# RADIOGENOMICS FRAMEWORK FOR ASSOCIATING MEDICAL IMAGE FEATURES WITH TUMOUR GENETIC CHARACTERISTICS

THE UNIVERSITY OF SYDNEY

Faculty of Engineering

The University of Sydney

A thesis submitted to fulfil requirements for the degree of Doctor of Philosophy

Tian Xia

September 2023

# Statement of Originality

This is to certify that to the best of my knowledge, the content of this thesis is my own work. This thesis has not been submitted for any degree or other purposes.

I certify that the intellectual content of this thesis is the product of my own work and that all the assistance received in preparing this thesis and sources have been acknowledged.

Tian Xia

Date: 27/Sep/2023

# Authorship Attribution Statement

Chapter 4 of this thesis was previously published as T.Xia, A. Kumar, M. Fulham, D. Feng, Y. Wang, E. Kim, Y. Jung, and J. Kim, “Fused feature signatures to probe tumour radiogenomics relationships." *Scientific Reports 12, no. 1 (2022): 2173*. I wrote the manuscript, conducted data processing, data analysis and prepared figures and tables.

Chapter 5 of this thesis was previously published as s T. Xia, X. Fu, M. Fulham, Y. Wang, D. Feng, and J. Kim, “CT-based Radiogenomics Framework for COVID-19 Using ACE2 Imaging Representations”, *Journal of Digital Imaging. 2023 Aug 8:1-1.* I wrote the manuscript, conducted data processing, data analysis and prepared figures and tables.

Chapter 6 of this thesis was in preparation to submit as T. Xia, M. Fulham, Y. Wang, D. Feng, and J. Kim, “Deep multimodal Network for Predicting MGMT Promotor Region Methylation Status Using Tumour Heterogeneity Information from MR Imaging”, *IEEE J. Biomed. Health Inform. (JBHI), 2023.* I wrote the manuscript, conducted data processing, data analysis and prepared figures and tables.

In addition to the statements above, in cases where I am not the corresponding author of a published item, permission to include the published material has been granted by the corresponding author.

Tian Xia Signature: Date: 27/09/2023

As supervisor for the candidature upon which this thesis is based, I can confirm that the authorship attribution statements above are correct.

Professor Jinman Kim Signature: Date: 27/09/2023

# Abstract

Significant progress has been made in the understanding of human cancers at the molecular genetics level and it is providing new insights into their underlying pathophysiology. This progress has enabled the subclassification of the disease and the development of targeted therapies that address specific biological pathways. However, obtaining genetic information remains invasive and costly. Medical imaging, on the other hand, is a non-invasive technique that captures important visual characteristics (i.e. image features) of abnormalities and plays an important role in routine clinical practice. Recent advancements in computerised medical image analysis have enabled quantitative approaches to extract image features that can reflect tumour genetic characteristics, leading to the emergence of the growing field of 'radiogenomics'. Radiogenomics investigates the relationships between medical imaging features and tumour molecular characteristics, and it enables the derivation of imaging surrogates (radiogenomics features) to genetic biomarkers that can provide an alternative approach to non-invasive and accurate cancer diagnosis. The identification of image features that are associated with tumour genetic characteristics is crucial for providing accurate radiogenomics analysis.

This thesis presents a new framework that combines several novel methods for radiogenomics analysis that associates medical image features with tumour genetic characteristics, with the main objectives being: i) to provide a comprehensive characterisation of tumour image features that reflect underlying genetic information; ii) introduce a method that identifies and extracts radiogenomics features encoding common pathophysiological information across different diseases, overcoming the dependence on large volumes of

annotated datasets for radiogenomics analysis; and iii) develop a method that quantifies radiogenomics features from multi-modal imaging data and accounts for unique information encoded in tumour heterogeneity sub-regions. The present radiogenomics methods advance radiogenomics analysis and contribute to improving research in computerised medical image analysis.

# Acknowledgement

I would like to convey my sincere gratitude to my supervisor, Professor David Dagan Feng, for his invaluable inspiration, patience, guidance, enthusiasm, and unwavering support. Although words cannot fully capture my appreciation, I believe it is appropriate to cite the following passage:

“为天地立心，为生民立命，为往圣继绝学，为万世开太平。

To ordain conscience for Heaven and Earth. To secure life and fortune for the people. To continue lost teachings for past sages. To establish peace for all future generations.”

- Zhang Zai, Song dynasty, China

I wish to extend my profound gratitude to my supervisor, Professor Jinman Kim, for his exceptional support and mentorship throughout my research journey. His unwavering encouragement, vast knowledge, and provision of the opportunity to conduct my study have been invaluable to me. I shall forever remain grateful for his guidance and support, and I find it fitting to quote the following in recognition of his outstanding contribution:

“If the people prosper, how can the king not prosper with them? And if the people do not prosper, how may the king prosper without them?”

- King Sejong, Joseon Wangjo, Korea

To my wife, Jingshu Zhao:

“春天黎明很美；夏季夜色迷人；秋光最是薄暮；冬景尽在清晨。

春は夜がほのぼのと明けようとする頃；夏は夜；秋は夕暮れ；冬は早朝”

- The Pillow Book (Makura no Sōshi), Sei Shōnagon, Heian period, Japan

I would like to thank all the people whom I had the privilege to work with. Special thanks go to my co-supervisor Dr. Ashnil Kumar and Dr. Michael Fulham. I greatly appreciate their helpful advice and contribution towards my research.

My thanks also go to Dr. Younhyun Jung, Dr. Euijoon Ahn, Dr. Lei Bi, Dr. Michael de Ridder, Dr. Xiaohang Fu, (grand master) Hoijoon Jung, Yifan Liu, Yige Peng, MingJian Li, Shijia Zhou, Zimo Huang, Mingxiao Tu and other former and current colleagues and friends in the Biomedical Data Analysis and Visualisation (BDAV) lab.

Lastly, to myself:

“周虽旧邦，其命维新”

- Classic of Poetry (Shijing), Zhou Dynasty, China

# Contents

# List of Publications

The following publications were produced during the Ph.D. candidature. Many of the listed publications are based on the work presented in this thesis. Three first author journal articles and three conference articles are listed. Publications marked with the star symbol (*) presents major contributions to this thesis.

## Published or Accepted:

1. ***Xia T**, Kumar A, Fulham M, Feng D, Wang Y, Kim EY, Jung Y, Kim J. Fused feature signatures to probe tumour radiogenomics relationships. Scientific Reports. 2022 Feb 9;12(1):2173.
2. ***Xia, T.,** Fu, X., Fulham, M., Wang, Y., Feng, D., and Kim, J., "CT-based Radiogenomics Framework for COVID-19 Using ACE2 Imaging Representations", Journal of Digital Imaging. 2023 Aug 8:1-1.
3. Huang, Z., **Xia, T**., Kim, J., Zhang, L. and Li, B., "Combining CNN With Pathological Information for the Detection of Transmissive Lesions of Jawbones From CBCT Images." In 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) (pp. 2972-2975). IEEE.
4. Wang, C., **Xia, T**., Yang, J. Y. H., and Kim, J., "A Three-Stage Self Supervised Deep Learning Network for Automatic Calcium Scoring of Cardiac Computed Tomography Images," 2022 International Conference on Digital Image Computing: Techniques and Applications (DICTA), Sydney, Australia, 2022, pp. 1-7.

# Under Preparation, Review or Revision:

5. ***Xia, T**., Fulham, M., Wang, Y., Feng, D., and Kim, J., "Deep multimodal Network for Predicting MGMT Promotor Region Methylation Status Using Tumour Heterogeneity Information from MR Imaging", IEEE J. Biomed. Health Inform. (JBHI), 2023. (Under Internal Review)
6. Liu, Y., **Xia, T.,** Feng, D., and Kim, J., "Deep Radiogenomics with Intra-tumoral Heterogeneity Information for Predicting MGMT Promoter Region Methylation in Glioblastoma", IEEE J. Biomed. Health Inform. (JBHI), 2023. (Under Internal Review)

# List of Figures

# List of Tables

# Chapter 1. Introduction

## 1.1 Overview:

In modern healthcare, the ability to identify and extract imaging surrogates that can act as genetic biomarkers from cancer imaging is an important need. The development of optimal imaging surrogates has the potential to advance personalised and targeted medicine, thereby improving cancer diagnosis and therapy. This thesis aims to improve the methods used to identify and extract imaging surrogates for genetic biomarkers in the field of medical imaging analysis. Specifically, this thesis presents the outcomes of investigating the associations between cancer imaging characteristics, also known as image features, and their underlying genetic information using advanced medical image analysis methods.

This thesis describes research conducted to explore two key hypotheses: *what image features are considered important for representing the underlying genetic characteristics of tumours?* and *how can image features from various imaging modalities complement one another for cancer diagnosis?*

## 1.2 Motivation

Precision medicine is a personalised approach to disease prevention, treatment, and care that takes into account an individual's unique variability in genes, environment, and lifestyle [1]. Instead of employing a one-size-fits-all approach, it attempts to personalise medical procedures to the unique characteristics and needs of individual patients. Precision medicine enables the categorisation of individuals into disease-susceptible cohorts and predicts their responses to a particular therapy, therefore improving disease prognosis [2].

Cancers are a collection of disease categories characterised by a highly complex, diverse, and uncontrolled growth and spread of aberrant cells, which are one of the near-term focuses of precision medicine [1]. Cancer is a serious global health concern and is accountable for approximately 10 million deaths annually; this is despite substantial improvements in cancer research and therapy. The current standards of cancer patient management heavily rely on accurate diagnosis which may involve a combination of sophisticated medical procedures such as symptom evaluation during patient screening, non-invasive imaging for locating the disease, and examination of tissue specimens through histopathology.

Recent advances in understanding human cancers at the molecular genetic level have provided insights into their underlying pathophysiology. Such progress has aided in the subclassification of the disease and facilitated the development of therapies that target specific biological pathways. For example, therapies that target epidermal growth factor receptor (EGFR) in non-small cell lung cancer (NSCLC) and $O^6$-methylguanine methyltransferase (MGMT) in Glioblastoma (GBM) patients have demonstrated improved clinical outcomes [3, 4].

Obtaining genetic information is becoming an increasingly valuable complement to the existing diagnostic pipeline. It has the potential to provide optimal cancer treatment strategies at the level of individual patients [1]. The acquisition of genetic information, however, requires

adequate tumour tissue samples obtained from core biopsies that sample only a part of the tumour; further limitations are that biopsies are invasive and expensive [5]. These difficulties are compounded by human cancers exhibiting strong phenotypic and genetic heterogeneity, in which the abnormality develops at distinct sites with genetic variances [6].

In addition to invasive procedures to obtain genetic information, medical imaging is a non-invasive technique that plays a vital role in routine clinical practice by capturing important visual characteristics of abnormalities. For instance, medical images can encode visual characteristics that describe cancer phenotypic traits, such as location, size, texture, and shape. These visual characteristics, commonly referred to as "image features," have been proven invaluable for disease diagnosis and oncologic research [7].

Recent advancements in computerised medical image analysis have resulted in the development of "radiomics," a high-throughput and quantitative approach to extracting image features utilising machine learning-based algorithms and large volumes of medical data. Radiomics features have been demonstrated to offer additional insights into tumour phenotypic traits that complement traditional medical image analysis methods, which rely on explicitly designed image features based on existing medical knowledge, such as tumour location and shape. In addition, radiomics allows automated extraction of image features that can facilitate the characterisation of tumour phenotypic variations and provides insights into the patterns of disease spread and prognosis. Recent studies have demonstrated the capabilities of radiomics features for medical image analysis tasks, including the detection of tumours, characterisation of cancer subtypes, and prediction of treatment response [8, 9].

Therefore, combining medical images and genetic information can offer a more comprehensive understanding of disease biology, improve disease diagnosis, and enhance the development of personalised treatment plans [10]. These findings contributed to the development of a novel research field known as ‘radiogenomics’, whose objective is to

investigate the statistical relationships between the image features of a tumour and its underlying genetic characteristics [11]; these features are hereby denoted as ‘radiogenomics features’. Radiogenomics enables the derivation of imaging surrogates to genetic biomarkers and provides an alternative approach that contributes to a non-invasive and accurate cancer diagnosis [12-14].

Radiogenomics research has demonstrated the feasibility of deriving radiogenomics features by statistically correlating radiomics features with the underlying genetic characteristics of tumours in non-small cell lung cancer (NSCLC) patients [15]. In another study, Gevaert et al. [16] have shown that radiogenomics features can serve as non-invasive imaging surrogates to genetic biomarkers for the NSCLC prognosis. In another study, Kickingereder et al. [17] developed a radiogenomics-based approach and found that radiogenomics features are associated with biological information related to hypoxia / angiogenesis transcriptome signature in glioma patients. Early attempts have also been made to predict clinical outcomes in hepatocellular carcinoma using radiogenomics features [18, 19]. In these studies, image features such as tumour size, shape, and texture information were utilised to derive radiogenomics features.

Despite recent progress in radiogenomics for deriving imaging surrogates to genetic biomarkers, there is still a reliance on tumour image features during statistical analysis. Furthermore, the current state-of-the-art radiogenomics approach usually involves using a single category of image features explicitly designed based on prior human knowledge (also known as ‘handcrafted features’) or those generated through machine learning techniques. As a result, the derivation of radiogenomics features is often restricted and may not fully capture the genetic characteristics that reflect specific cancer types or variants among patient cohorts. An alternative approach to address the challenges of radiogenomics analysis is to utilise both categories of image features during statistical analysis, as handcrafted and machine-learning

derived features have been shown to encode complementary information about the disease [20]. Furthermore, the utilisation of both categories of image features may enable the derivation of unique radiogenomics features that have the potential to demonstrate stronger associations to genetic characteristics and establish additional non-invasive imaging biomarkers.

Another significant challenge in radiogenomics is the limited availability of annotated medical datasets, particularly for rare and emerging diseases [21]. The scarcity of annotated medical datasets restricts radiogenomics analysis to comprehensively extracting and deriving radiogenomics features that incorporate tumour genetic information. To address this challenge, an alternative approach is to use multi-disease analysis to leverage imaging and genetic information from related datasets where the diseases share similar pathophysiology. Multi-disease analysis enables the derivation of radiogenomics features that have the potential to serve as a versatile imaging surrogate to genetic biomarkers for diseases that share similar pathophysiology information, such as signalling pathways.

Furthermore, human cancers exhibit strong heterogeneity in imaging, histological, and genetic characteristics within and among patients. For instance, intra-tumoral heterogeneity in glioblastoma (GBM) can manifest as regions composed of diverse tissue necrosis, cystic degeneration, and haemorrhage at the imaging level [22-24]. The heterogeneous landscape of human cancers presents challenges for medical image analysis. Existing methods focus on processing and analysing imaging characteristics within entire tumour regions while neglecting the complementary information encoded in distinct image modalities. Radiogenomics has the potential to address the challenge by deriving radiogenomics features that leverages information from various tumour heterogeneous regions and integrates complementary information from distinct imaging modalities. Radiogenomics, therefore, demonstrates the potential to provide unique insights into tumour pathophysiology and facilitate medical image analysis tasks.

## 1.3 Aims and Objectives

In this thesis, I propose a radiogenomics framework that aims to address the following challenges in the field of medical image analysis:

1. The comprehensive characterisation of tumour imaging representations to reflect its underlying genetic information.
2. The limited availability of annotated medical datasets, particularly for rare and emerging diseases for radiogenomics analysis.
3. The comprehensive characterisation of tumour heterogeneity information for radiogenomics analysis.

The proposed radiogenomics framework incorporates the following methods that address these challenges:

1. I propose the fused feature signature ($FF_{Sig}$), a selection of image features that encodes complementary cancer imaging visual characteristics using a combination of handcrafted and deep learning-based image feature extraction techniques. The $FF_{Sig}$ encodes complementary imaging characteristics of tumours and identifies more radiogenomics relationships with a broader range of genes related to important biological functions. The proposed $FF_{Sig}$ is robust and generalisable across different datasets and allows the identification and extraction of important radiogenomics features that may facilitate cancer diagnosis and treatment planning.
2. I propose a novel framework to derive radiogenomics features linked to the genetic characteristics of one disease, which can be applied to another disease through multi-disease analysis. Furthermore, the proposed framework identifies and extracts radiogenomics features that encode common pathophysiological information across

different diseases, thus overcoming the dependence on large volumes of annotated datasets for radiogenomics analysis.

3. I propose a novel radiogenomics method that quantifies radiogenomics features from multi-modal imaging data and accounts for unique information encoded in tumour heterogeneity sub-regions. Furthermore, the framework leverages the complementary knowledge from the extracted radiogenomics features and identifies the relevant features for various medical image analysis tasks.

## 1.4 Thesis Structure

The remainder of this thesis is structured as follows:

Chapters 2 and 3 orient the readers by providing theoretical background knowledge of the thesis. Chapter 2 provides an overview of medical imaging and image processing techniques. Chapter 3 presents an overview of the state-of-the-art methods in radiogenomics analysis. In particular, it provides a summary of gaps from recent radiogenomics research.

Chapters 4, 5 and 6 provide detailed contributions to this thesis. We describe the Fused Image Feature Signature in chapter 4. In chapter 5, we describe the radiogenomics with multi-disease analysis. Chapter 6 presents the novel radiogenomics method for multi-modality images with tumour heterogeneity and its application.

Finally, chapter 7 summarises the contribution of this thesis and indicates directions for future research in the domain of radiogenomics analysis.

# Chapter 2. Medical Imaging and Image Processing

## 2.1 Overview

Medical imaging is integral tool to our contemporary healthcare systems for accurate cancer diagnosis and staging [25]. It provides a rapid and non-invasive assessment of morphological features across various cancers [7] through diverse imaging modalities. In addition, medical image processing techniques enable the extraction of valuable information that can provide insights into distinct aspects of patients' diseases. This chapter introduces medical imaging and the underlying theoretical principles of image processing techniques.

## 2.2 Definitions

Digital images are composed of numerical picture elements known as *pixels*. Pixel resolution refers to the number of pixels present in an image and can be denoted as either a single number or the number of pixels in each dimension.

Several medical imaging techniques, such as computed tomography (CT) and magnetic resonance imaging (MR), involve sampling 2D images along a third spatial axis to construct a 3D image. These medical imaging techniques are commonly referred to as volumetric images or image volumes. Three-dimensional volumetric images can capture spatial relationships between three-dimensional pixels, also known as voxels. Spatial resolutions of pixels and voxels describe the level of information captured by their individual elements. For instance, a

spatial resolution of 10.00mm × 10.00mm × 10.00mm means that a voxel depicts a region with a volume of 1000.00mm$^3$.

*Contrast resolution* refers to the intensity range or set of intensities for the red, green, and blue (RGB) channels that can be identified in grayscale and colour pictures. A relatively low contrast resolution can be interpreted as pixels/voxels with comparable and difficult-to-distinguish intensities in an image. Figure 2.1 depicts an illustration of a 2D picture and a 3D volumetric image. Pixels and voxels are represented as 2D and 3D grid arrays, respectively.

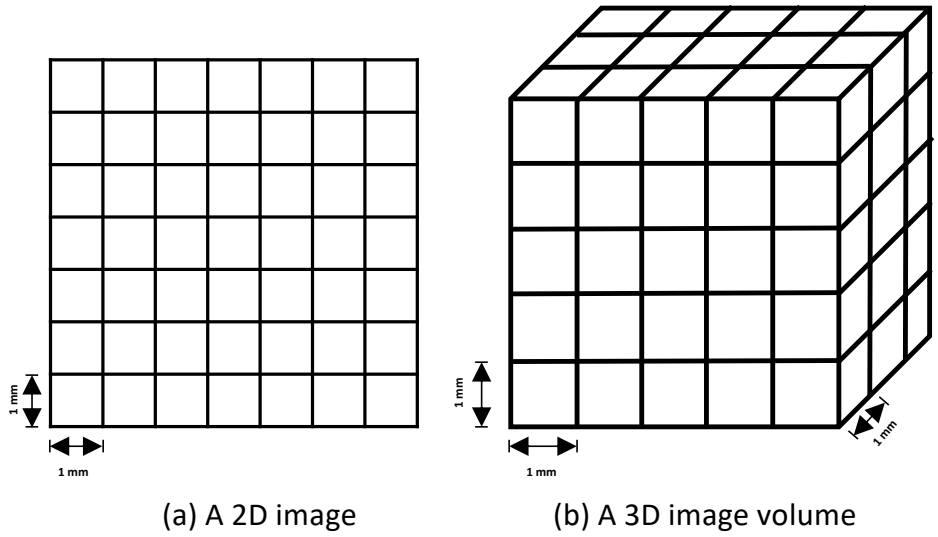


Figure 2.1. An illustration of pixels, voxels, and spatial resolutions in medical images: (a) a 2D image with pixel resolution of 7 × 7 and a spatial resolution of 1mm × 1mm. (b) a 3D image volume with voxel resolution of 5 × 5 × 3 and a spatial resolution is 1mm × 1mm × 1mm.

In medical image processing, a region of interest (ROI) frequently refers to a collection of pixels in 2D images that denotes an area that encodes important information to a particular domain or application in disease detection. The corresponding term for 3D volumetric pictures is the volume of interest (VOI), which refers to the emphasised regions that convey clinically useful information.

## 2.3 Medical Imaging Modalities

Modern healthcare utilises various medical imaging modalities for cancer diagnosis and treatment, with the goal of capturing various visual characteristics of the abnormality [26, 27]. Medical image modality refers to specific imaging techniques, such as X-ray and CT, used to generate images. Each imaging modality has unique physical principles, imaging parameters, and technical characteristics, which produce distinct images for various clinical applications. Medical image modalities can be categorised into two classes based on their physical principles and their ability to visualise information related to abnormalities [28]: (i) Anatomical and (ii) Functional (frequently referred to as 'physiological') imaging. Anatomical imaging modalities encode and visualise information about the structure and morphology of tissues and organs in 2D or 3D images. Anatomical medical images enable physicians to interpret and evaluate disease conditions for diagnostic purposes and for monitoring treatment responses [9]. In contrast, functional imaging allows the assessment of the metabolic and physiological status of the ROI, thereby facilitating the identification of abnormalities, such as tumours [29].

Common anatomical medical image modalities include X-ray, computed tomography (CT), magnetic resonance imaging (MRI), ultrasound and optical imaging. Common functional imaging includes single-photon emission computed tomography (SPECT), positron emission tomography (PET), functional MRI (fMRI), and diffusion tensor imaging (DTI). These imaging techniques generate a single type of image or image volumes and are referred to as single-modality medical imaging.

In addition to using single-modality medical images, multimodal imaging techniques have become a standard diagnostic procedure for cancer diagnosis in clinical practice and medical research. For example, PET-CT scanners, which stack a PET scanner on top of a CT scanner, provide complementary anatomical and functional information (e.g. metabolic activity) of the abnormality. Similarly, multimodal MRI utilises various types of MRI

sequences to provide a comprehensive assessment of the abnormality being studied, thereby enabling more accurate diagnosis and treatment planning [30].

This subsection provides an overview of the single- and multi-modality imaging techniques relevant to deriving radiogenomics features and their application in radiogenomics analysis.

### 2.3.1 Computed Tomography (CT)

Computed tomography (CT) is among the most commonly used medical imaging modalities for cancer detection, diagnosis, and prognosis [31, 32]. CT imaging employs a sequence of X-ray images with varying rotational angles and produces cross-sectional images (also referred to as image slices). These cross-sectional images were then combined to form 3D image volumes. CT imaging provides detailed visualisation of internal organs such as bones and soft tissues. Figure 2.2 shows an axial view CT image of the lung field obtained from a patient diagnosed with Non-small Cell Lung Cancer (NSCLC).

