ID N. 32971

UNIVERSITÀ CAMPUS BIO-MEDICO DI ROMA

DEPARTMENT OF ENGINEERING

HUMANITAS UNIVERSITY

DEPARTMENT OF BIOMEDICAL SCIENCES

Italian National Ph.D. in Artificial Intelligence
Health and Life Sciences
XXXVII Cycle

Deep Learning-Driven Classification of NSCLC Histological Subtypes Using PET and CT Imaging

Supervisors

Prof. Letterio Salvatore Politi

Prof. Arturo Chiti

Prof. Paolo Soda

Candidate

Fatih Aksu

December, 2025

To my family and friends.

Abstract

Lung cancer remains the leading cause of cancer-related deaths worldwide, with Non-Small Cell Lung Cancer (NSCLC) accounting for approximately 85% of all cases. Among its subtypes, Adenocarcinoma (ADC) and Squamous Cell Carcinoma (SQC) are the most prevalent, each associated with distinct prognoses and treatment pathways. Accurate histological subtype classification is thus critical for effective personalized therapy. However, current standard procedures rely on invasive biopsies, which may be unsuitable or risky for certain patients. Moreover, biopsy samples are often limited in size or affected by tumor heterogeneity, leading to diagnostic uncertainty. These limitations have motivated the search for non-invasive, imaging-based methods to predict histological subtypes.

To address this challenge, this thesis explores a series of deep learning strategies that leverage radiological imaging, particularly Computed Tomography (CT) and Positron Emission Tomography (PET) scans, to develop robust and accurate classification systems. The work systematically addresses key barriers in this domain, including limited dataset sizes, data privacy concerns, and the challenge of effectively integrating multimodal information.

The first part of the thesis investigates whether triplet networks, a type of metric learning model, can overcome the challenge of learning from small datasets, which is a common issue in medical imaging due to privacy concerns, annotation costs, and limited patient availability. Unlike conventional deep networks that rely on softmax classifiers and learn directly from individual samples, triplet networks learn by modeling the relationships between samples, specifically by comparing an anchor to both a similar (positive) and a dissimilar (negative) example. This relational learning framework effectively increases the number of informative training instances and enhances the model's ability to generalize. On a dataset of 87 NSCLC patients, triplet networks significantly outperformed these standard models, demonstrating their ability to learn discriminative representations even under limited data conditions.

Although triplet networks improve learning under limited data conditions, increasing the amount of training data remains critical for further enhancing model performance. However, acquiring additional data from external sources is often hindered by concerns related to patient privacy, data ownership, and regulatory compliance. To overcome these challenges, the second part of the thesis introduces a federated learning framework combined

with triplet loss. This approach enables multiple clinical institutions to collaboratively train a model without sharing patient data, thereby preserving privacy. Comparative experiments demonstrate that this federated learning method outperforms both isolated local training and federated models trained with conventional softmax loss, highlighting its potential to improve generalization while maintaining data confidentiality.

Having addressed the issues of small datasets and data sharing, the thesis then turns to multimodal learning, aiming to further improve classification performance by combining structural and metabolic information from CT and PET images. The third part introduces MINT, a novel Multi-stage INTermediate fusion architecture. MINT fuses PET and CT features at multiple stages of the feature extraction process, allowing the network to capture complementary cues across different levels of abstraction while preserving spatial correlations. The model is benchmarked against unimodal baselines, early and late fusion strategies, and the only other existing intermediate fusion method for this task. MINT achieves the highest performance across all comparisons, validating the advantage of intermediate fusion in leveraging multimodal imaging data for subtype classification.

Finally, while each of the previous contributions focused on developing specialized models, the final part of the thesis explores the potential of foundation models in this domain. These large-scale pretrained models have shown promise in generalizing across tasks with minimal fine-tuning. The study evaluates three 3D medical foundation models trained on diverse datasets and compares their performance against specialized architectures developed specifically for NSCLC subtype classification. Using a dataset of 714 patients, all foundation models outperformed the task-specific baselines, highlighting their potential for enabling accurate and data-efficient solutions in clinical imaging tasks.

Collectively, the contributions of this thesis address critical challenges in building non-invasive, accurate, and scalable systems for NSCLC histological subtype classification. By integrating strategies for learning from limited data, protecting privacy, fusing multimodal information, and harnessing pretrained models, the proposed approaches pave the way toward more accessible and precise diagnostic tools in lung cancer care.

Contents

List of Figures

List of Tables

Acronyms

ADC	Adenocarcinoma
AI	Artificial Intelligence
ANN	Artificial Neural Network
AUROC	Area Under Receiver Operating Characteristic Curve
BRISQUE	Blind/Referenceless Image Spatial Quality Evaluator
CLIP	Contrastive Language-Image Pretraining
CNN	Convolutional Neural Network
CT	Computed Tomography
DL	Deep Learning
EHR	Electronic Health Record
FC	Fully Connected
FDG	Fluorodeoxyglucose
FL	Federated Learning
GAP	Global Average Pooling
GTV	Gross Tumor Volume
HFL	Horizontal Federated Learning
HU	Hounsfield Units
kNN	k-Nearest Neighbors
LCC	Large Cell Carcinoma

LoRA	Low-Rank Adaptation
LR	Logistic Regression
LSTM	Long Short-Term Memory
MDL	Multimodal Deep Learning
ML	Machine Learning
MRI	Magnetic Resonance Imaging
NB	Naive Bayes
NOS	Not Otherwise Specified
NSCLC	Non-Small Cell Lung Cancer
PCA	Principal Component Analysis
PET	Positron Emission Tomography
RBF	Radial Basis Function
RF	Random Forest
ROI	Region of Interest
SCC	Small Cell Carcinoma
SQC	Squamous Cell Carcinoma
SUV	Standard Uptake Value
SVM	Support Vector Machine
TFL	Transfer Federated Learning
TSM	Triplet Selection Method
VFL	Vertical Federated Learning
VOI	Volume of Interest

Chapter 1

Introduction

1.1 NSCLC Histological Subtype Classification

Lung cancer is a leading cause of cancer-related deaths globally, with estimated age-adjusted incidence and mortality rates of 23.6 and 16.8 per 100,000 people, respectively [96]. Non-Small Cell Lung Cancer (NSCLC) accounts for 85% of primary lung cancers, with Adenocarcinoma (ADC) and Squamous Cell Carcinoma (SQC) being the most common subtypes [101]. The two primary histological subtypes not only have different biological characteristics and outcomes, but also different responses to targeted therapies and immunotherapies [20, 17]. In the context of early-stage NSCLC, a full histological examination of the primary tumour prior to surgery may be omitted in cases where there is a significant risk of biopsy-related complications and a compelling clinical indication of malignancy based on imaging and clinical findings. However, an accurate pathological diagnosis of the primary tumour is essential to determine prognosis and select the most effective therapeutic strategies in patients with clinical stage I-III disease [77]. Traditional methods of identifying these subtypes rely on tissue biopsy and histopathological examination, which are invasive and can carry significant risks for patients [26]. Moreover, such techniques often struggle with accuracy due to challenges like small tumor size, the tumor location near the lung's edges or critical structures, and the diverse characteristics of tumors, which can lead to inconsistent results [43]. These challenges, along with the limitations of current invasive diagnostic methods and the need to avoid such procedures, drive the search for non-invasive approaches to accurately classify NSCLC histological subtypes.

Positron Emission Tomography (PET)/Computed Tomography (CT) using the [^{18}F] Fluorodeoxyglucose (FDG) tracer plays a pivotal role in the diagnosis and management of lung cancer, with most patients undergoing this imaging modality prior to the initiation of treatment [93]. By integrating the metabolical imaging capabilities of PET with the detailed anatomical imaging from CT, PET/CT offers enhanced precision in tumor staging, signif-

icantly improving the detection and localization of loco-regional pathological lymph nodes and distant metastases [7]. Moreover, various subtypes of NSCLC exhibit differing characteristics in these radiological images. However, the limited specificity of these features makes it difficult for radiologists to accurately differentiate between NSCLC subtypes [56].

1.2 Artificial Intelligence in Medical Imaging

In recent years, the rapid advancement of Artificial Intelligence (AI), particularly Deep Learning (DL) algorithms, has revolutionized numerous fields, with computer vision experiencing a remarkable transformation in both capability and precision. The success of AlexNet [61] in the ImageNet competition [82] in 2012 marked a historical milestone, establishing Convolutional Neural Network (CNN)s as the go-to architecture for image classification. The subsequent evolution of CNN architectures, such as VGG [87], ResNet [44], MobileNet [48], DenseNet [49], and EfficientNet [94] etc., has further propelled advancements. Following this foundational progress, CNNs have been widely adopted across a diverse range of core tasks, such as image classification, object detection, and semantic segmentation. These techniques are foundational to numerous applications across diverse domains, including autonomous driving, manufacturing, agriculture, security, and, notably, medicine.

In the medical domain, computer vision plays a critical role in supporting clinical decision-making by automatically analyzing imaging data. Applications of computer vision in medicine are broad and span nearly all clinical disciplines. In radiology, automated systems assist in detecting abnormalities such as fractures in X-rays [81], pulmonary embolisms in CT angiography [88], or brain lesions in Magnetic Resonance Imaging (MRI) scans [83]. In gastroenterology, computer vision techniques are applied to endoscopic and colonoscopic images to detect and classify abnormalities such as polyps, ulcers, or bleeding, supporting early diagnosis and reducing oversight during procedures [37]. Ophthalmology has benefited significantly from image-based classification models capable of identifying diabetic retinopathy [53], macular degeneration [76], or glaucoma [108] from retinal fundus images. Dermatology employs AI models to classify skin lesions from dermoscopic images [52], aiding in the early detection of melanoma and other skin conditions.

Within this broad landscape, cancer imaging has emerged as a particularly active and impactful area of research in computer vision. Medical image classification, in particular, plays a central role in oncology by enabling the automated identification of malignancies, assessment of tumor grade, and estimation of patient prognosis from imaging data. These applications span a wide range of cancer types and imaging modalities. For example, deep learning models have been developed to detect breast cancer from mammograms [84], classify brain tumors using MRI [83], and identify liver lesions from CT [40] or ultrasound [25] images.

In lung cancer, chest CT scans are widely used not only for detecting nodules but also for assessing their malignancy [109] and guiding biopsy decisions [91]. PET imaging, on the other hand, provides valuable metabolic information that supports tumor grading and treatment response evaluation [104].

1.3 Challenges and Research Questions

Considering the challenges and limitations of current invasive diagnostic methods for histological subtype classification, along with the growing success of deep learning techniques in medical image analysis, this thesis investigates whether NSCLC subtypes can be classified from non-invasive imaging modalities such as CT and PET using deep learning methods. In doing so, it addresses several key challenges prevalent in medical image analysis, including data scarcity, privacy constraints, and the complexity of effectively integrating multimodal information, and further explores the potential of large-scale pretrained foundation models for the classification of NSCLC histological subtypes. The research dimensions corresponding to these challenges are summarized in Figure 1.1 and are discussed in detail in the following paragraphs.

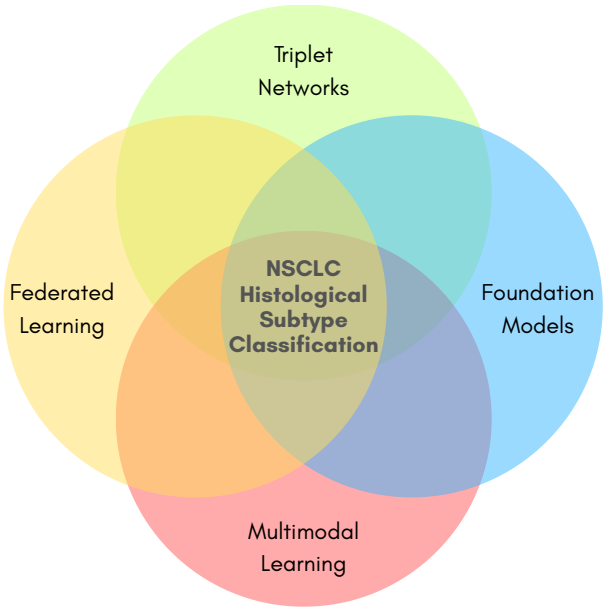

Figure 1.1: Overview of the research dimensions addressed in this thesis.

Data Scarcity

RQ1: Can triplet networks improve classification performance under limited data conditions?

The first challenge is data scarcity, which is a prevalent issue in medical imaging due to privacy constraints, high annotation costs, and limited patient availability. DL models typically require large and diverse datasets to achieve robust generalization, yet in medical contexts, collecting such data is often infeasible due to ethical, logistical, and institutional limitations. To address this problem, **we hypothesize that triplet networks, a form of metric learning model, can improve classification performance compared to traditional architectures**. Our hypothesis stems from observing that, unlike conventional deep networks, which learn directly from individual samples using a softmax classifier, triplet networks adopt a relational learning approach. In this framework, each sample (referred to as the anchor) is compared with a similar sample (positive) and a dissimilar sample (negative). By explicitly modeling these relationships, the network effectively learns the relative similarity between instances, rather than focusing solely on individual labels. This strategy increases the number of informative training examples derived from the same dataset and enhances the model's ability to generalize to unseen data.

Privacy Constraints

RQ2: Can federated learning with triplet loss maintain performance while ensuring data privacy?

The second challenge is the limited access to external data, which restricts the ability to train more robust DL models. Although triplet networks can partially reduce the effects of small datasets, the lack of sufficient training data still limits model generalization. In medical applications, sharing data across institutions is often restricted due to patient privacy concerns, data ownership issues, and regulatory compliance. To address this problem, **we hypothesize that integrating federated learning with triplet networks can preserve privacy without compromising performance.** Federated learning is a collaborative training approach in which multiple institutions train a shared model locally on their own data, while only model updates (and not the patient data itself) are exchanged and aggregated. This enables the model to learn from a larger and more diverse dataset while maintaining privacy and performance.

Integrating Multimodal Information

RQ3: How can CT and PET features be effectively integrated for NSCLC subtype classification?

The third challenge addressed in this thesis is effectively integrating complementary information from multiple imaging modalities. In NSCLC, CT scans provide structural information about the tumor and surrounding tissues, while PET scans capture metabolic activity that may reflect tumor heterogeneity or disease progression. Combining these modalities has the potential to offer a richer and more informative representation than either modality alone. However, the main challenge in multimodal learning is effectively leveraging this complementary information to improve predictive performance. To address this challenge, **we hypothesize that MINT, a novel Multi-stage INTermediate fusion architecture, can exploit the strengths of both imaging modalities to improve classification performance compared to unimodal models and conventional fusion strategies.** Unlike traditional early or late fusion methods, which combine features only at the input or decision level, MINT fuses PET and CT features at multiple stages throughout the feature extraction process. This multi-stage integration enables the network to capture complementary information at different levels of abstraction while maintaining spatial consistency across modalities.

Foundation models

RQ4: Can large-scale pretrained foundation models effectively predict histological subtypes in NSCLC?

The final part of the thesis explores the potential of large-scale pretrained foundation models for NSCLC histological subtype classification. These models, trained on massive and diverse datasets, have demonstrated strong generalization capabilities with minimal fine-tuning. This study evaluates several 3D medical foundation models and compares them with the task-specific architectures developed in this thesis to assess their effectiveness on this specialized clinical task.

1.4 Chapter Organization

The rest of this thesis is organized as follows.

Chapter 2 presents a comprehensive review of related work, focusing on imaging-based methods for NSCLC subtype classification. It covers existing approaches involving radiomics and deep learning.

Chapter 3 describes the datasets used throughout the thesis. It includes details on data acquisition, patient demographics, and imaging modalities.

Chapter 4 addresses RQ1 by investigating the use of triplet networks to improve model performance under limited data conditions. This chapter introduces a metric learning framework that models relationships between samples rather than relying on standard softmax classifiers.

Chapter 5 addresses RQ2 by introducing a federated learning approach that combines triplet loss with decentralized training. This framework enables multiple institutions to collaboratively train models without exchanging patient data, thereby addressing data privacy concerns.

Chapter 6 addresses RQ3 by proposing a novel multi-stage intermediate fusion architecture that integrates PET and CT features during multiple stages of feature extraction. This approach improves classification by capturing complementary spatial and semantic information across modalities.

Chapter 7 addresses RQ4 by evaluating the use of large-scale foundation models pretrained on diverse medical imaging datasets. It benchmarks their performance against the task-specific models, highlighting their generalization capabilities.

Chapter 8 concludes the thesis by summarizing key findings, discussing limitations, and suggesting potential directions for future research.

1.5 Related Publications

- **Fatih Aksu**, Fabrizia Gelardi, Arturo Chiti, and Paolo Soda, "Early Experiences on Using Triplet Networks for Histological Subtype Classification in Non-Small Cell Lung Cancer," *IEEE 36th International Symposium on Computer-Based Medical Systems (CBMS)*, 2023.

- **Fatih Aksu**, Ermanno Cordelli, Fabrizia Gelardi, Arturo Chiti, and Paolo Soda, "Enhancing NSCLC Histological Subtype Classification: A Federated Learning Approach Using Triplet Loss," *3rd International Workshop on Artificial Intelligence for Healthcare Applications (AIHA) at the International Conference on Pattern Recognition (ICPR)*, 2024.

- **Fatih Aksu**, Fabrizia Gelardi, Arturo Chiti, and Paolo Soda, "Toward a Multimodal Deep Learning Approach for Histological Subtype Classification in NSCLC," *7th Workshop on Artificial Intelligence Techniques for BioMedicine and HealthCare (AIBH) at the IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2024.

- Valerio Guarrasi, **Fatih Aksu**, Camillo Maria Caruso, Francesco Di Feola, Aurora Rofena, Filippo Ruffini, and Paolo Soda, "A systematic review of intermediate fusion in multimodal deep learning for biomedical applications," *Image and Vision Computing*, 2025.

- **Fatih Aksu**, Fabrizia Gelardi, Arturo Chiti, and Paolo Soda, "Multi-stage intermediate fusion for multimodal learning to classify non-small cell lung cancer subtypes from CT and PET," *Pattern Recognition Letters*, 2025.

- **Fatih Aksu**, Fabrizia Gelardi, Arturo Chiti, and Paolo Soda, "NSCLC histological subtype classification from CT scans using generalist 3D medical foundation models," *IEEE International Conference on Healthcare Informatics (ICHI)*, 2025.

Chapter 2

Related Work

Understanding the histological subtypes of NSCLC is crucial for determining prognosis and tailoring treatment strategies. Over the years, a wide range of computational approaches have been proposed to address this classification problem using medical imaging. These approaches can broadly be categorized into traditional radiomics pipelines and modern deep learning frameworks. This chapter reviews representative methods from both domains, with particular emphasis on deep learning approaches, which constitute the main focus of this thesis, highlighting the evolution from hand-crafted feature extraction to end-to-end neural networks, and the shift from unimodal to multimodal imaging analysis.

2.1 Radiomics Approaches

Radiomics, a quantitative approach that extracts a large number of features from medical images, has been extensively investigated in conjunction with traditional Machine Learning (ML) techniques for non-invasively characterizing lung cancer histopathology. These methods typically involve segmenting the Region of Interest (ROI), extracting diverse features (e.g., shape, first-order statistics, texture, wavelet-based), selecting the most informative features, and then training ML classifiers.

Several studies have focused on classifying NSCLC subtypes using CT-derived radiomics. Ferreira Junior et al. [57] utilized CT-based radiomic features encompassing size, shape, intensity, and texture, coupled with ML models like k-Nearest Neighbors (kNN), and Radial Basis Function (RBF)-based Artificial Neural Network (ANN), and Naive Bayes (NB) to predict lung cancer histopathology. Liu et al. [68] also employed CT radiomics, evaluating four different ML classifiers, including Random Forest (RF), Logistic Regression (LR), LR with L1 regularization, and LR with Principal Component Analysis (PCA) in their comparative study for differentiating ADC and SQC.

The scope of radiomics has extended to multi-subtype classification as well. Liu et al. [69]

developed the "SLS model" to distinguish between ADC, SQC, Large Cell Carcinoma (LCC), and Not Otherwise Specified (NOS) using CT radiomics. Their approach involved a comprehensive pipeline comprising the synthetic minority oversampling technique, L2,1-norm minimization, and Support Vector Machine (SVM)s. Similarly, Khodabakhshi et al. [58] proposed a high-dimensional CT radiomics signature for phenotyping NSCLC subtypes, including ADC, SQC, LCC, and NOS. They employed feature selection techniques such as a wrapper algorithm and multivariate adaptive regression splines, followed by multivariable multinomial logistic regression.

The integration of multimodality imaging, particularly PET/CT, has shown promise in enhancing radiomic analyses. Han et al. [43] performed an extensive evaluation using PET/CT images to differentiate NSCLC histologic subtypes (ADC vs. SQC). They extracted a large set of 688 features and systematically compared ten feature selection methods with ten different ML models (including LR, SVM, RF, NB, kNN, AdaBoost, and XGBoost), providing a detailed landscape of model performance. Yan and Wang [106] also leveraged combined CT and FDG-PET radiomics for the histological diagnosis of solitary pulmonary nodules, differentiating between ADC, SQC, and metastases.

