitary pulmonary nodules. *Medicine*, *98*(11), e14813.

Teramoto, A., Fujita, H., Takahashi, K., Yamamuro, O., Tamaki, T., Nishio, M., & Kobayashi, T. (2014). Hybrid method for the detection of pulmonary nodules using positron emission tomography/computed tomography: a preliminary study. *International journal of computer assisted radiology and surgery*, *9*, 59-69.

Tizhoosh, H. R., Diamandis, P., Campbell, C. J., Safarpoor, A., Kalra, S., Maleki, D., Riasatian, A., & Babaie, M. (2021). Searching images for consensus: can AI remove observer variability in pathology? *The American journal of pathology*, *191*(10), 1702-1708.

Tong, G., Li, Y., Chen, H., Zhang, Q., & Jiang, H. (2018). Improved U-NET network for pulmonary nodules segmentation. *Optik*, *174*, 460-469.

Torigian, D. A., Huang, S. S., Houseni, M., & Alavi, A. (2007). Functional imaging of cancer with emphasis on molecular techniques. *CA: a cancer journal for clinicians*, *57*(4), 206-224.

Townsend, D. W. (2008). Positron emission tomography/computed tomography. Seminars in nuclear medicine,

Townsend, D. W., Carney, J. P., Yap, J. T., & Hall, N. C. (2004). PET/CT today and tomorrow. *Journal of Nuclear Medicine*, *45*(1 suppl), 4S-14S.

Trotter, J., Pantel, A. R., Teo, B.-K. K., Escorcia, F. E., Li, T., Pryma, D. A., & Taunk, N. K. (2023). Positron emission tomography (PET)/computed tomography (CT) imaging in

radiation therapy treatment planning: a review of PET imaging tracers and methods to incorporate PET/CT. *Advances in radiation oncology*, *8*(5), 101212.

Umapathy, V. R., Raj, R. D. S., Yadav, S., Munavarah, S. A., Anandapandian, P. A., Mary, A. V., Padmavathy, K., & Akshay, R. (2023). Perspective of artificial intelligence in disease diagnosis: a review of current and future endeavours in the medical field. *Cureus*, *15*(9).

Umbehr, M. H., Muentener, M., Hany, T., Sulser, T., & Bachmann, L. M. (2013). The role of 11C-choline and 18F-fluorocholine positron emission tomography (PET) and PET/CT in prostate cancer: a systematic review and meta-analysis. *European urology*, *64*(1), 106-117.

Vaishya, R., Javaid, M., Khan, I. H., & Haleem, A. (2020). Artificial Intelligence (AI) applications for COVID-19 pandemic. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, *14*(4), 337-339.

Van Schil, P. (2007). *Lung metastases and isolated lung perfusion*. Nova Publishers.

Vlahos, I., Stefanidis, K., Sheard, S., Nair, A., Sayer, C., & Moser, J. (2018). Lung cancer screening: nodule identification and characterization. *Translational Lung Cancer Research*, *7*(3), 288.

Wang, C., Elazab, A., Wu, J., & Hu, Q. (2017). Lung nodule classification using deep feature fusion in chest radiography. *Computerized Medical Imaging and Graphics*, *57*, 10-18.

Wang, C., Liu, Y., Wang, F., Zhang, C., Wang, Y., Yuan, M., & Yang, G. (2024). Towards reliable and explainable AI model for pulmonary nodule diagnosis. *Biomedical Signal Processing and Control*, *88*, 105646.

Wang, S., Kang, B., Ma, J., Zeng, X., Xiao, M., Guo, J., Cai, M., Yang, J., Li, Y., & Meng, X. (2021). A deep learning algorithm using CT images to screen for Corona Virus Disease (COVID-19). *European radiology*, *31*, 6096-6104.

Wang, X., Deng, X., Fu, Q., Zhou, Q., Feng, J., Ma, H., Liu, W., & Zheng, C. (2020). A weakly-supervised framework for COVID-19 classification and lesion localization from chest CT. *IEEE Transactions on Medical Imaging*, *39*(8), 2615-2625.

Wang, X., Jiang, L., Li, L., Xu, M., Deng, X., Dai, L., Xu, X., Li, T., Guo, Y., & Wang, Z. (2021). Joint learning of 3D lesion segmentation and classification for explainable COVID-19 diagnosis. *IEEE Transactions on Medical Imaging*, *40*(9), 2463-2476.

Weber, W. A., & Figlin, R. (2007). Monitoring cancer treatment with PET/CT: does it make a difference? *Journal of Nuclear Medicine*, *48*(1 suppl), 36S-44S.

Xiang, L., Qiao, Y., Nie, D., An, L., Lin, W., Wang, Q., & Shen, D. (2017). Deep auto-context convolutional neural networks for standard-dose PET image estimation from low-dose PET/MRI. *Neurocomputing*, *267*, 406-416.

Xiao, Z., Liu, B., Geng, L., Zhang, F., & Liu, Y. (2020). Segmentation of lung nodules using improved 3D-UNet neural network. *Symmetry*, *12*(11), 1787.

Xue, S., & Abhayaratne, C. (2023). Region-of-interest aware 3D ResNet for classification of COVID-19 chest computerised tomography scans. *Ieee Access*, *11*, 28856-28872.

Yadav, S. S., & Jadhav, S. M. (2019). Deep convolutional neural network based medical image classification for disease diagnosis. *Journal of Big data*, *6*(1), 1-18.

Yang, M., Huang, X., Huang, L., & Cai, G. (2023). Diagnosis of Parkinson's disease based on 3D ResNet: The frontal lobe is crucial. *Biomedical Signal Processing and Control*, *85*, 104904.

Yu, J., Yang, B., Wang, J., Leader, J., Wilson, D., & Pu, J. (2020). 2D CNN versus 3D CNN for false-positive reduction in lung cancer screening. *Journal of Medical Imaging*, *7*(5), 051202-051202.

Yu, X., Wang, J., Hong, Q.-Q., Teku, R., Wang, S.-H., & Zhang, Y.-D. (2022). Transfer learning for medical images analyses: A survey. *Neurocomputing*, *489*, 230-254.

Zandehshahvar, M., van Assen, M., Maleki, H., Kiarashi, Y., De Cecco, C. N., & Adibi, A. (2021). Toward understanding COVID-19 pneumonia: A deep-learning-based approach for severity analysis and monitoring the disease. *Scientific Reports*, *11*(1), 11112.

Zeidman, I. (1957). Metastasis: a review of recent advances. *Cancer research*, *17*(3), 157-162.

Zhang, K., Liu, X., Shen, J., Li, Z., Sang, Y., Wu, X., Zha, Y., Liang, W., Wang, C., & Wang, K. (2020). Clinically applicable AI system for accurate diagnosis, quantitative measurements, and prognosis of COVID-19 pneumonia using computed tomography. *Cell*, *181*(6), 1423-1433. e1411.

Zhao, K., Wang, C., Shi, F., Huang, Y., Ma, L., Li, M., & Song, Y. (2021). Combined prognostic value of the SUVmax derived from FDG-PET and the lymphocyte-monocyte ratio in patients with stage IIIB-IV non-small cell lung cancer receiving chemotherapy. *BMC cancer*, *21*, 1-13.

Zhong, Z., Kim, Y., Zhou, L., Plichta, K., Allen, B., Buatti, J., & Wu, X. (2018). 3D fully convolutional networks for co-segmentation of tumors on PET-CT images. 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018),

Zuidhof, G. (2017). Full preprocessing tutorial. *URL https://www. kaggle. com/gzuidhof/full-preprocessing-tutorial*.

APPENDIX

The detailed demographic information of patients whose ^{18}F FDG PET/CT images were used in this study are presented in the tables A, B, and C.

Table A. Demographic information for benign nodule class patients.