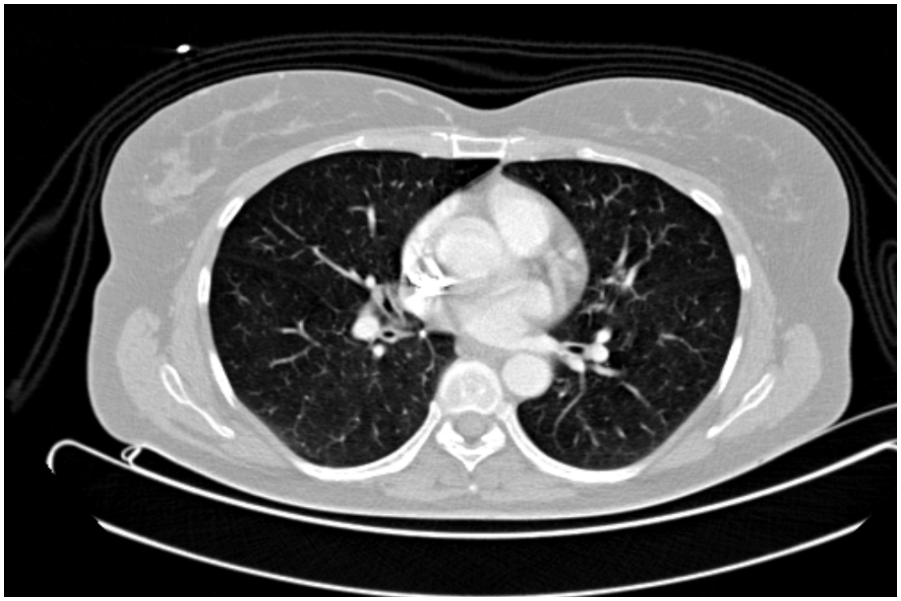

Figure 2.2. A CT scan depicting the lung field of a patient diagnosed with Non-small Cell Lung Cancer (NSCLC). (Data description detailed in Chapter 5).

## 2.3.2 Magnetic Resonance (MR) Imaging

Magnetic resonance (MR) imaging is a non-invasive imaging technique that utilises magnetic fields and radiofrequency pulses to generate images of the internal organ. MR images are produced by detecting the signals emitted by protons in the body following their excitation by a magnetic field and radiofrequency pulses. Different tissue types can be emphasised by alternating the acquisition parameters, such as sequences of radiofrequency pulses, to generate MR images with distinctive appearances.

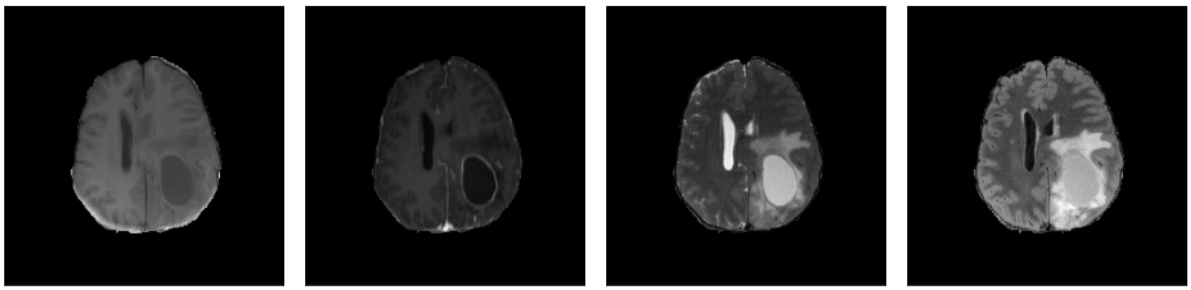


Figure 2.3. Illustration of four common MRI sequences acquired from the same subject diagnosed with glioblastoma (GBM).

Common MRI sequences include: i) T1-weighted (T1w), which produces images with good contrast between fat, muscle, and other soft tissues; ii) Gadolinium-enhanced T1-weighted (T1-Gd, also known as T1CE), which highlights the contrast between enhancing tissues, such as vascular structures; iii) T2-weighted (T2w), which produces images with good contrast between fluid-filled spaces and other soft tissues, and iv) T2-fluid-attenuated inversion recovery (FLAIR), which suppresses fluid signals to enhance the visualisation of pathological tissue. Figure 2.3 shows MR images generated with various sequences. Additionally, multiple MR sequences may be combined to obtain complementary information about the tissues being imaged; such combined MR imaging sequences are referred to as multimodal MR [33]. The

utilisation of multimodal MR images enables obtaining different types of contrast, which can aid in the better characterisation and diagnosis of various pathological conditions.

## 2.4 Machine Learning, Deep Learning and Convolutional Neural Networks

Machine learning is a well-established research field and has been widely utilised for medical image analysis. The capacity of machine learning to quantitatively extract image features from input medical images enables the development of computerised medical image analysis algorithms for various clinical applications. However, the effectiveness of machine learning-based medical image analysis algorithms is constrained by the necessity of domain expertise and meticulous engineering to facilitate the learning and transformation of the input data [34]. Deep learning is a machine learning method that enables the learning of data representations using multiple abstraction levels. Among deep learning techniques, convolutional neural networks (CNNs) are particularly adept at extracting image features from medical images to reveal intricate underlying representations using deep networks [35]. This chapter focuses on deep learning and convolutional neural networks (CNNs) as they constitute the current state-of-the-art medical image analysis. Relevant background knowledge and CNN architectures will be presented in the following sections.

### 2.4.1 Machine Learning

Machine learning is a major area of computer science that has found extensive application in modern healthcare systems, particularly for medical image processing and analysis by incorporating meticulously designed mathematical models, machine learning systems have been employed to perform disease recognition [36], medical image classification [20], medical image retrieval [37] and tumour segmentation [38]. In contrast to rule-based algorithms, which

rely on a predefined set of rules and conditions to make decisions or predictions based on the input data, machine learning-based techniques are based on creating models that can learn and make decisions based on patterns and relationships discovered within the input data. Therefore, it is crucial for a machine learning model to recognise and differentiate patterns and characteristics within the input data. This involves developing a model that can identify and extract the relevant information from the data and then employ it to produce accurate predictions or classifications [39].

The training process for a machine learning model typically involves employing algorithms that enable the model to adjust its parameters based on the feedback it receives from the training data. Throughout the iterative training process, the model learns to recognise and classify patterns and provide improved accuracy. The learning processes to train a machine learning model are typically classified into three categories: i) supervised learning, which learns features that are relevant to the labelled data; ii) unsupervised learning, which discovers patterns and relationships in unlabelled data; and iii) reinforcement learning, which involves learning through trial and error to maximise a reward function designed for specific tasks. While unsupervised and reinforcement learning has demonstrated strong potential across various medical image processing and analysis tasks, our thesis centres on employing labelled medical imaging data through supervised training methods.

### 2.4.2 Supervised Learning:

Supervised learning is one of the most utilised medical image processing and analysis approaches. It involves training a model with annotated datasets, enabling the model to learn and comprehend the intrinsic features of input data and generate predictions on corresponding labels [40]. Supervised learning models have been demonstrated in clinical practice to aid

radiologists and clinicians in disease diagnosis, prognosis and predicting treatment responses [41]. In supervised learning, a dataset is usually split into the following parts:

- Training set: the largest subset of the entire dataset utilised to train a machine learning model.
- Validation set: a smaller, distinct subset of the entire dataset used to assess the model's performance during the training process and prevent overfitting. During validation, the predicted output is compared with the actual data labels from the validation set, which provides an assessment of the model's performance on unseen data and enables the tuning of model parameters.
- Test data: A small subset of the dataset withheld and employed to evaluate the model's performance once training is complete. This subset is not included during training and is used to assess the model's ability to generalise to unseen data. In addition, the test data provides an assessment of the trained model's performance on new examples, which is important for ensuring the model's performance in practical applications.

### 2.4.3 Artificial Neural Network

Artificial Neural Networks (ANNs) are a machine learning technique that draws inspiration from the biological neural networks in the human brain [42]. ANNs consist of networks of interconnected nodes or neurons, which process and transmit information through a series of weighted connections. These connections allow the network to learn complicated, non-linear relationships from the input data, enabling ANNs to recognise patterns and make predictions. Furthermore, the structure and training approach of ANNs can be adjusted based on the complexity of the problem being solved, making them versatile and powerful tools for various medical image processing and analysis applications.

Neurons serve as the fundamental processing units in ANNs. Figure 2.4 depicts the structure of a single neuron that receives multiple input values and produces a single output value. The inputs to the neuron are represented by a vector $X$ with $n$ elements, and the weights associated with each input are represented by a corresponding vector $W$. The neuron's output is determined by a mathematical function that combines the inputs and weights, as shown in equation 2.3.1. The weights are adjusted during the training process to optimise the network's performance on the given task.

$$\mathrm{L}(W, X) = \sum_{X=0}^{n} f(x_i, w_i) + \text{bias} \tag{2.3.1}$$

where $f(x_i, w_i)$ is the ANN function that generates the output value.

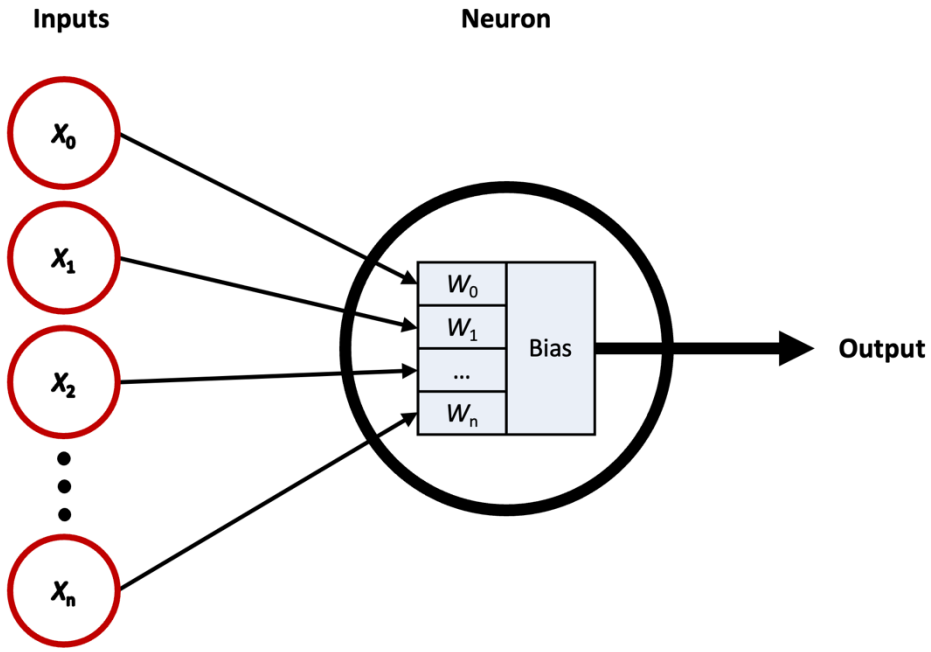


Figure 2.4. An illustration of the structure of a single neuron in ANNs.

The training process of ANNs can be illustrated using a single-neuron model. During the training process, the features of each element in the training dataset are extracted and fed into the neuron as a feature vector. In a classification task involving two classes, neurons in ANNs process the input feature vector by multiplying it with its internal weights. The resulting product is compared to a threshold value to make a class prediction. This process is repeated

for each element in the training dataset, and the weights are adjusted iteratively to improve the model's performance.

In addition to using a single neuron model to train ANNs, the training dataset is typically divided into batches to allow for weight adjustment after each batch. The process of training the model using all the batches in the training dataset is known as an epoch. The weights associated with each neuron are also adjusted after each epoch, which allows the model to learn from the entire training dataset iteratively. The number of epochs required to train a model depends on the complexity of the problem being solved and the size of the training dataset. The model's weights are typically adjusted over hundreds of training epochs to achieve robust performance for classification tasks.

### 2.4.4 Deep Learning

Deep learning is a class of machine learning techniques [43] that involves using computation models with multiple processing layers to learn the internal representation of input data at numerous levels of abstraction [44]. Deep learning models employ a hierarchical learning approach, where the output of one layer is fed as input to the next layer, thereby allowing the model to learn increasingly complex features and patterns in the data [39].

Deep learning techniques utilise the backpropagation of error technique, or "backpropagation," to train multilayer models for supervised learning. During backpropagation, the gradient of the cost function (also known as loss function or error function) of the neural network is calculated with respect to its weights, and the gradient is propagated backwards through the network. This allows the network to adjust its weights to minimise the difference between the predicted and actual output. In contrast to conventional methods that calculate the gradient of the cost function for each layer separately, backpropagation enables more efficient computation of gradients in neural networks.

The implementation of deep learning techniques utilises specialised graphics processing units (GPUs) to accelerate the training process by 10 to 20 times compared to traditional training approaches on standard central processing units (CPUs) [39]. This is because deep learning models involve large volumes of computations, and GPUs are better suited for performing calculations in parallel. Recent advances in deep learning have led to a state-of-the-art performance in various domains, including the visual object recognition [39], speech recognition [45], and medical image processing and analysis [46].

### 2.4.5 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are a specific type of ANNs that are designed to process input data in the form of multidimensional arrays, such as coloured 2D images composed of 2D arrays for each RGB (colour) channel. In contrast to traditional ANNs with fully connected layers, where each input neuron is connected to every output neuron and leads to overfitting and slow learning, CNNs utilise localised connections and shared weight parameters. This approach diminishes the number of parameters necessitated for learning, consequently enhancing the model's performance when processing unobserved data.

CNNs comprise a series of stages with specialised layers and unique functions. The building blocks of CNNs typically consist of three specialised layers: convolutional, pooling, and activation layers, such as the rectified linear unit (ReLU) layer [43]. The convolutional layers employ filters to extract features from the input data, whilst the pooling layers reduce the size of the feature maps and improve the computational efficiency of the network. The activation layers introduce nonlinearity into the network and enable it to learn complex patterns and relationships in the data.

Convolutional layers in CNNs are composed of organised units in the form of feature maps. Each unit is connected to local patches from the previous layer through a set of weights

known as filter banks, as illustrated in Figure 2.5. The output of the convolutional layer is typically obtained by summing the products of the input values and the corresponding weights in the filter bank, which is then transmitted through a non-linear activation layer such as the ReLU layer e.g. $f(x) = max(0, x)$.

A filter bank is shared by all units contained within a single feature map. This architecture enables the convolutional layers to detect local conjunctions of features from the preceding layer. Local values are frequently highly correlated and invariant to the location within the input data. Pooling layers in CNNs are specifically designed to merge features in spatial proximity that share semantic similarities into one. The underlying principle behind pooling layers is to detect the position of motifs typically formed by highly correlated features through a coarse-grained approach. For instance, a typical pooling layer calculates the maximum value of a local patch in one or more feature maps. By merging features that share semantic similarities, pooling layers reduce the dimensions of the representations and create invariance to minor distortions and shifts in the input data. By combining convolutional and pooling layers, CNNs can learn spatial features and develop robust representations of variations in the input data. In a standard CNN architecture, layers are typically stacked in convolution, ReLU activation, and pooling order, followed by fully connected layers, as depicted in Figure 2.6.

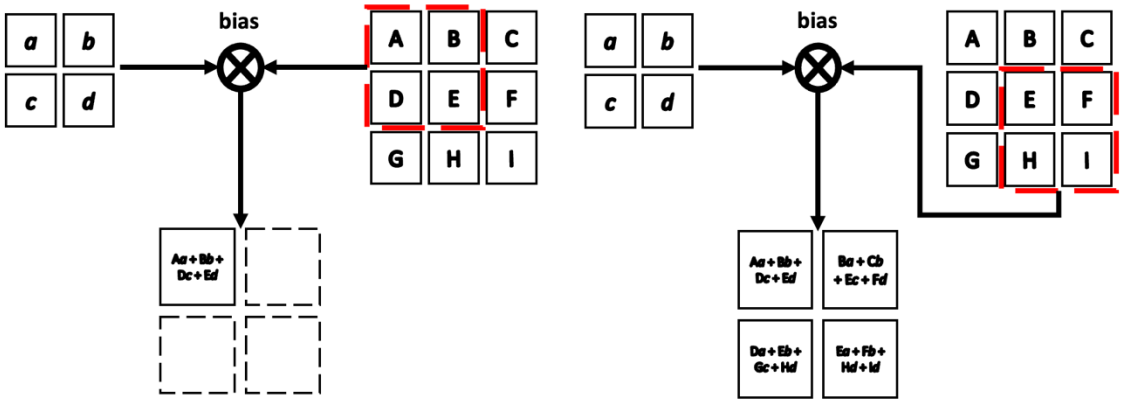


Figure 2.5. An illustration a 2D convolution operation, which involves a 3 × 3 input image convolved with a 2 × 2 filter, resulting in a 2 × 2 output feature map.

In the subsequent subsection, we will present an overview of ResNet. This widely utilised CNN architecture has demonstrated state-of-the-art performance in imaging object recognition and is particularly relevant to our research contributions.

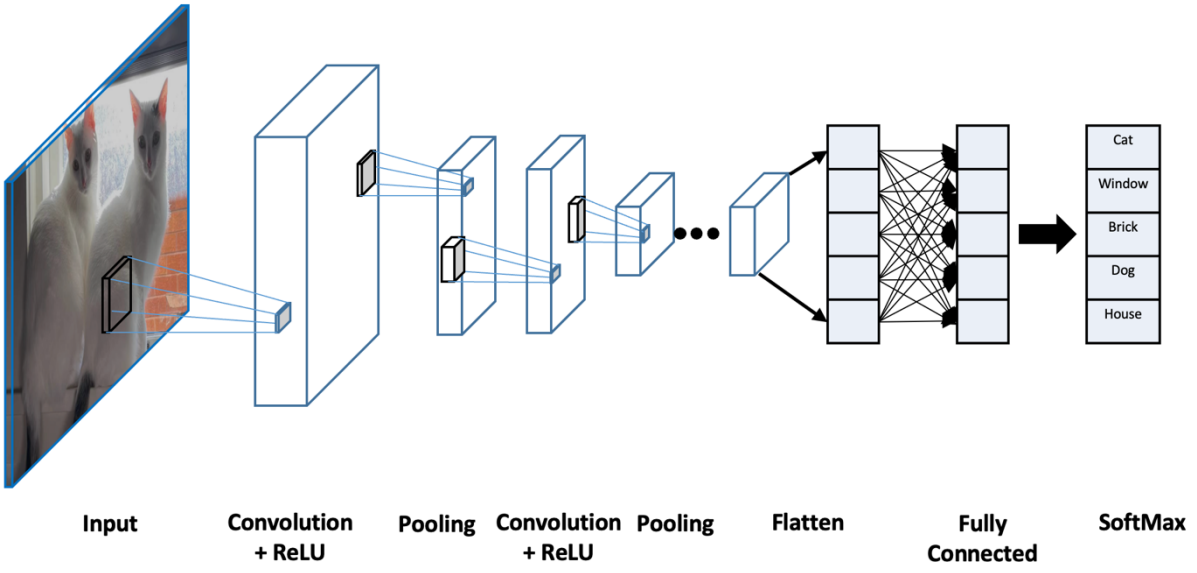


Figure 2.6. CNNs with multiple convolutional layers. At each layer, filters are applied to the input image, and the resulting feature maps are used as inputs to the subsequent convolutional layer.

## 2.4.6 Deep Residual Networks (ResNet)

It has been shown in multiple studies, such as EfficientNet, that the depth of CNNs is a crucial factor in enhancing the learning outcome for image object classification tasks with the ImageNet dataset [47-50]. However, in CNN architectures employing conventional layer stacking, the learning accuracy may rapidly decrease as additional layers are added. This performance decline was not attributable to overfitting but rather to an optimisation issue involving the mapping of identities for the new layers.

ResNet was introduced to tackle the problem of learning accuracy degradation in deep CNN architectures by utilising residual mapping to fit the successive stacked layers [51]. ResNet utilises residual representations and incorporates shortcut connections with gating functions for the identity mapping [52-56]. The residual function and identity mapping through shortcuts can be expressed as follows:

$$f(x) = \mathcal{H}(x) - x \tag{2.3.2}$$

In the equation above, $f(x)$ represents the residual function, obtained by subtracting the input $x$ of the first layer of stacked layers from the underlying mapping $\mathcal{H}(x)$. Residual learning propels the weights of multiple non-linear layers towards zero to achieve identity mapping, which can be expressed as:

$$y = f(x, W_i) + x \tag{2.3.3}$$

The variables $y$ and $x$ refer to the output and input of the stacked layers, respectively. In contrast, $f(x, W_i)$ represents the function that denotes the residual mapping that needs to be learned. The inclusion of shortcut connections in ResNet does not augment the number of network parameters or computational complexity. Figure 2.7 illustrates a ResNet building block that incorporates residual learning and shortcut connections.

ResNet has exhibited exceptional performance in image object recognition, achieving a top-5 error rate of 4.49% with a 152-layer ResNet architecture. Furthermore, it is both robust

and generalisable for various recognition tasks and, as a result, has been extensively utilised in the medical image analysis [51, 57, 58].

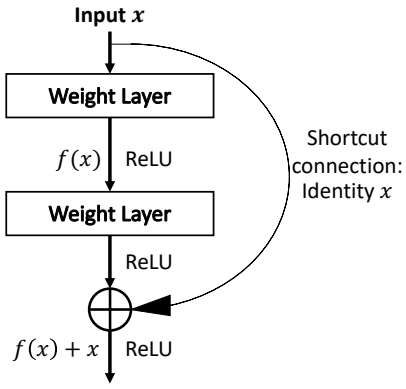


Figure 2.7. A building block of ResNet.

## 2.4.7 Domain Adaptation for Transfer Learning

Typical applications of machine learning techniques involve training models to learn the internal representations of data from a specific domain. This approach is based on the assumption that the training dataset and real-world data, which will be used in future applications, exhibit the same feature space with a similar distribution [59].

When the distribution of feature spaces in a dataset changes, the machine learning model typically requires retraining to adapt to the newly collected data. However, this assumption may not hold in all cases, as most machine learning methods rely on large, annotated training datasets specific to the target domain. Consequently, machine learning applications are considerably constrained in domains such as medical image analysis, where acquiring labelled training data is costly and resource intensive.

Transfer learning is an alternative approach to training a machine learning model that eliminates the requirement for large quantities of domain-specific datasets. This technique involves transferring the knowledge obtained from one large database to facilitate learning applications on a smaller but related dataset. In the context of neural network applications,

transfer learning refers to an approach that accelerates learning a particular domain by employing and transferring the weights derived from a network that was trained for a related source task [60]. Figure 2.8 depicts an illustration of the transfer learning approach.

Transfer learning allows CNNs to be trained and applied to datasets from distinct but related domains, thereby reducing the need for a large amount of domain-specific training data. The differences between such domains can be classified into two categories: (i) those with different feature spaces or (ii) those with the same feature space but different marginal probability distributions [59]. Furthermore, the representation and similarity between two domains can be evaluated using techniques such as A-distance, which are commonly used in transfer learning applications [61, 62].

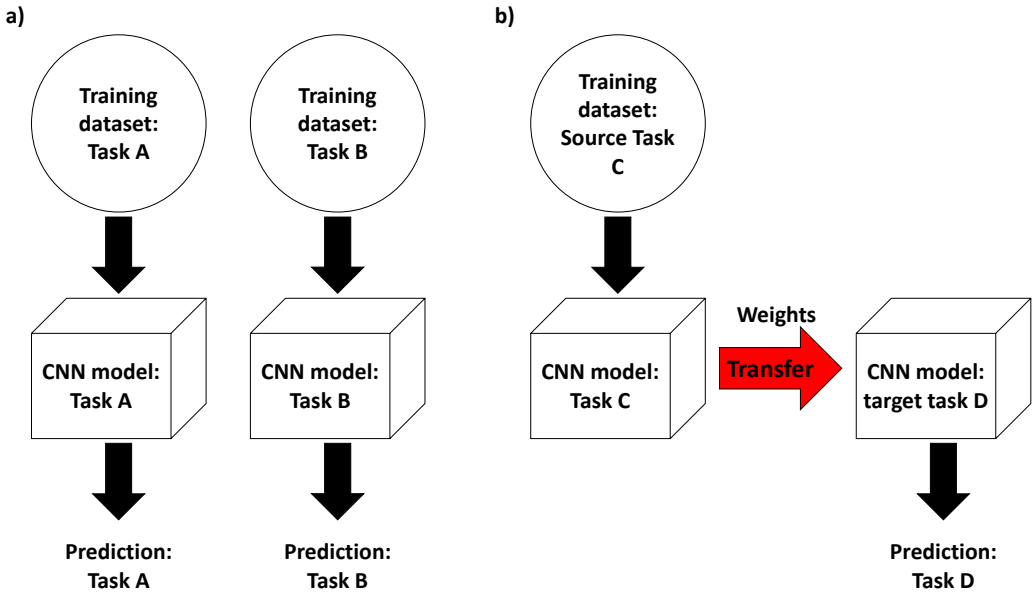


Figure 2.8. Traditional and transfer learning-based approaches for CNN models. a) Traditional machine learning. b) Transfer learning-based approach for CNN model.