These studies collectively demonstrate the significant role of radiomics combined with diverse ML strategies in extracting valuable diagnostic information from standard medical images. While effective, these pipelines typically rely on manual feature engineering and have limited capacity to capture hierarchical patterns in imaging data, motivating a transition towards automated feature learning approaches.

2.2 Deep Learning Approaches

As an evolution from traditional radiomics, deep learning techniques have gained increasing attention for their ability to learn hierarchical representations directly from imaging data. These methods, predominantly based on CNNs, eliminate the need for handcrafted features by automatically extracting high-level abstractions that are optimized for the classification task. Recent efforts have focused on designing robust architectures for both unimodal (CT or PET) and multimodal (PET/CT) inputs to improve subtype discrimination. A summary table presenting key details such as the dataset, input and network characteristics, evaluation strategy, and performance of each deep learning-based study is provided in Table 2.1. Note that the performance metrics reflect the results reported in the respective papers.

2.2.1 CT-based

In CT-based deep learning studies, various network designs have been explored to improve classification accuracy for NSCLC subtypes. Chaunzwa et al. [21] employed a VGG-16 based

Table 2.1: Summary of the deep learning approaches. The performance metrics show the results reported in the respective papers. AUC: Area under the receiver operating characteristic curve, ACC: Accuracy, ADC: Adenocarcinoma, SQC: Squamous Cell Carcinoma, SCC: Small Cell Carcinoma, NOS: Not Otherwise Specified.

Study	**Year**	**Modalities**	**Classes**	**Dataset**	**Dataset size**	**Input type**	**Network type**	**Evaluation strategy**	**Performance metrics**
Chaunzwa et al. [21]	2021	CT	ADC, SQC, Other	Private	311 patients	Tumor slice	VGG-16	Hold-out test set	0.71 AUC, 0.69 ACC
Marentakis et al. [74]	2021	CT	ADC, SQC	Subset of Radiomics [2]	102 patients	Whole slice w/ tumor	4 CNNs + LSTM	Five random hold-out sets	0.78 AUC, 0.74 ACC
Liu et al. [68]	2021	CT	ADC, SQC	Private	126 patients	Tumor volume	CapsuleNet	Four random hold-out sets	0.85 AUC, 0.81 ACC
Gao et al. [33]	2023	CT	ADC, SQC	Private + Subsets of Radiomics [2] & Radiogenomics [9]	499 patients	Tumor slice	Multi-view autoencoder-based	5-fold cross-validation	0.81 AUC, 0.77 ACC
Li et al. [64]	2022	CT	ADC, SQC	Private + Subsets of Radiomics [2] & Radiogenomics [9]	574 patients	Tumor slice	Autoencoder-based	5-fold cross-validation & external test	0.85 AUC, 0.80 ACC
Chen et al. [22]	2023	CT	ADC, SQC	Subsets of Radiomics [2], Radiogenomics [9], Lung3 [1]	402 patients	Tumor slice	ResNet-based multi-task	Hold-out & external test	0.84 AUC, 0.81 ACC
Fathalla et al. [30]	2022	CT	ADC, SQC, NOS	Subsets of Radiomics [2], Radiogenomics [9], Lung3 [1]	589 patients	Lung volume	3D residual CNN	Hold-out test set	0.87 AUC, 0.94 ACC
Tomassini et al. [98]	2023	CT	ADC, SQC	Subsets of Radiomics [2], Radiogenomics [9], Lung3 [1], LUAD [6]	368 patients	Lung volume	ConvLSTM	External test set	0.97 AUC, 0.92 ACC
Han et al. [43]	2021	CT, PET	ADC, SQC	Private	1419 patients	Tumor slice	VGG-16	Hold-out test set	0.90 AUC, 0.84 ACC
Zhao et al. [111]	2024	CT, PET	ADC, SQC	Private	189 patients	Tumor slice	7 CNNs	Hold-out test set	0.77 AUC, 0.76 ACC
Jacob and Menon [54]	2022	CT, PET	ADC, SQC, SCC	Subset of Lung-PET-CT-Dx [65]	N/A	Whole slice	Shallow CNN	Hold-out test set	0.92/0.96 AUC, 0.95 ACC
Qin et al. [79]	2020	CT, PET	ADC, SQC, SCC	Private	397 patients	Whole volume w/ tumor	3D DenseNet-based	10-fold cross-validation	0.92 AUC, 0.72 ACC
Barbouchi et al. [10]	2023	CT, PET	ADC, SQC, SCC	Subset of Lung-PET-CT-Dx [65]	25 patients	Whole slice	DETR	Hold-out test set	0.98 AUC, 0.96 ACC

model to classify histological subtypes (ADC, SQC, and others) from cropped 2D tumor slices, highlighting the potential of DL for learning representative features. They evaluated their approach on a private dataset, the Boston Lung Cancer Survival (BLCS) dataset, which includes 311 patients. Using a hold-out test set, they compared the performance of an end-to-end trained VGG-16 model with traditional ML classifiers applied to features extracted from the final layers of VGG-16. Experiments were conducted on both a two-class setting (ADC vs SQC) and a three-class setting (ADC vs SQC vs others). Their results showed that for binary classification, the end-to-end trained VGG-16 achieved an Area Under Receiver Operating Characteristic Curve (AUROC) of 0.71 and an accuracy of 0.69, surpassing the performance of the traditional ML classifiers. Further exploring architectural innovations, Marentakis et al. [74] combined an Inception model with a Long Short-Term Memory (LSTM) network to capture spatial coherence across CT slices, achieving improved performance in distinguishing ADC and SQC. Their study also benchmarked several state-of-the-art CNNs, including AlexNet, ResNet101, Inceptionv3, and InceptionResnetv2, finding that an LSTM combined with Inception yielded superior results compared to standalone CNNs and even expert radiologists. The experiments were conducted on a subset of the NSCLC-Radiomics [2] dataset, comprising 48 ADC and 54 SQC patients. Using five random hold-out permutations (each with a 50% training set and a 50% test set), the model achieved an AUROC of 0.78 and an accuracy of 0.74. Liu et al. [68] investigated the use of Capsule Networks (CapsNet), which are designed to better delineate global image characteristics by encoding relative spatial relationships of image elements. Their comparative study on a single-center private dataset with 126 patients (72 ADC and 54 SQC) found that CapsNet outperformed traditional CNNs and several radiomics models in discriminating ADC and SQC subtypes. They used cropped 3D tumor volumes as network input. Using four random hold-out permutations (each with a 75% training set and a 25% test set), the model achieved an average AUROC of 0.85 and an accuracy of 0.81. Researchers have also developed more sophisticated DL frameworks. Gao et al. [33] proposed a multi-view feature decomposition method where features from axial, coronal, and sagittal CT views are decomposed into common and specific features using an attention mechanism and constrained by a feature similarity loss. This approach aims to obtain comprehensive tumor representations by efficiently integrating these common features. Their model was evaluated on a dataset combining private data with subsets of two public datasets (NSCLC-Radiomics [3] and NSCLC-Radiogenomics [9]), comprising a total of 499 patients. Using 5-fold cross-validation, they achieved an average AUROC of 0.81 and an accuracy of 0.77. Li et al. [64] introduced a Reconstruction-Assisted Feature Encoding Network (RAFENet). This model employs a shared encoder and a task-aware encoding module with cross-level non-local blocks to learn multi-level task-specific features. A semantic consistency loss function, encompassing feature and prediction consistency, was de-

signed to enhance model regularization during the image reconstruction auxiliary task. The model was trained on a dataset combining subsets of NSCLC-Radiomics [3] and NSCLC-Radiogenomics [9], and evaluated on a private dataset. It achieved an AUROC of 0.85 and an accuracy of 0.80 on the external test set. Chen et al. [22] presented a multi-task learning model for histologic subtype classification. Their model incorporates a classification branch and a staging branch that share initial feature extraction layers and are trained simultaneously. The model used 2D cropped tumor patches as input. It was trained on a combined dataset of NSCLC-Radiomics [3] and NSCLC-Radiogenomics [9], and evaluated on both an internal test set derived from these two datasets and an external test set, Lung3 [1]. The model achieved an AUROC of 0.84 and an accuracy of 0.81 on the internal test set, and an AUROC of 0.73 and an accuracy of 0.71 on the external test set. Fathalla et al. [30] developed DETECT-LC, a multi-stage computational model that integrates both 3D deep learning and textural radiomics for lung cancer staging and tumor phenotyping. Their pipeline involves a radiomics preparatory stage for unsupervised slice selection using Haralick features and k-means clustering, followed by an ALT-CNN-DENSE Net for classification. The DETECT-LC model was evaluated for its ability to classify NSCLC carcinoma subtypes (ADC, SQC, and NOS) and for lung cancer staging, demonstrating superior performance compared to baseline models like ResNet-50 and InceptionV3. The model was evaluated separately on three datasets: NSCLC-Radiomics [3], NSCLC-Radiogenomics [9], and LUNG3 [1]. Each dataset was divided into training, validation, and testing sets, with approximately 55% of samples used for training, 15% for validation, and 30% for testing. The proposed model outperformed its competitors, achieving AUROCs and accuracies of 0.83 and 0.96, 0.89 and 0.93, and 0.88 and 0.92 on the three datasets, respectively. Tomassini et al. [98] developed LUCY, an on-cloud decision-support system, which implemented and compared two end-to-end neural networks where the core layer is a ConvLSTM layer, to characterize ADC and SQC from thorax CT scans. These scan-based approaches, including those using ConvLSTMs or time-distributed 2D CNNs with recurrent layers, aim to analyze volumetric information in more detail than slice-based methods. The models were trained on a subset derived from a combination of three publicly available datasets: NSCLC-Radiomics [3], LUNG3 [1], and TCGA-LUAD [6], and evaluated on a subset of another public dataset, NSCLC-Radiogenomics [9]. The ConvLSTM-based model achieved an AUROC of 0.97 and an accuracy of 0.92.

Overall, CT-based deep learning studies reveal a steady progression from simple CNNs to more elaborate architectures designed to incorporate spatial context, multitask learning, and volumetric representation, all contributing to more accurate and clinically relevant subtype predictions.

2.2.2 PET/CT-based

While CT imaging alone provides valuable anatomical detail, combining it with PET allows for the inclusion of functional information that is critical for tumor characterization. This has led to a growing number of studies employing PET/CT inputs in deep learning models for NSCLC subtype classification. Han et al. [43] conducted end-to-end classification experiments using a pre-trained VGG-16 model and compared its performance with ten traditional machine learning models trained on 50 radiomic features selected by ten different feature selection methods. Their study was based on a private dataset comprising 867 ADC and 552 SQC patients for ADC–SQC classification. The experiments used merged CT and PET images as input. The VGG-16 model achieved an AUROC of 0.90 and an accuracy of 0.84 on a hold-out test set, outperforming all other machine learning models. Zhao et al. [111] focused on discriminating NSCLC pathological subtypes (ADC vs SQC) using 18F-FDG PET/CT images. They conducted a comparative study of seven different deep learning models (Shufflenet, VGG16, Googlenet, Inception v3, Resnet50, Densenet201, and Mobilenet v2), evaluating their performance and exploring correlations between deep learning-extracted features and traditional radiological parameters like tumor size and Standard Uptake Value (SUV)max. The models were trained using cropped 2D tumor patches as input. On a private dataset comprising 189 patients, all seven models were evaluated, and MobileNetv2 was found to outperform the others, achieving an AUROC of 0.77 and an accuracy of 0.76. Jacob and Menon [54] proposed a novel shallow CNN for the pathological categorization of lung carcinoma (ADC, SQC, and Small Cell Carcinoma (SCC)) using slices extracted from multimodality CT and PET images, employing a Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) for slice selection. After applying BRISQUE, 1,744 PET/CT images were selected from a total of 760,818 images in the Lung-PET-CT-Dx [65] dataset. The images were divided into training and testing sets. On the internal test set, the model achieved AUROC scores of 0.92 and 0.96 for the ADC and SQC classes, respectively, with an overall accuracy of 0.95. Beyond standard CNNs, researchers have investigated more advanced architectures and fusion techniques. Qin et al. [79] introduced a model for fine-grained lung cancer classification (ADC, SQC, and SCC) from PET and CT images that incorporates a multidimensional attention mechanism. This mechanism, featuring channel-wise and spatial attention within densely connected convolutional networks, was designed to effectively fuse complementary information from the two imaging modalities, leading to more robust feature representations. The model was evaluated on a private dataset comprising 397 patients using a 10-fold cross-validation strategy and achieved an average AUROC of 0.92 and an accuracy of 0.72. Further pushing architectural boundaries, Barbouchi et al. [10] utilized the DETR transformer for the simultaneous detection and classification of lung cancer (ADC, SQC, and SCC) from PET/CT images. Their work demonstrated that

transformer models, known for capturing global relationships in data, can outperform traditional CNN architectures like ResNet, VGG, and DenseNet in this domain. The model was evaluated on 1,160 images extracted from 25 patients in the Lung-PET-CT-Dx dataset [65]. On an internal hold-out test set, it achieved an AUROC of 0.98 and an accuracy of 0.96. These PET/CT-based studies further demonstrate how multimodal imaging inputs can enhance classification accuracy and provide complementary information not always discernible from a single modality. The integration of anatomical and functional data through deep learning pipelines continues to gain traction, offering promising avenues for more robust and generalizable models in NSCLC histopathological classification.

2.3 Limitations of Existing Approaches

A notable commonality in the literature is that, except for the studies by Fathalla et al. [30], Tomassini et al. [98], and Jacob and Menon [54], most works require tumor annotations. These studies either use tumor patches or volumes as input, which necessitates knowing tumor boundaries, or slices containing tumors, which requires identifying which slices include tumors. Tumor annotation and segmentation are costly and time-consuming processes. To address this, in the works presented in Chapters 6 and 7, we utilized 3D lung volumes where the lungs were segmented automatically using an existing segmentation algorithm.

Another challenge arises from the use of 2D images. The data split between training and test sets must be performed at the patient level to prevent data leakage. Unfortunately, many studies do not explicitly state whether this was done, and some appear to suffer from leakage. To avoid this issue, in the studies using 2D patches, as in Chapters 4 and 5, we ensured that all patches from the same patient were assigned to the same set.

A further limitation is that most works employ an internal hold-out test set for evaluation, even though the dataset sizes are small compared to other AI domains. This limits assessment of model generalizability to different cohorts. In all studies presented in this thesis, we applied 5-fold cross-validation to better evaluate generalization.

Furthermore, some works apply random undersampling to balance classes, as most datasets are naturally imbalanced according to the prevalence of subtypes. While this approach balances the classes, it complicates fair comparisons because results are based on a random subset, which is particularly problematic in hold-out evaluation. To address this, we utilized all images belonging to the ADC and SQC classes in each dataset, excluding only those with obvious defects, to produce results that more accurately reflect the entire dataset.

Finally, a major challenge in the literature is the lack of open-source implementations, which makes it difficult to reproduce the reported results. Among the studies presented in Table 2.1, only the work of Tomassini et al. [98] provides open-source code. To enable a fair

comparison with our methods, we re-implemented several studies from the literature and trained and tested them on the datasets used in this thesis. Specifically, we re-implemented the works of;

- Chaunzwa et al. [21] as a competitor at Chapter 5, since it employs a similar strategy to ours, using 2D tumor patches as input to a CNN for histological subtype classification;
- Fathalla et al. [30] and Tomassini et al. [98] at Chapters 6 and 7, as their methods also use 3D lung volumes without requiring tumor segmentations as ours;
- Qin et al. [79] at Chapter 6, as it represents the only intermediate fusion approach in the literature applied to histological subtype classification.

We observed that none of these methods achieved the performance reported in their respective papers, instead exhibiting degraded results when applied to our datasets.

Chapter 3

Datasets

To evaluate our proposed method, we employed four independent datasets comprising patients with NSCLC. These datasets include both institutional and publicly available collections, offering a diverse range of imaging protocols, annotation methods, and clinical characteristics. Each dataset provides paired CT and PET/CT images, along with varying levels of clinical and molecular data. A summary of the datasets, including the number of patients, available modalities, histological subtype distribution and the class imbalance ratio between subtypes, is provided in Table 3.1.

Table 3.1: Summary of the included datasets, showing the total number of patients, the distribution of histological subtypes, the class imbalance ratio , and the availability of imaging modalities. The imbalance ratio is computed as the number of instances of the majority class (ADC) divided by the number of instances of the minority class (SQC).

Dataset	**Total Patients**	**ADC**	**SQC**	**Imbalance Ratio**	**CT**	**PET**
Humanitas	423	312 (74%)	111 (26%)	2.81	✓	✓
NSCLC-Radiomics	203	152 (75%)	51 (25%)	2.98	✓	X
NSCLC-Radiogenomics	193	160 (83%)	33 (17%)	4.85	✓	✓
Lung-PET-CT-Dx	98	74 (76%)	24 (24%)	3.08	✓	✓

3.1 Humanitas

The patients are retrospectively selected from the institutional database of the IRCCS Humanitas Research Hospital[1] according to the following inclusion criteria [59]: pathological diagnosis of NSCLC, baseline [^{18}F]FDG PET/CT and surgical intervention were performed in this institution. Exclusion criteria were histology different from lung adenocarcinoma or squamous cell carcinoma and concomitant or previous cancers within 3 years of lung cancer

[1]The Ethics Committee of the Humanitas Clinical and Research Centre IRCCS approved the study on 2017/04/18 with the authorisation number 1751.

diagnosis. We collected demographic data, tumor information (histological subtypes, tumor grade, and pathological TNM staging [36]), and clinical outcomes (overall survival and progression-free survival) from the electronic medical record. Finally, the Volume of Interest (VOI) of the primary tumor lesion was obtained by manually adjusting on CT images the semiautomatic segmentation extracted from the PET, where we exploited the volume uptake by thresholding SUV using commercial software (PET VCAR, GE Healthcare, Waukesha, WI, USA). This threshold was heuristically set equal to 40% of the maximum SUV , following standard practice in PET-based tumor segmentation, where this value is commonly applied to delineate metabolically active tumor regions [55, 97] . Such an adjustment is needed to correct the position of the lesion if respiratory movements caused a mismatch between CT and PET images.

3.2 NSCLC Radiomics

NSCLC Radiomics [2] was retrospectively collected from the MAASTRO Clinic (Maastricht, The Netherlands), comprising 422 patients diagnosed with inoperable NSCLC and treated with either radical radiotherapy alone or in combination with chemotherapy. Inclusion criteria involved a histologic or cytologic confirmation of NSCLC and availability of pre-treatment FDG PET/CT scans acquired in radiotherapy position using a dedicated PET/CT simulator (Biograph, Siemens). Clinical data, including demographic information (mean age 67.5 years, 290 male, 132 female), tumor characteristics (UICC stages I–IIIb), and outcomes, were available for all patients. The Gross Tumor Volume (GTV) was manually delineated by a radiation oncologist on fused PET/CT images using a standard institutional protocol. Delineation was guided by fixed window settings and performed using commercial radiotherapy planning software (XiO, Computerized Medical Systems, St Louis, MO, USA). The data also included radiotherapy treatment details and follow-up clinical outcomes. All procedures were conducted in compliance with institutional guidelines, and ethical approval was waived due to the retrospective nature of the study [3]. We excluded the patients with histological types other than ADC or SQC, eventually, we included 203 patients in total, 152 diagnosed with ADC and 51 with SQC.

3.3 NSCLC Radiogenomics

NSCLC Radiogenomics [9] was retrospectively collected from two institutions (Stanford University Medical Center and Palo Alto Veterans Affairs Healthcare System), including 211 patients with NSCLC who underwent surgical treatment. Inclusion criteria required availability of preoperative CT and PET/CT scans, as well as excised tumor tissue samples.

Demographic and clinical data (mean age 68, 135 male, 76 female) were collected, including histological subtype, pathological TNM staging, and survival outcomes. Most patients had adenocarcinoma (n = 172), followed by squamous cell carcinoma (n = 35). Imaging acquisition protocols varied, with CT slice thickness ranging from 0.625 to 3 mm and PET images acquired with FDG doses of 138.9–572.3 MBq after an uptake time of 23–129 minutes. The primary tumor lesions were segmented in 3D from CT images and revised by expert thoracic radiologists. A subset of patients also had semantic annotations of tumor features and available RNA-seq or microarray-based molecular data. This dataset was designed to support radiogenomic analyses that investigate the association between imaging features, molecular profiles, and clinical outcomes [8]. We excluded the cases with histological types other than ADC or SQC and the ones that do not have both CT and PET scans. Eventually, we included 193 patients, 160 of them were diagnosed with ADC, and 33 of them with SQC.