Patient number	**patient sex**	**patient age**	**malignancy**
1	F	70	Breast Cancer
2	M	34	Hodgkin's Lymphoma
3	M	71	Breast Cancer
4	M	50	DLBL
5	F	61	Breast Cancer
6	M	56	Melanoma
7	F	26	Breast Cancer
8	F	60	Breast Cancer
9	M	66	Pancreatic cancer
10	F	54	Pancreas Adenocarcinoma
11	M	24	Hodgkin's Lymphoma
12	M	65	DLBL
13	M	45	Nasopharyngeal Carcinoma
14	M	60	Hodgkin's lymphoma
15	M	78	Prostate Cancer
16	F	37	Ovarian cancer
17	F	21	Hodgkin's Lymphoma
18	M	66	Pancreatic cancer
19	F	63	Hodgkin's lymphoma
20	M	16	Rhabdomyosarcoma
21	F	60	Hodgkin's Lymphoma
22	F	63	Unknown
23	F	53	Breast Cancer
24	F	71	Colon Carcinoma
25	F	53	Breast Carcinoma
26	F	65	Papillary thyroid carcinoma
27	M	67	Liposarcoma

28	F	61	Non-Hodgkin's Lymphoma
29	M	61	B-Cell Lymphoma
30	F	23	Hodgkin's Lymphoma
31	M	49	Hodgkin's Lymphoma
32	F	47	Breast Cancer
33	F	44	Breast cancer
34	F	57	Colon Cancer
35	F	65	Colon Cancer
36	M	14	Hodgkin's Lymphoma
37	F	72	Breast Cancer
38	M	18	Hodgkin's Lymphoma
39	M	43	Hodgkin's Lymphoma
40	M	62	Hodgkin's Lymphoma
41	F	59	Breast Cancer
42	F	25	Non-Hodgkin's Lymphoma
43	M	44	Urinary bladder carcinoma
44	M	24	Hodgkin's Lymphoma
45	M	47	Lymphoma
46	F	62	Colorectal Cancer
47	F	63	Coronary Arteries Disease
48	F	50	Colonic Cancer
49	M	23	Seminoma
50	F	52	Cervical Cancer
51	F	60	Hodgkin's Lymphoma
52	M	54	Rectal Cancer
53	M	38	Ocular Melanoma
54	M	59	Non-Hodgkin's Lymphoma
55	M	59	Gastric Cancer
56	M	42	Non-Hodgkin's Lymphoma
57	F	54	Breast Cancer

58	M	64	Colon Cancer
59	F	65	Gastric carcinoma
60	M	43	Hodgkin Lymphoma
61	F	49	Breast Cancer
62	F	26	Hodgkin disease
63	F	75	Colonic cancer
64	F	38	Hodgkin's Lymphoma
65	F	4	Neuroblastoma
66	F	36	Colonic Cancer
67	M	31	Hodgkin's Lymphoma
68	F	73	Rectal Cancer
69	M	18	Hodgkin's Lymphoma
70	F	47	N/A
71	M	83	Esophageal Cancer
72	F	44	Ovarian cancer
73	F	47	Non-Hodgkin Lymphoma
74	F	70	Ovarian cancer
75	F	37	Lymphoma
76	M	19	Hodgkin disease
77	F	42	Papillary Thyroid Cancer
78	F	55	Hodgkin's Lymphoma
79	F	32	Breast cancer
80	M	44	Cholangiocarcinoma
81	F	86	Colon cancer
82	F	68	Colonic Cancer
83	F	57	NonHodgkin's Lymphoma
84	M	33	Seminoma
85	F	44	Rectal Cancer
86	M	58	Nasopharyngeal tumor
87	F	48	Breast Cancer
88	M	60	Hodgkin's Lymphoma
89	M	39	Hodgkin's Lymphoma
90	F	44	Ovarian cancer

91	M	9	Hodgkin's Lymphoma
92	M	32	Renal Cell Carcinoma
93	M	30	Hodgkin's Lymphoma
94	M	44	Non-Hodgkin's Lymphoma
95	M	24	Orbital Cancer
96	M	54	GIST
97	F	19	Non-Hodgkin's Lymphoma
98	F	72	Breast Cancer
99	F	38	Hodgkin's Lymphoma
100	F	41	Breast Cancer
101	M	34	Hodgkin's Lymphoma
102	F	28	Hodgkin's Lymphoma
103	F	58	Hodgkin's Lymphoma
104	F	61	PTC
105	F	40	Hodgkin's Lymphoma
106	M	36	Hodgkin's Lymphoma
107	M	25	Hodgkin's Lymphoma
108	F	52	N/A
109	M	75	Esophagus Cancer
110	M	77	Renal Cell Carcinoma
111	F	40	Ovarian Cancer
112	M	37	Bladder Carcinoma
113	F	77	Non-Hodgkin's Lymphoma
114	M	36	Hodgkin's Lymphoma
115	F	36	Non Hodgkin Lymphoma (T Cell Rich B-Cell Lymphoma)
116	F	53	Breast Cancer
117	F	34	Hodgkin's Lymphoma
118	F	57	Colon Cancer
119	F	72	Mantle cell lymphoma
120	F	38	GIST

121	M	66	Squamous Cell carcinoma
122	M	67	Gastric Cancer
123	M	49	Hodgkin's Lymphoma
124	F	38	Breast Cancer
125	F	48	B-cell Lymphoma
126	M	49	Lymphoma
127	F	22	Ovarian Dysgerminoma and Gonadoblastoma
128	F	53	SCC of Tongue
129	M	57	Colon Cancer
130	M	62	Rectal Cancer
131	F	26	Colon Cancer
132	M	43	High Grade B-cell Lymphoma
133	F	60	Breast Cancer
134	M	32	Hodgkin's Lymphoma
135	M	67	Colon Cancer
136	F	71	Ovarian Cancer
137	M	82	Gastric Lymphoma
138	F	48	Breast Cancer
139	M	74	Non-Hodgkin's Lymphoma
140	M	84	Hodgkin's Lymphoma (mixed cellularity)
141	F	33	Ewing's Sarcoma
142	M	59	Non-Hodgkin's Lymphoma
143	M	81	Cancer of Rectum
144	M	60	Colon Cancer
145	M	43	Hodgkin's Lymphoma
146	F	55	Hodgkin's Lymphoma
147	M	69	Neuroendocrine tumor of Pancreas
148	M	93	Sarcoma
149	F	72	Hodgkin's disease

150	F	45	Gastric carcinoma
151	F	37	Breast Cancer
152	F	82	Colon Cancer
153	F	56	Ovarian Cancer
154	M	72	B-cell Lymphoma
155	F	82	Melanoma
156	M	63	Esophageal carcinoma
157	M	48	CLL
158	F	61	Breast Cancer
159	F	61	Breast Cancer
160	F	51	Gastric Lymphoma
161	F	76	Mucinous Carcinoma
162	F	50	Breast cancer
163	F	64	Rectal Cancer
164	M	60	Colon Cancer
165	M	70	Non-Hodgkin's Lymphoma
166	F	54	Cervix Cancer
167	M	43	non-Hodgkin's Lymphoma
168	M	24	Hodgkin's Lymphoma
169	F	32	Breast cancer
170	M	53	Hodgkin's Lymphoma
171	M	61	Mesothelioma
172	M	50	Hodgkin's Disease
173	M	48	Hodgkin's Lymphoma
174	M	21	Osteosarcoma
175	M	36	Hodgkin's Lymphoma
176	F	31	Breast Cancer
177	M	37	Hodgkin's Lymphoma
178	M	62	Unknown
179	M	55	Non-Hodgkin's Lymphoma
180	F	42	Breast Cancer
181	F	64	Mucinous Neoplasm of Appendix

182	M	51	Non-Hodgkin's Lymphoma
183	M	47	GIST
184	F	39	Squamous Cell Carcinoma (cervix)
185	F	60	Non-Hodgkin's Lymphoma
186	F	55	Ovarian Cancer
187	F	73	Endometrial Cancer
188	M	32	Hodgkin's Lymphoma
189	F	67	Malignant Ascites
190	F	57	Ovarian Cancer
191	F	47	Breast Cancer
192	F	27	Hodgkin's Lymphoma
193	F	38	Breast Cancer
194	F	72	Gastric Cancer
195	F	57	Pheochromocytoma
196	M	35	Hodgkin's Lymphoma
197	F	42	Hodgkin's Lymphoma
198	F	18	Hodgkin's Lymphoma
199	F	66	Malignant Ascites
200	F	21	Hodgkin Lymphoma
201	F	31	Hodgkin's Lymphoma
202	F	50	Unknown
203	M	20	Non-Hodgkin's Lymphoma
204	M	44	Papillary thyroid carcinoma
205	F	40	Hodgkin's Lymphoma
206	F	25	Papillary Thyroid Carcinoma
207	M	59	Hodgkin's Lymphoma
208	M	20	Papillary Thyroid Cancer
209	F	70	Ovarian Cancer
210	M	28	Hodgkin's Lymphoma