For CNNs applications in medical image analysis, there are two primary approaches to employ domain-adapted transfer learning: (i) using "off-the-shelf CNN" features; and (ii) using domain adaptation with fine-tuning techniques [63]. The term "off-the-shelf CNN" refers to

the procedure of utilising a CNN model, which was trained on a larger dataset from scratch as a feature extractor directly on a new dataset with a smaller volume. The parameters in the convolutional and fully connected layers are not changed. For classification tasks, the extracted features from the "off-the-shelf CNN" can be used to train a separate classifier, such as support vector machines or random forest classifier or to train only the classification layer of the model to fit the number of classes of the new dataset [63-66].

The process of updating pre-trained CNN models with new datasets from the desired domain using backpropagation is referred to as fine-tuning. Commonly, fine-tuning involves training only the last few convolutional layers, which is known as "shallow tuning" or all convolutional layers, commonly referred to as "deep tuning" [67]. Deep tuning is typically employed when the distances between domains are considered significant. Studies have demonstrated that fine-tuning may be as effective as training a CNN from scratch and is more robust to the size of the training dataset [72]. Therefore, fine-tuning has been applied to various medical image analysis tasks, such as the medical image modality classification [57]. An example of shallow tuning is illustrated in Figure 2.9.

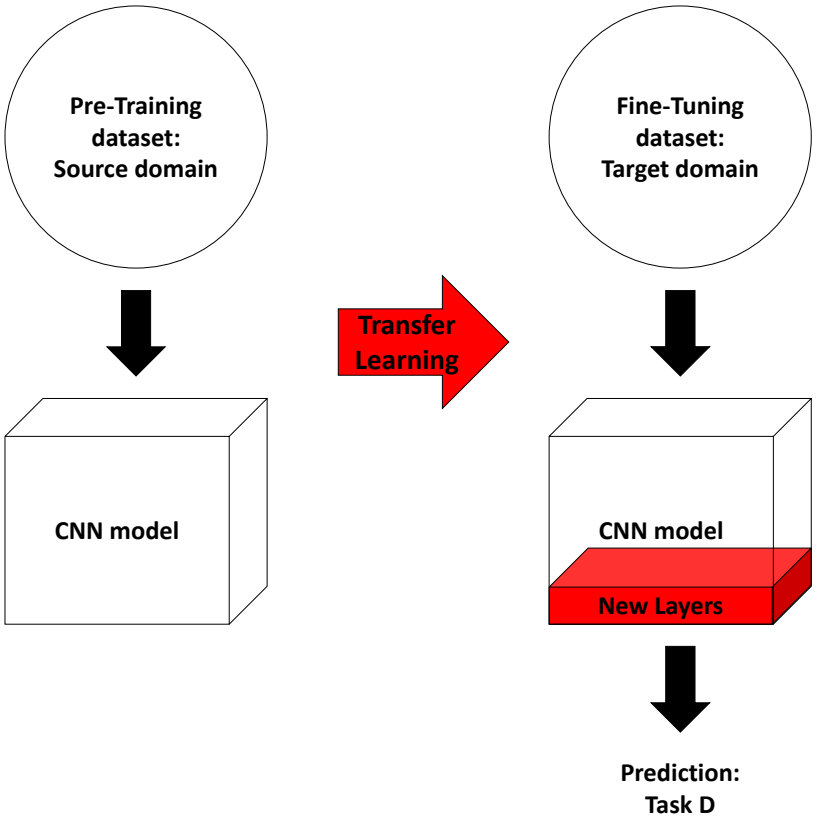


Figure 2.9. An overview of shallow tuning-based transfer learning approach.

## 2.5 Medical Image Processing

Medical image processing is a rapidly evolving research field that seeks to address the challenges emerging from clinical practices. By leveraging advanced computational techniques and artificial intelligence, this domain aims to improve the accuracy and efficiency of disease diagnosis, treatment planning, and patient monitoring. It encompasses various tasks such as image acquisition, pre-processing, segmentation, registration, and classification, all of which contribute to better understanding and interpretation of medical images. This chapter focuses on medical image segmentation and classification, as they constitute the core contributions of the thesis.

### 2.5.1 Medical Image Segmentation

This subsection provides an overview of various medical image segmentation methods relevant to our research.

Medical image segmentation involves partitioning an image into sections, each comprising a group of pixels collectively representing a region of interest (ROI) [68]. This technique is frequently employed to identify and emphasise ROIs, eliminating irrelevant image regions and reducing the complexity of the medical image analysis [69].

Manual delineation of medical images by experienced physicians is widely regarded as the gold standard in most clinical applications. However, this process relies on the visual inspection of imaging data, which is time-consuming and subject to errors based on the expertise of the pathologist [70].

To address this issue, computerised medical image analysis algorithms have been developed to facilitate the segmentation of medical images. Based on their input and techniques employed, existing computerised medical image segmentation algorithms can be classified into three categories: (i) semi-automated segmentation methods that require human interaction

during the segmentation process, (ii) supervised segmentation methods that train the algorithm using labels generated based manual delineations, and (iii) unsupervised segmentation methods that enable the algorithm to learn sophisticated patterns within the image to generate ROI segmentations without the need for human interaction or manual delineations.

Conventional medical image segmentation algorithms are generally centred on pixel analysis and commonly utilise a semi-automated approach. For example, region growing is a widely used method that necessitates predefined seed points and examines the adjacent pixels to recursively expand the region, encompassing pixels that share common characteristics [71]. Pixel thresholding is a commonly used segmentation technique that segregates pixels into clusters with similar characteristics based on a threshold value. This value can be either manually defined or automatically calculated using statistical properties of the image, such as the pixel intensities [72].

Recent developments in machine learning algorithms have enabled data-driven techniques for medical image analysis. Deep learning is one such method that employs Convolutional Neural Networks (CNNs) to learn sophisticated and abstract imaging features directly from large volumes of labelled training image data [43]. Deep learning-based medical image segmentation methods have demonstrated superior performance over traditional segmentation algorithms for various diseases.

For instance, Bi et al. proposed a semi-automatic technique for skin lesion segmentation, which utilises fully connected layers (FCN) to combine user inputs with high-level semantic information obtained by the deep learning model. Furthermore, In a recently published work, Primakov et al. introduced a fully automated deep learning-based method for detecting and segmenting non-small cell lung cancer (NSCLC), integrating volumetric CT scans from multiple institutions. However, although deep learning-based medical image segmentation algorithms have demonstrated greater performance than prior automatic and

semi-automatic approaches, few studies have shown improved outcomes than manual cancer delineation [73, 74].

### 2.5.1 Medical Image Classification

Medical image classification involves assigning a label or a category to a medical image based on the characteristics of the input image. Medical image classification plays an essential role in medical diagnosis, treatment planning and disease prognosis. However, the conventional methods for medical image classification require visual interpretation by radiologists, which can be time-consuming and subject to inter-observer variability. These challenges may be exacerbated by the need to identify quantitative morphological features of anomalies, such as tumour shape and cell counts [46].

Computerised medical image classification typically involves the following processes: i) image feature extraction, ii) image feature selection, and iii) image classification [75]. The image feature extraction step involves quantifying imaging characteristics that are informative and relevant to the labels to be assigned. These characteristics can be quantified through handcrafted features or by adopting deep learning-based approaches that automatically learn the features from the training data. Image feature selection refers to the process of selecting a subset of relevant features from previously extracted image features to reduce the dimensionality of the image data, which can improve the efficiency and accuracy of the machine learning algorithm. Conventional image feature selection methods comprise i) the filter methods that rank the features based on their relevance to the task and select the top-ranked features; ii) the wrapper methods that use the machine learning algorithm to quantify the relevance of a subset of features; and iii) embedded methods that incorporate feature selection into the machine learning algorithm's training process. Finally, in the classification

step, the selected image features are fed into a machine learning model, which assigns a label or a category to the image.

In contrast to conventional medical image classification techniques, deep learning-based approaches enable end-to-end automatic learning of image features relevant to the corresponding labels. Furthermore, unlike conventional methods, these approaches do not require explicit feature engineering [76]. As a result, deep learning-based medical image classification techniques have achieved state-of-the-art performance for a variety of tasks, including lesion classification [20], tumour grading [77], and disease subtype classification [78].

# Chapter 3. Radiogenomics

## 3.1 Overview

In this chapter, I present a literature review of the current research in radiogenomics across various clinical applications. The review focuses on the state-of-the-art techniques used in radiogenomics to extract image features that exhibit statistical associations with genetic information. This also includes radiogenomics applied across different imaging modalities. Furthermore, I summarise identified gaps in the existing literature to provide insights for future research.

## 3.2 Overview of the Radiogenomics Pipeline

As introduced in previous chapters, advances in understanding human cancers' genetic characteristics have provided insights into their underlying pathophysiology, aided in the subclassification of the disease, and facilitated the development of therapies that target specific biological pathways.

Gene expression profiling, on the other hand, necessitates the acquisition of sufficient tumour tissue samples via core biopsies, which only capture a fraction of the tumour and can be both invasive and expensive. Consequently, the comprehensive evaluation of all tumours via gene expression profiling is not always feasible and may be further exacerbated by the substantial phenotypic and genetic heterogeneity observed across different tumour sites [6].

In contrast to invasive procedures, medical imaging is a non-invasive technique that plays a crucial role in routine clinical practice by capturing crucial visual characteristics of

abnormalities. Recent advancements in computerised medical image analysis have led to the development of "radiomics," a high-throughput and quantitative approach that utilises machine learning-based algorithms and large volumes of medical data to extract image features. Radiomic features have been shown to provide valuable supplementary insights into tumour phenotypic traits and complement traditional medical image analysis methods. For instance, Yang et al. proposed a radiomics-based methodology that leverages machine learning models for predicting the overall survival of NSCLC patients [79]. The authors found that radiomics features, including statistical descriptors and tumour textural information, are associated with NSCLC patients' survival time. In another study by Aerts et al., the author demonstrated that radiomics features, such as tumour shape features, are associated with the tumour's underlying gene expression patterns [15].

The studies above have made significant contributions to the growing research field of "radiogenomics," which aims to investigate the association between medical imaging features and genetic characteristics. Radiogenomics features can be determined by identifying image features that have statistically significant associations with the gene expression [10, 15, 21]. Radiogenomics presents opportunities for the non-invasive assessment of important molecular characteristics that contribute to tumour development. Recent studies have shown promising results in identifying prognostic image-genomics biomarkers for several cancers, such as hepatocellular carcinoma (HCC) and renal cell carcinoma (RCC). An et al. [80] reported that radiogenomics features are associated with the mammalian target of rapamycin (mTOR) pathway gene activity in HCC, where the mTOR signalling pathway governs cellular activities and offers opportunities for targeted anti-tumour treatment. Lee et al. [81] identified a collection of radiogenomics features that predict postsurgical metastases in patients with pathological stage T1 RCC.

In contrast to conventional imaging features, radiogenomics features have been shown to provide unique insights into intratumor heterogeneity, which can be linked to clinical outcomes. However, it is important to emphasise that the majority of radiogenomics research at present is constrained to publicly accessible datasets, specifically those featured in The Cancer Imaging Archive (TCIA) [82] and The Cancer Genome Atlas [83]. The limited amount and quality of these datasets present significant challenges for radiogenomics research, particularly with respect to feature selection due to insufficient clinical metadata and annotations.

The upcoming subsections delve into the utilisation of image features and the corresponding image feature extraction techniques in various settings. Figure 3.1 illustrate the fundamental steps involved in a radiogenomics framework. The present discussion focuses on extracting image features from two frequently utilised imaging modalities, namely CT and MR imaging, and their associated application in radiogenomics.

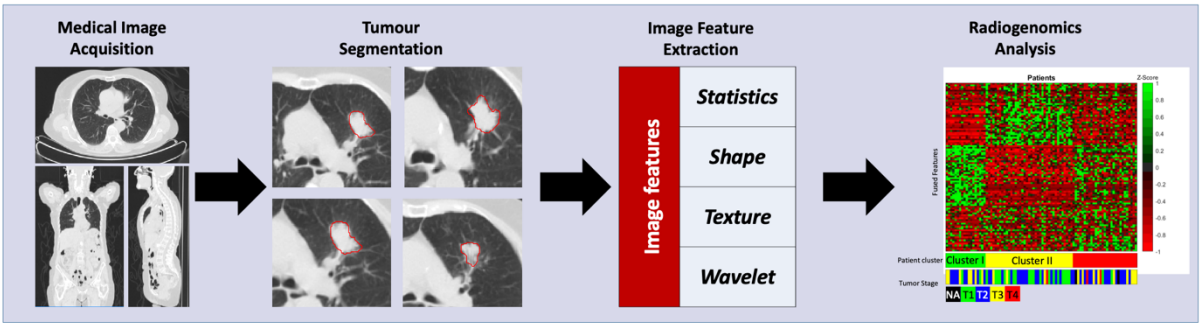


Figure 3.1. The overview of a radiogenomics framework that employs conventional handcrafted image features.

## 3.3 Gene Expression Profiling

Although visual inspection of tumour histopathology is commonly employed to establish a cancer diagnosis, it is susceptible to diagnostic errors, particularly in cases where tumour

classification is difficult due to morphological properties that are visually indistinguishable. Such diseases include diffuse large B-cell lymphoma and breast cancer [84, 85].

Gene expression profiling is a method used to evaluate the activity of genes by examining the process by which genetic information is transformed into functional gene products, such as proteins. Gene expression analysis at various transcription levels can provide a comprehensive view of diverse biological functions that can be identified using computational and statistical methods [86]. Gene expression analysis provides insights that can facilitate predicting the clinical outcome for cancer patients and developing therapies that target specific biological pathways [3, 87]. In this subsection, I will describe the core concept of gene expression analysis and its application in clinical applications relevant to our contribution.

A *gene* is a specific section of deoxyribonucleic acid (DNA) strands that contain genetic information. These codes serve as the foundation for the process of ribonucleic acid (RNA) synthesis, which occurs via transcription. During translation, RNA molecules are capable of synthesising proteins that encode a broad range of biological functions. However, human genetic codes are susceptible to mutations, which are irreversible alterations to genetic elements that can be introduced by a range of factors, such as damage to DNA from environmental effects. Mutations can lead to numerous variations in human DNA, which may ultimately result in alterations in both gene function and behaviour.

Clinical studies have demonstrated that genetic mutations may play an important role in the prognosis of numerous human cancers [88-91]. Gene expression profiling enables the identification of genes with abnormal expressions in tumour tissue samples, thereby providing valuable insights into potential therapeutic options [92]. In addition, gene expression profiling enables the acquisition of patient-specific genetic information, including variations in the genetic information at the individual level. This information can ultimately be utilised in the precision medicine [93].

## 3.4 Image Feature Extraction

The famous adage by Bartlett, "A picture is worth a thousand words," epitomises the abundance of information encompassed within a singular image. Similar to human behaviour, computerised image analysis techniques extract and select information from images to address tasks from various domains.

As briefly discussed in Chapter 2, image feature extraction is a technique that locates and transforms image-encoded information into appropriate image representations for various applications [94-96]. Within the field of image analysis, extracted image features can be categorised as either global, which are computed from the entire image, or local, which are derived from specific ROIs. The statistical, geometric, textural, and structural representations of a picture are commonly beneficial for a broad spectrum of image analysis applications.

Textural features represent a crucial characteristic for the identification of objects and ROIs in an image. The texture is an inherent property of surfaces present in an image, regardless of the modality [97], and contains important information regarding the structural configuration of surfaces and their interrelations with the surrounding environment of the ROIs.

Texture-based features provide a precise depiction of the uniformity of local spatial variations in the pixel intensity [98]. Texture analysis is integral to computerised applications such as image segmentation [99], classification [100], and pattern recognition [101]. Texture-based features commonly employed in these applications include Haralick and wavelet features [102, 103]. Haralick features are derived from the distribution of co-occurring pixel values within a specified spatial neighbourhood, thereby providing a statistical summary of the relative distribution of gray levels in the image [97].

Wavelet-based texture feature extraction has also played a significant role in various image analysis tasks, including content-based image retrieval [104], segmentation [105], and classification [106]. Wavelet features are extracted by decomposing image signals at various resolutions using wavelet orthonormal functions [107]. The orthogonal multi-resolution representation offers a hierarchical framework that presents image characteristics in a coarse-to-fine approach, thereby enabling the extraction of contextual and low-level image features [106, 107].

Colour-based features represent one of the most widely adopted image representations as they are resilient to background complexities and are independent of image size and orientation [108]. Colour features provide both global and local image representations with diverse applications. Colour representations such as colour indexing, colour descriptors, and compact colour moments have been employed in applications such as image retrieval [109-111] and scene recognition [112]. Colour histograms offer insights into the global colour distribution and ranges and have been utilised in applications such as face detection [113] and bleeding detection in the medical domain [114]. Moreover, the use of colour-based image representations in the three-channel domain provides complementary characteristics to the image compared to single-channel grayscale images [115].

Shape-based features provide descriptions that quantify the shape of ROIs in a manner consistent with human perception of specific tasks [116]. Shape representations capture the geometric intricacies of ROIs within an image and are invariant to translation, rotation, scaling, and resistance to noise [108]. Shape-based features have been extracted and applied using multiple approaches. For 2D images, shape-based features have been represented as point sets [117], outline curves [118], and shock graphs [119]. Such features have been utilized for image object classifications [120], recognition [121], and content-based image retrieval in diverse domains [122].

In 3D volumetric images, commonly used shape features involve the use of spherical harmonics that decompose the 3D object into orientation-invariant and descriptive information [123]. Other approaches for extracting 3D shape-based features include the analysis of object surface curvature and the correlograms of objects from various viewpoints [124]. Such 3D shape-based features have found widespread application in 3D model search engines [123], classification tasks [125] and content-based retrieval [126].

In addition to the image features that quantify the statistical information of pixels (referred to as "agnostic features"), another set of features has commonly been utilised to represent high-level tumour characteristics based on human understanding (referred to as "semantic features"), such as tumour shape, size, and necrosis [21]. Unlike agnostic features, semantic features are frequently used by radiologists to describe tumours, with prior knowledge of their prognostic value in the cancer treatment [127]. The application of semantic features in radiogenomics studies has demonstrated their ability to identify prognostic imaging biomarkers [16] and predict gene expression patterns in hepatocellular carcinoma [128]. As a group, agnostic and semantic features quantify the image representation of tumour phenotypic characteristics based on established human knowledge of cancer physiology. To distinguish such a set of image features from AI-generated image descriptors, agnostic and semantic features are henceforth referred to as "handcrafted" features. Handcrafted image features have constituted a vital component of image-based CAD systems [129] in numerous clinical applications, including disease detection and classification [130, 131], enhancing diagnostic performance [132], and ROI segmentations [133].

## 3.5 Summary of Gaps in Radiogenomics Research

Recent advancements in the domain of radiogenomics involve the use of handcrafted image features or deep learning models that aim to extract a wider range of image features with the potential to derive radiogenomics relationships. Although both handcrafted and deep learning-derived image features have demonstrated their ability to encode complementary information in various medical imaging analysis tasks [20], radiogenomics studies have not yet exploited both categories of image features when deriving radiogenomics relationships.

Another challenge in the domain of radiogenomics is the scarcity of annotated data. Radiogenomics studies rely on medical images and their corresponding genetic information obtained from the same patient. Previous studies have extrapolated the conclusions drawn from genetic analyses of a particular disease to another disease that displays comparable genetic characteristics [134, 135]. However, radiogenomics studies have yet to adopt such approaches leverage imaging and genetic data from diseases with similar characteristics when deriving radiogenomics relationships.

Numerous radiogenomics investigations entail the use of multiple image modalities for diverse applications across various cancer types. The diversity of radiogenomics research has motivated studies to employ images with multiple modalities and explore correlations with genetic information. However, published radiogenomics studies have yet to explore the potential benefits of exploiting complementary information encoded in distinct tumour regions and from different modalities [136].

In the following sections, detailed explanation of these gaps in radiogenomics research are described with outline of the proposed methods to address them.

### 3.5.1 Radiogenomics with Fused Feature Signature

Image feature ensemble algorithms offer the opportunity to leverage handcrafted and deep learning-derived image features to extract complementary visual characteristics and provide additional information for medical image analysis. Feature fusion is a common ensemble technique that integrates both categories of image features to produce a more comprehensive image representation of the problem. There have been applications of feature fusion to improve in a range of medical image analysis tasks. Kooi et al. [20] proposed a computer-aided detection system for mammography by using handcrafted and deep image features to quantify image features. The deep features were found to be prone to misclassifying benign abnormalities as tumours because both share similar visual characteristics. In their study, handcrafted features complemented deep features by introducing information that is more difficult for deep features to learn, such as the location and surrounding structures of tumours, thereby increasing the detection performance when compared with using a single category of ETs. Hagerty et al. [137] demonstrated that using both categories of image features improved melanoma classification with increased area under the curve (AUC) of receiver operator characteristics (RUC). The handcrafted features quantified medically meaningful image features such as lesion colour distribution and atypical pigment network and were complementary to deep features that quantified the low-level descriptive image features. Although these ensemble methods demonstrate notable advantages, to the best of our knowledge, the ensemble feature method has yet to be investigated for radiogenomics analysis.

The employment of both handcrafted and deep image features offers opportunities to derive more radiogenomics relationships that may encode unique information. The fusion and extraction of both handcrafted and deep image features will also bring advantages for future image-genomic research. Chapter 4 details a novel radiogenomics method that aims to address such a challenge. Specifically, I propose the fused feature signature ($FF_{Sig}$), a selection of image

features that encodes complementary cancer imaging visual characteristics using a combination of handcrafted and deep learning-based image feature extraction techniques. The $FF_{Sig}$ encodes complementary imaging characteristics of tumours and identifies more radiogenomics relationships with a broader range of genes related to important biological functions. The proposed $FF_{Sig}$ is robust and generalisable across different datasets and allows the identification and extraction of important radiogenomics features that may facilitate cancer diagnosis and treatment planning.

### 3.5.2 Radiogenomics with Multi-disease Analysis

Current radiogenomics investigations rely on extensive amounts of multidimensional data, which can be challenging to obtain and require precise annotations. To mitigate the challenge of insufficient well-annotated datasets, multi-disease analysis has been employed in prior studies, particularly in the case of emerging diseases such as COVID-19 [134, 135]. The utilisation of multi-disease radiogenomics analysis presents opportunities for the identification of radiogenomics features by capitalizing on shared pathophysiological characteristics between two diseases. In addition, multi-disease radiogenomics analysis has the potential to develop alternative approaches or techniques for radiogenomics that are more effective, feasible, and applicable. However, the potential of such adaptations in radiogenomics applications has not been rigorously investigated.

Chapter 5 details a novel radiogenomics method to derive radiogenomics features that are associated with genetic characteristics of one disease, which can be applied to another disease through multi-disease analysis. The proposed radiogenomics method identifies and extracts radiogenomics features that encode common pathophysiological information across

different diseases, thus overcoming the dependence on large volumes of annotated datasets for radiogenomics analysis.

### 3.5.3 Radiogenomics with Tumour Heterogeneity and Multi-modal Imaging Data

Human cancers exhibit strong heterogeneity both within the tumour and among patients. Current radiogenomics investigations have endeavoured to comprehensively extract image features that reflect both anatomical and functional aspects of tumours using multimodal medical imaging, including PET-CT. For example, Nair et al. proposed a radiogenomics approach that employs image features extracted from PET-CT images of NSCLC patients. The authors discovered that textural information from PET-CT images can identify tumours with EGFR mutation and could serve as a surrogate imaging biomarker for NSCLC pre-treatment assessment and prognosis in the precision therapy [138].