3.4 Lung-PET-CT-Dx

Lung-PET-CT-Dx [65] includes retrospectively acquired CT and PET/CT images from patients who underwent standard-of-care biopsy following suspicion of lung cancer. Subjects were categorized based on histopathological diagnosis into adenocarcinoma, squamous cell carcinoma, small cell carcinoma, or large cell carcinoma. Demographic data were not disclosed, but imaging was performed under standardized conditions, including fasting and FDG PET protocols (mean dose 295.8 $\pm$ 64.8 MBq; mean uptake time 70.4 $\pm$ 24.9 minutes). PET/CT acquisition covered the region from the base of the skull to mid-femur, with image reconstruction using TrueX TOF. CT images had a resolution of 512 $\times$ 512 pixels with 1 mm slice thickness, and PET images were reconstructed at 200 $\times$ 200 pixels and matched slice thickness. Tumor locations were annotated using bounding boxes by five thoracic radiologists (at least two with over 15 years of experience) via a multi-reader verification process. Annotations are provided as XML files in PASCAL VOC format. This dataset was curated to support algorithm development for medical image analysis and tumor localization in multimodal lung cancer imaging. As before, we included data only from ADC and SQC cases that have both CT and PET images corresponding to 74 ADC and 24 SQC cases, 98 in total.

Chapter 4

Triplet Networks for Histological Subtype Classification in NSCLC

4.1 Introduction

Most of the works on histological subtype classification using deep neural networks adopt, in the final classification stage, the well-established softmax classifier, and in most cases, small private datasets are used. Since most of the well-known DL techniques require a generous amount of data, there is a need to develop learning approaches that are able to cope with small training data. As this is an issue impacting many different domains, in the literature several solutions have been proposed so far, such as dropout regularization, data augmentation, or the use of simpler networks, etc. [15]. Although introduced in different domains and not for addressing data scarcity, siamese networks [16] could be a viable learning paradigm to learn when few data are available thanks to the loss function employed, as detailed in the next subsections. Despite that, their potential over small datasets has been investigated by a few [31, 32, 50]. On these grounds, this study investigates the use of triplet networks [46], an improved version of siamese networks [16], to classify histological subtypes in NSCLC on a small dataset, and also compares their performance against plain deep networks employing the well-established cross entropy loss in conjunction with softmax activation in the output layer.

4.1.1 Triplet Networks

Siamese Neural Networks were first proposed by Bromley et al. in 1993 for a signature verification task [16]. The main idea is to give two inputs to two identical neural networks with shared weights computing the corresponding features. Then, a function computes the distance between these two feature vectors, and the network weights are updated to minimize

the *contrastive loss* [24], which is a well-established loss function in this domain. It aims not only to minimize the distance between the two inputs when they belong to the same class but also to increase their distance by at least up to a manually defined margin when they belong to different classes. This implies creating a feature space where the instances of the same class are close to each other, whereas those of different classes are far away, then, any classifier can be trained on that space to predict the class label, such as the kNN [110, 62], the SVM [46, 89], etc. Next, given a test sample, one of the two networks computes the sample feature vector and such a classifier is used to label the unknown sample.

Hoffer and Ailon in [46] proposed an improved version of siamese networks named *triplet networks*. Rather than using two identical networks, the authors used three identical networks with three inputs building up the input triplet: two belong to the same class whilst the other belongs to a different class. The first input is referred to as the *anchor*, the second input has the same class label and is named *positive*, and the third belonging to a different class is named *negative*. Triplet networks are usually trained using the *triplet loss* [86], which will be introduced in Section 4.2.

Triplet networks have been successfully used in many domains, such as face recognition [86, 24], person re-identification [23, 45], and object tracking [19, 28], to name a few. In general, all these applications are characterized by large training sets with few instances for each class. Nevertheless, it is worth noting that there are few applications in healthcare and even fewer in cancer research. For instance, Utkin et al. [102] used this approach to diagnose lung cancer from chest CT scans, Ghoneim et al. [35] classified histological images of breast cancer, Le Vuong et al. [63] used it for colorectal cancer grading from pathology images, and Battle et al. [11] classified types of skin cancer using clinical and dermatological images.

Despite this research, the potential of triplet networks to learn under a limited training set is still a topic deserving more investigation and, in this respect, the next section presents our proposal in this field.

4.2 Methods

In this section, we present how to exploit triplet networks to classify histological subtypes of NSCLC from CT images, when a limited number of training samples are available. Figure 4.1 shows the general architecture of our approach distinguishing between the training and classification phases, each described in the following subsections.

4.2.1 Training stage

In the upper panel of Figure 4.1, from the left to the right, given a batch B containing M images, each denoted by x, a CNN extracts M feature vectors $f(x) \in \mathbb{R}^d$, which are

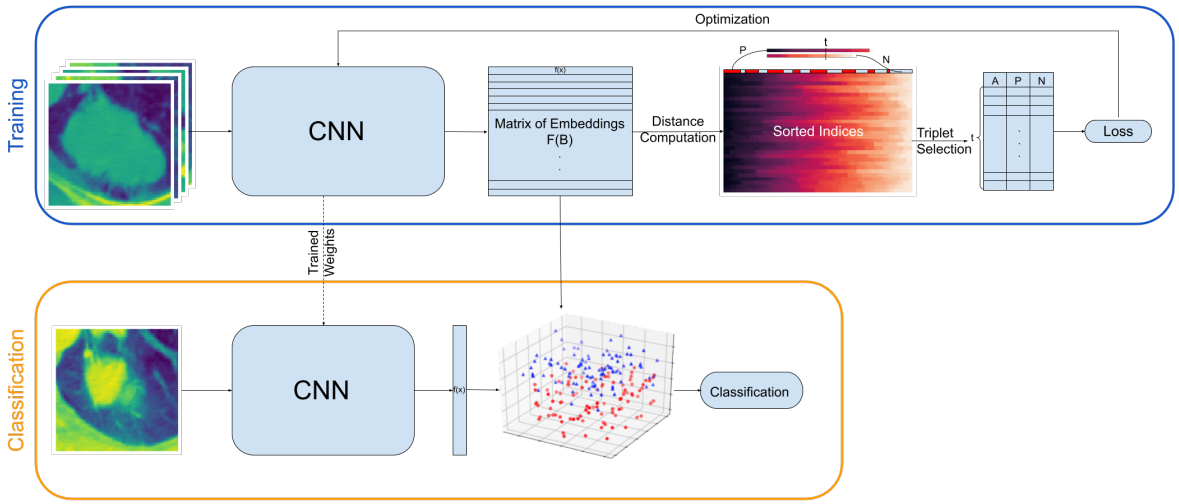

Figure 4.1: Overall framework of the proposed method. The upper panel illustrates the training phase of the triplet networks, while the lower panel depicts the inference phase. During the training phase, a CNN extracts embeddings from a batch of tumor patches (denoted as F(B)). The pairwise distances between these embeddings are then computed and sorted to identify the most and least similar samples. Based on these distances and the triplet selection strategy, anchor, positive, and negative samples (denoted as A, P, and N) are selected to compute the triplet loss and optimize the network. During the inference phase, the embedding of a query patch is extracted using the trained network, and the query is classified using a kNN classifier trained on the embeddings of the training set, which are obtained using the same network.

collected in $F(B) \in \mathbb{R}^{M \times d}$. Then, we compute the matrix $E \in \mathbb{R}^{M \times M}$ that collects the pairwise Euclidean distances among the feature vectors of the samples in the batch. As the next step, another matrix $I \in \mathbb{R}^{M \times M}$ consisting of the indices of the elements in the matrix E, sorted row-wise in an ascending manner, is created. This matrix I is used to obtain positive and negative lists where we choose positive and negative instances to compute the triplet loss. The details of the triplet selection methods are given in the following subsection.

The triplet loss aims to pull the positive image through the anchor while pushing the negative by a margin $m \in \mathbb{R}^+$, as depicted in Figure 4.2. Formally, let us introduce the following notation:

- superscript ^{a} applied to x specifies that such an image has been selected as an anchor;
- P is the set of images in B that belong to the same class of x^a, i.e., $P = \{\forall x \in B | y(x) = y(x^a) \wedge x \neq x^a\}$;
- N is the set of images in B that do not belong to the same class of x^a, i.e., $N = \{\forall x \in B | y(x) \neq y(x^a)\} = B - (P \cup \{x^a\})$;
- x_j^P and x_k^N denote the jth and kth element of P and N, respectively.

Given x^a, x_j^P and x_k^N, we, therefore, say that the training process of the triplet networks reaches its goal if:

$$\|f(x^a) - f(x_j^P)\|_2 + m < \|f(x^a) - f(x_k^N)\|_2 \tag{4.1}$$

This intuitive idea can be formalized in the following formulation of the training loss:

$$L = \frac{1}{M}\frac{1}{t}\sum_{i=1}^{M}\sum_{(j,k)=(q_P,q_N)}^{(q_P+t,q_N+t)}\left[\|f(x_i^a) - f(x_j^P)\|_2 - \|f(x_i^a) - f(x_k^N)\|_2 + m\right]_+ \tag{4.2}$$

where $t = min(|P| - q_P, |N| - q_N, T)$, q_P and q_N are any natural number in $[1, |P|]$ and $[1, |N|]$, respectively, and T is a hyper-parameter lying in $[1, \lfloor\frac{M-1}{2}\rfloor]$.

Triplet selection

It is worth noting that a brute force approach that computes all the possible triplets may result in several computations satisfying Equation 4.1, meaning that the distance between the anchor and the negative sample is already larger than the distance between the anchor and the positive sample plus the margin. Such triplets do not contribute to the loss (Equation 4.2), resulting in zero gradient updates. Randomly selecting triplets suffers from the same limitation, as the vast majority of random combinations are non-violating and yield zero loss. As noted by several studies [45, 86], effective triplet mining is crucial for fast convergence and model performance. For this reason, this study experimentally evaluates three non-trivial triplet selection strategies that emphasize triplets violating the constraint, thereby enabling more informative gradient updates and faster convergence.

In the case of the first selection method, denoted as Triplet Selection Method (TSM)1 in the following, from $F(B)$ we select the t easiest positives and t hardest negatives for each anchor. The easiest positive from P is the closest positive image to x^a according to the distances computed in E. The hardest negative from N is the closest negative image to x^a, again according to E. The rationale lies in observing that it is easy to pull a positive that is already close to the anchor, whereas it is hard to push a negative that is too close to it. Furthermore, q_P and q_N are now equal to zero.

In the case of the second selection method, named TSM2, we introduce an approach where the difficulty of selected positive instances dynamically varies during training. Indeed, the first t positives may lie too close to the anchor after some epochs, thus making it very easy to satisfy Equation 4.1. To overcome this limitation, positives started to be selected from the easiest, but when loss decreases below a threshold during training, we select them among the harder ones. Therefore, in Equation 4.2, q_P is set to zero at the beginning and it is increased by one when the loss is smaller than a predefined threshold; q_N is always equal to zero.

In the case of the last method, denoted as TSM3, both the positive and negative selection

dynamically varies during the training. In the beginning, they both start with the easiest instances and then they both get more difficult if the loss is less than a certain threshold. To work with Equation 4.2 the negative list is reversed, i.e., it is sorted in a descending manner where the easiest negatives are at the beginning of the list and the hardest ones are at the end. In Equation 4.2, both q_P and q_N are set to zero at the beginning, and they are incremented by one until the loss exceeds the threshold.

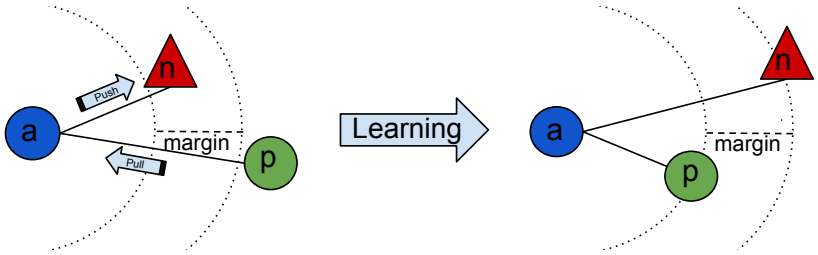

Figure 4.2: Triplet loss learning process, where a, p, and n denote the anchor, positive, and negative samples, respectively. During training, the network learns to minimize the distance between embeddings of instances with the same label (a and p) while maximizing the distance between embeddings of instances with different labels (a and n), effectively shaping a feature space where samples of the same class are closer to each other than those of different classes by at least a margin.

4.2.2 Classification stage

The lower panel of Figure 4.1 shows the classification stage: from left to right, given a query image x, the trained CNN extracts the corresponding feature vector $f(x) \in \mathbb{R}^d$. Then, this feature vector $f(x)$ is placed into a feature space learnt by the CNN during the training process and where the minimization of the loss function should have separated samples of the different classes. The test instance is then classified using the kNN classifier since this paradigm outperforms other approaches as reported in the literature [46, 89]. Let us remark that x is an image and, since we are dealing with CT scans, each patient has several slices: hence, for each patient in the test set, we set the final classification by majority voting on predictions of all the slices in the CT volume.

4.3 Experimental Configuration

The approach proposed in Section 4.2 is an end-to-end learning framework that includes the uses of CNNs. In this respect, here we worked with 4 different architecture families, which have proved to have promising results in many biomedical applications [39]. They are ResNet-50 [44], MobileNetV2 [85], VGG-16 [87] and GoogLeNet [92]. All the CNNs were pretrained on the ImageNet dataset, and each network is fine-tuned for 50 epochs with a

batch size of 32, also applying standard basic geometric augmentation techniques to 2D CT images extracted from the tumour VOI. Since the available data and computational resources were limited, 2D tumour patches were used to reduce computational cost while increasing the number of training samples, ensuring a more efficient and stable training process. Indeed, images are randomly flipped in the horizontal direction then they are randomly rotated in ± 5 degrees in the training phase. Furthermore, images are resized into 64×64 and normalized in $[0, 1]$ before being fed to the network. Adam optimizer is used for optimization with a starting learning rate of 0.001. The learning rate is decreased every 10 epochs by a factor of 0.1. Note that the loss defined in Equation 4.2 is calculated as the average of t triplets for each input image in the batch. They are accumulated during backpropagation until the whole batch is processed, then, weights are updated.

Let us now turn the attention to the competitors used. Triplet networks are compared against the same plain four deep architectures where the output layer uses the softmax activation. All the parameters are kept the same except the loss function: in fact, the cross-entropy loss is used instead of triplet margin loss. Class weights are applied in the loss function to mitigate the class imbalance problem. In the following, we shortly use the softmax loss to refer to the use of cross-entropy loss function and softmax activation.

In this work, we used the CT scans of 87 adult patients from the Humanitas dataset who had segmentation information to extract the tumor patches. We performed all the experiments using stratified 5-fold cross-validation, repeated 5 times; note that folds were defined at the patient level to prevent any bias, i.e., slices of the patient can be only in the training or test set. To ensure a fair comparison across methods, the same patient-level splits were consistently used in all experiments.

4.4 Results

We performed a wide set of experiments that use, as mentioned, 4 CNNs, and 3 triplet selection methods with 5 different numbers of triplets (1, 2, 4, 8 and 16). For each combination between the triplet selection method and the number of triplets, we tested 5 different numbers of neighbours for the kNN (1, 3, 5, 7 and 9). Considering also the 25 repetitions, we ran 7500 experiments for triplet networks, plus another 100 runs for the plain networks. For space reasons, we, therefore, omit to report an analytical table with all the results, and we deepen the results by means of a statistical comparison between the performance scores, measured in terms of AUROC averaged out the 25 runs.

In this section, we try to answer the following four issues:

1. If the triplet loss performs better than the softmax loss;

2. If any network architecture outperforms the others;

3. If any triplet selection method surpasses the others;

4. If there is a number of triplets t that perform better than the others.

Table 4.1: Win-loss comparison between triplet networks and plain CNN. The numbers in parentheses specify the wins/losses where $p < 0.1$.

Model	**Wins**	**Loss**
ResNet-50	74 (30)	1 (0)
MobileNetV2	47 (25)	28 (0)
VGG-16	43 (12)	32 (1)
GoogLeNet	64 (18)	11 (0)

To deepen the first issue, triplet loss is compared with softmax loss. To aggregate the results, we calculated in how many experiments the triplet loss performs better than the softmax loss, i.e., how many times the average AUROC of the triplet loss is larger than the corresponding value of the softmax loss for each network architecture. In this respect, Table 4.1 reports the number of wins and losses for each network (no tie was reported). Furthermore, we determine how many times the performances are statistically different according to the Wilcoxon signed-rank test, setting $p = 0.1$, conducted pairwise across the 25 folds. The corresponding values are reported in parenthesis in each cell of the table. The results show that the triplet approach outperforms the softmax loss for all the models; furthermore, while the wins are mostly statistically significant, losses are not, except in one case. These observations suggest that the triplet approach is more robust to data scarcity than the softmax approach.

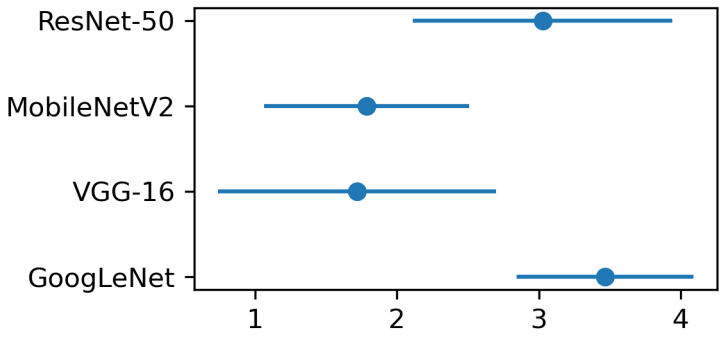

Figure 4.3: Average and standard deviation of the scores attained by the four models among all the experiments.

Let us now turn our attention to the second issue. In this respect, we compared all the models one against the other by fixing the model and comparing the experiments of

each parameter combination. In other words, for each combination of the triplet selection method, the number of triplets, and the number of neighbours, we first scored all the models between 1 to 4 according to the average AUROC (the better the AUROC, the larger the score), and then we calculated the mean and standard deviation of the scores. The results are shown in Figure 4.3: we note that GoogLeNet is the best model, whereas the ResNet-50 ranks second; VGG-16 and MobileNetV2 are both equally worse than the others to a large extent. Furthermore, we verified with the Friedman test that such performances are statistically different ($p = 1.4 \cdot 10^{-22}$); then, the Nemenyi test, a posthoc test used for multiple comparisons [27], confirms that ResNet-50 and GoogLeNet are significantly better than MobileNetV2 and VGG-16 ($p \leq 10^{-3}$). All statistical tests were conducted pairwise on the average results obtained across different experiments. With $p = 0.157$, the same test points out that GoogLeNet and ResNet-50 are not statistically different. We deem that the results can be expected: indeed, on the one side the MobileNetV2 was originally designed to be trained on computationally limited platforms, and it is smaller than GoogLeNet and ResNet-50, a feature that usually results in lower performance of this model with respect to the other two [18]. On the other side, it is likely that the large number of weights of the VGG-16 makes harder the fine-tuning. Furthermore, the GoogLeNet and the ResNet-50 were proposed to reduce the computational burden of deep neural networks while obtaining state-of-the-art performances, addressing also issues given by overfitting and degradation problem.

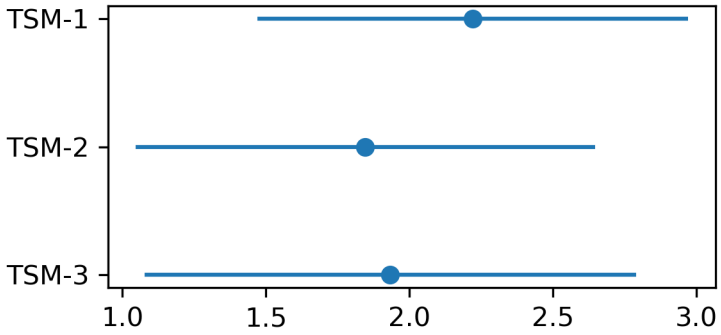

Figure 4.4: Average and standard deviation of the scores attained by the three triplet selection methods among all the experiments.

To answer the third and fourth issues, we proceeded as before, fixing triplet selection methods and the number of triplets, respectively. In the former case, Figure 4.4 shows the average and the standard deviation of the ranks attained by the three selection methods, whereas the Friedman test suggests that there exist results statically different ($p = 0.016$). The following Nemenyi posthoc test reveals that the difference is only significant between the TSM1 and TSM2 ($p = 0.013$), whilst it returns $p = 0.157$ when comparing TSM1 and TSM3.

These results that selecting the t easiest positives and t hardest negatives for each anchor should be preferred with respect to other solutions that, dynamically select the positive and negative instances, makes harder the minimization of the loss function. When investigating the number of triplets, the Friedman test returns $p = 0.931$, clearly suggesting that the use of different values of t does not provide different results.

4.5 Conclusion

This study focused on triplet deep neural networks to experimentally investigate if they performed better than plain deep neural networks using the well-known cross-entropy loss in conjunction with softmax activation when a limited training dataset is available, a limitation of any AI approach that frequently occurs in healthcare. Our exhaustive experiments with several different implementations of triplet networks demonstrated that they are a viable option to be used under these circumstances. Furthermore, this work offers two other insights into the use of triplet networks: ResNet-50 and GoogleNet are more suitable backbone networks to be used, and also, a triplet selection method where the first t easiest positive and negative instances are selected should be preferred. A possible direction for future work is to improve triplet mining by adding model-based difficulty measures. For example, a second stage could select samples that the classifier misclassifies or predicts with low confidence. Combining embedding-based and model-based criteria could help identify harder examples and improve generalization. Another promising direction is to explore more imbalance handling strategies, such as focal-like losses, adaptive re-weighting, or class-aware sampling, to further improve robustness under skewed data distributions. Although in this work we found that triplet networks are a better approach than conventional methods, the size of the dataset is still a limitation. To overcome these issues, in the next chapter, we investigated federated learning approaches to exploit additional datasets while preserving privacy.