211	F	55	Papillary Thyroid Cancer
212	F	52	Colon Cancer
213	M	64	FUO
214	M	27	Hodgkin Lymphoma
215	F	58	Rectosigmoid Adenocarcinoma
216	M	19	Hodgkin Lymphoma
217	M	59	GIST
218	F	6	Adrenocortical Neoplasm
219	M	55	Sacral Malignant Tumor
220	F	33	Hodgkin's Disease
221	F	47	R/O Paget Schroetter
222	F	24	Papillary Thyroid Cancer
223	F	35	Non-Hodgkin's Lymphoma
224	M	44	Colon Cancer
225	F	44	Non-Hodgkin's Lymphoma
226	M	61	Leukemia/Lymphoma
227	M	57	GIST
228	F	42	Breast Cancer
229	F	30	Hodgkin's Lymphoma
230	F	71	Melanoma
231	F	45	Non-Hodgkin's Lymphoma
232	F	63	Papillary Thyroid Cancer
233	M	54	Adenocarcinoma of Appendix
234	M	70	Sezary Syndrome
235	M	11	Hodgkin's Lymphoma
236	M	45	Colon Cancer

237	F	24	Carcinomatous Meningitis
238	M	37	Hodgkin Lymphoma
239	F	48	Papillary Thyroid Cancer
240	F	34	Non-Hodgkin's Lymphoma
241	F	29	Possibility of Breast Cancer
242	F	49	Neuroendocrine Tumor of the Duodenum
243	M	16	Possibility of Lymphoma
244	F	35	Hodgkin's Lymphoma
245	M	15	Hodgkin's Lymphoma
246	M	45	Hodgkin's Lymphoma
247	M	14	Hodgkin's Lymphoma
248	M	34	Papillary Thyroid Carcinoma
249	F	65	Hodgkin's Lymphoma
250	F	47	Neuro Endocrine Tumor
251	M	40	Non-Hodgkin's Lymphoma
252	F	32	Breast Cancer
253	M	37	Hodgkin's Lymphoma
254	F	49	Colon Cancer
255	F	39	Sarcoma
256	M	45	Hodgkin's Lymphoma
257	F	41	Krukenberg Tumor
258	F	60	R/O of MM
259	F	35	Hodgkin's Lymphoma
260	F	62	Cervix Cancer
261	F	66	Sarcoma
262	F	77	Tongue SCC
263	M	35	Hodgkin's Lymphoma

264	M	61	Melanoma
265	F	28	Hodgkin's Lymphoma
266	M	53	Bladder and Prostate Cancer
267	M	59	Renal Cancer
268	F	21	DLBL Lymphoma
269	F	61	Pancreas Cancer
270	M	38	Hodgkin's Lymphoma
271	F	50	Breast Cancer
272	F	19	Hodgkin's Lymphoma
273	F	42	Breast Cancer
274	M	17	Hodgkin Lymphoma
275	M	70	Colon Adenocarcinoma
276	M	58	Rectal Cancer
277	M	79	Squamous Cell Carcinoma of the Right Ear
278	M	72	Gastric Cancer
279	F	72	Rectal Cancer
280	M	50	Seminoma
281	M	58	Colon Adenocarcinoma
282	M	57	Unknown
283	M	62	Colon Adenocarcinoma
284	M	52	Giant Cell Tumor of the Left Fibula
285	F	39	Papillary Thyroid Carcinoma
286	M	32	PNET
287	F	29	Papillary Thyroid Cancer
288	M	33	Colon cancer
289	F	36	Breast Cancer
290	F	57	Ovarian Cancer
291	F	40	Rheumatoid Arthritis with unexplained persistent elevated ESR

292	M	65	Non-Hodgkin's Lymphoma
293	M	63	MALT Lymphoma
294	F	54	Breast Cancer
295	M	68	Colon Cancer
296	F	36	Non-Hodgkin's Lymphoma
297	M	51	Colon Cancer
298	F	14	Hodgkin's Lymphoma
299	M	45	Adrenal Cancer (According to the referring physician's note)
300	M	68	Brain Non-Hodgkin's Lymphoma
301	M	31	Hodgkin Lymphoma (Nodular Sclerosis)
302	F	37	Breast Leukemia/Lymphoma
303	M	11	Hodgkin Lymphoma
304	F	50	Squamous Cell Carcinoma of Esophagus
305	F	30	Hodgkin Lymphoma
306	M	35	Non-Hodgkin's Lymphoma
307	F	65	Gastric Cancer
308	F	56	Ulnar Tumor (Malignant round cell tumor with Malignant large B-cell lymphnoma)
309	F	67	Uterine Cancer
310	M	18	Hodgkin's Lymphoma
311	M	72	Low grade Non-Hodgkin's Lymphoma
312	M	41	Hodgkin's Lymphoma

313	M	65	Thyroid Cancer
314	F	32	Hodgkin's Lymphoma
315	M	51	Papillary Thyroid Cancer
316	F	52	Hodgkin's Disease
317	F	52	Cutaneous Lymphoma
318	F	46	Colon Cancer
319	M	36	Colon Cancer
320	M	33	Non-Hodgkin's Lymphoma
321	F	28	Hodgkin's Lymphoma
322	M	55	Renal Cell Cancer
323	F	48	Breast Cancer
324	F	73	Colon Cancer
325	M	64	Colon Adenocarcinoma
326	F	46	Breast Cancer
327	M	69	Gastric Cancer
328	M	54	Brain Non-Hodgkin's Lymphoma
329	F	52	Breast Cancer
330	M	36	Hodgkin Lymphoma
331	M	61	Brain metastases (SCC)
332	F	57	Renal Cell Carcinoma
333	M	43	Non-Hodgkin's Lymphoma
334	M	68	Colon Cancer
335	F	30	Breast Cancer
336	M	34	Hodgkin's Lymphoma
337	F	68	Breast Cancer
338	M	52	Small Cell Lymphocytic Lymphoma
339	M	23	Thyroid Cancer
340	M	33	Non-Hodgkin's Lymphoma
341	F	50	Uterine Sarcoma

342	M	35	Left Testicular Cancer
343	F	84	Colon Cancer
344	M	38	Hodgkin's Lymphoma
345	F	53	Breast Cancer
346	M	46	Diffuse Large B-Cell Lymphoma
347	F	53	Non-Hodgkin's Lymphoma
348	F	54	Breast Cancer
349	M	9	Hodgkin's Lymphoma
350	F	63	Ovarian Cancer (serous adenocarcinoma)
351	F	41	R/O Malignancy
352	F	54	Takayasu Arteritis
353	M	38	Seminoma
354	F	24	Hodgkin's Lymphoma
355	F	74	Endometrial Cancer
356	M	51	Colon Cancer
357	F	79	Breast Cancer
358	M	31	Hodgkin Lymphoma (Nodular Sclerosis)
359	M	29	Diffuse Large B-Cell Lymphoma
360	F	30	Hodgkin Lymphoma
361	F	55	Plasma Cell Tumor
362	M	55	Appendix Adenocarcinoma
363	M	40	Seminoma
364	F	73	Breast Cancer
365	F	24	Papillary Thyroid Carcinoma
366	M	55	Esophageal Squamous Cell Carcinoma
367	M	46	Non-Hodgkin Lymphoma