Another radiogenomics strategy for quantifying tumour heterogeneity entails the application of handcrafted features extracted from different tumour regions. For instance, Li et al. developed a radiogenomics approach to predict MGMT promoter region methylation that utilises handcrafted features obtained from various GBM tumour regions using a combination of T1w, T1CE, T2w, and T2-FLAIR images [139].

While radiogenomics approaches that quantify tumour heterogeneity information using handcrafted features demonstrated value in medical image analysis tasks, published studies that employ deep learning-based techniques are unable to exploit the complementary information encoded in distinct tumour regions and from different modalities [136]. Chapter 6 describes a radiogenomics method that aims to address such a challenge. The proposed framework

quantifies the radiogenomics features from diverse image modalities and accounts for unique information encoded in tumour heterogeneity sub-regions.

# Chapter 4. Radiogenomics with Fused Image Feature Signature

In this chapter, I present fused feature signature ($FF_{Sig}$), a selection of image features from both handcrafted and deep learning-derived image features (e.g., transfer learning and fine-tuning of deep learning models) for identifying radiogenomics relationships (RRs). Previous radiogenomics studies mainly relied on a single category of image feature extraction techniques (ETs); these are (i) handcrafted ETs that encompass visual imaging characteristics, curated from knowledge of human experts and, (ii) deep ETs that quantify abstract-level imaging characteristics from large data. Prior studies, therefore, failed to leverage the complementary information that are accessible from fusing the ETs. $FF_{Sig}$ improves upon conventional radiogenomics research by exploiting the complementary information available from fusing both handcrafted and deep learning-derived image features. To validate the performance of the proposed $FF_{Sig}$, experiments are conducted using two public datasets that contain CT images from patients with NSCLC. Experimental results demonstrate that $FF_{Sig}$ encodes complementary imaging characteristics of tumours and identifies more RRs with a broader range of genes related to important biological functions.

## 4.1 Contributions

The $FF_{Sig}$ has the following contributions:

- The $FF_{Sig}$ is constructed by leveraging the complementary information from handcrafted and deep features, followed by a multi-step feature selection scheme. The

$FF_{Sig}$ can determine radiogenomics relationships with gene expressions and gene ontology (GO) terms.

- The proposed $FF_{Sig}$ identified more and unique radiogenomics relationships with gene expressions and gene ontology (GO) terms when compared to the use of handcrafted or deep features independently.
- The proposed $FF_{Sig}$ is exclusively associated with biological functions that are related to protein kinase activities that play crucial roles in the tumourigenicity of NSCLC.
- Experimental results demonstrate the potentials for the proposed $FF_{Sig}$ to identify important radiogenomics relationships that may facilitate cancer diagnosis and treatment in the future.

The contribution of this chapter is aligned with the aim 1 stated in section 1.3 of Chapter 1 and addresses the gap 3.5.1 identified in Chapter 3. Furthermore, this chapter provides an extensive analysis and comparison of state-of-the-art radiogenomics analysis pipelines, which builds upon the literature review presented in Chapter 3.

# 4.2 Materials and Methods

## 4.2.1 Overview of the framework

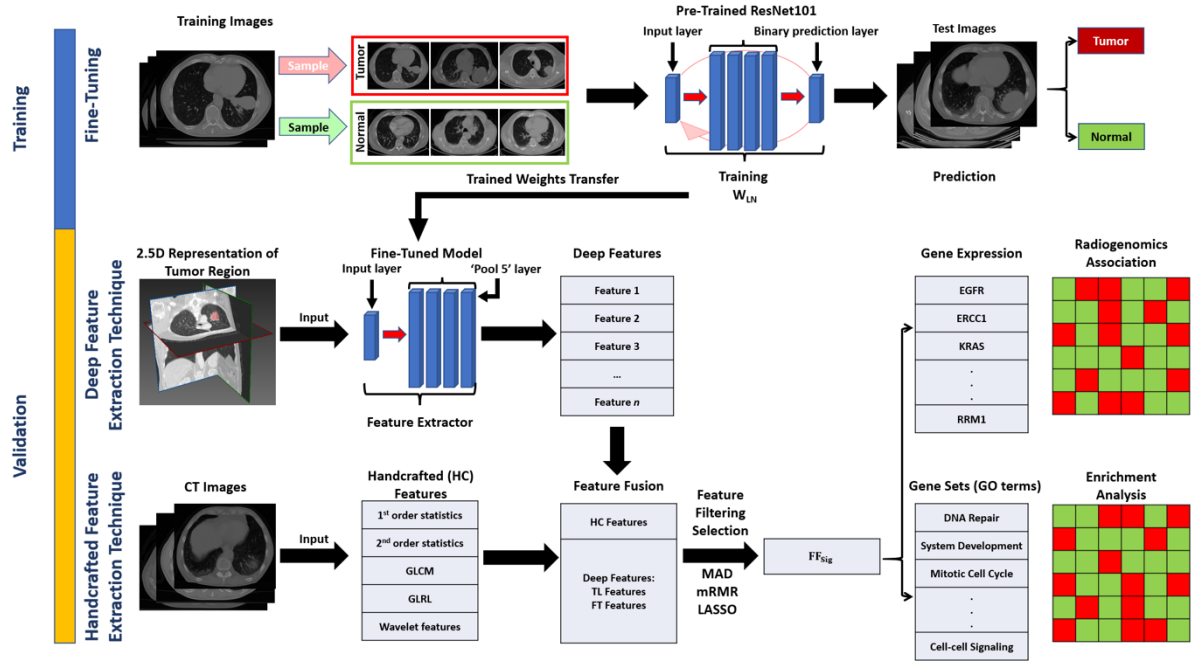


Figure 4.1. The workflow for generating the $FF_{Sig}$ and the identification of RRs with genes and GO terms.

Figure 4.1 presents an overview of the experimental design for constructing $FF_{Sig}$. Handcrafted and deep learning-based image features extraction techniques are used to extract handcrafted (denote as HC), transfer learning (denote as TL) and fine tuning (denote as FT) features from delineated tumour ROIs from CT image volumes. HC features are extracted from the CT image volume directly. TL and FT features are extracted from a 2.5D representation of the CT data around the tumour centroid. The extracted HC, TL and FT features are fused into a feature matrix using concatenation. The $FF_{Sig}$ is generated by applying a multi-step feature selection procedure involving median absolute deviation (MAD), minimum redundancy maximum relevance (mRMR), and least absolute shrinkage and selection operator (LASSO) generalised linear model. RRs are determined by using Spearman rank correlation between $FF_{Sig}$ and the averaged gene expressions. RRs between image features signatures and GO terms are determined by using GSEA. For evaluation purposes, the same multi-step feature selection

procedure is applied to HC, TL and FT features. The resulting collections of image features are denoted as $HC_{Sig}$, $TL_{Sig}$ and $FT_{Sig}$, respectively.

### 4.2.2 NSCLC Radiomics-Genomics Dataset

We used the public NSCLC Radiomics-Genomics dataset [140] from the Harvard University, and we refer to this dataset as the 'NRG-H'. The dataset was sourced from the Cancer Imaging Archive (TCIA)[82]. The NRG-H is a pre-processed and de-identified dataset. The creator of the dataset has indicated that the collection and processing of the dataset were conducted according to national laws and guidelines and approved by the appropriate local trial committee at Maastricht University Medical Center (MUMC1), Maastricht, The Netherlands. The dataset comprises 89 patients (29 W, 60 M; age range 37 – 85 years) with histologically confirmed NSCLC with T stage (T1-T4) [141].

All patients had a CT scan of the thorax / upper abdomen. CT scan slice thickness was between 1.5mm and 5mm. Gene expression information was acquired using the Rosetta/Merck human RSTA custom Affymetrix 2.0 microarray (Affymetrix HuRSTA-2a520709). Gene expression values were normalised using the RMA algorithm 5 in Bioconductor. Gene expression information was accessed via the Gene Expression Omnibus (GEO)[142]. The primary tumours were delineated by an experienced medical imaging specialist (M.F., more than 20 years of experience), slice-by-slice, on trans-axial image slices using open source software (Medical imaging Interaction Toolkit (MITK); version 2016.11 [143]). We excluded three patients (all men) because there were lung collapses distal to a proximal tumour and the extent of the tumour could not be reliably identified. Delineations were independently validated by a second clinician (E.K., 7 years of experience).

### 4.2.3 NSCLC-Radiogenomics Dataset

The NSCLC-Radiogenomics dataset reported by Bakr et al[144] from the Stanford University is a pre-processed and de-identified dataset, and we refer to this dataset as 'NRG-S'. The creator of the dataset has indicated that the collection and processing of the dataset were conducted under IRB approval from Stanford University and the Veterans Administration Palo Alto Health Care System. The NRG-S dataset comprises CT images and RNA-Seq data from 117 subjects (29 W ,88 M; age range 46 – 85 years) with histologically confirmed NSCLC with T stage (Tis, T1-T4).

All patients had a CT scan from the apex of the lung to the adrenal gland in supine position. CT scan thickness was between 0.625mm and 3mm. Detailed scanning parameters, such as the manufacturer attributes are specified in the DICOM headers. Total RNA was extracted from the tissue and analysed with RNA sequencing technology. Gene expression information was processed using the STAR algorithm[145] and Cufflinks version 2.0.2 [146]. Gene expression information was accessed via the Gene Expression Omnibus (GEO)[142]. Primary tumours were segmented using an unpublished automatic segmentation algorithm on the axial image slices for all 117 subjects. Segmentations were viewed by a thoracic radiologist (M.K.) with more than 5 years of experience and edited as necessary using ePAD. An additional thoracic radiologist (A.N.L.) reviewed and approved the final segmentations.

### 4.2.3 Image Features

We employed a set of standard HC feature extraction technique that are implemented in the pyradiomics framework to quantify HC features [147, 148]. For each patient, we extracted a well-documented set of 431 HC features from CT volumes [149, 150]. These 431 HC features comprised the following: (a) first-order statistics, describing the distribution of voxel

intensities; (b) shape and size that are geometric descriptors of tumoural 3D characteristics such as compactness and surface area; (c) textural or co-occurrence matrix features to illustrate the spatial distribution of the voxel intensities and, (d) first order statistics and textural features of the wavelet decompositions of the raw imaging data.

Deep image feature was extracted by using a ResNet-101 backbone that was pre-trained on ImageNet ILSVRC challenge data. To adopt the ResNet-101 model with pre-trained weight and to recognise the features in the NSCLC CT data, we fine-tuned it for the task of identifying CT images that contained tumours. The 86 subjects from the NRG-H dataset were divided into two groups: a training set that comprises imaging data from 69 patients and, a testing set that comprises imaging data from 17 patients. Subjects in the training and testing groups were randomly selected. We implemented a 5-fold cross-validation strategy on the training set of 69 patients to fine-tune the ResNet-101 model. The fine-tuning process of the ResNet-101 model involved 300 epochs of training using stochastic gradient descent with a momentum of 0.9 and a batch size of 5. The Learning rate was set at $1 \times 10^{-3}$, with L2Regularization set at 0.001. For every 100 epochs, the learning rate decreased by the factor of 0.1. TL and FT features were extracted from the 'pool5' layer of the ResNet-101 model.

We used a feature fusion strategy that concatenates the HC, TL and FT feature together to generate a feature matrix across the patients [151]. We hence applied a multi-step image feature selection scheme that aims to: i) reduce the dimensionality of the concatenated feature matrix; ii) remove image features that are redundant or irrelevant to the histology classification of tumours; and iii) identify a set of image features that are most relevant to the histology characteristics of patients. We used the median absolute deviation (MAD) as an indication for these features as it measures the variability across features and is robust against outliers in the concatenated feature matrix. Features that have poor variability and dispersion across patients were removed. The second stage reduced the dimensionality of the remaining features by

removing those that are redundant or irrelevant to the histology characteristics of patients. The histology characteristics of each patient were categorised into one the following classes: (1) squamous cell carcinoma, (2) adenocarcinoma and (3) other types including non-Small cell and Not otherwise specified (NOS). We used mRMR, a widely-adopted approach for feature selection, to produce a subset of features with high biological relevance [152]. A total of 100 features were selected using the mRMR method, taking into consideration of the number of patients as well as the original dimensionality of the feature matrix [153]. The last stage of feature selection employed LASSO regularisation for generalised linear models to identify the set of remaining image features that are most relevant to the histology characteristics of patients. We also applied the multi-stage image feature selection process to the HC, TL and FT features individually for comparison. The resulting image feature signatures are hereafter denoted as '$HC_{Sig}$', '$TL_{Sig}$' and '$FT_{Sig}$', correspondingly.

### 4.2.4 Radiogenomics Analysis

We used the following process to remove genes that had low variance, entropy and absolute expression value because such genes showed poor variability and dispersion, and therefore may not reflect the differences in the underlying tumour biology. We firstly removed genes with a variance of less than one-quarter percentile, as such genes may not reflect changes in tumour biological behaviours. The averaged gene expression was filtered to remove the genes with a variance of less than one-quarter percentile across all patients. The remaining genes were then filtered to remove genes that have an absolute expression level in the lowest quarter percentile of the gene expression; genes with low absolute expression were removed because they are prone to errors due to large quantisation or spot hybridisation. Finally, gene expressions were filtered to remove the genes with an entropy value that is less than the quarter percentile; genes

with low entropy are considered to be consistently expressed across patients and may not reflect the variance in tumour biological characteristics [154].

We determined RRs between the $FF_{Sig}$ with the averaged gene expressions using the Spearman rank correlation. We also employed functional enrichment analysis to enrich radiogenomics relationships with GO terms. We used 1,046 gene sets from the C5 collection of MSigDB [155], which categorise the following GO terms: molecular function, cellular component and biological process. The gene list was generated by ranking the radiogenomics associations for each of the features from $FF_{Sig}$ in descending order. Gene sets that include between 15 and 500 contributing genes were selected for the enrichment analysis as was the standard protocol in prior work [147]. The determined RRs were then assessed using a pre-ranked functional enrichment analysis. In this process, the radiogenomics relationships between $FF_{Sig}$ and gene expressions were sorted to provide a ranked gene list based on the strength of the Spearman rank correlation. We used the pre-ranked gene list to perform GSEA, which derives the association between the provided ranked gene list and GO terms by testing the enrichment of each annotated term iteratively in a linear model. The enriched radiogenomics relationships with GO terms can be quantified by calculating normalised enrichment scores (NES) based on the number of genes. To ensure that only significantly associated genes were used for functional enrichment analysis, RRs with p-value $< 0.001$ were selected and ranked and serve as input to the functional enrichment analysis with GO terms. The same procedure was applied to the $HC_{Sig}$, $TL_{Sig}$ and $FT_{Sig}$ for comparative experiments.

### 4.2.5 Evaluation Strategy

We evaluated the performance of $FF_{Sig}$ by: (i) determining if the proposed $FF_{Sig}$ can encode complementary medical image visual characteristics when compared with other image feature signatures; ii) determining if the proposed $FF_{Sig}$ is relevant to the tumour T stage by using the

$\chi 2$ test of independence; iii) assessing the distribution of RRs with genes; iv) assessing the distribution of RRs with GO terms; iv) determining if the proposed $FF_{Sig}$ can identify exclusive RRs with genetic biomarkers of NSCLC and GO terms that are related to NSCLC.

# 4.3 Results

## 4.3.1 Image Signatures and T Stage

The HC, TL and TF features were significantly associated with the T stage parameters (T1-T4) across patient clusters. The $\chi 2$ test statistics for HC, TL and TF features with T stage parameters are $p < 2.9 \times 10^{-4}$, $p < 5.0 \times 10^{-3}$ and $p < 4.8 \times 10^{-2}$, respectively. For image signatures, $FF_{Sig}$ was significantly associated with primary tumour T stages ($\chi 2$ test, $p < 4.0 \times 10^{-2}$). None of the $HC_{Sig}$, $TL_{Sig}$ or $FT_{Sig}$ is found to be significantly associated with primary tumour T stages, their $\chi 2$ test statistics are $p > 0.8$, $p > 6.0 \times 10^{-2}$ and $p > 0.5$, respectively. Figure 4.2 illustrate the relationships among $FF_{Sig}$, T stages and patient clusters from the NRG-H dataset.

## 4.3.2 RRs between Image Feature Signatures and Genes

A total of 11318 gene expression remained from the NRG-H dataset to establish radiogenomics associations. Figure 4.3a. represents the distribution of RRs that were determined between the averaged gene expression values of 11,318 individual genes and $FF_{Sig}$, $HC_{Sig}$, $TL_{Sig}$ and $FT_{Sig}$. $FF_{Sig}$ identified the highest number of RRs at 5039 and correlated with the highest number of genes at 3881. $HC_{Sig}$ identified 1193 RRs with 886 genes. $TL_{Sig}$ identified 3816 RRs with 3297 genes. $FT_{Sig}$ identified 2089 RRs with 2008 genes. Figure 4.4a. details the distribution of unique genes that were associated with $FF_{Sig}$, $HC_{Sig}$, $TL_{Sig}$, and $FT_{Sig}$. Among the 3,881 unique genes that were associated with the $FF_{Sig,}$ 1,896 unique genes cannot be associated with any of the $HC_{Sig}$, $TL_{Sig}$, and $FT_{Sig}$. In contrast, a total number of 3,269 unique genes were associated with one of the $HC_{Sig}$, $TL_{Sig}$, and $FT_{Sig}$, but were not correlated with the $FF_{Sig}$.

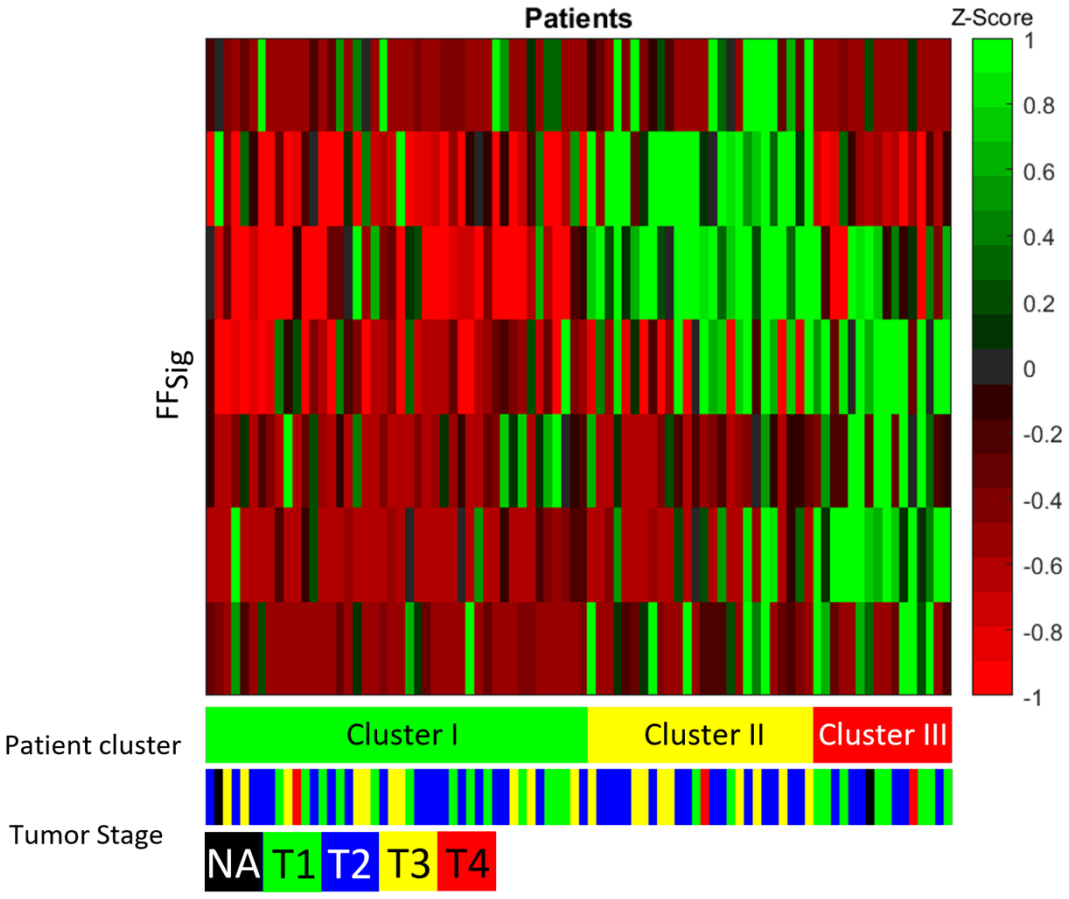


Figure 4.2. Heatmap of the $FF_{Sig}$ across patient clusters with corresponding T stage from the NRG-H dataset.

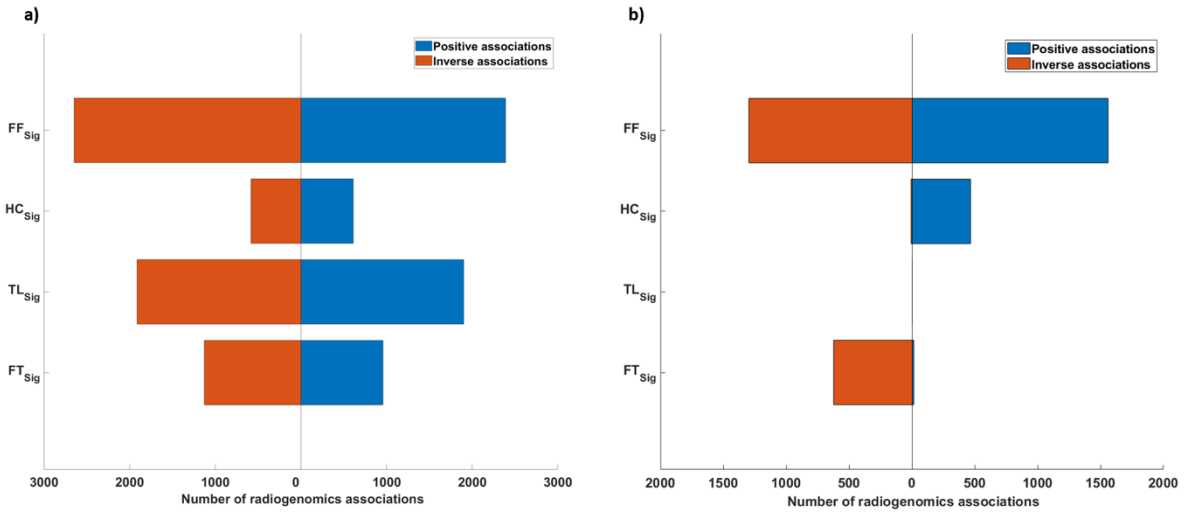


Figure 4.3. The distribution of RRs between feature signatures and: a) gene expression value of the processed genes (n = 11,318) from the NRG-H dataset. b) gene expression value of the processed genes (n = 2,993) from the NRG-S dataset.

Table 4.1 compares the strengths of all RRs that were determined using the $FF_{Sig}$ against those determined using $HC_{Sig}$, $TL_{Sig}$ and $FT_{Sig}$. Our results show stronger RRs are identified between the $FF_{Sig}$ and genes, when compared with $HC_{Sig}$ and $TL_{Sig}$, in the inverse direction. Figure 4.5 illustrates the distribution of RRs that were determined between image feature signatures of the $FF_{Sig}$, $HC_{Sig}$, $TL_{Sig}$, $FT_{Sig}$ with the gene expression value from the key genetic biomarkers of EGFR for NSCLC [156]. Our result shows that the $FF_{Sig}$ and $FT_{Sig}$ were inversely correlated with EGFR expression. In contrast, $HC_{Sig}$ is shown to be the only positive RRs with EGFR. Notably, $FT_{Sig}$ shows to derive more and stronger inverse RRs with EGFR when compared with the $FF_{Sig.}$ In addition, our result shows that ERCC1, a key genetic biomarker for NSCLC, is exclusively correlated with a single feature from the $FF_{Sig,}$ where the same feature showed inverse RRs with EGFR previously.