Chapter 5

Enhancing NSCLC Histological Subtype Classification: A Federated Learning Approach Using Triplet Loss

5.1 Introduction

The power of analyzing vast amounts of data is crucial for AI's performance, but limited data availability, especially in the medical domain, poses challenges. Various methods address this issue, including dropout regularization, data augmentation, and simpler network architectures [15]. In Chapter 4, we explored triplet networks for limited data scenarios, showing their potential over conventional deep networks using cross-entropy loss and softmax activation. While data augmentation offers benefits like class balancing and overfitting mitigation, it has drawbacks. It may not accurately reflect true data distributions in medical centers, potentially leading to skewed feature representations. Models might overfit to augmented data, hindering real-world generalization [51]. Geometric transformations can introduce artifacts, compromising model performance [34]. Consequently, healthcare scenarios prioritize training with real datasets, often by aggregating multiple datasets. However, this centralized approach raises data privacy concerns, as medical centers typically have strict protective policies [103]. A promising solution is Federated Learning (FL), a distributed machine learning framework that allows multiple models to extract insights from local data while maintaining privacy [60]. In FL, clients keep their sensitive data confidential, periodically sharing only updated model weights with a central server, enabling collaborative learning without compromising data security. Despite these advantages, FL presents challenges that can affect model performance. Variations in data distributions across institutions, known as data heterogeneity, may hinder convergence and reduce generalization. As a result, models

trained in federated settings might show lower performance compared to centralized training, especially when data imbalance or domain shifts exist among clients.

Given these challenges, our study explores the integration of triplet loss within a federated learning framework. This approach aims to address the issue of limited data availability by effectively leveraging information from external sources while maintaining strict privacy standards. To evaluate the effectiveness of our proposed method, we conduct a comprehensive comparison against two alternative scenarios: (1) a local training approach, where the model is trained exclusively on data from a single institution, and (2) the widely-used combination of cross-entropy loss and softmax activation in the output layer. Through these comparisons, we seek to demonstrate the potential benefits of our federated learning approach with triplet loss in preserving data confidentiality while maintaining model performance.

5.2 Methods

Our approach leverages federated learning and triplet loss to enhance NSCLC subtype classification while maintaining data privacy. Triplet networks are explained in detail in Chapter 4, particularly in Sections 4.1.1 and 4.2.1, while the concept of federated learning is elaborated in Section 5.2.1. Section 5.2.2 outlines our proposed framework, detailing three critical components: preprocessing steps, training methodology, and inference procedure. Section 5.2.3 provides a comprehensive overview of our experimental setup, including hyperparameter selection, input transformation techniques, data augmentation strategies, and validation methodology.

5.2.1 Federated Learning

There are three main paradigms of FL: Horizontal Federated Learning (HFL), Vertical Federated Learning (VFL), and Transfer Federated Learning (TFL). In HFL, the server combines contributions from models across different clients, assuming that all models share the same structure. This approach is applicable when customer datasets exhibit significant overlap in characteristics but minimal overlap in samples. However, this specific data structure is not always present [66]. In contrast, VFL is an innovative approach within federated learning, designed for situations where data are vertically partitioned and distributed among several parts. VFL improves the representation of samples by incorporating features from different parts, thus improving the model's capability. Unlike HFL, VFL is typically used in business collaborations between companies [107]. Finally, TFL is an advanced approach that integrates FL with the principles of transfer learning. In TFL, multiple clients collaborate to train machine learning models using their local data without sharing it, while at the same time exploiting knowledge from related activities or domains. This method is particularly

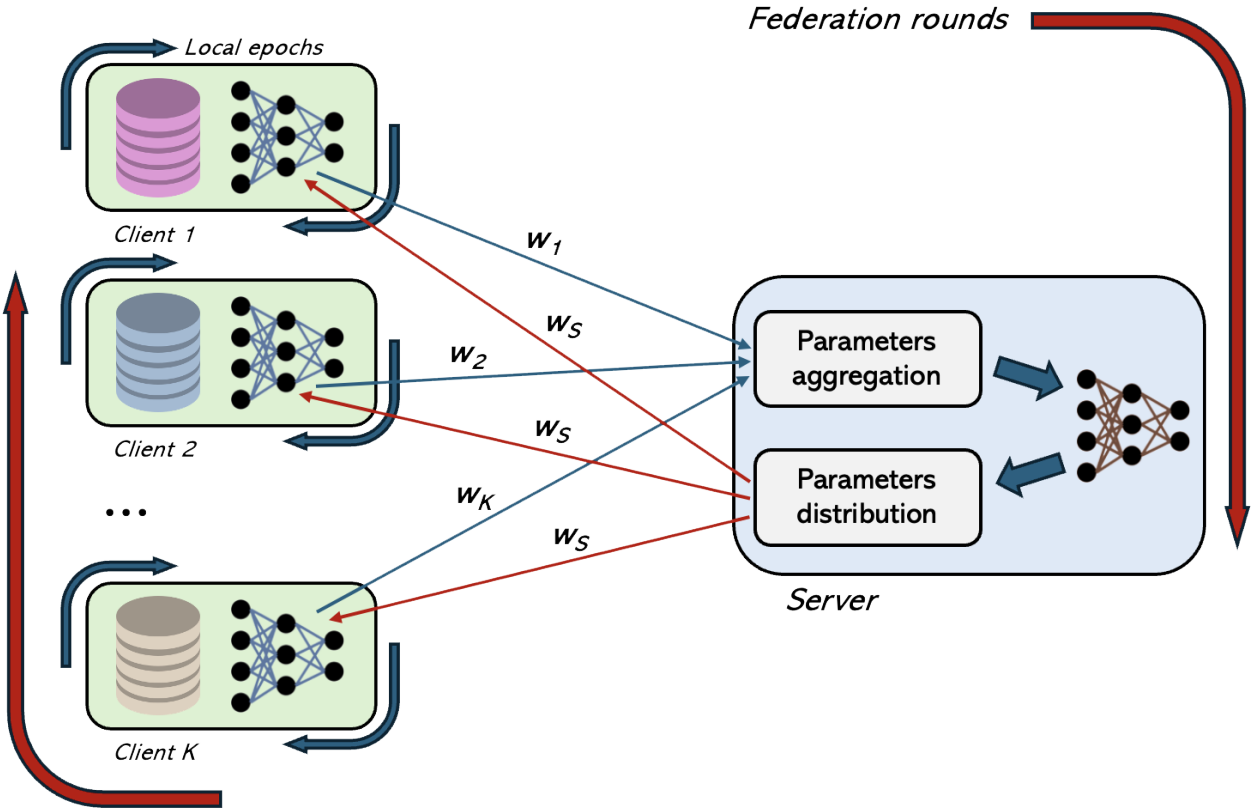

Figure 5.1: Horizontal federated learning framework. In this framework, all clients and the central server share the same network architecture. After a predefined number of local training epochs at each client, the model weights of each client (denoted as w_k) are sent to the server. The server aggregates these weights and distributes the aggregated weights (w_s) back to all clients. This process is repeated for a predefined number of federation rounds.

advantageous in scenarios where data are isolated between different organisations or domains, allowing the transfer of learnt knowledge to improve model performance [29]. In this manuscript, we propose to revisit triplet networks with FL mechanism and the paradigm that better describes such an integration is HFL. Therefore, this section gives a detailed explanation on how it works before the presentation of the final pipeline.

Let $\boldsymbol{w}$ denote the global model parameters initialized on the server side. Each client k (with $k = 1, 2, ..., K$) has its local and private dataset D_k containing N_k samples each, and receive an exact copy of the server model, denoted as local model. The local objective function for a client k can be defined as:

$$L_k(\boldsymbol{w}) = \frac{1}{N_k} \sum_{(x_i, y_i) \in D_k} \mathcal{L}(f(x_i, \boldsymbol{w}), y_i) \tag{5.1}$$

where $\mathcal{L}$ is the loss function, f indicates the model function, x_i denotes the input data, and y_i denotes the corresponding labels. Next, each client k runs a number of local training epochs to update its local model parameters $\boldsymbol{w_k}$ by updating them using gradient descent:

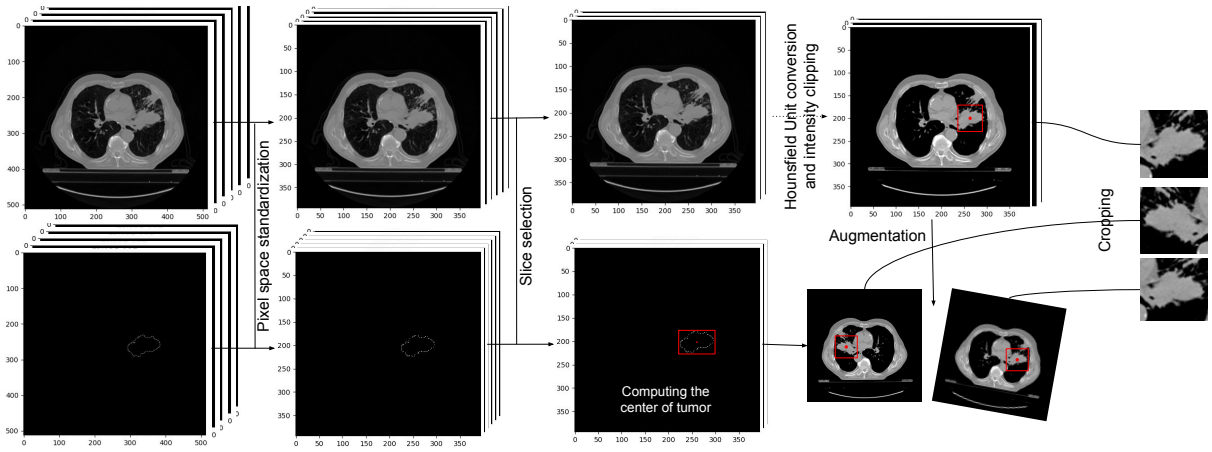

Figure 5.2: Key steps of the preprocessing pipeline.

$$\boldsymbol{w_k} = \boldsymbol{w_k} - \eta \nabla L_k(\boldsymbol{w_k}) \tag{5.2}$$

At the end of local training, each client sends its updated model parameters $\boldsymbol{w_k}$ to the server which aggregates them through averaging to form the new global model parameters $\boldsymbol{w_s}$:

$$\boldsymbol{w_s} = \frac{1}{K} \sum_{k=1}^{K} \boldsymbol{w_k} \tag{5.3}$$

Finally, the server updates the global model and sends $\boldsymbol{w_s}$ back to the clients, repeating the process for several communication rounds until convergence. A schematic version of HFL is depicted in Figure 5.1.

5.2.2 Overall framework

The following subsections detail our approach, which consists of three key components: preprocessing, training methodology, and inference procedure. This structure provides a comprehensive overview of our experimental framework.

Preprocessing

We employed a series of preprocessing steps on raw DICOM images prior to network input, drawing upon established practices in medical image analysis and our previous research experiences. Our approach employs 2D tumor patches; hence, from 3D raw scans, we only included the slices where the tumor is present, identified using existing segmentation data. The key stages of the preprocessing pipeline are illustrated in Figure 5.2. First, we converted the pixel intensities into Hounsfield Units (HU)s and standardized the photometric interpre-

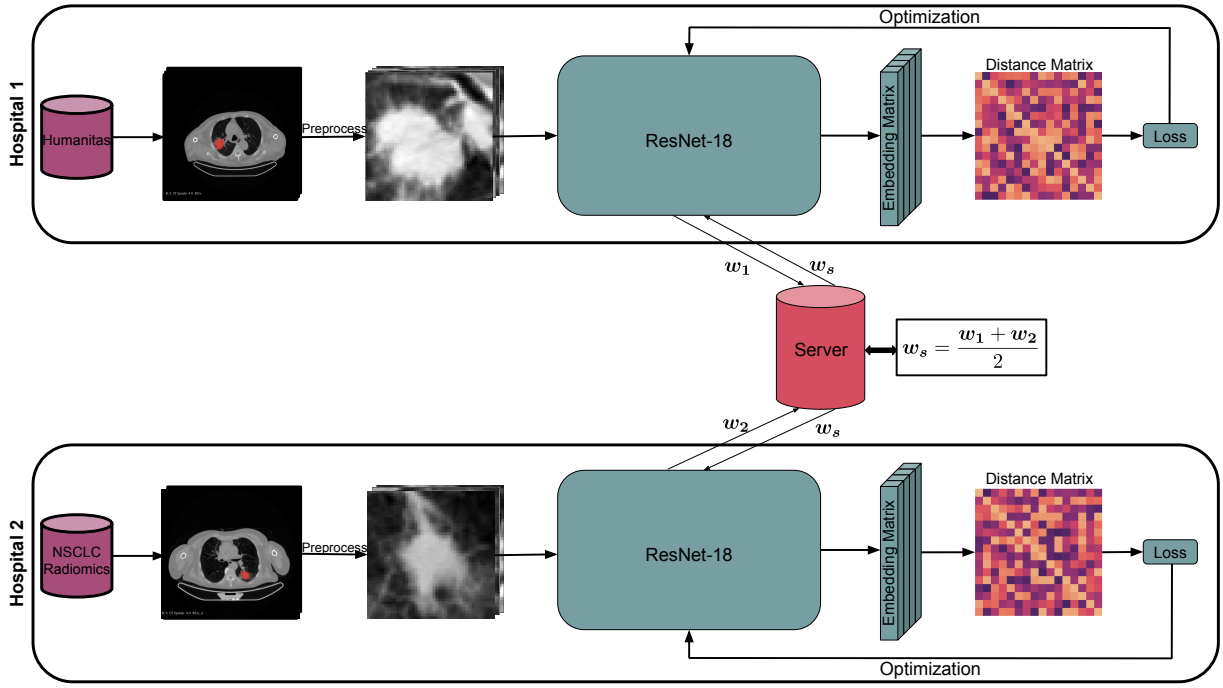

Figure 5.3: Training pipeline. The proposed approach integrates triplet networks with HFL. Each hospital and the central server share an identical network architecture, specifically ResNet-18. During the local training phase, each model is trained independently using triplet loss on its respective local dataset (Humanitas and NSCLC-Radiomics). Triplet selection and loss computation are performed as described in our previous work, presented in Chapter 4.2.1. After several local training epochs, both hospitals send their model weights (w_1 and w_2) to the server. The server aggregates these weights by averaging them and returns the aggregated weights (w_s) to each client.

tation to ensure that higher intensities represent brighter regions in all scans, promoting consistency across different CT scanners and protocols. We then applied linear interpolation to the CT scans to achieve uniform pixel spacing, adopting the most commonly observed dimension in our dataset, 0.977×0.977. Meanwhile, we converted the segmentation coordinates into pixel data and interpolated them using the nearest neighbor algorithm to preserve the same dimension with CT slices. From these pixel data, we extracted bounding boxes by calculating the minimum and maximum coordinates of the tumor pixels in each slice. We excluded slices where the area of the bounding box was lower than 100 pixels to eliminate slices with negligible tumor presence. Furthermore, we calculated the center points of each bounding box to use later in cropping after augmentations. Before augmentation, we clipped the pixel intensities to the range of [-600, 400] HUs, which covers the typical values for soft tissues and tumors while excluding air and dense bone. Finally, we normalized the clipped intensities to the range [0, 1] to standardize the input for our neural network.

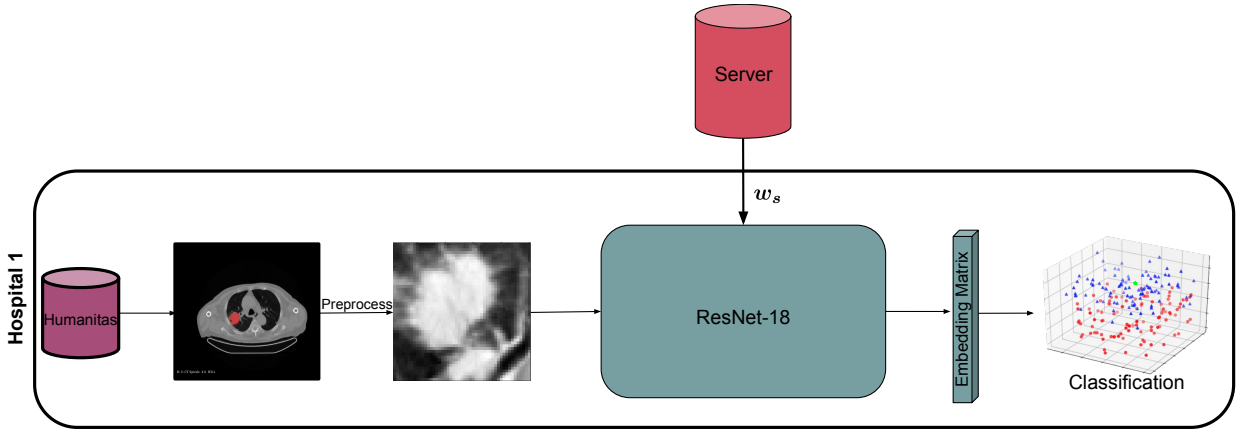

Figure 5.4: Inference procedure. During inference, the final network weights (w_s) are retrieved from the server. The embeddings of a query sample are extracted using this network and subsequently classified with kNN.

Training

Figure 5.3 demonstrates the training procedure of the proposed method in the case of two hospitals. Each client has a local copy of the network, trained locally with triplet loss. First, preprocessed images are fed into a CNN to extract deep features. Specifically, we employed a modified version of ResNet-18 [44], a well-known CNN architecture, where we removed the final classification layer to focus on feature extraction. After obtaining the feature vectors, we computed their pairwise Euclidean distances within each batch. This distance matrix serves as the basis for selecting positive and negative instances for each anchor image. In our implementation, we adopted a strategy of selecting the "easiest" positive and negative instances. This means we chose the same-label instance (positive) and the opposite-label instance (negative) with the minimum distances to each anchor image. While selecting positives, we excluded the patches that belong to the same patient as the anchor. This choice was made to prevent the model from learning patient-specific features rather than generalizable disease characteristics. By selecting positives from different patients, we ensure that the model focuses on identifying common patterns across patients, which is crucial for developing a robust and generalizable diagnostic tool. Finally, we calculate the loss, which is explained in detail in Section 4.2.1 using these selected instances. After all the clients are trained for a predefined number of epochs, the weights are shared with the server, where they are averaged and returned. This process is repeated for several rounds until convergence.

Inference

During the inference step, which is depicted in Figure 5.4, the final model weights are gathered from the server to be used in the classification of new instances. The process then

proceeds as follows: the CNN extracts a d-dimensional feature vector $f(x)$ from a given query image x. This feature vector is then mapped into a learned feature space, where the CNN's training process has ideally separated different classes through loss function minimization. Classification of the test instance is performed using a kNN classifier, a method chosen for its documented superiority in related literature [46, 89]. Given that we're dealing with CT scans, where each patient has multiple slices, the final classification for a test patient is determined by majority voting across predictions for all slices containing tumor in their CT volume. This approach leverages the full information available in the CT scan to make a robust classification decision.

5.2.3 Experimental Configuration

The training process follows a federated learning approach, consisting of 50 rounds. In each round, the networks are trained locally for 5 epochs before the local weights are sent to the server, averaged, and distributed back to all networks. This results in a total of 250 epochs of training. We use a batch size of 64 and the AdamW optimizer [72] with a weight decay of 0.001 and an initial learning rate of 0.01. The learning rate is reduced by a factor of 0.1 every 50 epochs. Our model is based on a modified ResNet-18 architecture, initialized with weights pre-trained on ImageNet. To adapt the model to our specific task, the first three layers remain frozen while only the fourth layer is fine-tuned. For the triplet loss function, we set the margin parameter to 0.5. For the cross-entropy loss, we used class weights to mitigate the class imbalance problem. During the inference, we selected k as 5 in kNN.

To augment our datasets, we applied geometric transformations to the preprocessed images before feeding them into the network. The augmentation process involved four distinct transformations: horizontal flipping, vertical flipping, rotation, and translation. For rotation, we randomly selected an angle between -90° and 90°, while for translation, we shifted the images by a random number of pixels ranging from 0 to 10 in each dimension. Importantly, the rotation was performed around the tumor center, which had been previously calculated during the preprocessing step detailed in Section 5.2.2. Following these transformations, we extracted a 64 × 64 pixel patch centered on the tumor. To ensure a balanced and robust training set, we applied one or more of these transformations iteratively to the image patches, resulting in 5,000 augmented patches for each class in the training dataset.

In all experiments, we employed stratified 5-fold cross-validation at the patient level. This means that all slices from a single patient are contained within either the training or test set, but never both. This approach helps to prevent any bias that could arise from having data from the same patient in both the training and test sets. To ensure a fair comparison across methods, the same patient-level splits were consistently used in all experiments.

Table 5.1: Average results of the 5 folds. In bold, are the largest scores of each metric. *Proposed method.