368	M	47	Diffuse-Large B-Cell Lymphoma
369	F	22	Hodgkin Lymphoma
370	M	55	Renal Cell Cancer
371	F	18	Hodgkin Disease
372	M	55	Unknown; Evaluation of generalized pain
373	F	63	R/O Malignancy
374	F	49	Gastric Cancer
375	M	69	Gastric Cancer
376	M	66	Colon Cancer
377	F	27	Rectal Cancer
378	M	72	Colon Cancer
379	M	48	Rule out Malignancy
380	F	48	Breast Cancer
381	M	36	Hodgkin Lymphoma
382	M	70	Colon Adenocarcinoma
383	F	62	Breast Cancer
384	M	30	Seminoma
385	F	41	Gastric Cancer
386	M	21	Hodgkin Lymphoma
387	M	58	Colon Adenocarcinoma
388	F	63	Uterine Cancer
389	F	34	Thyroid Cancer
390	M	62	Colon Adenocarcinoma
391	F	38	Hodgkin Disease
392	F	36	Breast Cancer
393	F	40	Rheumatoid Arthritis with unexplained persistent elevated ESR
394	M	63	MALT Lymphoma
395	F	34	Non-Hodgkin’s Lymphoma
396	M	13	Lymphoma (Plasmoblastic Type)

397	F	34	Gastric Cancer, Carcinomatous meningitis
398	M	35	Non-Hodgkin's Lymphoma
399	M	17	Hodgkin Lymphoma
400	F	69	Malignant Melanoma
401	M	70	Bladder Cancer
402	F	46	Breast Cancer
403	M	28	Choroidal Melanoma
404	M	72	Hepatocellular Carcinoma
405	F	51	Colon Cancer
406	F	53	Breast Carcinoma
407	M	35	Large B-cell Lymphoma
408	M	61	Unknown
409	M	82	Gastric Lymphoma
410	M	70	Gastric Adenocarcinoma
411	M	50	Colon Cancer
412	M	34	Hodgkin's Lymphoma
413	F	34	Adrenocortical Carcinoma
414	M	43	Colon Adenocarcinoma
415	F	62	Duodenal Cancer
416	M	62	Multiple Myeloma
417	F	50	Colon Cancer
418	F	67	Gastric Lymphoma
419	M	50	Medullary Thyroid Carcinoma
420	M	56	Hodgkin-Lymphoma
421	F	21	Rectal Cancer
422	M	72	Transitional Cell Carcinoma of Bladder
423	F	21	Colon Cancer

424	F	29	Colon Cancer
425	F	70	Breast Cancer
426	M	71	Gastric Cancer
427	F	50	Ovarian Cancer
428	M	65	Unknown
429	F	60	Gastric Cancer
430	M	51	Non-Hodgkin Lymphoma (Posterior Tongue)
431	F	60	Large B cell Lymphoma
432	F	33	Hodgkin's Lymphoma
433	M	55	Solitary Pulmonary Nodule
434	M	43	Gastric Cancer
435	M	59	Non-Hodgkin's Lymphoma
436	F	47	Carcinoid of Small Intestine + Breast Cancer
437	M	59	Colon Cancer
438	F	50	Breast Carcinoma
439	F	12	Anterior Mediastinal Mass on CT images
440	M	43	Leiomyosarcoma
441	F	55	Cervix cancer
442	M	69	Colon Cancer
443	F	36	Ovarian Cancer
444	F	53	Follicular Thyroid Cancer
445	M	83	Laryngeal Cancer
446	M	58	Nasopharyngeal carcinoma
447	F	64	T-cell Lymphoma
448	M	39	Hodgkin's disease
449	M	62	Adrenocortical carcinoma

450	F	46	Non-Hodgkin Lymphoma (DLBL)
451	F	73	Non-Hodgkin's Lymphoma (Diffuse Large B-Cell Lymphoma)
452	F	56	Colon Cancer
453	M	60	Pancreas Cancer
454	F	51	Hodgkin's Lymphoma
455	F	34	Hodgkin's Lymphoma
456	M	55	Renal Cell Carcinoma
457	M	61	Solitary Pulmonary Nodule
458	F	50	Colon Cancer
459	F	75	Breast Carcinoma
460	M	48	Hodgkin's Lymphoma
461	F	80	Unknown (Possibility of Renal Cell Carcinoma)
462	F	55	Hodgkin's Lymphoma
463	F	30	Hodgkin's Lymphoma
464	M	65	Unknown primary malignancy
465	F	56	Colon Cancer
466	M	46	Non-Hodgkin Lymphoma (NHL)
467	F	62	Breast Carcinoma
468	M	31	Hodgkin's Lymphoma
469	F	61	Non-Hodgkin's Lymphoma (Diffuse Large B-Cell Lymphoma)
470	F	48	Brain Masses
471	M	17	Mixed Germ Cell Tumor of Testis
472	F	64	Non-Hodgkin's Lymphoma

473	M	82	Gastric Lymphoma
474	F	25	Hodgkin's Lymphoma
475	F	21	Hodgkin's Lymphoma
476	M	43	N/A
477	F	50	Breast Cancer
478	M	30	Non-Hodgkin's Lymphoma
479	M	66	Laryngeal Squamous Cell Carcinoma
480	M	33	Non-Hodgkin Lymphoma
481	F	82	Colon Cancer
482	M	59	Papillary Thyroid Carcinoma
483	F	61	Breast Cancer
484	M	70	Non-Hodgkin's Lymphoma
485	M	53	Pancreatic Adenocarcinoma
486	F	74	Colon Cancer
487	M	80	DLBL
488	M	47	N/A
489	M	24	Hodgkin's Lymphoma
490	F	25	Hodgkin Lymphoma
491	M	48	Hodgkin's Lymphoma
492	M	62	Pulmonary Nodule
493	M	54	Renal cell Carcinoma
494	M	41	Hodgkin's Lymphoma
495	M	43	Hodgkin's Lymphoma
496	F	26	Colon Cancer
497	M	22	Hodgkin-Lymphoma
498	M	37	Hodgkin's Lymphoma
499	F	60	Hodgkin's Disease
500	M	56	Rectal Adenocarcinoma
501	M	54	Non- Hodgkin Lymphoma

502	F	59	Ovarian Cancer
503	F	65	Papillary Thyroid Carcinoma
504	M	32	Hodgkin's Lymphoma
505	M	71	Colon Cancer
506	M	36	Adenocarcinoma of Rectum
507	M	30	Hodgkin's Lymphoma
508	F	41	Colon Cancer
509	F	55	Breast Cancer
510	M	48	Rectal Cancer
511	M	67	Colon Cancer
512	M	73	Large B-Cell Lymphoma
513	F	63	Non-Hodgkin' Lymphoma
514	M	58	Cancer of Testis
515	F	53	Breast Cancer
516	F	34	Hodgkin's Lymphoma
517	F	23	Hodgkin's Lymphoma
518	M	49	Hodgkin's Lymphoma
519	M	43	Large B-cell Lymphoma
520	F	56	Rectal Cancer
521	F	62	Breast Cancer
522	F	75	Colon Cancer
523	F	32	Hodgkin's Lymphoma
524	M	59	Multiple Myeloma
525	M	83	Rectal Cancer
526	M	42	Hodgkin Disease
527	M	43	Rectal Cancer
528	M	70	Colon cancer
529	F	47	Renal Cell Carcinoma
530	F	54	Colon Cancer
531	F	62	Scleroderma and Rheumatoid Arthritis

532	M	64	Nasopharyngeal Cancer
533	M	46	Diffuse large B-cell lymphoma
534	F	22	Diffuse Large B Cell Lymphoma
535	F	54	Papillary Thyroid Cancer
536	M	60	Diffuse Large B Cell Lymphoma
539	F	42	Uterus Cancer
540	M	54	Hodgkin's Lymphoma

Table B. Demographic information for malignant nodule class patients.