Using the NRG-S dataset, a total of 2,993 gene expression remained to establish radiogenomics associations. Figure 4.3b. represents the distribution of RRs that were determined between the averaged gene expression values of 2,993 individual genes and $FF_{Sig}$,

$HC_{Sig}$, and $FT_{Sig}$. Figure 4.4b. details the distribution of unique genes that were associated with $FF_{Sig}$, $HC_{Sig}$, $TL_{Sig}$, and $FT_{Sig}$. Table 4.2. compares the strengths of all RRs that were determined using the $FF_{Sig}$ against those determined using $HC_{Sig}$, and $FT_{Sig}$. Our validation results show that the $FF_{Sig}$ did not identify stronger RRs with genes, when compared with $HC_{Sig}$ and $TL_{Sig}$, in both statistical directions.

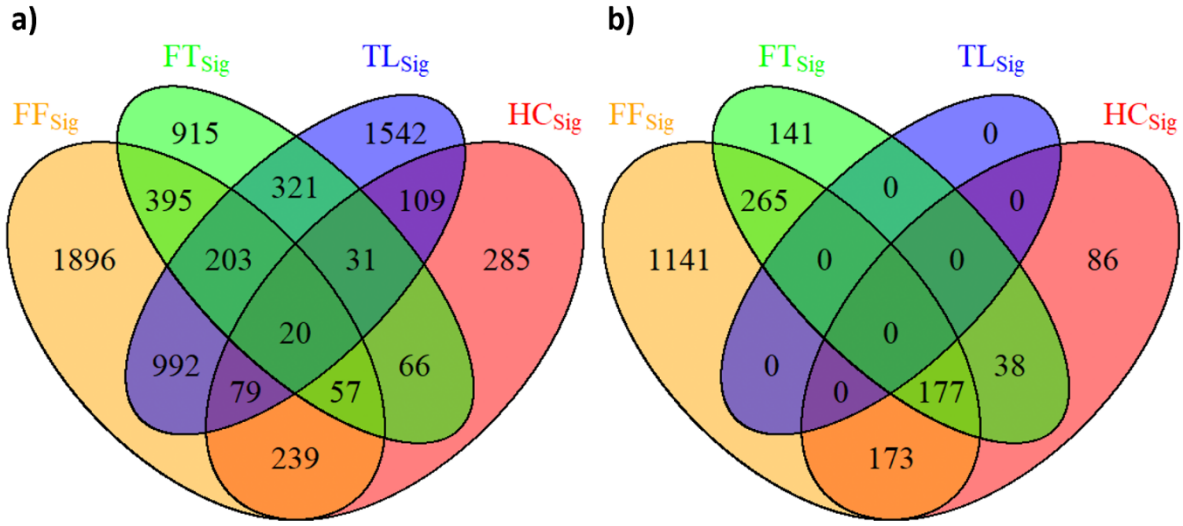


Figure 4.4. Venn diagram shows the distribution of unique genes that were associated with $FF_{Sig}$, $HC_{Sig}$, $TL_{Sig}$, and $FT_{Sig}$: a) generated using the NRG-H dataset. b) generated using the NRG-S dataset.

Table 4.1. Two-Sample t-Tests that assess the strengths of all RRs constructed using the $FF_{Sig}$ with $HC_{Sig}$, $TL_{Sig}$ and $FT_{Sig}$, in both statistical directions on the NRG-H dataset.

| **Strength of Positive RRs (Two-Sample t-Test)** | | | |
|---|---|---|---|
| Feature signature | **$HC_{Sig}$** | **$TL_{Sig}$** | **$FT_{Sig}$** |
| $FF_{Sig}$ | $p > 0.2$ | $p > 0.7$ | $p > 0.3$ |
| **Strength of Inverse RRs (Two-Sample t-Test)** | | | |
| $FF_{Sig}$ | $p < 1 \times 10^{-3}$ | $p < 1 \times 10^{-2}$ | $P > 0.6$ |

Table 4.2. Two-Sample t-Tests that assess the strengths of all RRs constructed using the $FF_{Sig}$ with $HC_{Sig}$ and $FT_{Sig}$, in both statistical directions on the NRG-S dataset.

| **Strength of Positive RRs (Two-Sample t-Test)** | | |
|---|---|---|
| Feature signature | **$HC_{Sig}$** | **$FT_{Sig}$** |
| $FF_{Sig}$ | $p > 0.8$ | $p > 0.2$ |
| **Strength of Inverse RRs (Two-Sample t-Test)** | | |
| $FF_{Sig}$ | $p > 0.3$ | $P > 0.08$ |

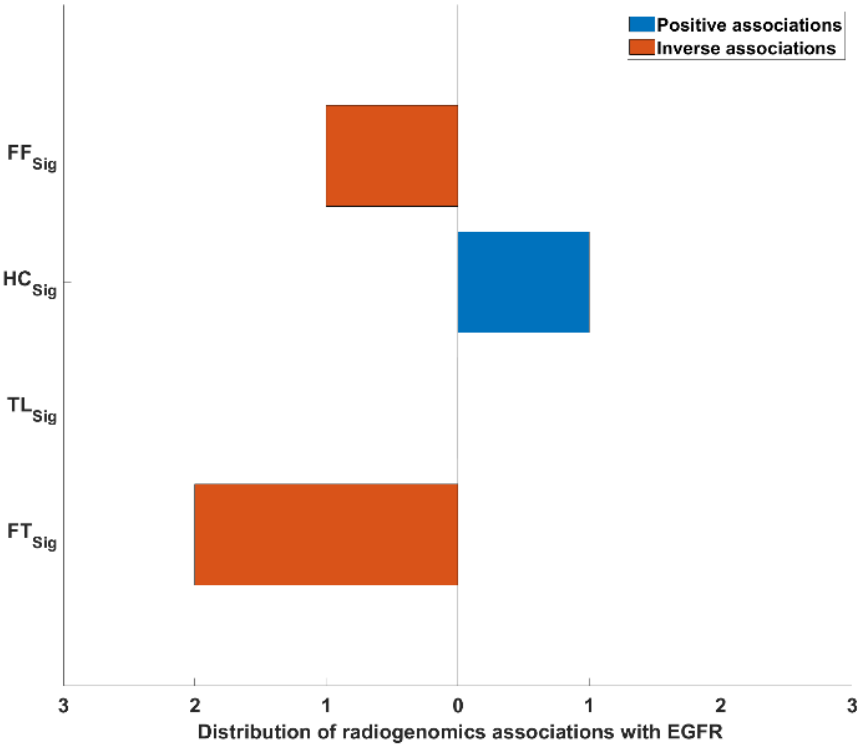


Figure 4.5. The distribution of RRs between the $FF_{Sig}$ with the key genetic biomarker of EGFR from the NRG-H dataset, in comparison to $HC_{Sig}$, $TL_{Sig}$ and $FT_{Sig}$.

## 4.3.2 RRs between Image Feature Signatures and GO Terms

From our experiments using the NRG-H dataset, $FF_{Sig}$ determined RRs with the highest number of GO terms at 244. $HC_{Sig}$ determined RRs with 62 GO terms $TL_{Sig}$ determined RRs with 246 GO terms. $FT_{Sig}$ determined RRs with 129 GO terms. Figure 4.6a. details the distribution of

GO terms that were associated with image feature signatures of $FF_{Sig}$, $HC_{Sig}$, $TL_{Sig}$, and $FT_{Sig}$. Among the 244 GO terms that have RRs with by $FF_{Sig}$, 122 GO terms were exclusively enriched; these GO terms account for 50% of the total enriched GO terms or 13.8% of the total 1,046 GO terms.

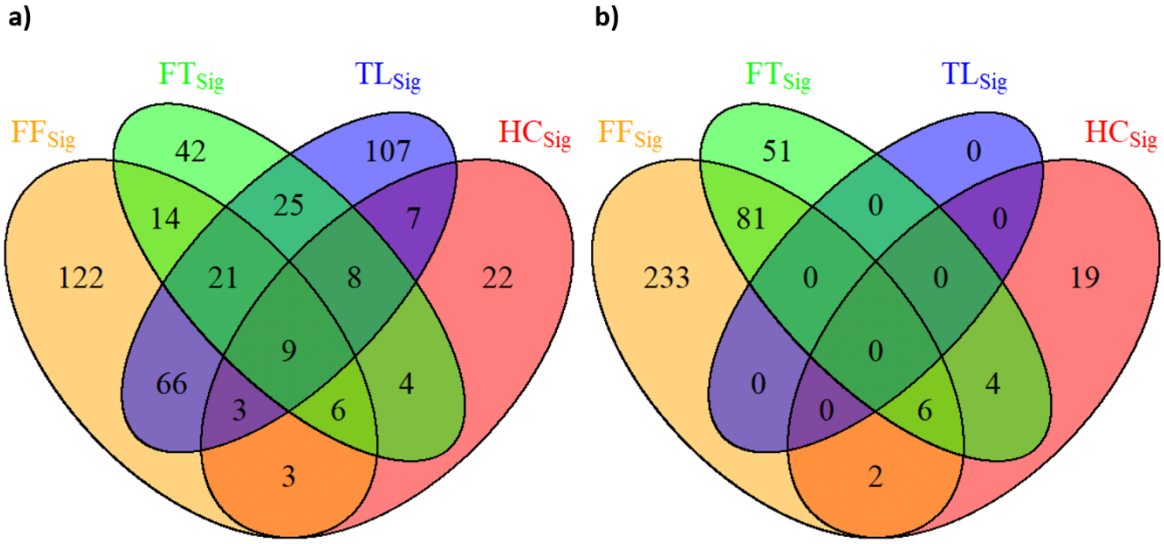


Figure 4.6. Venn diagram shows the distribution of GO terms that were associated with image feature signatures of $FF_{Sig}$, $TL_{Sig}$, $FT_{Sig}$ and $HC_{Sig}$: a) generated using the NRG-H dataset. b) generated using the NRG-S dataset.

Table 4.3 shows the GO terms with the highest NES. Notably, $FF_{Sig}$ determined RRs with GO terms that exhibit distinct patterns relating to the biological functions and cellular behaviours: i) 3 GO terms were related to lumen structures including organelle, nuclear and membrane; ii) 2 GO terms were reflecting biosynthesis processes that involve glycoprotein or macromolecule iii) 3 GO terms were related to the response mechanism to viruses, other organism or biotic stimulus. and other types of stimulus processes. In comparison, our results also show that $TL_{Sig}$ determined RRs with 4 GO terms that are associated with fraction activities. In addition, $FT_{Sig}$ determined RRs with GO terms that are related to enzyme activities. In contrast, $HC_{Sig}$ determined RRs with GO terms are shown to be without overlaps in their biological functionalities.

Table 4.5 shows the comparison between GO terms that have exclusive RRs with $FF_{Sig}$ and those GO terms that are restricted to have RRs with $FF_{Sig}$. Among the GO terms with the highest NES, our result shows clusters of biological functions and cellular behaviours that have exclusive RRs with the $FF_{Sig}$: i) 3 GO terms were related to kinase activities for transmembrane receptor protein and tyrosine kinase; ii) 2 GO terms were related to metabolism activities; The identical 3 GO terms were most enriched by $FF_{Sig}$ and related to the virus response mechanism. In contrast, our result shows 2 groups of related biological functions among the GO terms that were restricted to $FF_{Sig.}$ Such GO terms are related to fraction processes and enzyme activities.

From our validation experiment on the NRG-S dataset, functional gene enrichment analysis reveals that $FF_{Sig}$ determined RRs with the highest number of GO terms at 322. $HC_{Sig}$ determined RRs with 31 GO terms. $TL_{Sig}$ determined RRs with 0 GO terms. Figure 4.6b. details the distribution of GO terms that were associated with image feature signatures of $FF_{Sig,}$ $HC_{Sig}$, $TL_{Sig}$, and $FT_{Sig}$. $FT_{Sig}$ determined RRs with 142 GO terms. Among the 322 GO terms that have RRs with by $FF_{Sig,}$ 233 GO terms were exclusively enriched; these GO terms account for 72.4% of the total enriched GO terms or 22.3% of the total 1,046 GO terms.

Table 4.3. The GO terms that have RRs with $FF_{Sig}$, $HC_{Sig}$, $TL_{Sig}$ and $FT_{Sig}$ with positive and negative associations from the NRG-H dataset.

| **$FF_{Sig}$** | **NES** | **$HC_{Sig}$** | **NES** |
|---|---|---|---|
| Organelle Lumen | 2.43 | Extracellular Region | 1.74 |
| Nuclear Lumen | 2.22 | Regulation of Transferase Activity | 0.60 |
| Membrane Enclosed Lumen | 2.19 | Transferase Activity Transferring Phosphorus Containing Groups | 0.58 |
| Glycoprotein Biosynthetic Process | 1.98 | Protein Kinase Activity | 0.58 |
| Macromolecule Biosynthetic Process | 1.94 | Stress Activated Protein Kinase Signalling Pathway | 0.58 |
| Response to Virus | -1.98 | Carbohydrate Metabolic Process | -0.99 |
| Cell Cell Signalling | -1.98 | Phosphoric Monoester Hydrolase Activity | -0.99 |
| Response to Other Organism | -2.00 | Phosphoric Ester Hydrolase Activity | -1.01 |
| Anatomical Structure Morphogenesis | -2.01 | Alcohol Metabolic Process | -1.02 |
| Response to Biotic Stimulus | -2.01 | Hydrolase Activity Acting on Ester Bonds | -1.02 |
| **$TL_{Sig}$** | | **$FT_{Sig}$** | |
| Cell Fraction | 2.17 | Anatomical Structure Morphogenesis | 1.85 |
| Membrane Fraction | 2.03 | Enzyme Regulator Activity | 1.80 |
| Phosphoric Ester Hydrolase Activity | 2.02 | Enzyme Activator Activity | 1.79 |
| Soluble Fraction | 1.96 | Enzyme Linked Receptor Protein Signalling Pathway | 1.77 |
| Insoluble Fraction | 1.96 | Membrane Fraction | 1.73 |
| Homophillic Cell Adhesion | -1.66 | Extracellular Region Part | -1.23 |
| Sulfuric Ester Hydrolase Activity | -1.67 | Extracellular Space | -1.23 |
| Nervous System Development | -1.68 | Phosphorylation | -1.24 |
| Regulation of Anatomical Structure Morphogenesis | -1.68 | Lipase Activity | -1.25 |
| Cell Surface | -1.99 | Female Pregnancy | -1.27 |

Table 4.4. The GO terms that have RRs with $FF_{Sig}$, $HC_{Sig}$ and $FT_{Sig}$ with positive and negative associations from the NRG-S dataset.

| **$FF_{Sig}$** | **NES** | **$HC_{Sig}$** | **NES** | **$FT_{Sig}$** | **NES** |
|---|---|---|---|---|---|
| Perinuclear Region of Cytoplasm | 2.62 | Sensory Perception | 1.80 | Intracellular Protein Transport | 2.62 |
| Nervous System Development | 2.58 | Monooxygenase Activity | 1.78 | Establishment of Protein Localisation | 2.61 |
| Membrane Organisation and Biogenesis | 2.45 | Oxygen Binding | 1.78 | Macromolecule Localisation | 2.61 |
| Intercellular Junction | 2.05 | Electron Transport (GO 0006118) | 1.75 | Protein Localisation | 2.54 |
| Tight Junction | 1.96 | Neurological System Process | 1.70 | Protein Transport | 2.52 |
| Kinase Activity | -2.00 | Second Messenger Mediated Signalling | -0.77 | Soluble Fraction | -1.61 |
| Endoplasmic Reticulum | -2.11 | Establishment and or Maintenance of Cell polarity | -0.77 | Organelle Lumen | -1.62 |
| Nuclear Lumen | -2.19 | Regulation of Catalytic Activity | -0.77 | Nucleolus | -1.65 |
| Organelle Lumen | -2.84 | cAMP Mediated Signalling | -0.77 | Nuclear Lumen | -1.67 |
| Membrane Enclosed Lumen | -3.06 | G Protein Signalling Adenylate Cyclase Activating Pathway | -0.77 | Membrane Enclosed Lumen | -1.71 |

Table 4.4. shows the GO terms with the highest NES. Notably, $FF_{Sig}$ determined RRs with GO terms that exhibit distinct patterns relating to the cellular structure: i) 3 GO terms were related to lumen structures including organelle, nuclear and membrane; ii) 2 GO terms that reflect the cell junction. In comparison, $FT_{Sig}$ determined RRs with GO terms that are related to cellular structures, protein transportation and localisation. $HC_{Sig}$ determined RRs with GO terms that are related to signalling pathways, such as cAMP mediated signalling and second messenger mediated signalling.

Table 4.6. shows the comparison between GO terms that have exclusive RRs with $FF_{Sig}$ and those GO terms that are restricted to have RRs with $FF_{Sig}$. Among the GO terms with the highest NES, our validation results show a cluster of biological functions and cellular behaviours that have exclusive RRs with the $FF_{Sig}$: 3 GO terms were related to peptidase activity; ii) 2 GO terms that reflect the cell junction. In contrast, our result shows 2 groups of

related biological functions among the GO terms that were restricted to $FF_{Sig.}$ Such GO terms are related to the intrinsic components of organelle membranes and metabolic processes.

Table 4.5. The GO terms that have the highest NES and exclusively RRs with $FF_{Sig}$ (left) and the GO terms that are restricted to have RRs with $FF_{Sig}$ (right), experimented on the NRG-H dataset.

| **$FF_{Sig}$ Exclusive** | **NES** | **$FF_{Sig}$ Restricted** | **NES** |
|---|---|---|---|
| Transmembrane Receptor Protein Kinase Activity | 1.61 | Soluble Fraction | 1.96 |
| Protein Tyrosine Kinase Activity | 1.60 | Insoluble Fraction | 1.96 |
| Transmembrane Receptor Protein Tyrosine Kinase Activity | 1.53 | Enzyme Regulator Activity | 1.80 |
| Generation of Precursor Metabolic and Energy | 1.47 | Enzyme Activator Activity | 1.79 |
| Phospholipid Metabolic Process | 1.42 | Molecular Adaptor Activity | 1.73 |
| RNA Processing | -1.85 | Generation of Neurons | -1.66 |
| Organ Morphogenesis | -1.95 | Homophilic Cell Adhesion | -1.67 |
| Response to Virus | -1.98 | Sulfuric Ester Hydrolase Activity | -1.67 |
| Response to Other Organism | -2.00 | Regulation of Anatomical Structure Morphogenesis | -1.68 |
| Response to Biotic Stimulus | -2.01 | Cell Surface | -1.99 |

## 4.4 Discussion

Our main findings are that our $FF_{Sig}$: i) encoded complementary medical image's visual characteristics when compared with other image feature signatures; (ii) determined a greater number of RRs with a greater number of genes; (iii) determined RRs with distinct GO terms; (iv) determined exclusive RRs with genetic biomarkers of NSCLC and GO terms that are related to NSCLC and (v) is robust and generalisable for determining RRs when validated on NRG-S.

From our experiments using the NRG-H dataset, the $FF_{Sig}$ comprises 7 image features that are complementary to image features that were selected in the $HC_{sig}$, $TL_{sig}$, and $FT_{sig}$. Image features that are included in the $FF_{Sig}$ can be traced back to the 6,144-dimensional TL features. This finding indicates that the multi-step feature selection scheme prioritised a set of complementary image features that are relevant to the histological characteristics while reducing the overall redundancy in the information captured. This finding suggests that the $FF_{Sig}$ encodes unique medical imaging visual characteristics when compared with other image signatures. The $FF_{Sig}$ was the only feature signature that produced a significant association ($p < 0.05$) with the T stage. The $HC_{Sig}$, $TL_{Sig}$, and $FT_{Sig}$ did not have any association with the T stage, despite the fact that the $FF_{Sig}$ was selected from the HC, FT, and TL features. Our results showed that the semantic information that is encoded in the HC features and the abstract-level information that are encoded in the TL and FT features contributed towards the selection of features in $FF_{Sig}$. This finding implies that the association between $FF_{Sig}$ and T stage occurred because the $FF_{Sig}$ leveraged complementary information using both HC and deep ETs.

Table 4.6. The GO terms that have the highest NES and exclusively RRs with $FF_{Sig}$ (left) and the GO terms that are restricted to have RRs with $FF_{Sig}$ (right), experimented on the NRG-S dataset.

| **$FF_{Sig}$ Exclusive** | **NES** | **$FF_{Sig}$ Restricted** | **NES** |
|---|---|---|---|
| Perinuclear Region of Cytoplasm | 2.62 | Positive Regulation of Metabolic Process | 1.93 |
| Membrane Organisation and Biogenesis | 2.45 | Positive Regulation of Cellular Metabolic Process | 1.90 |
| Intercellular Junction | 2.05 | Neurite Development | 1.90 |
| Tight Junction | 1.96 | Steroid Hormone Receptor Signalling Pathway | 1.89 |
| Apical Junction Complex | 1.94 | Cellular Lipid Catabolic Process | 1.88 |
| Serine Type Peptidase Activity | -1.57 | cAMP Mediated Signalling | -0.77 |
| Serine Hydrolase Activity | -1.58 | G Protein Signalling Adenylate Cyclase Activating Pathway | -0.77 |
| Serine Type Endopeptidase Activity | -1.60 | Intrinsic to Golgi Membrane | -0.88 |
| Peptidase Activity | -1.75 | Intrinsic to Organelle Membrane | -0.93 |
| Endopeptidase Activity | -1.76 | Integral to Organelle Membrane | -0.93 |

The $FF_{Sig}$ determined a greater number of RRs with a greater number of genes when compared with the other image feature signatures. The $FF_{Sig}$ was also correlated with EGFR. One potential explanation for our finding is that the $FF_{Sig}$ encodes the imaging characteristics of the tumour that can reflect the underlying molecular characteristics of NSCLC [157]. The $FF_{Sig}$ has also determined stronger inverse RRs with a range of genes when compared to $HC_{Sig}$ and $TL_{Sig}$. There was no stronger positive RRs with genes when compared with the $HC_{Sig}$, $TL_{Sig}$ and $FT_{Sig}$. The reason for this is because the $FF_{Sig}$ did not incorporate any image feature that was learned from scratch from the raw data using deep ETs; the FT components were the closest and as stated previously were aligned with the non-medical TL features. We suggest that positive RRs may appear when deep ETs are directly trained from scratch on the NRG-H CT data.

In addition, from our experiments using the NRG-H dataset, the $FF_{Sig}$ determined RRs with a distinctive collection of GO terms with higher NES when compared to the other image feature signatures. A higher NES of GO terms is typically the result of a stronger correlation between the image feature signatures and the affiliated genes that contribute to the GO term and, RRs with a greater number of affiliated genes that contribute to the GO term. Notably, GO terms with the highest NES consist of a range of biological functions that relate to cellular structures. It has been reported that abnormalities in cellular structures are related to the development of NSCLC [158]. $FF_{Sig}$ has shown to determine RRs with more GO terms when compared with $HC_{Sig}$ and $FT_{Sig}$. A potential explanation for this finding is that the $FF_{Sig}$ determined RRs with a greater number of unique genes. These genes may be affiliated with a greater range of biological functions and therefore provide opportunities for $FF_{Sig}$ to determine RRs with more and unique GO terms. We note that while the $TL_{Sig}$ determined RRs with a

higher number of GO terms, these are generally related to normal human anatomical information rather than the subtle disease processes related to the primary tumour. This finding is evidenced by the most enriched GO terms, such as "Regulation of Anatomical Structure Morphogenesis", as shown in table 4.3.

From our experiments using the NRG-H dataset, $FF_{Sig}$ determined exclusive RRs with a group of GO terms that consist of a range of biological functions that are related to protein kinase activities, such as "Transmembrane Receptor Protein Kinase Activity". Atypical kinase and its activities have been reported previously as an oncogene in NSCLC [159], which play a crucial role in cell growth and tumourigenesis that may be observable in medical images [160]. In contrast, GO terms that are restricted to have RRs with $FF_{Sig}$ include, for example, "Soluble Fraction" and "Enzyme Regulator Activity". A potential explanation is that the specific enzyme activities and fractions cannot be depicted by CT images and hence cannot be quantified by the $FF_{Sig}$.