Method	Accuracy	AUROC	Sensitivity	Specificity	Gmean
*Triplet (Federated learning)	**0.664**	**0.664**	0.739	0.550	**0.630**
Softmax (Federated learning)	0.579	0.610	0.586	**0.567**	0.572
Triplet (Local dataset only)	**0.664**	0.654	**0.793**	0.468	0.604
Softmax (Local dataset only)	0.629	0.632	0.726	0.477	0.582
Chaunzwa et al. [21]	0.600	0.581	0.680	0.478	0.568

5.3 Results

Our study utilized two datasets: Humanitas, which served as our main source of data, and NSCLC Radiomics to increase the amount of data used to train the model. We conducted a series of experiments to evaluate: 1) the effects of employing an external dataset trained in a federated setting against using the local dataset alone, and 2) the effectiveness of triplet loss compared to traditional cross-entropy loss combined with softmax activation. For brevity, we will refer to the latter simply as softmax loss from this point forward. Our assessment included training the same network (ResNet-18) with both triplet and softmax loss on the private dataset. We also incorporated the NSCLC Radiomics dataset [2] into training using a federated learning approach. Furthermore, we compared our method to work by Chaunzwa et al. [21], who employed a comparable approach using 2D tumor patches as input to a CNN (specifically VGG-16 pre-trained on ImageNet) for histological subtype classification. Since the open-source code of Chaunzwa's method is not available, the model was reimplemented based on the details provided in the paper and was retrained and tested on the same dataset using the same fold configuration as in our experiments. We observed that Chaunzwa's method exhibited lower performance compared to the results reported in the original study. For evaluation, we employed five distinct metrics: accuracy, AUROC, sensitivity, specificity, and the geometric mean of sensitivity and specificity (Gmean).

Table 5.1 presents the results of our experiments. The first two rows report the results of experiments where the models are trained in a federated setting using both datasets. Specifically, the first row shows the results of the experiments where the models are trained with triplet loss, while the second row presents those with softmax loss. The subsequent two rows report the results of experiments where the model is trained on the local dataset only. The third row corresponds to the experiment with triplet loss, and the fourth row to that with softmax loss. Finally, the last row presents the results of the competitor model [21], which is trained on the local dataset only.

As shown in Table 5.1, the proposed method achieves the highest AUROC and Gmean scores while sharing the top accuracy with the other triplet approach trained solely on

the local dataset. To address the first objective, we evaluated the effect of incorporating an external dataset through federated learning compared to using only the local dataset. When applying the softmax loss, performance declined across key metrics, reflecting the challenges of leveraging external data without direct data sharing. In contrast, with triplet loss, performance was preserved. Although small gains were observed (1% in AUROC and 3% in Gmean), these differences were not statistically significant ($p > 0.05$). This result indicates that triplet loss maintains stable performance when trained on decentralized data, supporting its suitability for privacy-preserving collaborative learning. Regarding the second objective, we compared triplet loss against softmax loss within the federated framework. Triplet loss led to notable improvements, including an 8% increase in accuracy, a 5% increase in AUROC, a 5% increase in sensitivity, and a 6% increase in Gmean, with only a 2% decrease in specificity. The pairwise Wilcoxon signed-rank test across folds confirmed that these differences were significant ($p < 0.05$) for accuracy, AUROC, and sensitivity. Finally, the competitor model underperformed, with accuracy 6% lower, AUROC 8% lower, and Gmean 6% lower than our proposed method, ranking it lowest in AUROC and Gmean and second lowest in accuracy among all approaches. The Wilcoxon test confirmed that all these differences were statistically significant ($p < 0.05$).

These results demonstrate the robustness of our proposed method, which employs triplet loss in a federated learning setting. The effectiveness of triplet loss over softmax loss was evident, with improvements in almost all evaluated metrics. This highlights the potential of triplet networks in combination with federated learning to preserve privacy while maintaining performance. Notably, this advantage was not observed with softmax loss, underscoring the importance of loss function selection in federated learning contexts.

5.4 Conclusion

Our research demonstrates the effectiveness of combining federated learning with triplet loss preserving data privacy while maintaining model performance. The proposed method consistently outperformed alternative approaches across key metrics, particularly in AUROC and Gmean scores, while maintaining competitive accuracy. In the federated setting, triplet loss preserved the performance in the main evaluated metrics compared to softmax loss. Importantly, our findings highlight the synergistic effect of combining federated learning with triplet loss. This approach demonstrated robustness in leveraging external datasets. In contrast, models trained with softmax loss showed decreased performance in federated settings, emphasizing the critical role of loss function selection in federated learning contexts. However, a limitation of this study is that we did not systematically investigate how performance changes with varying amounts of training data. Future research should evaluate the impact

of gradually reducing training samples to clarify whether the superiority of the triplet-based approach is primarily pronounced when training data is limited or driven by differences in how the models train and stabilize between loss functions. Controlled experiments on progressively smaller subsets would provide deeper insights into the capabilities and robustness of the triplet networks.

While increasing the amount of data is essential, enriching the informational content of the data is equally important. In medical imaging, different modalities capture distinct and complementary aspects of the underlying pathology, and integrating them can provide a more comprehensive representation of the disease. Therefore, in the following chapter, we investigate multimodal learning approaches by incorporating PET imaging in addition to CT, aiming to assess the benefits of integrating complementary modalities for more accurate and robust classification.

Chapter 6

Multi-stage intermediate fusion for multimodal learning to classify NSCLC subtypes from CT and PET

6.1 Introduction

Most deep learning approaches for medical imaging use single modalities, but combining multiple data sources can improve performance. This approach, known as Multimodal Deep Learning (MDL), has shown promising results in healthcare applications [38]. MDL techniques can be categorized into three main methods: early, intermediate, and late fusion. Early fusion combines features at the raw data level, which can lead to the loss of unique modality-specific traits, while late fusion occurs at the decision level and may overlook deeper interactions between modalities. In contrast, intermediate fusion integrates data at the feature extraction stage, offering a more effective combination of modality-specific characteristics.

Table 6.1 presents a subset of Table 2.1, highlighting studies in the literature that employ PET and CT images for histological subtype classification, including key information such as fusion method, input type, evaluation strategy, and performance metrics reported in the respective papers. Detailed information about these studies can be found in Table 2.1 and Section 2.2.2. As seen in Table 6.1, only one approach [79] employs 3D input, while all other studies use either tumor patches or whole slices, thereby losing spatial coherence between slices. Only [79] performs 10-fold cross-validation; all other studies use a hold-out test set, making it difficult to assess the generalizability of these models. None of the studies provide the source code, limiting reproducibility. Finally, four out of five studies [79] employs an early fusion strategy to combine CT and PET images, integrating pixel values

from both modalities through image registration. Since this combination happens at the raw data level, it may result in the loss of modality-specific characteristics. As an alternative, intermediate fusion merges data during the feature extraction process. This can help preserve the unique information from each modality. This method works well for complex biomedical data [38]. MDL models can learn the non-linear relationships between modalities, which may help interpret the complementary information from both CT and PET. Despite potential benefits, only one study [79] has utilized intermediate fusion, indicating that this approach remains relatively unexplored in the context of histological subtype classification. Qin et al. [79] proposed two DenseNet-based CNNs to separately extract features from CT and PET images, followed by a gated multimodal unit to fuse these features, and a fully connected layer to classify histological subtypes. It is worth noting that such an approach, which first extracts deep features using one backbone network per modality, and then merges them at a single point before the classification head, is widely used in the literature [38]. However, this method overlooks the potential benefits of spatial correlations between the modalities, as the features are reduced to a vector at the end of the feature extraction backbones. Given that CT and PET images are acquired simultaneously, there is a high degree of spatial correlation between them, despite slight misalignments caused by respiratory motion.

Table 6.1: Subset of Table 2.1 showing the studies in the literature on multimodal deep learning methods for histological subtype classifications using PET and CT scans. The performance metrics are those reported in the respective papers and not the results of our reimplementations.

Study	**Fusion Method**	**Input type**	**Evaluation strategy**	**Performance metrics**
Qin et al. [79]	Intermediate	Whole volume w/ tumor	10-fold cross-validation	0.92 AUC, 0.72 ACC
Han et al. [43]	Early	Tumor slice	Hold-out test set	0.90 AUC, 0.84 ACC
Jacob & Menon [54]	Early	Whole slice	Hold-out test set	0.92 AUC, 0.95 ACC
Barbouchi et al. [10]	Early	Whole slice	Hold-out test set	0.98 AUC, 0.96 ACC
Zhao et al. [111]	Early	Tumor slice	Hold-out test set	0.77 AUC, 0.76 ACC

On these grounds, we propose MINT, a novel Multi-stage INTermediate fusion approach in MDL applied to histological subtype classification in NSCLC. Our approach fuses knowledge extracted by CNNs from CT and PET images at different levels of abstraction, enabling gradual integration across multiple layers of the feature hierarchy. To demonstrate the advantages of our intermediate fusion approach, we evaluated its performance against several

benchmarks. First, we compared it to unimodal models, which we implemented using the individual branches of our proposed multimodal model, as well as relevant models from the literature. Additionally, we compared our approach with other fusion strategies, specifically early and late fusion techniques. Finally, we benchmarked our model against the only existing study [79] utilizing an intermediate fusion of CT and PET images for histological subtype classification.

6.2 Methods

We propose an end-to-end deep learning classification pipeline that takes raw CT and PET images as input and predicts the histological subtypes. The entire framework is depicted in Figure 6.1. Our pipeline starts with a pre-processing step shown in panel (a), where the raw scans are prepared for processing by the network. The corresponding details are presented in Section 6.2.1. After the pre-processing, the images are fed into a custom 3D multimodal multi-fusion network for classification (panel (b) of Figure 6.1), with architectural details explained in Section 6.2.2. It utilizes a multi-stage fusion mechanism, in which fusion occurs repeatedly throughout the network, beginning in the early layers where feature maps still retain 3D spatial information. This approach allows us to leverage the spatial correlation between modalities. At each stage of our architecture, features are fused and then redistributed across the unimodal backbones within the same fusion block. This strategy enables the network to extract modality-specific features, effectively harnessing the complementary information provided by both imaging modalities in greater detail.

6.2.1 Pre-processing

We applied several preprocessing steps to the raw CT and PET scans before feeding them to the network. CT and PET scans contain large volumes of data that extend beyond the lungs. Since tumors represent only a small portion of each scan, this extra information makes it difficult for DL models to learn meaningful features. Additionally, different imaging machines produce scans with different characteristics, which complicates the learning process.

To address these challenges, we developed a preprocessing pipeline that standardizes the scans and focuses on the lung regions. First, we standardized the photometric interpretation to correct scans where higher intensity values represented darker regions. Second, we converted CT intensities to HU and PET intensities to SUV. Third, we applied linear interpolation to normalize slice thickness and pixel spacing across all scans. We set the xyz dimensions to 0.977 mm $\times$ 0.977 mm $\times$ 3.27 mm, which were the most common values in our dataset. Fourth, we aligned the CT and PET scans to ensure they matched spatially. Fifth, we used an established segmentation algorithm [47] to segment the lungs from CT

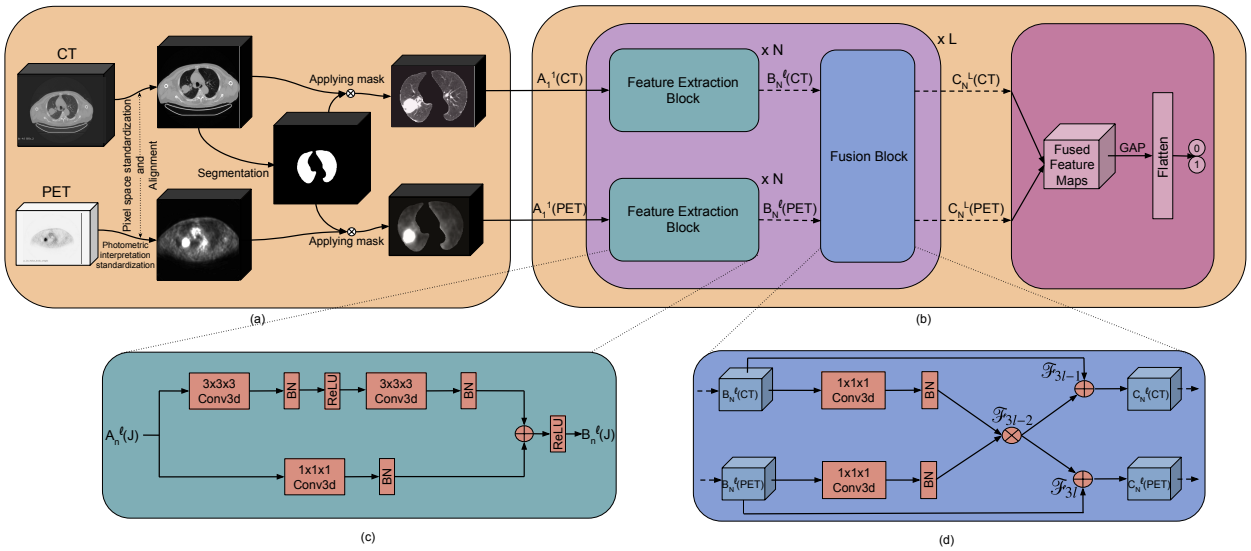

Figure 6.1: Overall framework of the proposed method. (a) Pre-processing. (b) Proposed multimodal convolutional architecture, where N indicates the number of feature extraction blocks in a stage, and L represents the number of stages in the model. GAP stands for Global Average Pooling, and the outputs 0 and 1 correspond to SQC and ADC, respectively. (c) Detailed schema of a feature extraction block: $A_n^l(J)$ and $B_n^l(J)$ represent the input and output of the nth block of the lth layer in the branch J, with J corresponding to either CT or PET. (d) Detailed schema of a fusion block: $B_N^l(CT)$ and $B_N^l(PET)$ are the inputs to the fusion block from the CT and PET branches, respectively, while $C_N^l(CT)$ and $C_N^l(PET)$ are the outputs of the corresponding branches; $\mathscr{F}_i$ represents a fusion point.

images and applied the resulting masks to both CT and PET scans. Finally, we clipped pixel intensities to [-1024, 1024] for CT and [0, 20] for PET, then normalized all values to the range [0, 1] to standardize the input for the DL models.

6.2.2 Network architecture

We designed a network architecture to extract and fuse features from both CT and PET images simultaneously. This design allows the modalities to inform each other throughout feature extraction. The overall network is organized in L *stages*, represented in violet in panel (b) of Figure 6.1, each performing feature extraction and fusion. Within each stage, we have N feature extraction blocks per modality represented in dark green in the figure. After extracting the features, each stage has one fusion block, shown in blue in the figure, that merges these features. We use L and N to highlight the network's modular design, which allowed us to test and find the best number of stages and blocks during our experiments. We provide the specific details, including network parameters and training procedure, in Section 6.2.3. At the end of this modular architecture, we concatenate the fused feature maps from both modalities along the channel dimension. This latent space is then passed

through the classification head. The head uses a Global Average Pooling (GAP) layer to average the spatial dimensions and generate a feature vector. A final fully connected output layer with two neurons then makes the final predictions.

Feature extraction block

The feature extraction blocks extract deep feature maps using the basic block of well-established 3D ResNet architecture [100], which has proven effective across a wide range of domains. This block consists of a main branch and a residual branch, enabling the construction of deeper networks by mitigating the vanishing gradient problem. Consequently, it allows us to increase the number of feature extraction blocks (N) and stages (L) in our implementation. As depicted in panel (c) of Figure 6.1, the main branch begins with a $3 \times 3 \times 3$ convolutional layer, followed by a batch normalization layer and a ReLU activation function. This is succeeded by another $3 \times 3 \times 3$ convolutional layer, followed again by a batch normalization layer. In parallel, the residual branch of the first feature extraction block in each stage includes a $1 \times 1 \times 1$ convolutional layer, followed by batch normalization. This convolution ensures the spatial dimensions align with the main branch, especially when the main branch reduces the spatial dimensions. The outputs of the two branches are merged through element-wise summation, then a ReLU activation function is applied. As shown in panel (b) of Figure 6.1, the output of a block, $B_n^l(J)$, can be passed either to the next feature extraction block, $A_{n+1}^l(J)$ (since there could be N block) or to a fusion block, B_N^l. In the initial block of each stage, the first convolutional layer uses a stride of 2 and produces $2 \times C$ feature maps, where C represents the number of feature maps in the final layer of the preceding stage. This convolutional layer reduces the spatial dimensions of the input, addressing the absence of pooling layers throughout the network. Additionally, it enhances the representational capacity of the extracted features by doubling the number of feature maps. The subsequent convolutional layers within the stage use a stride of 1 and also produce $2 \times C$ feature maps, maintaining the increased depth of the feature maps while preserving spatial dimensions. Hence, the parallel branch serves as a skip connection since the spatial dimensions are preserved.

Fusion block

Our core idea is that CT and PET images provide complementary information that can improve feature extraction for each patient. To achieve this, we design each fusion block to combine data from both modalities using element-wise multiplication. We also introduce two residual branches to incorporate the fused features back into the original unimodal data.

As illustrated in panel (d) of Figure 6.1, we denote as $B_N^l(CT)$ and $B_N^l(PET)$, the feature maps corresponding to the outputs of the previous basic blocks, which extract the features from CT and PET branches, respectively. These feature maps are first passed through a $1 \times 1 \times 1$ convolutional layer with an output feature map size of 1, squeezing the feature maps along the channel dimension and yielding a single feature map for each modality. After a batch normalization step applied to both modalities, we introduce an element-wise multiplication between the two feature maps. The resulting fused feature map is then added to the input feature maps, $B_N^l(CT)$ and $B_N^l(PET)$ with element-wise summation, producing the output maps $C_N^l(CT)$ and $C_N^l(PET)$. We formalize this fusion process by the following equations:

$$Fusion = BN(f_1^1(B_N^l(CT))) \otimes BN(f_1^1(B_N^l(PET))) \tag{6.1}$$

$$C_N^l(CT) = B_N^l(CT) \oplus Fusion \tag{6.2}$$

$$C_N^l(PET) = B_N^l(PET) \oplus Fusion \tag{6.3}$$

where BN denotes the batch normalization, and f_k^c represents a convolutional layer with a channel size of c and a kernel size of $k \times k \times k$. The symbols $\oplus$ and $\otimes$ are used to indicate element-wise addition and element-wise multiplication, respectively.

6.2.3 Network configuration

In the previous section, we reported that the network consists of N feature extraction blocks and L stages so that it is possible to identify the best configuration for a given task. To this end, we performed a grid search by varying both N and L in the range $[1, 5]$. We applied stratified 5-fold cross-validation after shuffling the three datasets used in this work: Humanitas, NSCLC-Radiogenomics, and Lung-PET-CT-Dx. To ensure a fair comparison across methods, the same splits were consistently used in all experiments. The training, validation, and test sets comprised 60%, 20%, and 20% of samples, respectively. Straightforwardly, the architecture search was conducted on the validation set. During these experiments we trained the models for 100 epochs, with the learning rate reduced by a factor of 0.1 every 25 epochs, using the Adam optimizer with class weights and an initial learning rate equal to 0.001. We evaluated performance using accuracy, the AUROC, and the geometric mean of sensitivity and specificity (Gmean). In particular, we consider AUROC and Gmean since the a-priori class distribution is imbalanced. Indeed, the former focuses on the model's ranking ability, evaluating how well a model differentiates between the two classes regardless of the class distribution. The latter offers a balanced assessment of the two classes ensuring that the model performs well for both. The results of this grid search showed us that the best-performing architecture consists of three stages, with three blocks per stage.

Table 6.2: Detailed architecture of MINT. Each convolutional layer is denoted as (*kernel size, number of filters, stride*). FE represents the feature extraction blocks. The second column in each branch denotes the parallel branch within the feature extraction blocks. GAP: Global Average Pooling, FC: Fully Connected

Stage	Block	CT branch		PET branch	
1	FE 1	3×3×3, 16, 2 3×3×3, 16, 1	1×1×1, 16, 2	3×3×3, 16, 2 3×3×3, 16, 1	1×1×1, 16, 2
	FE 2	3×3×3, 16, 1 3×3×3, 16, 1		3×3×3, 16, 1 3×3×3, 16, 1	
	FE 3	3×3×3, 16, 1 3×3×3, 16, 1		3×3×3, 16, 1 3×3×3, 16, 1	
	Fusion	1×1×1, 1, 1		1×1×1, 1, 1	
2	FE 1	3×3×3, 32, 2 3×3×3, 32, 1	1×1×1, 32, 2	3×3×3, 32, 2 3×3×3, 32, 1	1×1×1, 32, 2
	FE 2	3×3×3, 32, 1 3×3×3, 32, 1		3×3×3, 32, 1 3×3×3, 32, 1	
	FE 3	3×3×3, 32, 1 3×3×3, 32, 1		3×3×3, 32, 1 3×3×3, 32, 1	
	Fusion	1×1×1, 1, 1		1×1×1, 1, 1	
3	FE 1	3×3×3, 64, 2 3×3×3, 64, 1	1×1×1, 64, 2	3×3×3, 64, 2 3×3×3, 64, 1	1×1×1, 64, 2
	FE 2	3×3×3, 64, 1 3×3×3, 64, 1		3×3×3, 64, 1 3×3×3, 64, 1	
	FE 3	3×3×3, 64, 1 3×3×3, 64, 1		3×3×3, 64, 1 3×3×3, 64, 1	
	Fusion	1×1×1, 1, 1		1×1×1, 1, 1	
Classifier		GAP, 2-d FC, softmax			

With reference to the number of feature maps throughout the network, we selected 16 feature maps for the first convolutional layer in the first block of the initial stage. Consequently, all convolutional layers within the first stage output 16 feature maps, while the second stage outputs 32, and the third stage outputs 64. As a result, the final feature vector, combining information from both modalities, consisted of 128 features. Table 6.2 details the architecture of MINT, and it provides a description of each layer within each block at every stage to clearly outline the network's structural components. Each stage consists of three consecutive feature extraction blocks (FE 1, FE 2, and FE 3), followed by one fusion block that integrates features from both the CT and PET branches. The branches are presented in two columns to highlight their parallel structure during feature extraction. At the end of the third stage, the classification head is implemented, consisting of a GAP layer, a 2D Fully Connected (FC) output layer, and a softmax activation function.