Patient number	**patient sex**	**patient age**	**malignancy**
1	F	58	Hodgkin's Lymphoma
2	F	52	Breast cancer
3	F	72	Follicular thyroid cancer
5	M	66	Rectal Cancer
6	M	57	Laryngeal cancer
7	M	75	Colon Cancer
8	M	60	Subglottic Cancer
9	M	66	N/A
10	M	53	Melanoma
11	M	66	Stomach cancer
12	M	68	CLL
13	M	70	Rectal Cancer
14	F	81	PTC
15	F	56	Colon Cancer
16	M	78	Hodgkin's lymphoma
17	F	43	Breast Cancer
18	F	44	Colon Cancer
19	F	55	Cancer with Unknown Origin
20	F	60	Breast Cancer

21	F	64	Breast Cancer
23	M	72	Adenocarcinoma of prostate
24	F	59	Diffuse Large B-Cell Lymphoma
25	F	17	Papillary Thyroid Carcinoma
26	F	30	Hodgkin's Lymphoma
27	M	74	Melanoma
28	M	39	Malignant Thymoma
29	F	61	Breast Carcinoma
30	F	31	Breast Cancer
31	F	49	Uterine Sarcoma
32	M	64	Colon Cancer
33	M	72	Pancreatic Cancer
34	F	45	Breast Cancer
35	F	51	Carcinoma of unknown primary
36	F	51	Thyroid Cancer
37	M	57	B-Cell Lymphoma
38	F	73	Sigmoid Adenocarcinoma
39	M	30	B-Cell Lymphoma
40	M	66	Colorectal Cancer
41	M	57	Colorectal Cancer
42	M	57	Laryngeal cancer
43	F	28	Hodgkin's Lymphoma
44	M	68	CLL
45	M	61	Colon Cancer
46	F	43	Breast Cancer
47	F	53	Leiomyosarcoma
48	F	29	Papillary Thyroid Carcinoma
49	M	64	Transitional Cell Carcinoma of Bladder
50	F	33	SCC of cervix

51	F	73	Breast Cancer
52	F	74	Breast Cancer
53	F	35	Breast Cancer
54	M	64	Follicular Thyroid Carcinoma
55	F	78	Evaluation of Incidental Pulmonary Nodule
56	F	55	Retroperitoneal Sarcoma
57	M	73	Rectal Carcinoma
58	F	68	Glossal SCC
59	M	76	Melanoma
60	M	63	Rectal Cancer
61	F	57	Colorectal Cancer
62	M	51	Seminoma
63	F	25	Ewing's Sarcoma
64	M	28	N/A
65	M	65	Colon Cancer
66	F	62	Evaluation of Single Pulmonary Nodule (SPN)
67	F	62	Colon Carcinoma
68	F	70	Squamous Cell Carcinoma of the Anal Canal
69	F	66	Colon Cancer
70	M	59	Colon Cancer
71	F	62	Colorectal Cancer
72	M	75	Colon Cancer
73	M	37	Non-Hodgkin Lymphoma
74	F	58	Rectal Cancer
75	M	75	Colon Cancer
76	M	64	Melanoma
77	F	58	Breast Cancer

78	M	44	Rectal Cancer
79	F	60	Unknown Primary
80	F	50	Adenoid Cystic Carcinoma of Salivary Gland
81	M	43	Melanoma
82	M	87	Melanoma
83	M	70	Rectal Cancer
84	F	53	Breast Cancer
85	F	51	Breast Cancer
86	F	51	Breast Cancer
87	F	56	Breast Cancer
88	F	44	Breast Cancer
89	M	42	Melanoma
90	F	48	Breast Cancer
91	M	66	Renal Cell Carcinoma
92	M	27	Hodgkin's Lymphoma
93	M	46	Rectosigmoid Adenocarcinoma
94	F	71	Breast Cancer
95	M	67	Papillary Thyroid Cancer
96	M	75	Colon Cancer
97	F	58	Colon Adenocarcinoma
98	F	58	Colon Cancer
99	F	38	Breast Cancer
100	F	75	Colon Cancer
101	M	59	Colon Cancer
102	M	42	Melanoma
103	F	51	Paragangeloima
104	M	69	Hodgkin's Lymphoma
105	F	59	Breast Cancer
106	F	58	Colon Adenocarcinoma
107	M	42	Hodgkin's Lymphoma

108	F	39	Spindle Cell Tumor
109	M	37	Colon Cancer
111	M	51	Clear Cell Carcinoma
112	M	54	Papillary Thyroid Carcinoma
113	F	53	Renal Cell Carcinoma
114	M	36	Rectal Cancer
115	M	83	N/A
116	F	62	Papillary Thyroid Carcinoma
117	F	44	Breast Cancer
118	M	48	Rectal Cancer
119	M	41	NET or Lymphoma
120	F	51	Breast Cancer
121	M	53	Sarcoidosis
122	M	53	Colon Cancer
123	F	75	Gastric Cancer
124	M	69	Colon Adenocarcinoma
125	M	51	Colon Cancer
126	M	56	Unknown Primary Malignancy
127	F	68	Papillary thyroid carcinoma
128	F	36	Hodgkin's Lymphoma
129	M	56	Unknown Primary Cancer
130	F	58	Papillary Thyroid Carcinoma
131	F	66	Endometrial Cancer
132	F	51	Carcinoma of unknown primary
133	F	61	Solitary Pulmonary Nodule
134	F	52	Colon Cancer
135	M	66	Gastric Cancer

136	M	40	Non-Hodgkin Lymphoma
137	M	69	Papillary thyroid carcinoma
138	F	57	Hurtle Cell Carcinoma of the Thyroid
139	M	19	Hodgkin's Lymphoma
140	M	58	Colon Cancer
141	F	63	Breast Cancer
142	M	65	Colon Carcinoma
143	F	52	Breast Cancer
144	F	63	Cervix Cancer
145	M	74	Papillary Thyroid Carcinoma
146	M	62	Gastric Cancer
147	F	35	PNET
148	M	25	T-Cell Lymphoma
149	M	73	Gastric Adenocarcinoma
150	F	44	Hodgkin’s Lymphoma
151	F	61	Undifferentiated Pleomorphic Sarcoma
152	M	54	Renal Cell Carcinoma
153	F	62	Papillary Thyroid Carcinoma
154	F	35	Hodgkin’s Lymphoma
155	M	78	Lingual SCC
156	F	15	PNET
157	M	67	Rectal Cancer
158	F	61	Colorectal Carcinoma
159	M	50	Neuroendocrine Tumor
160	M	23	Germ cell tumor
161	F	29	Hodgkin’s Lymphoma
162	M	62	Rectal Cancer
163	M	43	Hodgkin’s Lymphoma

164	M	33	rectal adenocarcinoma
165	M	5	Wilm's tumor + Rhabdomyosarcoma
166	F	36	Colon Cancer
167	M	79	Laryngeal Cancer
168	M	36	Testis Tratuma
169	F	57	Papillary thyroid carcinoma
170	F	57	Colon Cancer
171	F	25	Ewing's Sarcoma
172	M	65	Colon Cancer
173	M	50	Suspicious for Vasculitis
174	F	62	Colon Carcinoma
175	F	26	Hodgkin's Lymphoma
176	F	58	Breast Cancer
177	M	78	Prostate Adenocarcinoma
178	F	50	Leiomyosarcoma
179	F	37	Adrenocortical Carcinoma/ Breast Cancer
180	F	38	Diffuse Large B cell Lymphoma
181	M	59	Papillary Thyroid Carcinoma
182	M	41	N/A
183	M	37	Non-Hodgkin Lymphoma
184	M	63	Adrenal Cortical Carcinoma
185	F	34	Colon Cancer
186	M	57	Papillary Thyroid Carcinoma
187	F	54	Uterine Leiomyosarcoma
188	F	31	Breast Cancer

189	M	18	Dendritic Cell Sarcoma
190	F	58	Renal Cell Carcinoma
191	M	72	Pancreatic Cancer
192	F	24	Hodgkin’s Lymphoma
193	M	76	Melanoma
194	M	60	Colon Adenocarcinoma
195	M	63	Rectal Cancer
196	F	45	Breast Cancer
197	F	58	Colon Carcinoma
198	F	58	Rectal Cancer
199	F	35	Sarcoidosis
200	M	45	Non-Hodgkin's Lymphoma
201	M	74	Colon Cancer
202	M	39	Melanoma
203	F	17	Hodgkin's Lymphoma
204	F	54	Ovarian Cancer
205	F	52	Colon Cancer
206	F	62	Colon Cancer
207	F	53	Leiomyosarcoma
208	F	60	Colon Cancer
209	M	82	Colon Cancer
210	F	49	Breast Cancer
211	F	54	Papillary Urothelial Carcinoma of Bladder
212	M	67	Rectosigmoid Cancer
213	F	33	SCC of cervix
214	M	62	Metastatic Colon Adenocarcinoma
215	F	43	Breast Cancer
216	M	68	Colon Cancer
217	M	14	Non-Hodgkin Lymphoma
218	M	66	Gastric Cancer

219	M	70	Breast Cancer
220	M	6	Wilms' tumour
221	M	58	Colorectal Cancer
222	M	51	Rectal Cancer
223	M	65	DLBL
224	M	60	Gastric Cancer
225	F	65	Breast Cancer
226	M	51	Gastric Adenocarcinoma
227	M	69	Colon Cancer
228	F	30	Breast Cancer
229	M	36	Renal Cell Carcinoma
230	F	71	Papillary Thyroid Carcinoma
231	F	42	Hodgkin's Lymphoma
232	F	53	Rectal cancer
233	F	63	Rectosigmoid Cancer
234	M	68	Gastric Cancer
235	M	61	Multiple Pulmonary Nodules
236	F	55	Colon Cancer
237	F	67	Evaluation of Unknown Origin Disease
238	M	51	Seminoma
239	M	65	Pancratic Cancer
240	F	73	Breast Cancer
241	M	65	Gastric Cancer

Table C. Demographic information for suspicious nodule class patients.