Our validation experiments on the NRG-S dataset show that the $FF_{Sig}$ comprises 13 image features that are complementary to image features that were selected in the $HC_{sig}$, and $FT_{sig}$. Among the 13 image features, 12 can be traced back to the 6,144-dimensional FT features and the other feature can be traced back to a HC feature. Our results using NRG-S demonstrated that the $FF_{Sig}$ encoded complementary medical imaging visual characteristics. The validation results are consistent with our previous findings from the NRG-H dataset.

However, none of the $FF_{Sig}$, $HC_{Sig}$, nor $FT_{Sig}$ from the NRG-S dataset produced a significant association with the T stage. We attribute our findings to the different scanning parameters used in the NRG-S dataset, for example, slice thickness that ranges from 0.625 to 3mm. Such factors contribute to subtle imaging differences and have potential impacts on the feature extraction process.

In our validation study, $FF_{Sig}$ has determine a greater number of RRs with a greater number of genes when compared with the other image feature signatures. This result validates that the $FF_{Sig}$ is robust and generalisable in encoding the imaging characteristics of the tumour that can reflect the underlying molecular characteristics of NSCLC. However, in our validation study, using the NRG-S dataset, $FF_{Sig}$ did not identify stronger RRs with a range of genes when compared with $HC_{Sig}$ and $FT_{Sig}$. One potential explanation is that the $FF_{Sig}$ did not incorporate any image feature that was fine-tuned on the NRG-S dataset. Despite NRG-S dataset has many similarities to the NRG-H dataset, such as the type of disease, the distribution of patients' clinical parameters and their histopathology status are vastly different to the NRG-H dataset. We suggest that stronger RRs may appear when deep ETs are fine-tuned on the NRG-S dataset.

In our validation experiments, the $FF_{Sig}$ has also shown to determine RRs with a distinctive collection of GO terms with higher NES when compared to the other image feature signatures. Notably, our validation results share a high degree of similarity with our previous findings from experiments using the NRG-H dataset. For example, from both experiments, the proposed $FF_{Sig}$ determined RRs with GO terms such as 'Membrane Enclosed Lumen' and 'Organelle Lumen'. Interestingly, such RRs with GO terms that relate to lumen structures are in opposite statistical direction. We attribute this finding to the differences between the NRG-H and NRG-S datasets where their distribution of T stage parameters and histology sub-types, as they played important roles in the multi-stage feature selection scheme. Such findings further demonstrate the robustness and generalisability of our proposed $FF_{Sig}$ to determine RRs with GOs across different datasets.

Furthermore, in our validation experiments using the NRG-S dataset, $FF_{Sig}$ determined exclusive RRs with a group of GO terms that consist of a range of biological functions that are related to peptidase activity such as 'Endopeptidase Activity'. Previous study has shown that bombesin-like peptides and other neuropeptides are autocrine growth factors for both small

cell lung cancer (SCLC) and NSCLC [161]. Our validation results demonstrate the robustness and generalisability of our proposed $FF_{Sig}$ for determining GO terms that are related to NSCLC.

We recognise that a limitation of our study is the size of the dataset and that lack of knowledge about the patients' mutation status. This limits the ability to optimise deep ETs to quantify image features that are most relevant to the NSCLC. Another limitation of this study is the differences between the train dataset and the independent test dataset. The two datasets use different methods for gene expression profiling, and as such the NRG-H dataset has a greater amount of genetic information compared to the NRG-S dataset. The ideal situation would have been to utilise two datasets that use the same technology for gene expression profiling, but at the time of experimentation and to the best of our knowledge, no such public radiogenomics dataset existed. However, despite these differences we note that the NRG-S dataset shares similarity with the NRG-H dataset, such as the type of disease and histopathology subtypes, and these similarities mean that it is the closest dataset that can be used for independent validation.

The limited availability of the clinical parameters e.g., survival data in the datasets has restricted our study from designing a deep learning-based image feature selection scheme. We note that as more radiogenomics datasets becomes available in the future, a key area for radiogenomics studies is to investigate the feasibility for a data-driven method for image feature selection [162]. Another potential future direction for our study is to investigate deep learning-based gene expression level prediction. Such a deep model can encode imaging characteristics that are reflective towards changes in gene expression levels and therefore may provide more insights into RRs.

## 4.5 Conclusion

In this chapter, a novel framework for radiogenomics analysis was presented that used a selection of image features from handcrafted and deep ETs, named $FF_{Sig}$, to determine the RRs. The results show that the $FF_{Sig}$ encoded complementary medical image visual characteristics when compared with other image feature signatures. The $FF_{Sig}$ determined more RRs with genes and with a group of distinct GO terms. Results show that $FF_{Sig}$ is correlated with a key biomarker for NSCLC and GO terms that are related to tumour developments in NSCLC. Furthermore, the validation experiments demonstrate that the $FF_{Sig}$ is robust and generalisable in different dataset. The $FF_{Sig}$ has demonstrated its potentials to identify important RRs that may facilitate cancer diagnosis and treatment in the future.

# Chapter 5. Radiogenomics with Multi-disease Analysis

This chapter introduces a novel radiogenomics method that utilises multi-disease analysis to derive radiogenomics features associated with specific genetic characteristics from one disease that can be applied to another disease. The proposed framework considers radiogenomics features that encode similar pathophysiology underlying two distinct diseases, providing an alternative approach for radiogenomics analysis when there is a lack of annotated medical image datasets. By leveraging the similarities between diseases, the framework can facilitate the transfer of knowledge and provide unique insights to radiogenomics analysis for diseases with limited available data.

To validate the performance of the proposed radiogenomics method, experiments are conducted using multiple public datasets that contain CT images from patients with Coronavirus disease 2019 (COVID-19) and lung adenocarcinoma (LUAD). COVID-19 is caused by Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2). SARS-CoV-2 enters the body via angiotensin-converting enzyme 2 (ACE2), a membrane-bound aminopeptidase abundantly expressed in the lungs and heart. Altered ACE2 expression due to SARS-CoV-2 infection plays a crucial role in the pathogenesis of COVID-19 and may lead to life-threatening conditions such as acute respiratory distress syndrome (ARDS). Gene expression profiling, however, is not routinely performed for COVID-19 patient management. Comparable to COVID-19, LUAD patients also exhibit altered ACE2 expression and LUAD data are relatively abundant. In this chapter, a radiogenomics method is presented for deriving image features (ACE2-RGF) associated with angiotensin-converting enzyme 2 (ACE2)

expression data from LUAD patients. The ACE2-RGF is utilised as a surrogate biomarker for ACE2 expression, thereby providing an alternative approach for COVID-19 analysis. The ACE2-RGF demonstrated potential for classifying COVID-19 and COVID-19 critical illness identification. These findings provide unique insights for automated COVID-19 analysis and future research.

## 5.1 Contributions

The proposed ACE2-RGF has the following contributions:

- The ACE2-RGF is constructed by identifying radiogenomics features that exhibit statistical associations between ACE2 gene expressions and CT images from both COVID-19 and LUAD patients.
- The ACE2-RGF encodes a distinct set of image features when compared with conventional machine learning-based frameworks.
- The ACE2-RGF can differentiate between COVID-19 and normal subjects, and it can be combined with image features derived using conventional machine learning-based techniques to improve classification performance.
- The ACE2-RGF can also effectively identify COVID-19 patients with critical illness.
- The ACE2-RGF can be used as a biomarker for various applications, as shown for both COVID-19 classification and critical illness identification.

The contribution of this chapter aligns with the aim 2 stated in section 1.3 of Chapter 1 and addresses the gap 3.5.2 identified in Chapter 3. Specifically, this chapter proposes a novel radiogenomics method that utilises limited datasets to improve the accuracy and efficiency of radiogenomics analysis. Moreover, this chapter provides an extensive analysis and comparison

of state-of-the-art radiogenomics analysis pipelines, which builds upon the literature review presented in Chapter 3.

## 5.2 Materials and Methods

### 5.2.1 Materials

We compiled CT scans from multiple public datasets. For LUAD, we used 3 datasets from The Cancer Imaging Archive (TCIA) [163]: i) NSCLC Radiogenomics from Stanford University [164] ('NRG-S'), ii) NSCLC Radiomics-Genomics from Harvard University [165], ('NRG-H'), and, iii) NSCLC Radiomics from Harvard University [165] ('NR-H'). Only NSCLC patients with the LUAD subtype were included. The NRG-S dataset contained scans from 161 patients, 112 also had lung tumour segmentation and 49 had valid ACE2 expression data. One patient was removed from our study due to an exceptionally high ACE2 expression level. The gene expression data were generated with RNA-Seq. The NRG-H dataset comprised CT and gene expression values generated using microarray from 42 patients. There were no corresponding segmentations in the original dataset. We obtained tumour segmentations from an imaging specialist experienced in reading radiology and CT scans. In total, there were 254 LUAD CT scans; 91 also had tumour segmentations and ACE2 expression data. For examples of COVID-19 and normal patients, we used images from the China Consortium of Chest CT Image Investigation (CC-CCII) [166]. We downloaded all available data and 1,496 COVID-19 and 725 normal scans were studied.

## 5.2.2 Overview of the framework

In our framework, image features were extracted from the CT. The ACE2-RGF was determined by using Spearman rank correlation between ACE2 expressions and image features from the NRG-S and NRG-H datasets. ACE2-RGF was used to train a multiple logistic regression (MLR) classifier, which comprised a single fully connected hidden layer, and two output nodes corresponding to each class (e.g., COVID-19 and normal). The MLR classifiers were trained using LUAD images and were evaluated for their performance for COVID-19 classification and critical illness identification. An overview of our framework is outlined in Fig 5.1.

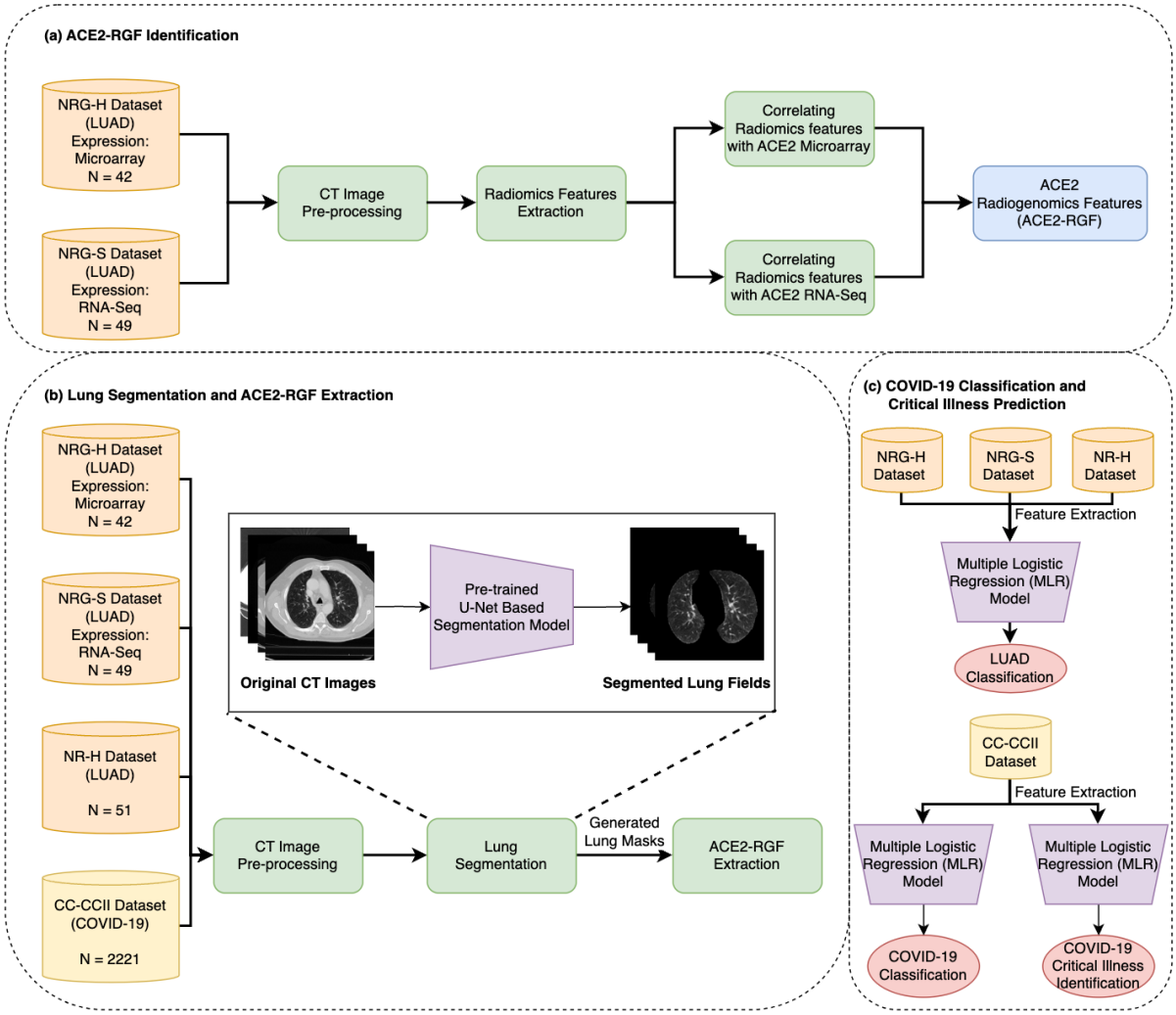


Figure 5.1. Our proposed radiogenomics method. It quantifies and identifies ACE2-RGF to construct a multiple logistic regression for classifying COVID-19 from normal subjects and identify critical illness from mild symptoms.

### 5.2.3 Radiogenomics Feature Extraction and Correlation Analysis

We extracted image features using the widely applied *pyradiomics* [167] Python package from the tumour regions of the images from the NRG-S and the NRG-H datasets, and from the segmented lung regions of all the available scans. A total of 1,288 features relating to shape, first order statistics, and texture were computed per scan volume. Features were extracted from the original images, derived images using Laplacian of Gaussian (LoG) filtering with 5 different sigma levels, and Wavelet decomposition with different combinations of low (denote as 'L') and high-pass (denote as 'H') filters on the X, Y and Z dimensions of the image. Shape features were computed only on the original inputs while all other features were extracted from the original and the derivatives. Shape characteristics included volume, surface area, and length. First order statistics, such as mean, kurtosis, and skewness, described the image intensity histogram. Texture features were quantified by means of grey level cooccurrence matrix (GLCM), grey level run length matrix (GLRLM), grey level size zone matrix (GLSZM), neighbouring grey tone difference matrix (NGTDM), and grey level dependence matrix (GLDM). GLCM [168] describes the spatial relationship between pixels of similar intensities. GLRLM [169] quantifies the length of consecutive pixels with the same intensity. GLSZM [170] depicts texture homogeneity and areas with the same grey-level. NGTDM [171] quantifies the difference between a pixel and its average neighbouring intensities. GLDM [172] represents the connectedness of similar grey-levels.

The extracted image features were associated with ACE2 gene expression using Spearman's rank correlation and assessed for significance and stability across the NRG-S and NRG-H datasets. Image features that were significantly correlated ($p < 0.05$) with ACE2 expressions across both datasets were selected and formed the ACE2-RGF.

### 5.2.4 Experiments

The proposed radiogenomics method was assessed by conducting two sets of experiments: i) ACE2-RGF classifying LUAD/normal and COVID-19/normal and, ii) ACE2-RGF classifying COVID-19/normal subjects, and in identifying critical illness subjects.

First, we derived ACE2-RGF from the NRG-S and NRG-H datasets according to their correlation to ACE2 gene profiles; these features were then used with MLR to measure their ability to classify LUAD/normal and COVID-19/normal subjects. Radiomics features were also extracted from the NRG-S and the NRG-H datasets. A variety of conventional feature selection techniques were employed to determine the best representative features for the tasks, including analysis of variance (ANOVA), mutual information [173], recursive feature elimination (RFE) [174] using a support vector classifier estimator, minimum redundancy maximum relevance (mRMR) [152], ReliefF [175], random forest with 100 estimators and Gini impurity, least absolute shrinkage and selection operator (Lasso) [176], Ridge, and Elastic Net [177] with an L1 ratio of 0.5. The resulting collections of selected image features are denoted as LUAD-RF. For instance, LUAD-RF$_{ANOVA}$ represents radiomics features extracted from LUAD subjects and was processed using the ANOVA feature selection technique. The performance of ACE2-RGF was compared to LUAD-RF and all extracted radiomics features ('LUAD-AF').

Next, the ACE2-RGF was used with MLR to measure its ability to separate COVID-19/normal. For this experiment, radiomics features were extracted from CC-CCII datasets. The same feature selection techniques were applied to the extracted radiomics features and the resulting collection of selected image features were denoted as COVID-19-RF. The performance of ACE2-RGF was compared to COVID-19-RF and all extracted radiomics features ('COVID-19-AF').

Lastly, our ACE2-RGF was used with MLR to measure its ability for identifying COVID-19 critical illness. For this experiment, radiomics features were also extracted from CC-CCII datasets. We followed the same feature selection procedure as for the extracted radiomics features and the resulting collection of selected image features were denoted as COVID-Crt-RF. The performance of ACE2-RGF was compared to COVID-Crt-RF and all extracted radiomics features ('COVID-Crt-AF').

5-fold cross-validation was performed for all experiments. We randomly sampled 250 examples each of LUAD and normal classes (500 in total), and further randomly divided the sample into training and validation sets with an 80/20 split, resulting in 200 examples for training and 50 for validation from each class. Identical patient splits were used for both methods and no subject were in both the training and validation sets of a fold. For the test set, all available COVID-19 patients and control subjects not chosen in the cross-validation sample were included. We evaluated our MLR models using performance metrics including accuracy (ACC), area under the ROC curve (AUC), $F_1$ score, $F_1$ score of only the positive (LUAD/COVID-19) class ($F_1$ POS), precision (PREC), recall (RECA), and specificity (SPEC). We define the best model based on the highest average score between $F_1$ and AUC on the validation set of its fold.

# 5.3 Results

## 5.3.1 ACE2-RGF for Classifying LUAD, COVID-19, and Normal Subjects

The ACE2-RGF had 12 features that were significantly correlated with the expression of the ACE2 gene (Table 5.1). These features were derived from the GLCM, GLRLM, GLSZM, and

GLDM, which are all descriptors of image texture. Eight of the 12 features related to textural "emphasis," which describes the proportion of various grey-level values and zones of varied sizes within an image. Notably, all 12 image features were extracted from the derived images using LoG filtering with sigma levels of 3 and 4.

Table 5.1. ACE2-RGF image features (12 features).

| **Feature Name** | **Origin** |
|---|---|
| log_sigma_3_0_mm_3D_glcm_Autocorrelation | GLCM |
| log_sigma_3_0_mm_3D_glcm_JointAverage | GLCM |
| log_sigma_3_0_mm_3D_glrlm_HighGrayLevelRunEmphasis | GLRLM |
| log_sigma_3_0_mm_3D_gldm_HighGrayLevelEmphasis | GLSZM |
| log_sigma_3_0_mm_3D_gldm_LargeDependenceLowGrayLevelEmphasis | GLSZM |
| log_sigma_4_0_mm_3D_glcm_Autocorrelation | GLCM |
| log_sigma_4_0_mm_3D_glcm_JointAverage | GLCM |
| log_sigma_4_0_mm_3D_glrlm_LongRunLowGrayLevelEmphasis | GLSZM |
| log_sigma_4_0_mm_3D_glrlm_LowGrayLevelRunEmphasis | GLRLM |
| log_sigma_4_0_mm_3D_gldm_HighGrayLevelEmphasis | GLSZM |
| log_sigma_4_0_mm_3D_gldm_LargeDependenceLowGrayLevelEmphasis | GLSZM |
| log_sigma_4_0_mm_3D_gldm_LowGrayLevelEmphasis | GLRLM |

Tables 5.2 and 5.3 compare the performance for LUAD-AF, ACE2-RGF, and LUAD-RF for classifying LUAD from normal subjects and classifying COVID-19 from normal subjects. LUAD-AF and LUAD-RF demonstrated superior performance than ACE2-RGF for classifying LUAD from normal patients. However, MLR classifiers showed substantial decreases in performance when LUAD-AF and LUAD-RF were used as inputs for COVID-19 classification. In contrast, MLR with ACE2-RGF showed consistent performance for classifying LUAD and COVID-19 from normal subjects.

Table 5.2. Performance of the MLR models for classifying LUAD from normal subjects using i) LUAD-AF, ii) ACE2-RGF, and iii) LUAD-RF. LUAD Radiomics features were extracted from the NRG-H and NRG-S datasets. ACE2-RGF was derived and extracted from the NRG-H and NR

| Input | ACC | AUC | $F_1$ | $F_1$ POS | PREC | RECA | SPEC |
|---|---|---|---|---|---|---|---|
| **LUAD-AF** | 0.99 | 1.00 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 |
| **ACE2-RGF** | 0.85 | 0.91 | 0.85 | 0.87 | 0.79 | 0.95 | 0.75 |
| **LUAD-RF$_{ANOVA}$** | 0.96 | 1.00 | 0.96 | 0.96 | 0.95 | 0.97 | 0.95 |
| **LUAD-RF$_{Mutual\ Info}$** | 0.97 | 1.00 | 0.97 | 0.97 | 0.97 | 0.98 | 0.97 |
| **LUAD-RF$_{RFE}$** | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| **LUAD-RF$_{mRMR}$** | 0.98 | 1.00 | 0.98 | 0.98 | 0.98 | 0.98 | 0.98 |
| **LUAD-RF$_{ReliefF}$** | 0.92 | 0.97 | 0.92 | 0.92 | 0.91 | 0.93 | 0.91 |
| **LUAD-RF$_{Forest}$** | 0.97 | 1.00 | 0.97 | 0.97 | 0.98 | 0.97 | 0.98 |
| **LUAD-RF$_{LASSO}$** | 0.99 | 1.00 | 0.99 | 0.99 | 1.00 | 0.99 | 1.00 |
| **LUAD-RF$_{Ridge}$** | 0.99 | 1.00 | 0.99 | 0.99 | 0.99 | 1.00 | 0.99 |
| **LUAD-RF$_{Elastic\ Net}$** | 0.99 | 1.00 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 |

Table 5.3. Performance of the MLR models for classifying COVID-19 from normal subject using i) LUAD-AF, ii) ACE2-RGF, and iii) LUAD-RF. Radiomics features were extracted from the NRG-S and NRG-H datasets. ACE2-RGF was derived and extracted from the NRG-H and NRG-S datasets.

| Input | ACC | AUC | $F_1$ | $F_1$ POS | PREC | RECA | SPEC |
|---|---|---|---|---|---|---|---|
| **LUAD-AF** | 0.28 | 0.70 | 0.25 | 0.09 | 0.99 | 0.05 | 1.00 |
| **ACE2-RGF** | 0.85 | 0.83 | 0.80 | 0.90 | 0.91 | 0.89 | 0.72 |
| **LUAD-RF$_{ANOVA}$** | 0.37 | 0.82 | 0.37 | 0.31 | 0.91 | 0.19 | 0.94 |
| **LUAD-RF$_{Mutual\ Info}$** | 0.38 | 0.83 | 0.37 | 0.31 | 0.90 | 0.20 | 0.94 |
| **LUAD-RF$_{RFE}$** | 0.28 | 0.52 | 0.24 | 0.09 | 1.00 | 0.04 | 1.00 |
| **LUAD-RF$_{mRMR}$** | 0.28 | 0.64 | 0.25 | 0.11 | 0.86 | 0.06 | 0.97 |
| **LUAD-RF$_{ReliefF}$** | 0.65 | 0.81 | 0.63 | 0.70 | 0.96 | 0.56 | 0.92 |
| **LUAD-RF$_{Forest}$** | 0.34 | 0.87 | 0.33 | 0.24 | 0.92 | 0.14 | 0.96 |
| **LUAD-RF$_{LASSO}$** | 0.30 | 0.55 | 0.28 | 0.15 | 1.00 | 0.08 | 1.00 |
| **LUAD-RF$_{Ridge}$** | 0.27 | 0.61 | 0.23 | 0.07 | 1.00 | 0.03 | 1.00 |
| **LUAD-RF$_{Elastic\ Net}$** | 0.28 | 0.69 | 0.24 | 0.09 | 0.98 | 0.05 | 1.00 |

## 5.3.2 MLR for COVID-19 Classification

For COVID-19 classification, radiomics features that were frequently selected by conventional feature selection techniques (Table 5.4) were exclusively derived from decomposed images using 3D wavelet decomposition with LLH filters. Notably, none of these wavelet features overlap to ACE2-RGF.