Given this network configuration, Figure 6.2 represents the overall fusion mechanism of our approach using the graph representation proposed in [38]. The nodes x_1 and x_2 represent

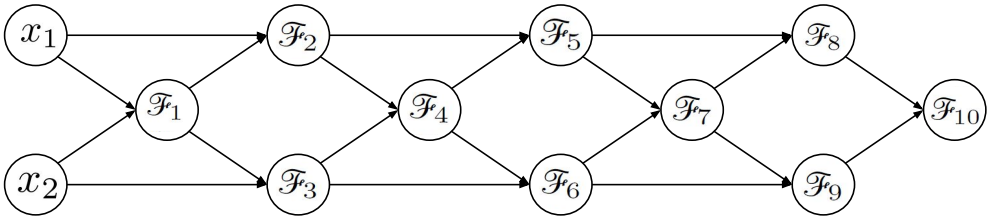

Figure 6.2: Graphical representation of the overall fusion mechanism. The nodes x_1 and x_2 represent the CT and PET inputs, respectively, while the nodes F_i denote the fusion points. This diagram illustrates only the flow of the fusion process and omits the architectural details between the fusion points.

the inputs of the network, $A_1^1(CT)$ and $A_1^1(PET)$ in Figure 6.1, respectively, while each $\mathscr{F}_i$ denotes a fusion occurring within the network. For visual clarity, we have omitted the multiple trainable layers between fusion points, representing only an arrow to indicate the fusion flow. In reality, these layers transform the inputs and extract deeper features. Still, according to [38], this process can be formalized by the following equations:

$$\mathscr{F}_{3l-2} = \otimes(\mathscr{F}_{3l-4}^{7}, \mathscr{F}_{3l-3}^{7}) \tag{6.4}$$

$$\mathscr{F}_{3l-1} = \oplus(\mathscr{F}_{3l-4}^{6}, \mathscr{F}_{3l-2}^{0}) \tag{6.5}$$

$$\mathscr{F}_{3l} = \oplus(\mathscr{F}_{3l-3}^{6}, \mathscr{F}_{3l-2}^{0}) \tag{6.6}$$

where $\mathscr{F}_i^j$ represents a fusion operation. $\mathscr{F}_{-1}$ and $\mathscr{F}_0$ are exceptions, as they represent the network inputs: x_1 and x_2 in Figure 6.2, which correspond to $A_1^1(CT)$ and $A_1^1(PET)$ in Figure 6.1, respectively. The subscript i is the fusion number while the superscript j is the number of trainable layers in which fusion inputs have been processed before the fusion. Finally, the symbols $\otimes()$ and $\oplus()$ represent element-wise multiplication and element-wise summation, respectively, and l is the stage index, which ranges from 1 to 3. The model is finalized with the following step of fusion:

$$\mathscr{F}_{10} = concat(\mathscr{F}_8^0, \mathscr{F}_9^0) \tag{6.7}$$

where $concat()$ indicates a concatenation operation.

6.3 Results

We conducted a series of experiments to evaluate the performance of the proposed model and to compare it against seven baseline methods. They are: i) four unimodal models that rely exclusively on either CT or PET imaging, ii) two alternative fusion strategies, i.e., early

Table 6.3: Average results across 5 folds, presented as mean (standard deviation). The highest scores for each metric are highlighted in bold.

	Model	#Params	Accuracy	AUROC	Gmean
Unimodal	CT branch	800k	.607 (.168)	.489 (.096)	.305 (.283)
	PET branch	800k	.624 (.206)	.465 (.153)	.329 (.109)
	DetectLC [30]	5.2M	.342 (.237)	.499 (.003)	.000 (.000)
	LUCY [98]	51.6M	**.762 (.008)**	.641 (.061)	.175 (.194)
Multimodal	Early fusion	800k	.655 (.164)	.452 (.078)	.224 (.218)
	Late fusion	1.6M	.657 (.109)	.513 (.116)	.342 (.244)
	Qin et al. [79]	2.5M	.539 (.164)	.421 (.073)	.280 (.171)
	MINT	1.6M	.724 (.030)	**.681 (.042)**	**.646 (.062)**

and late fusion methods, and iii) the only existing study utilizing intermediate fusion of CT and PET images for histological subtype classification [79].

In case i), we tested four different unimodal models. Two of them correspond to the CT and PET branches of our architecture, named as *CT Branch* and *PET Branch* in Table 6.3, respectively. Each branch consists of three stages with three blocks per stage, followed by a classification head, but no fusion layers. We selected the other two unimodal models from the literature: DetectLC [30] and LUCY [98], chosen because they classify histological subtypes using a 3D approach on CT lung volumes, similar to our unimodal approach in terms of input structure.

Case ii) tests early and late fusion by using the same unimodal architecture as before, i.e., two branches with three stages, each containing three blocks and a classification head at the end of each branch, but without any joint fusion blocks. To set up the early fusion, we merged the CT and PET images before feeding them into the network using element-wise multiplication, as in our multimodal approach. For late fusion, we again used the two separate unimodal branches and then averaged their output probabilities to make predictions during inference. Finally, in case iii) we compared our method with the only existing intermediate fusion approach for PET/CT histological subtype classification [79], which employs a single-fusion block that fuses the modalities after extracting individual feature vectors from each modality using two separate branches. Even though these branches are trained with a shared loss, the effect of each modality on the other remains at a high level of abstraction since the feature fusion occurs only once and before the classification head. In contrast, we have presented a multi-fusion method, where the fusions occur at various levels of the feature extraction hierarchy, preserving spatial correlations embedded in the feature maps and allowing for more extensive information sharing between modalities.

Table 6.3 presents the results attained by such seven competitors and by our method, reporting the average accuracy, AUROC, and Gmean scores computed across the five cross-

validation runs, along with the number of parameters each model has. The models selected from the literature were retrained and tested using the same dataset and fold configuration. For retraining, the published source code of LUCY [98] was used. Since the other two models [30, 79] do not have open-source implementations, they were reimplemented based on the information provided in their respective papers. Focusing on the results of the unimodal approaches, we notice that our multimodal method outperforms these competitors in all metrics, except for accuracy in the case of LUCY. We also observe that the two unimodal competitors drawn from the literature could not achieve the performance reported in their papers and attain the lowest Gmean scores, suggesting a bias toward one class. In particular, DetectLC collapses into a single class across all folds. Although LUCY demonstrated the highest accuracy and a comparable AUC score, its Gmean ranks as the second-worst: this suggests that it struggles to effectively predict the minority class, i.e., SQC in our dataset, and LUCY's high accuracy is likely a result of significant bias toward the majority class. To deepen this analysis we also run the Wilcoxon signed-rank test on the AUROC and Gmean scores, as these metrics better represent performance given the data's imbalance. The test was conducted pairwise across folds to assess whether the observed performance differences were statistically significant. In all pairwise comparisons for the Gmean score, our approach statistically differs from the unimodal approaches ($p < 0.05$). The same consideration holds for the AUROC score, except when comparing with LUCY ($p = 0.16$). It is also worth noting two unimodal baselines (CT and PET branches) are derived from our network and, hence, their comparison with our approach is equivalent to an ablation test. This observation, together with the previous ones, supports the consideration that neither modality alone captures the full range of meaningful features necessary for an effective classification.

Let us now turn our attention to the results of multimodal approaches in Table 6.3. We notice that our approach outperforms the other three in all metrics. The Wilcoxon signed-rank test shows that our performance statistically differs from all competitors for both AUROC and Gmean ($p < 0.05$), except for Gmean in the case of late fusion where we get $p = 0.0625$, which is close to the significance threshold. Furthermore, early fusion achieves a lower Gmean score than the unimodal backbones, suggesting that data-level fusion might even harm the classification model. While late fusion shows some improvement over early fusion, both methods still fall short of the proposed intermediate fusion approach, which demonstrates that fusion during the feature extraction process performs better than at the data or decision level. Finally, the sharp increase in all metrics compared to [79] demonstrates that our multi-stage voxel-wise fusion approach performs significantly better than a single-stage fusion of extracted features. This highlights the advantage of integrating features at multiple stages to better capture the complementary information between CT and PET modalities. Note that [79] could not achieve performance similar to what was

Table 6.4: Performance metrics for different kernels, filtered to include only those with 15 or more items.

Vendor	**Kernel**	**# Scans**	**Accuracy**	**AUROC**	**Gmean**
Siemens	B30f	19	.632	.629	.621
Siemens	B31s	231	.736	.658	.663
Siemens	B70f	78	.667	.738	.717
GE	Standard	355	.741	.667	.639

reported in their paper, which is also indicated in Table 2.1.

Turning our attention to efficiency, our model achieves these results with a modest parameter count of 1.6M, which is comparable to late fusion and significantly smaller than DetectLC (5.2M), LUCY (51.6M), and Qin et al. (2.5M). This highlights MINT's ability to balance efficiency while effectively leveraging multimodal information. Furthermore, MINT achieves an inference time of 185 ms per instance, which is low for clinical applications.

Furthermore, we evaluated our model's performance across different reconstruction kernels used in CT scanners. Since our dataset is composed of three real-world datasets, it naturally includes a variety of reconstruction kernels, each with distinct characteristics. Robustness to these variations demonstrates the model's ability to handle input variations and noise effectively. Reconstruction kernels significantly influence the magnitude and texture of noise [90]. Our merged dataset contains 10 different kernels from six CT models manufactured by two different vendors. However, as some kernels were used in only a small number of scans, we limited our analysis to those with at least 15 instances to ensure statistical reliability. Table 6.4 presents the results stratified by kernel type. We observe variations in all metrics across different kernels. However, in terms of Gmean, in all cases, our model achieves higher scores than the competing models reported in Table 6.3. Specifically, when applied to images reconstructed with the B70f kernel, it attains the highest Gmean score (0.717), followed by B31s (0.663). Even in the case of B30f (0.621), it still outperforms all competitors. Regarding AUROC, only in the case of B30f (0.629) does our model fall slightly below LUCY (0.641), whereas in all other cases, it surpasses all competing models. These findings demonstrate that despite variations in reconstruction settings, our model consistently performs well across different types of reconstructed images, highlighting its robustness and generalizability. To assess whether any statistically significant differences exist among kernel types, we conducted a Kruskal-Wallis test, which yielded a p-value of 0.3987. This result indicates no statistically significant differences between kernel types, confirming that our model is robust to variations in input noise.

Finally, we conducted an ablation study to assess the impact of multi-stage fusion. We modified the network, altering only the fusion points while maintaining the number of stages

Table 6.5: Results of fusion at different stages, presented as mean (standard deviation) across 5- folds.

Stage (l)	**Accuracy**	**AUROC**	**Gmean**
1	.573 (.059)	.570 (.064)	.560 (.032)
2	.702 (.080)	.628 (.067)	.493 (.278)
3	.653 (.092)	.604 (.085)	.559 (.087)

and feature extraction blocks. Specifically, instead of employing three fusion blocks (one at each stage), we evaluated configurations in which only one fusion block was placed at different stages. The results of these experiments are presented in Table 6.5, and they show that our multi-stage fusion approach outperforms single-stage fusion (i.e., fusion at only stage 1, 2, or 3) across all evaluation metrics.

6.4 Conclusion

In this work, we have presented a novel multimodal approach for histological subtype classification in NSCLC, utilizing an intermediate fusion method that integrates CT and PET images at various network depths. Our experiments show the effectiveness of this approach in comparison to unimodal baselines and other fusion techniques, being also able to handle the challenges posed by dataset imbalance. By harnessing the complementary information from both imaging modalities, we underscore the value of multimodal fusion in medical image analysis to provide a more comprehensive understanding of tumor characteristics.

While this chapter highlights the benefits of integrating multiple imaging modalities with specialized architectures, the next chapter turns to foundation models, a rapidly growing and broader area in AI. Unlike our prior work, which relied on task-specific architectures and multimodal inputs, this work explores whether large-scale pretrained models can provide strong generalization with minimal adaptation.

Chapter 7

NSCLC histological subtype classification from CT scans using generalist 3D medical foundation models

7.1 Introduction

DL models require large amounts of labeled data, but manual labeling is expensive and time-consuming. Most DL approaches in medical imaging use supervised learning, where models are trained on labeled data for specific tasks. These task-specific models, such as CNNs trained on tumor patches, slices, or entire volumes, have been successfully applied to various medical imaging modalities, including CT, PET, MRI, and X-ray. In cancer research, task-specific DL models have been employed for tumor detection, stage classification, prediction of overall survival outcomes, and histological subtype classification.

In the context of NSCLC histological subtype classification, early works primarily relied on traditional machine learning algorithms with hand-crafted radiomics features, while more recent studies have leveraged task-specific CNNs. Some trained their models using tumor patches [21, 64, 33, 4], others used entire CT slices [105, 78], and several approaches employed 3D CNNs [5, 30] or combined CNNs with LSTMs [99, 98] to enhance performance. To date, these methods have been designed for specific tasks and rely heavily on labeled data.

Foundation models offer an alternative approach. These models are trained on large, diverse datasets using self-supervised learning. They learn broad representations that can be adapted to different tasks through fine-tuning [14]. This approach is promising for medical imaging because it allows models to learn from extensive unlabeled data and then be fine-tuned on smaller labeled datasets.

To our knowledge, no study has yet investigated the use of foundation models for NSCLC

histological subtype classification using CT images. In this study, we aim to address this gap by evaluating three generalist medical foundation models pre-trained on 3D CT scans of diverse diseases and fine-tuning them on a dataset of 714 NSCLC patients. Additionally, we train and evaluate three task-specific models from the literature to provide a comparative baseline. This analysis allows us to assess the potential of foundation models in improving accuracy and efficiency for NSCLC subtype classification relative to conventional task-specific approaches.

7.2 Method

In this study, we investigated generalist vision foundation models trained on 3D CT scans for the task of NSCLC histological subtype classification and compared them with task-specific models that were explicitly designed, trained, and tested for this purpose. Section 7.2.1 introduces foundation models and elaborates on the specific models employed in this study. Section 7.2.2 provides an overview of the task-specific models and describes the architectures used in our experiments. Finally, Section 7.2.3 outlines the experimental setup and configurations.

7.2.1 Foundation models

In recent years, foundation models have gained significant attention. Foundation models are characterized by training on large and diverse datasets, enabling them to learn generalizable representations that can be adapted to many domains. These models are usually trained in two stages: pretraining and fine-tuning. During pretraining, these models often use self-supervised learning techniques, where they learn from large amounts of unlabeled data such as contrastive learning or masked image modelling. This helps them develop broad, transferable representations that require fewer labels. In the fine-tuning stage, the model is refined for specific tasks. Because of the knowledge gained in pretraining, fine-tuning typically requires only a small amount of labeled data, and in some cases, it can even be done without any task-specific labels.

One common approach in self-supervised learning is contrastive learning, which focuses on learning representations by comparing data samples. The core idea is to bring similar samples closer together in the representation space while pushing dissimilar ones farther apart. This approach typically relies on constructing positive pairs (e.g., augmentations of the same image) and negative pairs (e.g., different images) to train the model. By optimizing this contrastive objective, the model learns discriminative representations that are transferable to downstream tasks.

One notable example of contrastive learning is Contrastive Language-Image Pretraining (CLIP) [80], which extends this concept to multimodal data. CLIP learns joint representations of images and text by aligning them in a shared embedding space. It pairs each image with its corresponding caption as positive examples and uses unrelated image-caption pairs as negatives. This approach allows CLIP to learn powerful image and text features that can be effectively applied to a wide range of tasks. The architecture typically consists of two encoders: one for image encoding and one for text encoding. The embeddings generated by these encoders, are then aligned in a shared latent space, enabling the model to learn cross-modal relationships between the image and its textual description.

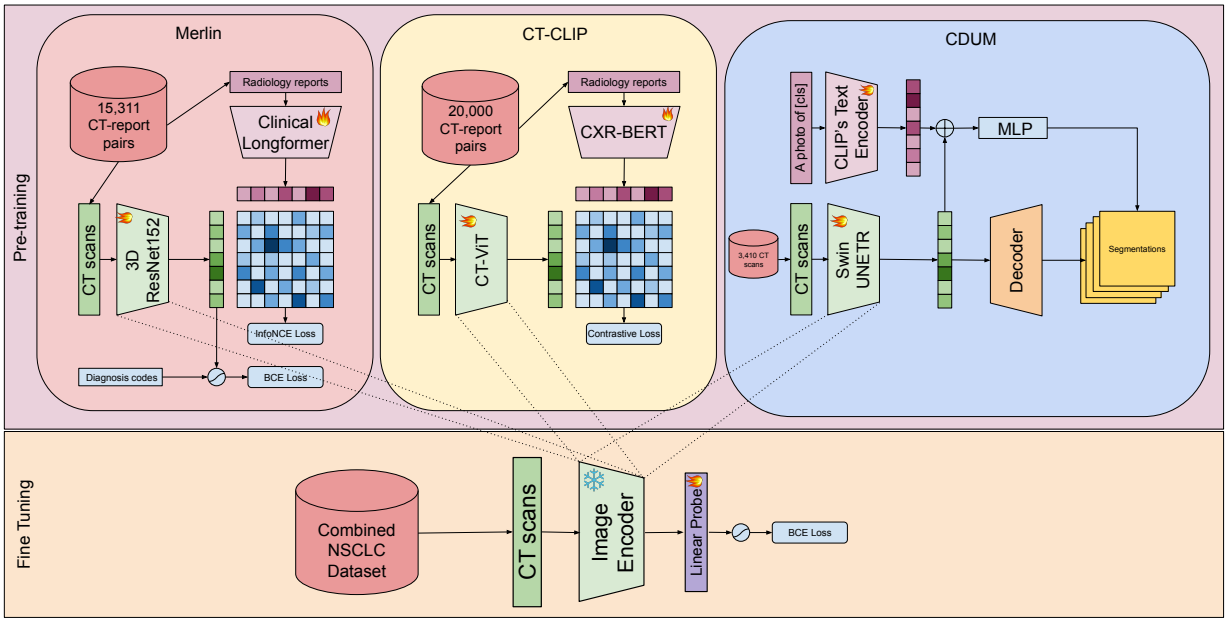

Figure 7.1: Pre-training and finetuning steps of each foundation model. MLP: Multi Layer Perceptron

In this study, we evaluated three foundation models, all trained on 3D CT scans. Figure 7.1 provides an overview of the pre-training and fine-tuning stages of these models. While the fine-tuning strategy remains consistent across all three models, the pre-training objectives and data sources differ substantially. The following paragraph describes each model in detail.

The first foundation model we employed is Merlin [12], a vision-language model. Merlin comprises an image encoder and a text encoder, specifically utilizing a 3D ResNet152 for image encoding and a clinical Longformer [67] for text encoding. The model was trained on 15,311 CT scans paired with Electronic Health Record (EHR) diagnosis codes and radiology reports. Training was conducted in two stages: first, binary cross-entropy loss was used with diagnosis codes, followed by contrastive training with radiology reports using the InfoNCE loss function. The model is evaluated on 752 individual tasks across six different task

types, including findings classification, phenotype classification, cross-modal retrieval, five-year chronic disease prediction, radiology report generation, and 3D semantic segmentation. The second foundation model is CT-CLIP [41], another vision-language model. They employed CT-ViT's image encoder [42], which is a vision transformer, as the image encoder and CXR-BERT [13] as the text encoder. CT-ViT is an autoregressive encoder-decoder network that utilizes all-to-all spatial and causal attention layers to encode CT tokens. The model was trained using contrastive learning on paired CT scans and radiology reports from 20,000 patients. The authors evaluated their model on two downstream tasks: multi-abnormality detection and case retrieval. The final foundation model is the CLIP-Driven Universal Model (CDUM) [70], which employs a training approach different from the previous two models. While it utilizes the text encoder of CLIP [80], it does not rely on a contrastive loss. Instead, the model uses text embeddings as a substitute for traditional one-hot class labels. As the vision encoder, it utilizes Swin UNETR [95], which consists of a Swin Transformer [71] to encode the 3D patches and a CNN-based decoder connected via skip connections, forming a U-shaped network. Designed for organ and tumor segmentation, CDUM was trained on 3,410 CT scans collected from 14 different datasets.