Patient number	**patient sex**	**patient age**	**malignancy**
1	F	40	Synovial Cell Sarcoma
2	M	62	N/A
3	F	66	Melanoma

4	F	66	Pancreatic Cancer
5	F	67	Colon Cancer and RCC
6	M	83	Gastric Cancer
7	F	56	NHL
8	M	60	Multiple Myeloma
9	F	48	Breast Cancer
10	F	36	Rectal Carcinoma
12	M	63	Non-Hodgkin's Lymphoma
13	F	52	Breast Cancer
14	M	65	Rectal Cancer
15	F	63	Pancreatic Mass
16	M	48	Rectal Cancer
17	F	51	Sigmoid colon cancer
18	M	63	Nasopharyngeal Cancer
19	M	76	Large B-cell Lymphoma
20	M	65	Rectal Cancer
21	F	41	Malignant melanoma
22	F	53	Breast Cancer
23	F	48	Breast cancer
24	M	44	Rectal Cancer
25	F	63	B Cell Lymphoma
26	F	31	Breast Cancer
27	M	36	Pancreatic Adenocarcinoma
29	F	60	Rectal Cancer
30	M	48	Rectal Cancer
31	M	32	Papillary Thyroid Carcinoma
32	M	28	Hodgkin's Lymphoma
33	F	75	Hodgkin's Lymphoma
34	M	54	Rectal Adenocarcinoma
35	F	50	Breast Cancer; status post right lumpectomy
36	M	63	CBD Adenocarcinoma

37	F	29	Chronic Pleuritis
38	M	65	Rectal Cancer
39	F	47	Breast Cancer
40	F	66	Melanoma
41	M	70	Gastric Cancer
42	F	58	Colon Adenocarcinoma
43	F	63	Ovarian Carcinoma
44	F	64	Serous Cyst Adenocarcinoma
45	M	86	Colon Cancer
46	M	62	Hodgkin's Lymphoma
47	F	56	Breast cancer
48	M	65	Leiomyosarcoma
49	F	59	Papillary Thyroid Carcinoma
50	F	34	Hodgkin's Lymphoma
51	M	21	Hodgkin`s Lymphoma
52	F	46	Breast Cancer
53	M	60	Neuroendocrine Carcinoma
54	M	63	Laryngeal SCC
55	F	50	Breast Cancer + Ovarian Cancer
56	F	51	Breast cancer
57	M	51	Hodgkin's Lymphoma
58	M	38	Hodgkin's Disease
59	M	60	Rectal Cancer
60	M	57	Follicular Lymphoma
61	F	55	Gastric Adenocarcinoma
62	F	66	Colon Carcinoma + Breast Carcinoma
63	M	56	Nonhodgkin's Lymphoma
64	F	37	N/A
65	F	60	Colon Cancer

66	M	53	Non-Hodgkin's Lymphoma
67	M	51	Nasal Cavity Melanoma
68	M	69	Thyroid Cancer
69	F	16	Hodgkin Lymphoma
70	F	35	Hodgkin's Lymphoma
71	F	77	Gastric cancer
72	F	55	Thyroid cancer
73	F	48	Breast Cancer
74	M	54	Non Hodgkin's Lymphoma
75	M	60	Colonic Cancer
76	F	57	Uterine cancer
77	F	79	Rectal Cancer
78	F	49	Ovarian cancer
79	F	69	Breast Cancer
80	F	65	Renal Cell Carcinoma
81	M	72	Melanoma
82	M	50	Vertebral Fibrous Tumor
83	F	54	Esophageal Cancer
84	F	71	Hodgkin's Lymphoma
85	M	50	Gastric Cancer
86	M	64	Colonic Cancer
87	M	66	Gastric Cancer
88	F	43	Breast Cancer
90	M	58	Non- Hodgkin's lymphoma
91	M	62	Melanoma
92	F	54	N/A
93	M	79	Malignant Melanoma
94	F	71	Metastatic Ovarian Mucinous Cystadenocarcinoma
95	F	54	Pancreas Adenocarcinoma

96	F	53	Breast Cancer
97	M	24	Hodgkin's Lymphoma
98	M	82	Colon Cancer
99	F	66	Medullary Thyroid Carcinoma
100	F	47	Rectal Cancer
101	F	38	PTC
102	F	45	Pulmonary Nodule
103	F	32	Adenocarcinoma of appendix+ovary
104	M	50	Vertebral Fibrous Tumor
105	M	48	Colon Cancer
106	M	43	Adenocarcinoma of Distal part of CBD
107	M	66	Colon Cancer
108	M	41	Colon Cancer
110	F	57	Metastatic Cancer
111	F	54	Breast Cancer
112	M	86	Colon Cancer
113	F	56	Solitary Pulmonary Nodule
114	F	55	Esophagus Cancer
115	F	56	Breast cancer
116	F	59	Papillary Thyroid Carcinoma
117	F	49	Ovarian cancer
118	F	61	Ovarian Cancer
119	M	70	Hodgkin Disease
120	F	61	Breast cancer
121	F	28	Non-Hodgkin's Lymphoma
122	F	64	Breast Cancer
123	F	50	Ovarian Cancer
124	F	69	Breast Cancer
125	M	58	Colon Adenocarcinoma

126	F	74	Plasmacytoma + Glassy Cell Carcinoma
127	M	71	Esophageal Cancer
128	F	45	Pulmonary Nodule
129	F	34	Colon Cancer
130	M	76	Hodgkin's Lymphoma
131	M	58	Gastric Cancer
132	F	43	Colon Cancer
133	M	65	Non-Hodgkin’s Lymphoma
134	F	53	B-Cell Lymphoma
135	M	59	Colon Cancer
136	M	73	Colon cancer
138	M	69	Papillary Urothelial Carcinoma
139	F	24	Non-Hodgkin’s Lymphoma
140	M	66	Pancreatic Cancer
141	M	59	Esophageal/Gastric Cancer
142	M	33	Hodgkin’s Lymphoma
143	M	65	Colon Cancer
144	M	72	Malignancy of Unknown Primary Origin
145	M	75	Papillary Thyroid Cancer
146	F	78	B-cell Lymphoma
147	M	52	Pancreatic Adenocarcinoma
148	M	55	Cholangiocarcinoma
149	M	53	Chronic Lymphoid Leukemia
150	F	56	Colon Cancer
151	M	60	Colon Cancer
152	M	61	Colon Cancer

153	F	36	Hodgkin's Lymphoma
154	F	26	Hocking Lymphoma
155	M	58	Rectal Cancer
156	M	47	Colon Cancer
157	F	53	Breast Cancer
158	M	38	Colon Cancer
159	F	39	Colon Cancer
160	M	73	RCC
161	M	47	Medullary Thyroid Carcinoma
162	F	72	Increased CA 15-3 Level
163	F	24	Hodgkin's disease
164	F	58	Rule out of Sarcoidosis
165	M	62	Rectal Adenocarcinoma
166	F	48	Rectal Cancer
167	M	46	Colorectal Cancer
168	F	43	Breast Cancer and Ovarian Cancer
169	M	32	Multiple Pulmonary Nodules, Possibility of Metastases
170	M	67	Pancreatic Cancer
171	M	51	Plasmacytoma
172	M	63	Colon Cancer
174	F	68	Cholangiocarcinoma
175	M	72	Rectal Adenocarcinoma
176	M	81	Hodgkin Lymphoma
177	F	62	Breast Cancer
178	F	42	Ovarian Cancer
179	F	63	Ovarian Cancer
180	F	60	Gastric Lymphoma
181	M	57	Colon Cancer
182	F	41	Breast Cancer
183	F	64	Papillary Thyroid Cancer