Table 5.4. Top 12 radiomics features that were frequently selected by conventional image feature selection techniques for COVID-19 classification.

| Feature Name | *Frequency of Selection (%)* |
|---|---|
| wavelet_LLH_firstorder_Maximum | 60.0 |
| wavelet_LLH_firstorder_Range | 54.3 |
| wavelet_LLH_glszm_HighGrayLevelZoneEmphasis | 48.6 |
| wavelet_LLH_glrlm_HighGrayLevelRunEmphasis | 45.7 |
| wavelet_LLH_glszm_SmallAreaHighGrayLevelEmphasis | 42.9 |
| wavelet_LLH_glcm_Autocorrelation | 42.9 |
| wavelet_LLH_gldm_HighGrayLevelEmphasis | 40.0 |
| wavelet_LHH_firstorder_Mean | 37.1 |
| wavelet_LLH_glrlm_ShortRunHighGrayLevelEmphasis | 37.1 |
| wavelet_LLH_gldm_SmallDependenceHighGrayLevelEmphasis | 31.4 |
| wavelet_LLH_glszm_GrayLevelVariance | 31.4 |
| wavelet_LLH_glrlm_LongRunHighGrayLevelEmphasis | 31.4 |

Table 5.5 presents the performance for COVID-19-AF, ACE2-RGF, and COVID-19-RF for classifying COVID-19 from normal subjects. Although ACE2-RGF did not achieve the highest performance for classifying COVID-19, the ACE2-RGF performed comparably or better in AUC, $F_1$ POS, accuracy, and recall when compared to a variety of COVID-19-RF. When ACE2-RGF was fused with COVID-19-RF, several MLR models showed improved performance for COVID-19 classification (Table 5.6). Notably, among the MLR models with improved performance, ACE2-RGF typically improved the $F_1$, $F_1$POS, and precision of those models.

Table 5.5. Performance of the MLR models for classifying COVID-19 from normal subject using i) COVID-19-AF, ii) ACE2-RGF, and iii) COVID-19-RF. COVID-19-AF Radiomics features were extracted from CT images of the CC-CCII dataset. ACE2-RGF was derived from the NRG-H and NRG-S datasets and was extracted from the CC-CCII dataset.

| Input | ACC | AUC | $F_1$ | $F_1$ POS | PREC | RECA | SPEC |
|---|---|---|---|---|---|---|---|
| **COVID-19-AF** | 0.94 | 0.99 | 0.94 | 0.94 | 0.97 | 0.91 | 0.97 |
| **ACE2-RGF** | 0.82 | 0.92 | 0.82 | 0.83 | 0.79 | 0.87 | 0.77 |
| **COVID-19-RF$_{ANOVA}$** | 0.89 | 0.94 | 0.89 | 0.88 | 0.90 | 0.87 | 0.90 |
| **COVID-19-RF$_{Mutual\ Info}$** | 0.88 | 0.95 | 0.88 | 0.88 | 0.89 | 0.88 | 0.89 |
| **COVID-19-RF$_{RFE}$** | 0.94 | 0.98 | 0.94 | 0.94 | 0.95 | 0.94 | 0.95 |
| **COVID-19-RF$_{mRMR}$** | 0.91 | 0.96 | 0.91 | 0.91 | 0.93 | 0.88 | 0.93 |
| **COVID-19-RF$_{ReliefF}$** | 0.64 | 0.69 | 0.63 | 0.64 | 0.63 | 0.66 | 0.61 |
| **COVID-19-RF$_{Forest}$** | 0.92 | 0.97 | 0.92 | 0.92 | 0.94 | 0.91 | 0.94 |
| **COVID-19-RF$_{LASSO}$** | 0.96 | 0.99 | 0.96 | 0.96 | 0.98 | 0.95 | 0.98 |
| **COVID-19-RF$_{Ridge}$** | 0.96 | 0.99 | 0.96 | 0.96 | 0.98 | 0.94 | 0.98 |
| **COVID-19-RF$_{Elastic\ Net}$** | 0.94 | 0.98 | 0.94 | 0.93 | 0.96 | 0.91 | 0.96 |

Table 5.6. Performance of MLR models for classifying COVID-19 subject from normal subjects. ACE2-RGF was fused with COVID-19-RF. COVID-19-RF Radiomics features were extracted from the CC-CCII dataset. ACE2-RGF was derived from the NRG-H and NRG-S datasets and was extracted from the CC-CCII dataset. Numbers in bold indicate improved performance from fusing ACE2-RGF with COVID-19-RF.

| Input | ACC | AUC | $F_1$ | $F_1$ POS | PREC | RECA | SPEC |
|---|---|---|---|---|---|---|---|
| **COVID-19-RF$_{ANOVA}$ + ACE2-RGF** | 0.88 | **0.95** | 0.88 | 0.88 | 0.89 | **0.88** | 0.89 |
| **COVID-19-RF$_{Mutual\ Info}$ + ACE2-RGF** | **0.90** | 0.95 | **0.90** | **0.89** | **0.91** | 0.88 | **0.91** |
| **COVID-19-RF$_{RFE}$ + ACE2-RGF** | **0.95** | 0.98 | **0.95** | **0.95** | **0.96** | 0.94 | **0.96** |
| **COVID-19-RF$_{mRMR}$ + ACE2-RGF** | 0.90 | 0.96 | 0.90 | 0.90 | 0.92 | 0.87 | 0.92 |
| **COVID-19-RF$_{ReliefF}$ + ACE2-RGF** | **0.91** | **0.96** | **0.91** | **0.91** | **0.92** | **0.90** | **0.92** |
| **COVID-19-RF$_{Forest}$ + ACE2-RGF** | 0.86 | 0.93 | 0.86 | 0.86 | 0.86 | 0.86 | 0.86 |

### 5.3.3 MLR for COVID-19 Critical Illness Identification

For COVID-19 critical illness identification, image features commonly selected using conventional feature selection techniques (Table 5.7) were derived from log and wavelet filters. Notably, none of these wavelet features overlapped ACE2-RGF. Table 5.8 presents the performance for COVID-Crt-AF, ACE2-RGF, and COVID-Crt-RF for identifying COVID-19 critical illness. Although ACE2-RGF did not achieve the greatest performance for COVID-19 critical illness identification, the gap between the top performing models and ACE2-RGF was within 5% in AUC.

Table 5.7. Top 12 radiomics features that were frequently selected by conventional image feature selection techniques for COVID-19 critical illness identification.

| Feature Name | *Frequency of Selection (%)* |
|---|---|
| wavelet_LLL_glcm_Correlation | 42.9 |
| wavelet_HLH_glcm_Idn | 37.1 |
| wavelet_HLL_firstorder_Kurtosis | 28.6 |
| wavelet_LLL_glcm_Idmn | 22.9 |
| wavelet_LLH_gldm_SmallDependenceLowGrayLevelEmphasis | 20.0 |
| wavelet_HLH_glcm_Idmn | 17.1 |
| wavelet_HLH_firstorder_Kurtosis | 17.1 |
| wavelet_LLH_glcm_JointAverage | 17.1 |
| log_sigma_4_0_mm_3D_gldm_SmallDependenceEmphasis | 17.1 |
| log_sigma_1_0_mm_3D_glcm_Idmn | 17.1 |
| wavelet_LLL_glcm_Imc2 | 17.1 |
| wavelet_HLL_glcm_Idn | 14.3 |

Table 5.8. Performance of MLR models for identifying COVID-19 critical illness using various feature selection methods. COVID-Crt-AF radiomics features were extracted from the CC-CCII dataset. ACE2-RGF was derived from the NRG-H and NRG-S datasets and was extracted from the CC-CCII dataset.

| Input | ACC | AUC | $F_1$ | $F_1$ POS | PREC | RECA | SPEC |
|---|---|---|---|---|---|---|---|
| **COVID-Crt-AF** | 0.81 | 0.88 | 0.80 | 0.78 | 0.75 | 0.82 | 0.80 |
| **ACE2-RGF** | 0.77 | 0.85 | 0.77 | 0.73 | 0.73 | 0.74 | 0.80 |
| **COVID-Crt-RF$_{ANOVA}$** | 0.81 | 0.89 | 0.80 | 0.76 | 0.80 | 0.73 | 0.87 |
| **COVID-Crt-RF$_{Mutual\ Info}$** | 0.81 | 0.88 | 0.80 | 0.76 | 0.79 | 0.74 | 0.86 |
| **COVID-Crt-RF$_{RFE}$** | 0.80 | 0.88 | 0.79 | 0.75 | 0.77 | 0.73 | 0.84 |
| **COVID-Crt-RF$_{mRMR}$** | 0.84 | 0.89 | 0.84 | 0.81 | 0.84 | 0.78 | 0.89 |
| **COVID-Crt-RF$_{ReliefF}$** | 0.48 | 0.46 | 0.45 | 0.34 | 0.38 | 0.33 | 0.60 |
| **COVID-Crt-RF$_{Forest}$** | 0.79 | 0.86 | 0.79 | 0.75 | 0.78 | 0.73 | 0.84 |
| **COVID-Crt-RF$_{LASSO}$** | 0.79 | 0.87 | 0.79 | 0.76 | 0.75 | 0.77 | 0.81 |
| **COVID-Crt-RF$_{Ridge}$** | 0.77 | 0.84 | 0.76 | 0.73 | 0.73 | 0.73 | 0.79 |
| **COVID-Crt-RF$_{Elastic\ Net}$** | 0.81 | 0.89 | 0.81 | 0.79 | 0.76 | 0.82 | 0.81 |

## 5.4 Discussion

Our main findings are that our framework can: i) encode ACE2-RGF imaging biomarkers using LUAD data, which are distinct to radiomics features extracted for COVID-19 classification and critical illness identification; ii) the ACE2-RGF can distinguish COVID-19 from normal subjects, and can be combined with COVID-19 RF to improve classification performance; iii) the ACE2-RGF can also effectively identify COVID-19 patients with critical illness and, iv) the ACE2-RGF can be used as a biomarker for various applications, as shown for both COVID-19 classification and critical illness identification.

The ACE2-RGF comprises 12 radiomics features (Table 5.1) that encodes textural information in CT images. Notably, none of the ACE2-RGF features were among the most

frequently selected features when compared with COVID-19-RF (Table 5.4) and COVID-Crt-RF (Table 5.7). The ACE2-RGF encoded texture descriptors are a 2D isotropic quantification of the second spatial derivative of an image, and they identify locations with rapid intensity changes within the CT image. Such ACE2-RGF encoded textural information were consistent to the CT findings reported in ARDS and COVID-19 [178, 179], including ground glass opacity, vascular enlargement and crazy-paving pattern. In contrast, the COVID-19-RF encoded statistical and texture features from decomposed images using 3D wavelet decomposition with LLH filters. In comparison, COVID-Crt-RF encoded a distinct collection of image features that were derived from decomposed images using a variety of low and high-pass filters, including LLL, LLH, HLL, and HLH filters and LoG filtered image with Gaussian sigma values at 1 and 4 mm. Our findings indicate that the presented radiogenomics method enabled the derivation of image features associated with ACE2 and encoded unique features regarding disease manifestation related to variations in ACE2 expression. In contrast, conventional machine learning-based approaches quantify and select image features that are optimized for particular tasks, thus may neglect important imaging representations related to the pathophysiology of the disease. This is owing to the possibility for multiple 'optimal' feature sets to be selected for a particular task, despite different feature sets may offer distinct information [180, 181].

When compared to LUAD-AF and LUAD-RF variants, our radiogenomics method derived ACE2-RGF demonstrated consistent performance for classifying LUAD (Table 5.2) and COVID-19 (Table 5.3) patients from normal subjects. MLR models using LUAD-AF and LUAD-RF demonstrated a substantial decline in performance for classifying COVID-19 patients from normal subjects. Our results show that our framework derived ACE2-RGF encoded imaging representations of pathophysiology information that are common to LUAD and COVID-19. Despite the ACE2-RGF having inferior performance when compared with

COVID-19-RF for separating COVID-19 patients from normal subjects (Table 5.5), the use of ACE2-RGF did not require identifying and extracting COVID-19-RF features. Our findings indicate that the ACE2-RGF encoded imaging representations are associated with alterations in ACE2 expression and are relevant to the pathophysiology of both LUAD and COVID-19. However, such information may not provide the optimal classification value that is specific to both LUAD and COVID-19.

Notably, MLR models trained with COVID-19-AF performed similarly to MLR models trained with multiple COVID-19-RF in classifying COVID-19 patients from healthy subjects (Table 5.2). Our findings suggest that despite radiomics features (COVID-19-AF) may encode distinctive information, these features have demonstrated their capability to classify COVID-19 when used collectively. In contrast, the conventional machine learning frameworks that quantify task-specific image features may neglect radiomics features that encode relevant information for classifying COVID-19, such as statistical and textural features using various LoG filters.

The classification performance for COVID-19 was enhanced when ACE2-RGF was fused with COVID-19-RF (Table 5.6). In contrast to COVID-19-RF, ACE2-RGF encoded distinct pathophysiological image features linked with COVID-19, and therefore is complementary to COVID-19-RF. Our results suggest that the conventional machine learning frameworks that quantify task-specific image features may neglect the underlying pathophysiology information of COVID-19 and its clinical manifestation due to altered ACE2 expression. For instance, the involvement of the lower respiratory tract in individuals with early-stage or moderate COVID-19 and the possibility of ARDS progression [182].

Our framework showed it could identify COVID-19 patients with critical illness. The performance of the MLR model trained with ACE2-RGF for identifying COVID-19 critical illness was similarly to that of models trained with COVID-Crt-RF (Table 5.8). Our findings

suggest that the ACE2-RGF may not contain imaging representations exclusive to COVID-19 critical illness status, but rather imaging characteristics associated with ACE2 expression alterations that are tied with the progression of COVID-19 critical illness [183]. Notably, the performance gap between ACE2-RGF and the best performing COVID-Crt-RF for identifying COVID-19 critical illness was less than the gap between ACE2-RGF and the best performing COVID-19-RF for COVID-19 classification. One explanation of our finding is that patients with COVID-19 critical illness commonly have multiple complications that are related or results of ACE2 and RAAS failure, such as ARDS [184, 185].

Our framework demonstrated potential to serve as an imaging biomarker for COVID-19 classification and COVID-19 critical illness identification using the same set of ACE2-RGF. We attribute this to the encoding of altered ACE2 expression in ACE2-RGF. Recent research has implicated the role of ACE2 in the infection, development, and clinical manifestations of COVID in the human body [186]. It is also suggested that ACE2 and its variants affect the binding of SARS-COV2 virus and hence the disease severity following COVID-19 infection [187]. Therefore, our framework has the potential to serve as a valuable biomarker that complements existing image-based frameworks and offer new research possibilities to derive additional features for future automated COVID-19 classification and critical illness identification.

In our study, we selected the use of traditional handcrafted image features relating to shape, first order statistics, and texture. These features are the routinely used to study radiogenomics due to its wide acceptability, comprehension and for its explainability. With the modern advances in deep learning feature extractors, it is able to extract complementary feature set to the handcrafted set, for example, in the recent work Xia et al. [10] in the analysis of lung cancer radiogenomics, they demonstrated that deep learning features produced unique features from the traditional set, but these features were not explainable e.g., they were not descriptive

or characterized. In this study, our focused is on analysing the ability to encode ACE2 features from CT images and offer explanation, and the traditional handcrafted feature sets were sufficient in this requirement. Our future work will investigate if deep learning features can complement our study.

A limitation of our study is the lack of ACE2 expression for the COVID-19 patients. This limits the ability to optimize the ACE2-RGF for COVID-19 classification and critical illness. We anticipate that with ACE2 expression data of COVID-19 patients, our model can be improved by identifying and selecting ACE2-RGF directly on COVID-19 imaging data.

## 5.5 Conclusion

This chapter introduces a novel radiogenomics method that utilises multi-disease analysis to derive radiogenomics features associated with similar pathophysiology underlying two distinct diseases. This approach provides an alternative method for radiogenomics analysis in the absence of annotated medical image datasets. The proposed framework has the potential to derive biomarkers for various applications, as demonstrated by its use for COVID-19 classification and critical illness identification. By leveraging the similarities between diseases, this framework can facilitate the transfer of knowledge and improve the accuracy and efficiency of radiogenomics analysis.

# Chapter 6. Radiogenomics with Tumour Heterogeneity and Multi-modal Imaging Data

This chapter introduces a novel radiogenomics method that captures radiogenomics features from multiple heterogeneous regions of the tumour and integrates radiogenomics features from multi-modal imaging data. The proposed framework leverages the complementary information encoded in distinct tumour heterogeneity regions and in distinct imaging modalities to facilitate radiogenomics analysis on human cancers.

To validate the proposed radiogenomics method's performance, experiments were conducted using a publicly available dataset that contained multi-modal MR images from patients with GBM. GBM patients typically receive treatment involving surgical resection, radiotherapy, and chemotherapy with alkylating agents such as temozolomide. The MGMT gene plays a crucial role in repairing DNA damage caused by alkylating agents, and its expression is regulated by a promoter region. Methylation of the MGMT promoter region leads to altered activity of the MGMT enzyme, which is relevant to GBM treatment response. Previous research has shown the potential of deep learning-based radiogenomics approaches for identifying diseases and predicting their prognosis and treatment response. However, these approaches typically focus on learning imaging features that depict the entire tumour and use a single MR sequence, such as T1, despite GBM tumours exhibiting strong heterogeneity and complementary information encoded in different MR sequences.

The experimental results demonstrate the effectiveness of the proposed framework in capturing radiogenomics features from multiple tumour regions and integrating multi-modal imaging data for improved radiogenomics analysis.

## 6.1 Contributions

The proposed radiogenomics method has the following contributions:

- The proposed radiogenomics method utilises multi-modal imaging data, and accounting for tumour heterogeneity sub-regions to achieve end-to-end classification.
- The proposed radiogenomics method can effectively extract, fuse, and learn complementary radiogenomics features that contribute to the identification of the relevant radiogenomics features for radiogenomics classification task.

The contribution of this chapter aligns with the aim 3 stated in section 1.3 of Chapter 1 and addresses the gap 3.5.3 identified in Chapter 3. The proposed radiogenomics method utilises multi-modal imaging data of diseases with complex tumour heterogeneity to improve radiogenomics analysis. This chapter provides an extensive analysis and comparison of state-of-the-art radiogenomics analysis pipelines, which builds upon the literature review presented in Chapter 3.

## 6.2 Materials and Method

### 6.2.1 Materials

MR sequences from the publicly available RSNA-ASNR-MICCAI BraTS 2021 challenge dataset is obtained for this study (hereby denote as 'BraTS21'). This dataset consists of two unique subsets prepared for specific tasks: the first subset is intended for segmentation analysis (also known as task 1), while the second subset is designed for MGMT promoter methylation classification (also known as task 2).

The segmentation dataset consists of 1251 subjects' 3D MR scans that is accessible in T1W, T1CE, T2W, and FLAIR sequences. Tumour sub-regions were generated by using an interactive segmentation pipeline that employs STAPLE [188] fusion of the best performing BraTS algorithms from prior competitions, specifically nnU-Net [189], DeepScan [190], and DeepMedic [191]. The results were then manually refined by neuroradiology specialists and validated by board-certified attending neuroradiologists with over 15 years of experience working with gliomas. The segmented tumour sub-regions comprise the following: i) enhancing tumour (ET), characterised by regions with both visibly avid and faint enhancement, ii) peritumoral oedematous/invaded tissue (ED), which is characterized by an abnormal hyperintense signal that encompasses the non-enhancing infiltrative tumour and includes vasogenic oedema in the peritumoral region, and iii) the necrotic tumour core (NCR).

The MGMT Promoter Methylation Dataset is comprised of 585 subjects with confirmed MGMT promoter region methylation status determined using pyrosequencing and next generation quantitative bisulfite sequencing of promoter CpG sites. Among the 585 subjects, 574 patients were found to be part of the segmentation dataset. In addition, three subjects were removed by the BraTS2021 challenge authority. Each patient's MR imaging data

is pre-processed and is accessible in T1W, T1CE, T2W, and FLAIR sequences with a dimension of 240×240×155.

### 6.2.2 Image pre-processing

The raw imaging data from the BraTS21 Dataset underwent a series of pre-processing steps. Firstly, the imaging data was rotated into the same plane and resized to have the same pixel spacing of 1×1×1 mm in the 3D dimension, based on the metadata retrieved from the DICOM header. Subsequently, for each patient, all sequences were co-registered using the SimpleITK image registration tool.

The T2W sequence served as the reference volume for image co-registration. Our co-registration process involved the utilisation of the Euler3DTransform, the Mattes-Mutual-Information cost function, the Gradient Descent optimiser, and linear interpolation. The registered data were rescaled to a dimension of 240 × 240 × 155 to match the pre-processed imaging data provided in the segmentation dataset. The average time for pre-processing was approximately 26 seconds per volume.

Subsequently, the segmentation masks of the tumour heterogeneity sub-regions underwent further processing to differentiate the following regions: the Whole Tumour (WT), the Enhancing Tumour (ET), and the Tumour Core (TC). WT encompasses the complete extent of the brain tumour, the ET subregion is represented by the mask sections identified as enhancing tumour, and the TC comprises the combined regions of both ET and Necrosis (NCR). Lastly, the pre-processed image volumes were rescaled to 100 x 100 x 100 to reduce GPU memory usage.

### 6.2.3 Deep Multi-sequence Multi-region Classification Model (DeepMMC)

Figure 6.1 depicts an overview of the proposed DeepMMC, which is designed to facilitate the classification of MGMT promoter region methylation status by extracting, fusing, and learning complementary imaging features from distinct tumour heterogeneity sub-regions and multiple MR sequences. DeepMMC splits MR sequences and processes each sequence independently using MRF modules that leverage 3D CNN backbones. Within the MRF module, deep radiogenomics features from each tumour heterogeneity sub-region are first extracted and processed separately and then fused with those deep radiogenomics features processed from other sequences.

Subsequently, the fused deep radiogenomics features are processed using the MSF module, which employs explicitly designed 2D CNN layers. The MSF module first investigates the relationships between deep radiogenomics features that were processed from MR sequences and then identifies the deep radiogenomics from various tumour heterogeneity sub-regions that are most relevant to the classification of MGMT promoter region methylation status.