All three models are specifically designed to handle 3D data, in contrast to many other foundation models for medical imaging, which are limited to processing 2D images. Working with 2D slices often leads to various problems, such as a loss of spatial context between adjacent slices, limiting the model's ability to capture complex anatomical relationships and spatial dependencies. These challenges highlight the advantage of using 3D models, making them the preferred choice for this study to maintain spatial coherence and accurately capture the full volume of interest.

7.2.2 Task-specific models

Task-specific models are specialized architectures designed to solve narrowly defined problems by aligning their designs and training pipelines with specific data and goals. However, their reliance on high-quality annotated datasets makes them challenging to train effectively in data-limited settings.

The typical pipeline for training task-specific models begins with curating an annotated dataset tailored to the specific task. This dataset is systematically divided into subsets designated for training, validation (development), and testing. The process then involves identifying and implementing a suitable model architecture. The selected architecture is trained using the training set while its hyperparameters and performance are fine-tuned through validation on the development set. Finally, the model is rigorously evaluated on the test set to ensure its robustness and reliability.

In this study, we reimplemented and evaluated two models from the literature which are

specifically designed for histological subtype classification in NSCLC using 3D CT scans in addition to our previous model, the CT branch of MINT which is presented in Chapter 6. Our model is a 10-layer ResNet-based 3D CNN with approximately 800k parameters. It comprises three convolutional blocks, each containing two branches: a main branch and a residual branch with a bottleneck structure. In the main branch, two consecutive 3×3×3 convolutional layers are followed by batch normalization layers. The first convolution uses a stride of 2 to reduce the spatial dimensions, and a ReLU activation function is applied after batch normalization. The residual branch includes a 1×1×1 convolutional layer with a stride of 2, followed by a batch normalization layer. These two branches are combined via element-wise summation, followed by a final ReLU activation. After processing through the three convolutional blocks, global average pooling is applied to the resulting feature maps, which are then connected to the output layer.

The second model we employed is DETECT-LC [30], an 11-layer 3D CNN with residual connections and approximately 5.20 million parameters. The first six layers consist of 3D convolutional layers, each followed by a ReLU activation function and an average pooling layer. Residual connections are present between each of these layers, enhancing gradient flow and mitigating vanishing gradient issues. Following the convolutional layers, the model includes four dense layers, culminating in an output layer. Additionally, a dropout layer is applied to the flattened feature vector after the convolutional layers to reduce overfitting.

The final task-specific model we employed is LUCY [98], a 3-layer ConvLSTM-based neural network with approximately 51.6 million parameters. The first layer consists of a 3×3 ConvLSTM layer with 8 filters, followed by a dropout layer to mitigate overfitting. This is succeeded by a dense layer with 128 neurons and a ReLU activation function, accompanied by another dropout layer. Finally, the network concludes with an output layer featuring a softmax activation function for classification.

Although the three algorithms employ different preprocessing steps, they all utilize 3D CT scans as input, with the lungs segmented beforehand. We strictly followed the preprocessing procedures described in the respective papers and trained each model using the parameters specified in their original implementations.

7.2.3 Experimental configuration

For this study, we combined three datasets, Humanitas, NSCLC-Radiogenomics, and Lung-PET-CT-Dx, totaling 714 subjects. For our experiments, we applied stratified 5-fold cross-validation, ensuring that the training, validation, and test sets comprised 60%, 20%, and 20% of the samples, respectively. Notably, the same fold configuration was maintained across all experiments to ensure consistency and comparability of results.

All three task-specific models were trained from scratch using the training set, adhering

strictly to the hyperparameters and configurations specified in their respective papers. The validation set was employed for early stopping, utilizing different metrics tailored to each approach: the maximum geometric mean of sensitivity and specificity for our own model, CT branch of MINT, maximum accuracy for DetectLC [30], and minimum loss for LUCY [98]. Finally, all three models were evaluated on the same test set to ensure a consistent comparison.

For the foundation models, we utilized their image encoders to extract embeddings from 3D CT scans. To achieve this, we froze the entire model and passed the input scans through the image encoders. The size of the embeddings generated by the image encoders varies across the foundation models: 512 for both Merlin and CT-CLIP, and 7680 for CDUM. Once the embeddings were obtained, we trained a linear probe to classify histological subtypes based on these embeddings. The linear layer was trained using the cross-entropy loss function, where class weights were applied to address class imbalance, and the AdamW optimizer [73]. Training was conducted for 200 epochs, with early stopping applied based on the best geometric mean of sensitivity and specificity. We performed experiments with various learning rates and batch sizes for each model, selecting the combinations that yielded the best performance on the validation set. The learning rates are set to 0.0001, 0.001, 0.001; and batch sizes are set to 32, 64, 128 for the linear probes of Merlin, CT-CLIP, and CDUM, respectively. We used the same training, validation, and test sets as in the task-specific model experiments to ensure consistency in the evaluation process.

Table 7.1: Average results across 5 folds, presented as mean (standard deviation).

	Model	**Sensitivity**	**Specificity**	**Gmean**	**Accuracy**
Foundation	Merlin [12]	.614 (.054)	.654 (.084)	.632 (.049)	.623 (.044)
	CT-CLIP [41]	.568 (.046)	.624 (.067)	.593 (.015)	.581 (.021)
	CDUM [70]	.582 (.071)	.684 (.123)	.627 (.051)	.606 (.050)
Specialized	Our approach	.690 (.282)	.343 (.289)	.305 (.283)	.607 (.168)
	DetectLC [30]	.200 (.400)	.800 (.400)	.000 (.000)	.342 (.237)
	LUCY [98]	.976 (.031)	.065 (.098)	.175 (.194)	.762 (.008)

7.3 Results

In this study, we conducted a series of experiments to evaluate the performance of three open-source, open-weights foundation models: Merlin [12], CT-CLIP [41], and CDUM [70]. To benchmark the performance of these foundation models, we also assessed three task-specific

models, CT branch of MINT, DetecLC [30], and LUCY [98], that have demonstrated promising results in histological subtype classification for NSCLC. The evaluation employed four metrics: sensitivity, specificity, the geometric mean of sensitivity and specificity (Gmean), and accuracy. While calculating these metrics, we designated ADC as the positive class and SQC as the negative class. Given the imbalance in the dataset used for our experiments, we placed particular emphasis on Gmean, as it provides a balanced assessment by accounting for both sensitivity and specificity.

Table 7.1 presents the results of our experiments. The first three rows display the performance of the foundation models; Merlin, CT-CLIP, and CDUM, respectively. The last three rows summarize the performance of the task-specific models in order: our previous model, DetectLC, and LUCY. We observed lower performance for DetectLC [30] and LUCY [98] compared to the results reported in their original studies. This suggests that these models face one of the key limitations discussed in Section 2.3, namely limited generalizability.

Focusing on the Gmean scores, we observe a clear distinction between the foundation models and the task-specific models. Merlin achieves the highest Gmean score of 0.632, followed by CDUM with 0.627. CT-CLIP attains a comparatively lower score of 0.593. However, the task-specific models show a sharp decline in Gmean. Specifically, the best-performing task-specific model in terms of Gmean is the CT branch of MINT, our own model presented in Chapter 6, with a score of 0.305, approximately 28% lower than the worst-performing foundation model, CT-CLIP. This provides strong evidence of the robustness of the foundation models, highlighting their ability to balance performance across both classes more effectively than the task-specific models. Furthermore, we observe relatively lower scores for LUCY and DetectLC, with Gmean values of 0.175 and 0, respectively, suggesting a bias toward one class. In particular, DetectLC collapses into a single-class prediction across all folds. This is evident when analyzing the sensitivity and specificity scores. While all three foundation models achieve balanced sensitivity and specificity, task-specific models exhibit an unbalanced distribution, indicating a bias toward one of the classes. Notably, LUCY achieves extremely high sensitivity but extremely low specificity, demonstrating a clear bias toward the positive class, ADC, which is the majority class in the dataset. This suggests that LUCY predominantly predicts ADC, leading to an imbalanced classification. Similarly, we observe that DetectLC collapses into a single-class prediction across all folds. It predicts only SQC in four folds, while in one fold, it predicts only ADC. In all cases, it yields a Gmean score of 0. Consequently, DetectLC has an average sensitivity of 0.2, an average specificity of 0.8, and an average Gmean of 0. It is well known that accuracy is not a reliable metric for evaluating model performance on imbalanced datasets such as ours [75]. This is evident in our results, as LUCY achieves the highest accuracy score while having the lowest specificity score.

Another interesting observation is that task-specific models, except for LUCY, exhibit much higher standard deviations across nearly all metrics. This further reinforces the evidence that foundation models are more robust and generalizable. Since task-specific models are trained from scratch in each fold, their outcomes depend heavily on the specific samples included in the training set. With limited data, such dependence can easily lead to overfitting and unstable performance. In contrast, since the foundation models are frozen and only a linear layer is fine-tuned with the new data, they offer consistent feature extraction. This approach prevents the models from being overly influenced by the small, fold-specific training sets, ensuring greater stability and reduced bias toward the fine-tuning dataset. For the case of LUCY, we observe that it is consistently biased toward the majority class. As a result, it shows lower standard deviations compared to the other models.

These results provide strong evidence for the advantages of foundation models in histological subtype classification, particularly when working with limited labeled data. The superior performance of foundation models can be attributed to their large-scale pre-training, which enables them to extract meaningful and generalizable features without being overly influenced by the small training sets. In contrast, task-specific models, which are trained from scratch, struggle to generalize due to their reliance on a limited dataset, leading to increased variability in performance. Furthermore, the diverse datasets used during pretraining allow foundation models to develop robust hierarchical representations that transfer effectively to new tasks. This makes them inherently more stable and less prone to dataset-specific biases compared to task-specific models, which are more susceptible to overfitting. Overall, these findings emphasize the importance of foundation models in scenarios with limited data and demonstrate their potential for improving the robustness of NSCLC subtype classification.

7.4 Conclusion

In this study, we evaluated foundation models, a class of deep learning models trained on large-scale unannotated data that have shown strong generalization capabilities across various domains. We compared their performance with task-specific models, which are trained from scratch for a specific task, to assess their effectiveness in NSCLC histological subtype classification. The results indicate that foundation models outperform task-specific models, mainly due to their ability to leverage pre-trained features that provide stable and generalizable representations even when only limited training data are available. In contrast, task-specific models, which rely entirely on the available data for training, face challenges in achieving comparable generalization. Overall, these findings highlight the advantages of foundation models as a robust and efficient approach for medical imaging applications where annotated data are limited.

Chapter 8

Conclusions

This thesis has addressed the pressing need for accurate and non-invasive histological subtype classification in NSCLC, with a particular focus on differentiating ADC and SQC, two subtypes that require distinct treatment strategies. Recognizing the limitations of current biopsy-based approaches, including invasiveness, sampling bias, and diagnostic uncertainty, the work presented here leverages radiological imaging and deep learning to develop more accessible and scalable diagnostic tools. To address this objective, the thesis was structured around four research questions (RQ1–RQ4), each targeting a specific challenge in developing effective and generalizable diagnostic models for NSCLC subtype classification. The key findings corresponding to each research question are summarized below.

RQ1: *Can triplet networks improve model performance under limited data conditions?*

The first core contribution of this thesis investigated the use of triplet networks, a form of metric learning, to enhance classification performance under the constraint of limited training data, a frequent challenge in medical imaging due to privacy concerns and small sample sizes. Our comprehensive experiments demonstrated that triplet networks, particularly when paired with architectures such as ResNet-50 and GoogleNet, outperformed traditional deep learning models trained with softmax loss. Furthermore, our findings showed that selecting the top-t easiest positive and negative samples within a triplet batch improves training effectiveness. These findings confirm that triplet networks effectively address RQ1 by improving model performance under limited data conditions.

RQ2: *Can federated learning with triplet loss maintain performance while ensuring data privacy?*

To scale the benefits across institutions while preserving patient confidentiality, the second major contribution introduced a FL framework using triplet loss. This privacy-preserving approach enabled collaborative model training across decentralized datasets without data sharing. The combination of FL and triplet loss consistently improved key performance metrics such as AUROC and Gmean compared to both locally trained models and

federated models using softmax loss. These findings provide a positive answer to RQ2, demonstrating that federated triplet learning maintains strong performance while fully preserving data privacy.

RQ3: *How can CT and PET features be effectively integrated for NSCLC subtype classification?*

The third part of the thesis tackled the challenge of multimodal integration by proposing MINT, a novel Multi-stage INTermediate fusion architecture designed to combine CT and PET information at multiple depths of the network. MINT successfully leveraged the complementary anatomical and metabolic information from both modalities and outperformed unimodal models, as well as early, late, and existing intermediate fusion strategies. These results provide a clear answer to RQ3, demonstrating that multi-stage intermediate fusion is an effective strategy for capturing complementary multimodal information.

RQ4: *Can large-scale pretrained foundation models effectively predict histological subtypes in NSCLC?*

The final contribution shifted focus to foundation models, exploring their applicability in NSCLC subtype classification. These large-scale pretrained models, known for their ability to generalize across tasks with minimal fine-tuning, were evaluated against task-specific architectures. Despite class imbalance and dataset limitations, foundation models exhibited superior performance, especially in low-data settings, due to their robust feature representations learned from diverse pretraining sources. The results provide an affirmative answer to RQ4, showing that foundation models effectively generalize to NSCLC subtype classification.

Across all contributions, this thesis has proposed and validated strategies that address the critical barriers to developing reliable, non-invasive diagnostic tools: limited data availability, privacy constraints, the need for multimodal integration, and the challenge of building models that generalize well in clinical settings. The proposed solutions, triplet networks, federated metric learning, intermediate fusion, and foundation model adaptation, offer complementary advances toward this goal.

Looking ahead, several avenues for future work remain. These include expanding datasets to better capture the diversity of NSCLC subtypes, exploring more sophisticated federation strategies such as VFL and TFL, and incorporating genomic data to enhance multimodal fusion. Additionally, fine-tuning foundation models using advanced techniques such as Low-Rank Adaptation (LoRA) or evaluating their zero-shot capabilities can further push the boundaries of data-efficient model development. Ultimately, the integration of these strategies moves us closer to realizing non-invasive, accurate, and personalized diagnostic systems that can be translated into routine clinical practice for lung cancer care.

Bibliography

[1] HJWL Aerts, E Rios Velazquez, RTH Leijenaar, C Parmar, P Grossmann, S Carvalho, J Bussink, R Monshouwer, B Haibe-Kains, D Rietveld, F Hoebers, MM Rietbergen, CR Leemans, A Dekker, J Quackenbush, RJ Gillies, and P Lambin. Data From NSCLC-Radiomics-Genomics, 2015.

[2] Hugo J W L Aerts, Leonard Wee, Emmanuel Rios Velazquez, Ralph T H Leijenaar, Chintan Parmar, Patrick Grossmann, Sara Carvalho, Johan Bussink, Ren Monshouwer, Benjamin Haibe-Kains, Derek Rietveld, Frank Hoebers, Michelle M Rietbergen, C Ren Leemans, Andre Dekker, John Quackenbush, Robert J Gillies, and Philippe Lambin. Data From NSCLC-Radiomics, 2019.

[3] Hugo JWL Aerts, Emmanuel Rios Velazquez, Ralph TH Leijenaar, Chintan Parmar, Patrick Grossmann, Sara Carvalho, Johan Bussink, René Monshouwer, Benjamin Haibe-Kains, Derek Rietveld, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nature communications*, 5(1):4006, 2014.

[4] Fatih Aksu, Fabrizia Gelardi, Arturo Chiti, and Paolo Soda. Early Experiences on using Triplet Networks for Histological Subtype Classification in Non-Small Cell Lung Cancer. In *2023 IEEE 36th International Symposium on Computer-Based Medical Systems (CBMS)*, pages 832–837. IEEE, 2023.

[5] Fatih Aksu, Fabrizia Gelardi, Arturo Chiti, and Paolo Soda. Multi-stage intermediate fusion for multimodal learning to classify non-small cell lung cancer subtypes from CT and PET, 2025.

[6] B. Albertina, M. Watson, C. Holback, R. Jarosz, S. Kirk, Y. Lee, K. Rieger-Christ, and J Lemmerman. The Cancer Genome Atlas Lung Adenocarcinoma Collection (TCGA-LUAD) (Version 4) [Data set], 2016.

[7] Gerald Antoch, Jorg Stattaus, Andre T Nemat, Simone Marnitz, Thomas Beyer, Hilmar Kuehl, Andreas Bockisch, Jörg F Debatin, and Lutz S Freudenberg. Non–

small cell lung cancer: dual-modality PET/CT in preoperative staging. *Radiology*, 229(2):526–533, 2003.

[8] Shaimaa Bakr, Olivier Gevaert, Sebastian Echegaray, Kelsey Ayers, Mu Zhou, Majid Shafiq, Hong Zheng, Jalen Anthony Benson, Weiruo Zhang, Ann NC Leung, et al. A radiogenomic dataset of non-small cell lung cancer. *Scientific data*, 5(1):1–9, 2018.

[9] Shaimaa Bakr, Olivier Gevaert, Sebastian Echegaray, Kelsey Ayers, Mu Zhou, Majid Shafiq, Hong Zheng, Weiruo Zhang, Ann Leung, Michael Kadoch, Joseph Shrager, Andrew Quon, Daniel Rubin, Sylvia Plevritis, and Sandy Napel. Data for NSCLC Radiogenomics Collection (Version 4) [Data set], 2017.

[10] Khalil Barbouchi, Dhekra El Hamdi, Ines Elouedi, Takwa Ben Aïcha, Afef Kacem Echi, and Ihsen Slim. A transformer-based deep neural network for detection and classification of lung cancer via pet/ct images. *International Journal of Imaging Systems and Technology*, 33(4):1383–1395, 2023.

[11] Michael Luke Battle, Amir Atapour-Abarghouei, and Andrew Stephen McGough. Siamese neural networks for skin cancer classification and new class detection using clinical and dermoscopic image datasets. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 4346–4355. IEEE, 2022.

[12] Louis Blankemeier, Joseph Paul Cohen, Ashwin Kumar, Dave Van Veen, Syed Jamal Safdar Gardezi, Magdalini Paschali, Zhihong Chen, Jean-Benoit Delbrouck, Eduardo Reis, Cesar Truyts, Christian Bluethgen, Malte Engmann Kjeldskov Jensen, Sophie Ostmeier, Maya Varma, Jeya Maria Jose Valanarasu, Zhongnan Fang, Zepeng Huo, Zaid Nabulsi, Diego Ardila, Wei-Hung Weng, Edson Amaro Junior, Neera Ahuja, Jason Fries, Nigam H. Shah, Andrew Johnston, Robert D. Boutin, Andrew Wentland, Curtis P. Langlotz, Jason Hom, Sergios Gatidis, and Akshay S. Chaudhari. Merlin: A Vision Language Foundation Model for 3D Computed Tomography, 2024.

[13] Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, et al. Making the most of text semantics to improve biomedical vision–language processing. In *European conference on computer vision*, pages 1–21. Springer, 2022.

[14] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, and Emma Brunskill. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[15] Lorenzo Brigato and Luca Iocchi. A close look at deep learning with small data. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 2490–2497. IEEE, 2021.

[16] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. Signature verification using a" siamese" time delay neural network. *Advances in neural information processing systems*, 6, 1993.

[17] Joshua D Campbell, Anton Alexandrov, Jaegil Kim, Jeremiah Wala, Alice H Berger, Chandra Sekhar Pedamallu, Sachet A Shukla, Guangwu Guo, Angela N Brooks, Bradley A Murray, et al. Distinct patterns of somatic genome alterations in lung adenocarcinomas and squamous cell carcinomas. *Nature genetics*, 48(6):607–616, 2016.

[18] Camillo Maria Caruso, Valerio Guarrasi, Ermanno Cordelli, Rosa Sicilia, Silvia Gentile, Laura Messina, Michele Fiore, Claudia Piccolo, Bruno Beomonte Zobel, Giulio Iannello, et al. A multimodal ensemble driven by multiobjective optimisation to predict overall survival in non-small-cell lung cancer. *Journal of Imaging*, 8(11):298, 2022.

[19] Miaobin Cen and Cheolkon Jung. Fully convolutional siamese fusion networks for object tracking. In *2018 25Th IEEE international conference on image processing (ICIP)*, pages 3718–3722. IEEE, 2018.

[20] Kari Chansky, Jean-Paul Sculier, John J Crowley, Dori Giroux, Jan Van Meerbeeck, and Peter Goldstraw. The International Association for the Study of Lung Cancer Staging Project: prognostic factors and pathologic TNM stage in surgically managed non-small cell lung cancer. *Journal of thoracic oncology*, 4(7):792–801, 2009.

[21] Tafadzwa L Chaunzwa, Ahmed Hosny, Yiwen Xu, Andrea Shafer, Nancy Diao, Michael Lanuti, David C Christiani, Raymond H Mak, and Hugo JWL Aerts. Deep learning classification of lung cancer histology using CT images. *Scientific reports*, 11(1):1–12, 2021.