184	M	52	Rectal Cancer
185	M	50	Colon Cancer
186	F	75	Pulmonary Nodules
187	F	63	Colon Cancer
188	F	62	Colon Cancer
189	F	65	Unknown
190	M	37	Pancreatic Adenocarcinoma
191	F	72	Breast Cancer
192	M	69	Gastric Cancer
193	M	66	Gastric Cancer (cardia with liver metastases)
194	M	43	Non-Hodgkin's Lymphoma
195	F	33	Gastric Cancer
196	M	61	Colon Cancer
197	F	50	Rectal Cancer
198	M	55	Hodgkin Lymphoma
199	M	64	Non-Hodgkin Lymphoma
200	M	64	Diffuse Large B-Cell Lymphoma
201	M	53	Suspected IgG4-related Disease/Sarcoidosis
202	M	29	Testis Teratoma
203	M	63	Colon Cancer and Multiple Myeloma
204	M	60	Rectal Cancer
205	M	31	Burkitt Lymphoma
206	F	47	Breast Cancer
207	F	64	Uterine Cancer
208	F	69	Colon Cancer
209	F	29	Hodgkin's Lymphoma
210	F	39	Breast Cancer
211	F	59	Cervix Cancer
212	M	64	Colon Cancer

213	M	16	Hodgkin Lymphoma
214	F	32	Non-Hodgkin Lymphoma
215	M	42	R/O of Takayasou
216	F	48	Rule of vasculitis
217	F	40	Hodgkin's Lymphoma
218	F	50	Breast Cancer
219	M	17	Hodgkin's Lymphoma
220	M	66	Gastric Cancer (cardia with liver metastases)
221	F	48	R/O Malignancy
222	F	50	Breast Cancer
223	M	49	Gastric Cancer
224	F	6	Neuroblastoma
225	M	64	Diffuse Large B-Cell Lymphoma
226	M	56	Colon Cancer
227	M	60	GIST of the Stomach
228	F	67	Colon Cancer
229	M	63	Colon Cancer and Multiple Myeloma
231	M	61	Gastric Cancer
232	F	71	Breast Cancer; Status post right mastectomy
233	F	66	Breast Cancer
234	M	31	Burkitt Lymphoma
235	F	26	Hodgkin Disease
236	F	40	Hodgkin's Lymphoma
237	F	44	Breast Cancer
238	F	62	Sarcoidosis
239	M	14	Non-Hodgkin's Lymphoma
240	F	37	Osteosarcoma
241	M	61	Esophageal Cancer
242	M	49	Esophageal Cancer
243	M	57	Cholangiocarcinoma

244	F	69	Breast Cancer
245	M	58	Ampulla of Vater Cancer
246	M	45	Colon Cancer
247	M	55	Colon Cancer
248	M	63	Cardia Cancer
249	F	66	Colon Cancer
250	M	53	Solitary Pulmonary Nodule
251	F	62	Colon Cancer
252	F	53	Hodgkin's Lymphoma
253	M	71	Rectal Cancer and Lymphoma
254	F	70	N/A
255	M	52	Maltoma
256	F	68	Ovarian Serous Adenocarcinoma
257	F	66	Breast Cancer
258	F	53	Renal Cell Carcinoma
259	M	38	Gastric Adenocarcinoma
260	F	77	Carcinosarcoma of Uterus
261	F	61	Hodgkin’s Lymphoma
262	M	66	Metastatic Neuroendocrine Tumor
263	M	52	Hodgkin’s Lymphoma/Restaging
264	M	60	Papillary Thyroid Carcinoma
265	F	63	Breast Cancer
266	F	27	Rectal Cancer
267	F	63	Breast Cancer
268	F	76	Colon Cancer
269	F	31	Melanoma
270	F	44	Breast Cancer

271	F	37	Orbit Mass
272	F	38	Unknown
273	M	35	Papillary Thyroid Carcinoma
274	F	63	Papillary carcinoma of thyroid gland
275	M	53	Papillary Thyroid Cancer
276	F	71	DLBCL
277	M	59	Colon cancer
278	M	58	Diagnosis
279	F	32	Non-Hodgkin' lymphoma
280	F	50	Breast Cancer
281	F	54	Breast Cancer
282	F	68	Colon Cancer
283	M	63	Esophageal Cancer
284	M	51	Gastric Cancer
285	M	16	Small Round Cell Tumor
286	F	71	Lymphoma
287	M	25	Osteosarcoma
288	F	80	Colon Cancer
289	M	63	Hodgkin's Lymphoma
290	M	61	Esophageal Cancer
291	F	14	Osteosarcoma
292	F	75	Abdominal Mass with Unknown Primary Origin
293	M	70	Esophagus Cancer
294	F	64	Breast Cancer
295	M	58	Melanoma
296	F	58	Rectosigmoid Cancer
297	F	50	Hodgkin's Lymphoma
298	F	73	Unknown Primary Malignancy

299	M	61	Rectal Cancer
300	F	53	Rectal Cancer
301	F	72	Rectal Carcinoma
302	M	41	Evaluation of malignancy
303	F	55	Thyroid cancer
304	M	31	Papillary Thyroid Carcinoma
305	M	62	Urinary Bladder Cancer
306	M	63	Gall bladder Adenocarcinoma
307	M	73	Esophageal Cancer
308	F	48	Breast Cancer
309	F	59	Breast Cancer
311	F	63	Breast Cancer
312	F	39	Cervix Cancer
313	M	52	Gastric Cancer
314	F	56	B Cell Lymphoma
315	M	46	Colon Cancer
316	F	53	Colorectal Cancer
317	F	34	Hodgkin’s Lymphoma
318	F	54	Uterine Cancer
319	F	54	Rectal Cancer
320	F	47	Breast Cancer
321	M	58	Colon Cancer
322	M	30	Germ-Cell Tumor of Testis
323	M	74	Gastric cancer
324	M	54	Non-Hodgkin’s Lymphoma
325	M	59	Colon Cancer
326	M	63	Renal Cell Cancer
327	M	66	Hepatic Cholangiocarcinoma
328	M	54	Metastatic Carcinoma
329	F	42	Breast Cancer

330	M	57	SCC of Tongue
331	F	48	Hodgkin Lymphoma
332	M	64	Pulmonary Nodule
333	F	40	Hodgkin's Lymphoma
334	M	76	Non-Hodgkin's Lymphoma
335	F	67	Breast cancer
336	F	48	Breast Cancer
337	F	55	Urotheliac Carcinoma
338	F	65	Colon Cancer
339	F	82	Hodgkin's Lymphoma
340	F	53	Breast Cancer
341	M	72	RCC
342	F	61	Breast cancer
343	F	43	Esophageal Cancer
344	M	62	Hodgkin's Lymphoma
345	F	53	Lymphoma
346	F	78	Colon Cancer
347	M	44	Rectal Cancer
348	F	37	Hodgkin's Lymphoma
349	F	54	Papillary Thyroid Cancer
350	F	63	Papillary carcinoma of thyroid gland
351	F	57	Uterine cancer
352	F	38	Colon Cancer
353	M	59	Gastric MALToma
354	F	34	Hodgkin Lymphoma
355	F	38	Rectal Cancer
356	F	69	Gastric Cancer
357	M	25	Hodgkin's Lymphoma
358	M	62	Large B-cell Lymphoma
359	F	60	Multiple Myeloma
360	M	42	Hodgkin's Lymphoma and Rectal Cancer

361	F	31	Hodgkin's Lymphoma
362	M	50	Colon Cancer
363	M	48	Rectal Cancer
364	M	68	Esophageal Squamous Cell Carcinoma
365	M	70	Non-Hodgkin's Lymphoma (DLBCL)
366	M	78	Hodgkin's Lymphoma
367	M	55	B Cell Lymphoma
368	M	57	Polyneuropathy
369	F	56	Non Hodgkin Lymphoma
370	M	57	Papillary Thyroid Carcinoma
372	F	48	Breast Cancer
373	F	65	Colon Cancer
375	M	68	Adenocarcinoma of colon
376	F	80	Breast Cancer
377	M	39	Hodgkin's Lymphoma
378	F	54	Gastric cancer
379	F	24	Non-Hodgkin's Lymphoma
380	M	66	Pancreatic Cancer
381	F	62	Evaluation of Single Pulmonary Nodule (SPN)
382	M	50	Colon Cancer
383	M	42	Hodgkin's Lymphoma and Rectal Cancer
384	M	65	Colon Cancer
385	F	51	Rectal Adenocarcinoma
386	M	63	Nasopharyngeal Cancer
387	M	54	Hemangiopericytoma
388	M	55	Cholangiocarcinoma
389	F	51	Breast Cancer

390	F	75	Ovarian Cancer
391	F	55	Colon Cancer
392	M	79	Adenocarcinoma of rectosigmoid
393	M	53	Squamous Cell Carcinoma of the Tongue
394	M	61	Colon Cancer
395	M	71	Metastatic Adenocarcinoma Of The Right Humerus
396	M	52	Colon Cancer
397	F	50	Breast Cancer
398	M	70	Unknown origin malignancy
399	F	30	PTC
400	F	52	Melanoma
401	F	54	Breast Cancer
402	F	75	Adenocarcinoma of common bile duct
403	F	65	Colon Adenocarcinoma
404	M	75	Melanoma
405	F	24	Hodgkin’s disease
406	F	58	Evaluation of Abdominal and Pelvic Mass
408	M	67	Colon Cancer
409	F	37	Hodgkin’s Lymphoma
410	M	53	Colon Cancer
411	M	39	N/A
412	M	58	Duodenal Adenocarcinoma
413	F	14	Hodgkin’s Lymphoma
414	M	4	Wilm’s tumor
415	M	79	SCC of the lip
416	M	29	Malignant Mixed Germ cell tumor of Tesis

417	M	31	Colon Cancer
418	M	68	Gastric Cancer
419	M	30	Germ Cell Tumor
420	F	48	Rectal Cancer
421	F	56	Non Hodgkin Lymphoma
422	F	24	Hodgkin's Lymphoma
423	F	54	Colon Cancer
424	F	66	Breast Cancer
425	M	41	GIST of stomach
426	F	54	Gastric cancer
427	M	28	Malignant Melanoma
428	M	34	Gastric Cancer
430	F	47	Breast Cancer
431	M	65	Colon Cancer
432	F	64	Gastric Adenocarcinoma
433	F	40	Endometrial Cancer
434	F	27	Hodgkin disease
435	M	70	Gastric Cancer
436	F	57	Cervix Cancer
437	M	42	Non-Hodgkin’s Lymphoma
438	M	56	Evaluation of Hypoglycemia
439	F	16	Ovarian Teratoma
440	F	47	Rectal Carcinoma
441	M	70	Gastric Cancer
442	F	49	Ovarian cancer
443	F	49	Breast Carcinoma
444	F	61	Ovarian Cancer
445	M	70	Hodgkin Disease
446	F	28	Non-Hodgkin’s Lymphoma
447	F	48	Colon Cancer

448	M	64	Non-Hodgkin's Lymphoma
449	F	39	Breast Cancer
450	F	61	Renal Cell Carcinoma
451	M	72	Breast Cancer
452	F	44	Pancreas Cancer
453	M	50	Gastrointestinal Stromal Tumor (GIST)
454	F	47	Uterine Cancer
455	F	53	Breast Cancer
456	M	76	Gastric Adenocarcinoma
457	M	72	Esophageal/Gastric Cancer
458	M	62	Metastatic Mucinous Adenocarcinoma of Testis
459	F	65	Breast Cancer
460	M	60	Gastric Cancer
461	F	26	Lymphoma
462	F	69	Carsinosarcoma (Malignant Mixed Mullerian Tumor)
463	F	71	Breast Cancer
464	F	63	Breast Cancer
465	F	38	Breast Cancer
466	F	45	Evaluation of Pulmonary Nodule
467	F	55	Solitary Pulmonary Nodule
468	F	66	SPN
469	M	64	Gastric Cancer
470	F	70	Papillary Thyroid Carcinoma
471	F	40	Uterine Leiomyosarcoma
472	F	69	Colon Cancer

473	F	59	Colon Cancer
474	M	73	Colon Cancer
475	F	40	Breast Cancer
476	M	69	Meningioma
477	M	8	Osteosarcoma
478	F	67	Gastric Cancer
479	M	72	Melanoma
480	F	61	Non-Hodgkin's Lymphoma
481	F	59	Breast Cancer
482	F	49	Breast Cancer
483	M	71	Non-Hodgkin's Lymphoma
484	M	36	Esophageal Carcinoma (SCC)
485	F	57	Breast Cancer
486	M	33	Seminoma
487	F	44	Rectal Musinous Adenocarcinoma
488	F	55	Uterine Cancer
489	M	60	Hodgkin's Lymphoma
490	M	72	Colon cancer
491	F	31	Low-grade Endometrial sarcoma
492	F	52	Breast Cancer
493	F	50	Breast Cancer
494	F	64	Ovarian Cancer
495	M	22	DLBL
496	F	31	Papillary Thyroid Carcinoma
497	F	32	Adenocarcinoma of appendix+ovary
499	F	59	Phyllodes tumors
500	F	52	Sigmoid Cancer
501	M	58	Gastric Cancer

502	F	70	Metastatic Breast Cancer
503	F	71	Rectal adenocarcinoma
504	M	60	Gastric Cancer
505	F	56	Non Hodgkin Lymphoma
506	F	63	B Cell Lymphoma
507	M	65	Rectal Cancer
508	M	86	Colon Cancer
509	M	62	Hodgkin's Lymphoma
510	F	55	Esophagus Cancer
511	M	65	Leiomyosarcoma
512	F	50	Malignant Melanoma
513	M	59	Papillary Thyroid Carcinoma
514	F	34	Hodgkin's Lymphoma
516	M	21	Hodgkin`s Lymphoma
517	F	51	Unknown Malignancy
518	F	41	Hodgkin' lymphoma
520	F	41	Colon Cancer
521	F	68	Rectal Cancer
522	M	53	Non-Hodgkin's Lymphoma
523	F	61	N/A
524	F	48	Breast Cancer
525	M	42	Seminoma
526	M	45	mediastinal and hilar lymphadenopathies
527	F	77	Colon Cancer
528	M	37	PTC
529	F	32	Breast Carcinoma
530	F	82	Breast Cancer
531	M	30	Hodgkin's lymphoma
532	F	67	Breast cancer
533	F	53	Colon Cancer
534	M	83	Bladder Carcinoma

535	F	53	Hodgkin Lymphoma
536	M	32	Osteosarcoma
537	F	27	Hodgkin's Lymphoma
539	M	59	Esophageal Carcinoma
540	M	54	Chondrosarcoma
541	F	20	Ewing Sarcoma
542	F	45	Hodgkin's Lymphoma
543	F	86	Diffuse large B cell lymphoma
544	M	54	Colon Cancer
545	M	31	Nasopharyngeal Adenoid Cystic Carcinoma
546	F	30	Non-Hodgkin's Lymphoma
547	F	30	Liposarcoma
548	M	62	Salivary Gland Ductal Carcinoma
549	M	66	Colon Cancer
550	M	53	Rectosigmoid Adenocarcinoma
551	F	70	Transitional Cell Carcinoma
552	F	73	Esophageal cancer
553	M	53	Bladder Cancer, Transitional Cell Carcinoma
554	F	58	Carcinoma of the Ampulla of Vater
555	F	59	Ampulla of Vater Carcinoma
556	M	39	Hodgkin's Lymphoma
557	M	46	Malignant Melanoma
558	M	73	Colon Cancer