[22] Kun Chen, Manning Wang, and Zhijian Song. Multi-task learning-based histologic subtype classification of non-small cell lung cancer. *La radiologia medica*, 128(5):537–543, 2023.

[23] Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang. Beyond triplet loss: a deep quadruplet network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 403–412, 2017.

[24] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE computer society*

conference on computer vision and pattern recognition (CVPR'05), volume 1, pages 539–546. IEEE, 2005.

[25] Hind Dadoun, Anne-Laure Rousseau, Eric de Kerviler, Jean-Michel Correas, Anne-Marie Tissier, Fanny Joujou, Sylvain Bodard, Kemel Khezzane, Constance de Margerie-Mellon, Hervé Delingette, et al. Deep learning for the detection, localization, and characterization of focal liver lesions on abdominal US images. *Radiology: Artificial Intelligence*, 4(3):e210110, 2022.

[26] C de Margerie-Mellon, C De Bazelaire, and E De Kerviler. Image-guided biopsy in primary lung cancer: Why, when and how. *Diagnostic and interventional imaging*, 97(10):965–972, 2016.

[27] Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine learning research*, 7:1–30, 2006.

[28] Xingping Dong and Jianbing Shen. Triplet loss in siamese network for object tracking. In *Proceedings of the European conference on computer vision (ECCV)*, pages 459–474, 2018.

[29] George Drosatos, Pavlos S Efraimidis, and Avi Arampatzis. Federated and transfer learning applications, 2023.

[30] Karma M Fathalla, Sherin M Youssef, and Nourhan Mohammed. DETECT-LC: A 3D deep learning and textural radiomics computational model for lung cancer staging and tumor phenotyping based on computed tomography volumes. *Applied Sciences*, 12(13):6318, 2022.

[31] Geovanni Figueroa-Mata and Erick Mata-Montero. Using a convolutional siamese network for image-based plant species identification with small datasets. *Biomimetics*, 5(1):8, 2020.

[32] Yu Fu, Peng Xue, Huizhong Ji, Wentao Cui, and Enqing Dong. Deep model with siamese network for viable and necrotic tumor regions assessment in osteosarcoma. *Medical Physics*, 47(10):4895–4905, 2020.

[33] Heng Gao, Minghui Wang, Haichun Li, Zhaodi Liu, Wei Liang, and Ao Li. A multi-view feature decomposition deep learning method for lung cancer histology classification. In *Fourteenth International Conference on Graphics and Image Processing (ICGIP 2022)*, volume 12705, pages 369–377. SPIE, 2023.

[34] Yang Gao, Fan Song, Peng Zhang, Jian Liu, Jingjing Cui, Yingying Ma, Guanglei Zhang, and Jianwen Luo. Improving the subtype classification of non-small cell lung cancer by elastic deformation based machine learning. *Journal of Digital Imaging*, 34(3):605–617, 2021.

[35] O Ghoneim, G Soliman, A Galal, and H Mahgoub. Breast cancer histological image classification using ensemble convolutional neural network and triplet loss. *IOSR J. Comput. Eng. Ser II II*, pages 30–42, 2021.

[36] Peter Goldstraw, Kari Chansky, John Crowley, Ramon Rami-Porta, Hisao Asamura, Wilfried EE Eberhardt, Andrew G Nicholson, Patti Groome, Alan Mitchell, Vanessa Bolejack, et al. The iaslc lung cancer staging project: proposals for revision of the tnm stage groupings in the forthcoming (eighth) edition of the tnm classification for lung cancer. *Journal of Thoracic Oncology*, 11(1):39–51, 2016.

[37] Premananth Gowtham, Mahesan Niranjan, and Anantharajah Kaneswaran. Automated gastrointestinal abnormalities detection from endoscopic images. In *2021 IEEE 16th International Conference on Industrial and Information Systems (ICIIS)*, pages 191–196. IEEE, 2021.

[38] Valerio Guarrasi, Fatih Aksu, Camillo Maria Caruso, Francesco Di Feola, Aurora Rofena, Filippo Ruffini, and Paolo Soda. A systematic review of intermediate fusion in multimodal deep learning for biomedical applications. *arXiv preprint arXiv:2408.02686*, 2024.

[39] Valerio Guarrasi and Paolo Soda. Multi-objective optimization determines when, which and how to fuse deep networks: An application to predict COVID-19 outcomes. *Computers in Biology and Medicine*, 154:106625, 2023.

[40] Sidra Gul, Muhammad Salman Khan, Asima Bibi, Amith Khandakar, Mohamed Arselene Ayari, and Muhammad EH Chowdhury. Deep learning techniques for liver and liver tumor segmentation: A review. *Computers in Biology and Medicine*, 147:105620, 2022.

[41] Ibrahim Ethem Hamamci, Sezgin Er, Furkan Almas, Ayse Gulnihan Simsek, Sevval Nil Esirgun, Irem Dogan, Muhammed Furkan Dasdelen, Omer Faruk Durugol, Bastian Wittmann, Tamaz Amiranashvili, Enis Simsar, Mehmet Simsar, Emine Bensu Erdemir, Abdullah Alanbay, Anjany Sekuboyina, Berkan Lafci, Christian Bluethgen, Mehmet Kemal Ozdemir, and Bjoern Menze. Developing Generalist Foundation Models from a Multimodal Dataset for 3D Computed Tomography, 2024.

[42] Ibrahim Ethem Hamamci, Sezgin Er, Anjany Sekuboyina, Enis Simsar, Alperen Tezcan, Ayse Gulnihan Simsek, Sevval Nil Esirgun, Furkan Almas, Irem Doğan, Muhammed Furkan Dasdelen, et al. Generatect: Text-conditional generation of 3d chest ct volumes. In *European Conference on Computer Vision*, pages 126–143. Springer, 2025.

[43] Yong Han, Yuan Ma, Zhiyuan Wu, Feng Zhang, Deqiang Zheng, Xiangtong Liu, Lixin Tao, Zhigang Liang, Zhi Yang, Xia Li, et al. Histologic subtype classification of non-small cell lung cancer using PET/CT images. *European journal of nuclear medicine and molecular imaging*, 48:350–360, 2021.

[44] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[45] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.

[46] Elad Hoffer and Nir Ailon. Deep metric learning using triplet network. In *Similarity-based pattern recognition: third international workshop, SIMBAD 2015, Copenhagen, Denmark, October 12-14, 2015. Proceedings 3*, pages 84–92. Springer, 2015.

[47] Johannes Hofmanninger, Forian Prayer, Jeanny Pan, Sebastian Röhrich, Helmut Prosch, and Georg Langs. Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology problem. *European Radiology Experimental*, 4:1–13, 2020.

[48] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[49] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[50] Bram Hunt, Eugene Kwan, Derek Dosdall, Rob S MacLeod, and Ravi Ranjan. Siamese neural networks for small dataset classification of electrograms. In *2021 Computing in Cardiology (CinC)*, volume 48, pages 1–4. IEEE, 2021.

[51] Guillermo Iglesias, Edgar Talavera, Ángel González-Prieto, Alberto Mozo, and Sandra Gómez-Canaval. Data augmentation techniques in time series domain: a survey and taxonomy. *Neural Computing and Applications*, 35(14):10123–10145, 2023.

[52] Imran Iqbal, Muhammad Younus, Khuram Walayat, Mohib Ullah Kakar, and Jinwen Ma. Automated multi-class classification of skin lesions through deep convolutional neural network with dermoscopic images. *Computerized medical imaging and graphics*, 88:101843, 2021.

[53] Md Mohaimenul Islam, Hsuan-Chia Yang, Tahmina Nasrin Poly, Wen-Shan Jian, and Yu-Chuan Jack Li. Deep learning algorithms for detection of diabetic retinopathy in retinal fundus photographs: A systematic review and meta-analysis. *Computer Methods and Programs in Biomedicine*, 191:105320, 2020.

[54] Chinnu Jacob and Gopakumar Chandrasekhara Menon. Pathological categorization of lung carcinoma from multimodality images using convolutional neural networks. *International Journal of Imaging Systems and Technology*, 32(5):1681–1695, 2022.

[55] Robert Jeraj, Tyler Bradshaw, and Urban Simončič. Molecular imaging to plan radiotherapy and evaluate its efficacy. *Journal of Nuclear Medicine*, 56(11):1752–1765, 2015.

[56] Binghu Jiang, Shodayu Takashima, Chie Miyake, Tomoaki Hakucho, Yoshiyuki Takahashi, Daisuke Morimoto, Hodaka Numasaki, Katsuyuki Nakanishi, Yasuhiko Tomita, and Masahiko Higashiyama. Thin-section ct findings in peripheral lung cancer of 3 cm or smaller: are there any characteristic features for predicting tumor histology or do they depend only on tumor size? *Acta radiologica*, 55(3):302–308, 2014.

[57] José Raniery Ferreira Junior, Marcel Koenigkam-Santos, Federico Enrique Garcia Cipriano, Alexandre Todorovic Fabro, and Paulo Mazzoncini de Azevedo-Marques. Radiomics-based features for pattern recognition of lung cancer histopathology and metastases. *Computer methods and programs in biomedicine*, 159:23–30, 2018.

[58] Zahra Khodabakhshi, Shayan Mostafaei, Hossein Arabi, Mehrdad Oveisi, Isaac Shiri, and Habib Zaidi. Non-small cell lung carcinoma histopathological subtype phenotyping using high-dimensional multinomial multiclass CT radiomics signature. *Computers in biology and medicine*, 136:104752, 2021.

[59] Margarita Kirienko, Luca Cozzi, Lidija Antunovic, Lisa Lozza, Antonella Fogliata, Emanuele Voulaz, Alexia Rossi, Arturo Chiti, and Martina Sollini. Prediction of

disease-free survival by the PET/CT radiomic signature in non-small cell lung cancer patients undergoing surgery. *European journal of nuclear medicine and molecular imaging*, 45:207–217, 2018.

[60] Jakub Konečnỳ, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016.

[61] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[62] Vijay Kumar B G, Gustavo Carneiro, and Ian Reid. Learning local image descriptors with deep siamese and triplet convolutional networks by minimising global loss functions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[63] Trinh Thi Le Vuong, Kyungeun Kim, Boram Song, and Jin Tae Kwak. Ranking loss: a ranking-based deep neural network for colorectal cancer grading in pathology images. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VIII 24*, pages 540–549. Springer, 2021.

[64] Haichun Li, Qilong Song, Dongqi Gui, Minghui Wang, Xuhong Min, and Ao Li. Reconstruction-assisted feature encoding network for histologic subtype classification of non-small cell lung cancer. *IEEE Journal of Biomedical and Health Informatics*, 26(9):4563–4574, 2022.

[65] Ping Li, Shuo Wang, Tang Li, Jingfeng Lu, Yunxin HuangFu, and Dongxue Wang. A Large-Scale CT and PET/CT Dataset for Lung Cancer Diagnosis [Data set], 2020.

[66] Quan Li, Xiguang Wei, Huanbin Lin, Yang Liu, Tianjian Chen, and Xiaojuan Ma. Inspecting the running process of horizontal federated learning via visual analytics. *IEEE Transactions on Visualization and Computer Graphics*, 28(12):4085–4100, 2021.

[67] Yikuan Li, Ramsey M Wehbe, Faraz S Ahmad, Hanyin Wang, and Yuan Luo. Clinical-longformer and clinical-bigbird: Transformers for long clinical sequences. *arXiv preprint arXiv:2201.11838*, 2022.

[68] Han Liu, Zhicheng Jiao, Wenjuan Han, and Bin Jing. Identifying the histologic subtypes of non-small cell lung cancer with computed tomography imaging: a comparative

study of capsule net, convolutional neural network, and radiomics. *Quantitative Imaging in Medicine and Surgery*, 11(6):2756, 2021.

[69] Jian Liu, Jingjing Cui, Fei Liu, Yixuan Yuan, Feng Guo, and Guanglei Zhang. Multi-subtype classification model for non-small cell lung cancer based on radiomics: Sls model. *Medical physics*, 46(7):3091–3100, 2019.

[70] Jie Liu, Yixiao Zhang, Jie-Neng Chen, Junfei Xiao, Yongyi Lu, Bennett A Landman, Yixuan Yuan, Alan Yuille, Yucheng Tang, and Zongwei Zhou. Clip-driven universal model for organ segmentation and tumor detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21152–21164, 2023.

[71] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[72] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[73] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019.

[74] Panagiotis Marentakis, Pantelis Karaiskos, Vassilis Kouloulias, Nikolaos Kelekis, Stylianos Argentos, Nikolaos Oikonomopoulos, and Constantinos Loukas. Lung cancer histology classification from ct images based on radiomics and deep learning models. *Medical & biological engineering & computing*, 59:215–226, 2021.

[75] Michael Owusu-Adjei, James Ben Hayfron-Acquah, Twum Frimpong, and Gaddafi Abdul-Salaam. Imbalanced class distribution and performance evaluation metrics: A systematic review of prediction accuracy for determining model performance in healthcare systems. *PLOS Digital Health*, 2(11):e0000290, 2023.

[76] Yifan Peng, Shazia Dharssi, Qingyu Chen, Tiarnan D Keenan, Elvira Agrón, Wai T Wong, Emily Y Chew, and Zhiyong Lu. DeepSeeNet: a deep learning model for automated classification of patient-based age-related macular degeneration severity from color fundus photographs. *Ophthalmology*, 126(4):565–575, 2019.

[77] PE Postmus, KM Kerr, M Oudkerk, S Senan, DA Waller, J Vansteenkiste, C Escriu, and S Peters. Early and locally advanced non-small-cell lung cancer (NSCLC): ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Annals of oncology*, 28:iv1–iv21, 2017.

[78] Jing Qi, Zhengqiao Deng, Guogui Sun, Shuang Qian, Li Liu, and Bo Xu. One-step algorithm for fast-track localization and multi-category classification of histological subtypes in lung cancer. *European Journal of Radiology*, 154:110443, 2022.

[79] RuoXi Qin, Zhenzhen Wang, LingYun Jiang, Kai Qiao, Jinjin Hai, Jian Chen, Junling Xu, Dapeng Shi, and Bin Yan. Fine-Grained Lung Cancer Classification from PET and CT Images Based on Multidimensional Attention Mechanism. *Complexity*, 2020(1):6153657, 2020.

[80] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[81] Guillaume Reichert, Ali Bellamine, Matthieu Fontaine, Beatrice Naipeanu, Adrien Altar, Elodie Mejean, Nicolas Javaud, and Nathalie Siauve. How can a deep learning algorithm improve fracture detection on x-rays in the emergency room? *Journal of Imaging*, 7(7):105, 2021.

[82] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[83] Soheila Saeedi, Sorayya Rezayi, Hamidreza Keshavarz, and Sharareh R. Niakan Kalhori. MRI-based brain tumor detection using convolutional deep learning methods and chosen machine learning techniques. *BMC Medical Informatics and Decision Making*, 23(1):16, 2023.

[84] Adyasha Sahu, Pradeep Kumar Das, and Sukadev Meher. Recent advancements in machine learning and deep learning-based breast cancer detection using mammograms. *Physica Medica*, 114:103138, 2023.

[85] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

[86] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[87] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[88] Shelly Soffer, Eyal Klang, Orit Shimon, Yiftach Barash, Noa Cahan, Hayit Greenspana, and Eli Konen. Deep learning for pulmonary embolism detection on computed tomography pulmonary angiogram: a systematic review and meta-analysis. *Scientific reports*, 11(1):15814, 2021.

[89] Ghada Sokar, Elsayed E Hemayed, and Mohamed Rehan. A generic OCR using deep siamese convolution neural networks. In *2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, pages 1238–1244. IEEE, 2018.

[90] Justin B Solomon, Olav Christianson, and Ehsan Samei. Quantitative comparison of noise texture across CT scanners from different manufacturers. *Medical Physics*, 39(10):6048–6055, 2012.

[91] Yohan Sumathipala, Majid Shafiq, Erika Bongen, Connor Brinton, and David Paik. Machine learning to predict lung nodule biopsy method using CT image features: A pilot study. *Computerized Medical Imaging and Graphics*, 71:1–8, 2019.

[92] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[93] Satoshi Takeuchi, Benjapa Khiewvan, Patricia S Fox, Stephen G Swisher, Eric M Rohren, Roland L Bassett, and Homer A Macapinlac. Impact of initial PET/CT staging in terms of clinical stage, management plan, and prognosis in 592 patients with non-small-cell lung cancer. *European journal of nuclear medicine and molecular imaging*, 41:906–914, 2014.

[94] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

[95] Yucheng Tang, Dong Yang, Wenqi Li, Holger R Roth, Bennett Landman, Daguang Xu, Vishwesh Nath, and Ali Hatamizadeh. Self-supervised pre-training of swin transformers for 3d medical image analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20730–20740, 2022.

[96] The International Agency for Research on Cancer (IARC). Global cancer observatory, 2024.

[97] Anil Tibdewal, Mangesh Patil, Shagun Misra, Nilendu Purandare, Venkatesh Rangarajan, Naveen Mummudi, George Karimundackal, Sabita Jiwnani, and Jaiprakash Agarwal. Optimal standardized uptake value threshold for auto contouring of gross tumor volume using positron emission tomography/computed tomography in patients with operable nonsmall-cell lung cancer: Comparison with pathological tumor size. *Indian Journal of Nuclear Medicine*, 36(1):7–13, 2021.

[98] Selene Tomassini, Nicola Falcionelli, Giulia Bruschi, Agnese Sbrollini, Niccolo Marini, Paolo Sernani, Micaela Morettini, Henning Müller, Aldo Franco Dragoni, and Laura Burattini. On-cloud decision-support system for non-small cell lung cancer histology characterization from thorax computed tomography scans. *Computerized Medical Imaging and Graphics*, 110:102310, 2023.

[99] Selene Tomassini, Nicola Falcionelli, Paolo Sernani, Agnese Sbrollini, Micaela Morettini, Laura Burattini, and Aldo Franco Dragoni. Cloud-YLung for non-small cell lung cancer histology classification from 3D computed tomography whole-lung scans. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 1556–1560. IEEE, 2022.

[100] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018.

[101] William D Travis, Elisabeth Brambilla, Andrew G Nicholson, Yasushi Yatabe, John HM Austin, Mary Beth Beasley, Lucian R Chirieac, Sanja Dacic, Edwina Duhig, Douglas B Flieder, et al. The 2015 world health organization classification of lung tumors: impact of genetic, clinical and radiologic advances since the 2004 classification. *Journal of thoracic oncology*, 10(9):1243–1260, 2015.

[102] Lev Utkin, Anna Meldo, Maxim Kovalev, and Ernest Kasimov. An ensemble of triplet neural networks for differential diagnostics of lung cancer. In *2019 25th Conference of Open Innovations Association (FRUCT)*, pages 346–352. IEEE, 2019.

[103] Serge Vaudenay. Centralized or decentralized? the contact tracing dilemma. Cryptology ePrint Archive; 2020/531, 2020.

[104] Lise Wei and Issam El Naqa. Artificial intelligence for response evaluation with PET/CT. In *Seminars in nuclear medicine*, volume 51, pages 157–169. Elsevier, 2021.

[105] Zhiwen Xu, Haijun Ren, Wei Zhou, and Zhichao Liu. ISANET: Non-small cell lung cancer classification and detection based on CNN and attention mechanism. *Biomedical Signal Processing and Control*, 77:103773, 2022.

[106] Mengmeng Yan and Weidong Wang. Development of a radiomics prediction model for histological type diagnosis in solitary pulmonary nodules: the combination of CT and FDG PET. *Frontiers in Oncology*, 10:555514, 2020.

[107] Liu Yang, Di Chai, Junxue Zhang, Yilun Jin, Leye Wang, Hao Liu, Han Tian, Qian Xu, and Kai Chen. A survey on vertical federated learning: From a layered perspective. *arXiv preprint arXiv:2304.01829*, 2023.

[108] Mohammad JM Zedan, Mohd Asyraf Zulkifley, Ahmad Asrul Ibrahim, Asraf Mohamed Moubark, Nor Azwan Mohamed Kamari, and Siti Raihanah Abdani. Automated glaucoma screening and diagnosis based on retinal fundus images using deep learning approaches: A comprehensive review. *Diagnostics*, 13(13):2180, 2023.

[109] Penghua Zhai, Yaling Tao, Hao Chen, Ting Cai, and Jinpeng Li. Multi-task learning for lung nodule classification on chest CT. *IEEE access*, 8:180317–180327, 2020.

[110] Kai Zhang, Shouliang Qi, Jiumei Cai, Dan Zhao, Tao Yu, Yong Yue, Yudong Yao, and Wei Qian. Content-based image retrieval with a Convolutional Siamese Neural Network: Distinguishing lung cancer and tuberculosis in CT images. *Computers in biology and medicine*, 140:105096, 2022.

[111] Hongyue Zhao, Yexin Su, Zhehao Lyu, Lin Tian, Peng Xu, Lin Lin, Wei Han, and Peng Fu. Non-invasively discriminating the pathological subtypes of non-small cell lung cancer with pretreatment 18F-FDG PET/CT using deep learning. *Academic Radiology*, 31(1):35–45, 2024.