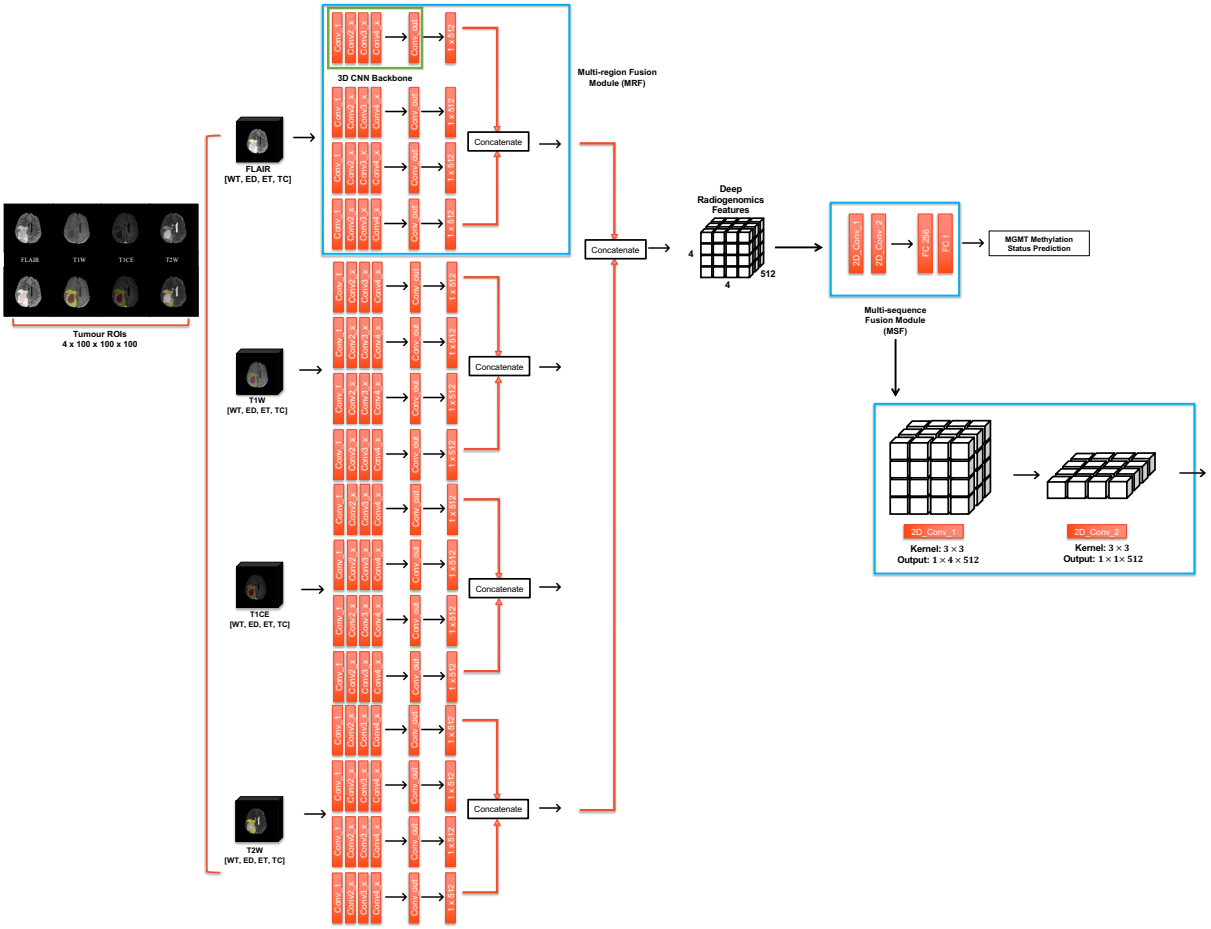


Figure 6.1. Overview of the proposed DeepMMC model. DeepMMC extracts, fuses, and learns complementary imaging features from distinct tumour heterogeneity sub-regions and multiple MR sequences using the MRF module. The output deep radiogenomics features were then processed using the MSF module to classify MGMT promoter region methylation status.

### 6.2.3.1 Multi-region Fusion Module (MRF)

The MRF module utilises a customised 3D CNN backbone that draws inspiration from the inception module [192]. The inception module has been applied to a variety of medical image analysis tasks, including the detection of COVID-19 and the classification of breast cancer [193, 194]. Specifically, the 3D CNN backbone consists of four branches of 3D CNN blocks, where each block comprises a 3D CNN layer, a 3D batch normalisation layer, a leaky Rectified Linear Unit (ReLU) layer, and a 3D max pooling layer. To prevent overfitting, we implemented

a Conv_out block immediately after the 3D CNN backbone, which reduces the number of features. The Conv_out block consists of a single 3D CNN layer with a kernel size of 1 × 1 × 1, a 3D batch normalization layer, a leaky ReLU layer, and a 3D adaptive max pooling layer.

In the MRF module, deep radiogenomics features are extracted from each tumour heterogeneity sub-region using independent 3D CNN backbones. These features are then processed and transformed into latent radiogenomics features. A total of 512 latent radiogenomics features are derived from each 3D CNN backbone. The latent radiogenomics feature from each tumour heterogeneity sub-region are fused together using concatenation, resulting in a feature vector with a size of 4 × 512. In a similar fashion, feature vectors that have been processed from each MR sequence are also fused together using concatenation to generate a three-dimensional fused feature vector with a size of 4 × 4 × 512.

### 6.2.3.2 Multi-sequence Fusion Module (MSF)

The MSF module consists of two explicitly designed 2D CNN blocks. Each CNN block comprises a 2D CNN layer, a 2D batch normalization layer, a leaky ReLU layer, and a 2D max pooling layer. The first CNN block (2D_Conv_1) is designed to investigate the relationship between latent radiogenomics features derived from MR sequences. This is accomplished by utilising a 2D CNN layer with a kernel size of 3 x 3 that processes latent radiogenomics features across the channel to which the fused feature vector was concatenated. The output from the 2D_Conv_1 is then fed into the second 2D CNN block (2D_Conv_2) to investigate the relationships between processed latent radiogenomics features and their corresponding tumour heterogeneity sub-regions. In the final step, a total of 512 deep radiogenomics features are derived from the MSF module. These features are then fed into the classification block, which is composed of two fully connected layers. Between these layers, we have inserted a single

leaky ReLU layer and a dropout layer with a probability of an element to be zeroed set at the default value of 0.5.

### 6.2.4 Experimental Setup

The proposed DeepMMC was assessed by conducting experiments of MGMT promoter region methylation status classification using the BraTS21 dataset. We firstly performed ablation studies to determine the contribution of the proposed MRF and MSF modules. Specifically, the classification performance from the following approaches were compared: i) DeepMMC that employ both MRF and MSF modules (denote as 'experimental setting A'), ii) DeepMMC models that only employ MRF modules using four MR sequences and four tumour heterogeneity sub-regions (denote as 'experimental setting B'), and iii) DeepMMC models that only employ MRF modules using selected MR sequences and tumour heterogeneity sub-regions (denote as 'experimental setting C').

We also compared the DeepMMC model with three recent works for classifying MGMT promoter region methylation status. These works are: i) a 3D Densenet-based approach that utilises image volumes of the whole brain from the T2W MR sequence [136], ii) the winning algorithm for the RSNA-MICCAI Brain Tumour Radiogenomics Classification challenge that uses image volumes of the whole brain from four MR sequences (T1W, T1CE, T2W and FLAIR), and iii) a deep learning pipeline that uses segmented tumour ROI from T1CE and FLAIR sequences for classifying MGMT promoter region methylation status [195].

The proposed DeepMMC model was implemented using the PyTorch framework (version 1.12.1) on a Linux operating system (Ubuntu, version 20.04) and trained on an NVIDIA GeForce 3090 GPU with 24GB of memory. The DeepMMC was trained from scratch for a maximum of 150 epochs with a batch size of 4. Weight decay coefficient was set at $1 \times$

$10^{-6}$. We utilised the Adam optimizer with a learning rate of 0.0001 to optimize the Binary Cross Entropy (BCE) with logits loss. The training process was configured to terminate if the BCE loss did not reduce in the previous 10 epochs.

For all experiments, we conducted a 5-fold cross-validation. The imaging and MGMT data from the 574 patients were partitioned into five different sets of training and testing datasets with an 80/20 split. We evaluated the performance of our MLR models using several performance metrics, including accuracy (ACC), area under the ROC curve (AUC), F1 score, precision, and recall. We defined the best model based on the highest average score between ACC and AUC on the testing set of its respective fold.

# 6.3 Results

## 6.3.1 Ablation study

Table 6.1. Performance of the DeepMMC model for classifying MGMT promoter region methylation status using i) MRF module that leverage selected MR sequence and tumour heterogeneity sub-regions, ii) MRF module that leverage all four MR sequences and tumour heterogeneity sub-regions available in the BraTS21 dataset, and iii) MRF and MSF modules that leverage all four MR sequences and tumour heterogeneity sub-regions.

| **Model** | **MR Sequence** | **Tumour Sub-regions** | **Accuracy** | **AUC** | **$F_1$** | **Precision** | **Recall** |
|---|---|---|---|---|---|---|---|
| MRF module | Four sequences | WT | 0.540 | 0.546 | 0.539 | 0.531 | 0.656 |
| | Four sequences | ED | 0.538 | 0.548 | 0.596 | 0.545 | **0.685** |
| | Four sequences | ET | 0.582 | 0.587 | 0.519 | **0.673** | 0.566 |
| | Four sequences | NCR | 0.571 | 0.568 | 0.511 | 0.474 | 0.590 |
| | FLAIR | Four regions | 0.585 | 0.582 | 0.599 | 0.577 | 0.648 |
| | T1W | Four regions | 0.545 | 0.554 | 0.571 | 0.543 | 0.654 |
| | T1CE | Four regions | 0.570 | 0.554 | 0.486 | 0.567 | 0.513 |
| | T2W | Four regions | 0.528 | 0.547 | 0.507 | 0.539 | 0.484 |
| MRF module | Four sequences | Four regions | 0.516 | 0.548 | 0.541 | 0.544 | 0.563 |
| DeepMMC (MRF and MSF) | Four sequences | Four regions | **0.587** | **0.600** | **0.615** | 0.590 | 0.646 |

Table 6.1 presents the results of the ablation study conducted on the proposed DeepMMC model for classifying MGMT promoter region methylation status using various combinations of MRF modules, MSF modules, MR sequences, and tumour heterogeneity sub-regions. Our findings indicate that the DeepMMC model, which incorporated both MRF and MSF modules, achieved the highest classification performance metrics, including an ACC of 0.587, an AUC of 0.6, and an F1 Score of 0.615. Additionally, the MRF models that utilised the FLAIR sequence and all four tumour heterogeneity regions yielded the second-highest ACC of 0.585 and the second highest F1 score of 0.599. The MRF models that utilised all four MR sequences and the ET region yielded the second-highest AUC of 0.587 and the highest precision of 0.673. Finally, the MRF models that utilised all four MR sequences and the ET region yielded the highest recall of 0.685.

Furthermore, our results show that the MRF module that utilised all four MR sequences and tumour heterogeneity sub-regions did not improve the classification performance metrics compared to the MRF modules that utilised selected MR sequences and tumour heterogeneity sub-regions. Notably, the performance differences between DeepMMC and MRF models that utilised selected MR sequences and tumour heterogeneity sub-regions were smaller than the differences between DeepMMC and the MRF module that utilised all four MR sequences and tumour heterogeneity sub-regions.

### 6.3.2 Comparison to the State-of-the-art

Table 6.2. MGMT promoter region methylation status classification performance between the proposed DeepMMC and existing classification methods.

| Model | MR Sequence | Tumour Sub-regions | Accuracy | AUC | $F_1$ | Precision | Recall |
|---|---|---|---|---|---|---|---|
| DeepMMC | Four sequences | Four regions | **0.587** | 0.600 | 0.615 | 0.590 | **0.646** |
| Faghani et al | T2W | Whole brain (WB) | 0.472 | 0.500 | 0.531 | **0.800** | 0.397 |
| Kaggle #1 | Four sequences | Whole brain (WB) | 0.530 | **0.602** | **0.630** | 0.776 | 0.554 |
| Chen et al | T1CE | WT | 0.498 | 0.558 | 0.420 | 0.487 | 0.435 |
| | FLAIR | WT | 0.514 | 0.539 | 0.257 | 0.303 | 0.224 |

Table 6.2 presents the performance of MGMT promoter region methylation status classification between the proposed DeepMMC and three existing classification methods. The DeepMMC achieved the highest ACC of 0.587 and also outperformed other methods in recall of 0.646. In contrast, the winning algorithm from the RSNA-MICCAI Brain Tumour Radiogenomics Classification challenge that leveraged imaging data from all four MR sequences achieved the highest AUC of 0.602 and F1 score of 0.63. The highest precision of 0.8 was achieved by the 3D Densenet-based approach that utilised image volumes of the whole brain from the T2W MR sequence. The differences between the highest and second-highest ACC, AUC, $F_1$ score, precision and recall were 0.057, 0.002, 0.015, 0.004, and 0.092, respectively. Notably, methods that leveraged all four MR sequences outperformed those that utilised a single MR sequence in every evaluation metric except precision.

## 6.4 Discussion

Our two main findings are that: i) MR sequences and tumour heterogeneity sub-regions encode distinct imaging features that collectively can reflect MGMT promotor region methylation

status, ii) the proposed DeepMMC that utilises MRF and MSF modules can effectively extract, fuse, and learn these complementary deep radiogenomics features, ultimately enhancing the classification performance of MGMT promoter region methylation status.

### 6.4.1 Ablation Study Analysis

Our ablation study results demonstrate that the proposed DeepMMC model utilising experimental setting A exhibited improved performance for classifying MGMT promoter region methylation status when compared to DeepMMC models using experimental setting B and C. Our findings suggest that distinct tumour heterogeneity sub-regions encode complementary information that may enhance the classification performance, as evidenced by our results showing that DeepMMC with experimental setting C that utilises the ED and ET regions yielded the highest precision and recall, respectively. Similar findings have been reported in published works. For instance, Carrillo et al found that MGMT promoter methylation status is associated with tumour ED regions and patient survival time [196]. Our results suggest that the proposed MRF and MSF modules are effective in extracting and learning relevant deep radiogenomics features for the classification of MGMT promoter region methylation status. This is evidenced by the performance differences between the DeepMMC models using experimental settings B and C, where the inclusion of the MSF module in experimental setting C resulted in improved ACC, AUC, and $F_1$ score compared to experimental setting B. We attribute this finding to the design of the proposed MRF and MSF modules, where deep radiogenomics features that are most relevant to the MGMT promoter region methylation status were identified using the explicitly designed 3D CNN back bones and 2D CNN blocks.

### 6.4.2 Comparison to the State-of-the-Art

Our proposed DeepMMC model utilises a novel approach that incorporates MR sequences and tumour heterogeneity sub-regions for end-to-end classification of MGMT promoter region methylation status. Previous deep learning-based methods have combined MR sequences and tumour heterogeneity sub-regions through techniques such as majority voting based on predicted probabilities obtained from using different MR sequences as input. In addition, existing methods typically utilise segmented tumour ROIs or the entire brain for the classification task. In contrast, the DeepMMC explicitly exploits the intrinsic relationships between MR sequences and tumour heterogeneity sub-regions and learns the relevant deep radiogenomics features from each MR sequence and tumour heterogeneity sub-region. Our results (Table 2) demonstrate that the proposed DeepMMC outperforms alternative approaches in terms of classification accuracy and recall. While the DeepMMC exhibited inferior performance in terms of AUC and $F_1$ score, the differences in performance between the DeepMMC and alternative approaches are marginal, with all metrics differing by less than 2%. However, the precision of DeepMMC was lower compared to alternative approaches. It is important to note that these approaches are prone to over-representing one class, which result in a high precision score but a lower recall score. Our findings can be attributed to both the design of the DeepMMC and the MRF and MSF modules, which enable the utilisation of distinct MR sequences and tumour heterogeneity regions for the classification of MGMT promoter region methylation status. The proposed DeepMMC model demonstrated the ability to learn complementary deep radiogenomics features from MR sequences and to identify the most relevant features for MGMT promoter region methylation status classification through an end-to-end training strategy.

We acknowledge a limitation in our study with regards to computational power and the lack of comprehensive gene expression profiles from the dataset. Due to the significant

computational demand for processing 3D imaging volumes, we were unable to implement a deep 3D CNN backbone in the proposed MRF module, which may limit the module's ability to learn, quantify, and fuse optimal image features most relevant to tumour heterogeneity. Moreover, the lack of comprehensive gene expression profiles from patients limits the proposed DeepMMC's ability to learn the relationships between various genetic information from different tumour heterogeneity sub-regions, which may contribute to identifying the optimal deep radiogenomics features that facilitate the classification of MGMT promoter region methylation status.

## 6.5 Conclusion

This chapter introduces a novel radiogenomics method that utilises multi-modal MR imaging data, including T1W, T1CE, T2W, and FLAIR sequences, and accounts for tumour heterogeneity sub-regions to achieve end-to-end classification. The proposed radiogenomics method demonstrated effective extraction, fusion, and learning of complementary radiogenomics features, ultimately enhancing the classification performance of MGMT promoter region methylation status. Experimental results indicate that the proposed radiogenomics method has the potential to complement existing image analysis algorithms by leveraging information from distinct multi-modal imaging data and tumour heterogeneity information, offering unique and additional insights for radiogenomics analysis and future research.

# Chapter 7. Conclusions and Future Works

## 7.1 Conclusions

Radiogenomics is a rapidly developing field that aims to identify and explore the associations between imaging and genetic information. This approach seeks to identify imaging surrogates for genetic biomarkers, with the potential to enhance existing medical image analysis techniques and optimise treatment planning for cancer patients. This research addressed several challenges and limitations in the field of radiogenomics to improve radiogenomics analysis.

In this thesis, a novel radiogenomics framework that combines several novel methods were developed and validated with the following contributions:

- A radiogenomics method that derives fused feature signature ($FF_{Sig}$) that leverages both handcrafted and deep image features to derive radiogenomics relationships with gene expression and GO terms;
- A radiogenomics method that utilises multi-disease analysis to derive radiogenomics features associated with similar pathophysiology underlying two distinct diseases;
- A radiogenomics method that utilises multi-modal imaging data, and accounting for tumour heterogeneity sub-regions for medical image analysis.

Chapter 2 of the thesis introduced the fundamental concepts of medical images, machine learning, and various medical image processing techniques. Chapter 3 discussed the basics of radiogenomics and identifies the gaps in the current research.

Chapter 4 presented a novel radiogenomics method that identifies $FF_{Sig}$, a set of image features derived from using both handcrafted and deep learning-based image feature extraction techniques, to identify radiogenomics relationships. Previous radiogenomics studies primarily relied on a single category of image feature extraction techniques, despite previous studies demonstrating that handcrafted and deep features encode complementary information. The proposed radiogenomics method and the $FF_{Sig}$ identified more relationships with genes and a distinct group of GO terms and correlated with a key biomarker for NSCLC and GO terms related to tumour development in NSCLC. Furthermore, validation experiments demonstrated that the radiogenomics method is robust and generalisable across different datasets. The proposed framework shows potential for identifying critical radiogenomics relationships that may improve cancer diagnosis and treatment in the future.

In chapter 5, a novel radiogenomics method that used multi-disease analysis to derive radiogenomics features associated with specific genetic characteristics from one disease that can be applied to another disease. The proposed framework leverages radiogenomics features that encode similar pathophysiology underlying two distinct diseases, providing an alternative approach for radiogenomics analysis when there is a lack of annotated medical image datasets. This framework is validated using several publically availble datasets that contain CT images from COVID-19 and LUAD patients. The proposed radiogenomics method demonstrated potential for classifying COVID-19 and COVID-19 critical illness identification. In addition, the framework demonstrates potential to facilitate the transfer of knowledge and provide unique insights to radiogenomics analysis for diseases with limited available data.

Chapter 6 introduced a novel radiogenomics method that captures radiogenomics features from multiple heterogeneous regions of the tumour and integrates radiogenomics features from multi-modal imaging data. The proposed framework leverages the complementary information encoded in distinct tumour heterogeneity regions and in distinct imaging modalities to facilitate

radiogenomics analysis on human cancers. Experimental results show that the proposed radiogenomics method can effectively extract, fuse, and learn complementary radiogenomics features that contribute to the identification of the relevant radiogenomics features for radiogenomics classification task.

## 7.2 Future Works

This thesis identifies several research directions that can further the development of radiogenomics analysis. The significant contributions of this thesis rely on machine learning and deep learning-based approaches for medical image processing. However, deep learning-based genetic information processing is an emerging field that enables the extraction of knowledge in addition to human knowledge. The opportunity for an end-to-end approach for deep learning-based methods to automatically extract and identify important information from both imaging and genetic data may provide opportunities for future radiogenomics research. For instance, the proposed radiogenomics method Tumour Heterogeneity and Multi-modal Imaging Data (detailed in chapter 6) may expand to include raw genetic information as part of the feature derivation process.

Recently, there have been significant engineering advancements that have led to the development of advanced PET-CT scanners with improved sensitivity such as the Siemen's Quadra Total Body scanner. These scanners can now image the entire body, from the vertex of the head to the upper thighs, simultaneously, with reduced doses of injected radioactivity. This breakthrough enables the imaging of physiology and pathophysiology across all the major body organs simultaneously [197]. This imaging technique allows the derivation of radiogenomics features across all abnormalities at distinct sites simultaneously, taking into account the metabolic activities from healthy organs, and enables the determination of personalised

radiogenomics relationships. Such an improvement can advance radiogenomics analysis and represents a significant step towards achieving precision medicine for human cancer diagnosis and treatment.

# Reference


1. Collins FS, Varmus H: **A new initiative on precision medicine**. *New England Journal of Medicine* 2015, **372**(9):793-795.
2. Friedman AA, Letai A, Fisher DE, Flaherty KT: **Precision medicine for cancer with next-generation functional diagnostics**. *Nature Reviews Cancer* 2015, **15**(12):747.
3. Mitsudomi T, Kosaka T, Yatabe Y: **Biological and clinical implications of EGFR mutations in lung cancer**. *International journal of clinical oncology* 2006, **11**(3):190-198.
4. Stupp R, Hegi ME, Mason WP, Van Den Bent MJ, Taphoorn MJ, Janzer RC, Ludwin SK, Allgeier A, Fisher B, Belanger K: **Effects of radiotherapy with concomitant and adjuvant temozolomide versus radiotherapy alone on survival in glioblastoma in a randomised phase III study: 5-year analysis of the EORTC-NCIC trial**. *The lancet oncology* 2009, **10**(5):459-466.
5. Marks L, Young S, Natarajan S: **MRI–ultrasound fusion for guidance of targeted prostate biopsy**. *Current opinion in urology* 2013, **23**(1):43.
6. Bedard PL, Hansen AR, Ratain MJ, Siu LL: **Tumour heterogeneity in the clinic**. *Nature* 2013, **501**(7467):355.
7. Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, Van Stiphout RG, Granton P, Zegers CM, Gillies R, Boellard R, Dekker A: **Radiomics: extracting more information from medical images using advanced feature analysis**. *European journal of cancer* 2012, **48**(4):441-446.

8. Karlo CA, Di Paolo PL, Chaim J, Hakimi AA, Ostrovnaya I, Russo P, Hricak H, Motzer R, Hsieh JJ, Akin O: **Radiogenomics of clear cell renal cell carcinoma: associations between CT imaging features and mutations**. *Radiology* 2014, **270**(2):464-471.

9. Evans PM: **Anatomical imaging for radiotherapy**. *Physics in Medicine & Biology* 2008, **53**(12):R151.

10. Xia T, Kumar A, Fulham M, Feng D, Wang Y, Kim EY, Jung Y, Kim J: **Fused feature signatures to probe tumour radiogenomics relationships**. *Scientific reports* 2022, **12**(1):1-15.

11. Rutman AM, Kuo MD: **Radiogenomics: creating a link between molecular diagnostics and diagnostic imaging**. *European journal of radiology* 2009, **70**(2):232-241.

12. Sailem HZ, Bakal C: **Identification of clinically predictive metagenes that encode components of a network coupling cell shape to transcription by image-omics**. *Genome research* 2017, **27**(2):196-207.

13. Su H, Shen Y, Xing F, Qi X, Hirshfield KM, Yang L, Foran DJ: **Robust automatic breast cancer staging using a combination of functional genomics and image-omics**. In: *Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE: 2015*. IEEE: 7226-7229.

14. Madabhushi A, Lee G: **Image analysis and machine learning in digital pathology: Challenges and opportunities**. *Medical image analysis* 2016, **33**:170-175.

15. Aerts HJ, Velazquez ER, Leijenaar RT, Parmar C, Grossmann P, Carvalho S, Bussink J, Monshouwer R, Haibe-Kains B, Rietveld D: **Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach**. *Nature communications* 2014, **5**(1):1-9.

16. Gevaert O, Xu J, Hoang CD, Leung AN, Xu Y, Quon A, Rubin DL, Napel S, Plevritis SK: **Non–small cell lung cancer: identifying prognostic imaging biomarkers by leveraging public gene expression microarray data—methods and preliminary results**. *Radiology* 2012, **264**(2):387-396.

17. Kickingereder P, Sahm F, Radbruch A, Wick W, Heiland S, Von Deimling A, Bendszus M, Wiestler B: **IDH mutation status is associated with a distinct hypoxia/angiogenesis transcriptome signature which is non-invasively predictable with rCBV imaging in human glioma**. *Scientific reports* 2015, **5**:16238.

18. Banerjee S, Wang DS, Kim HJ, Sirlin CB, Chan MG, Korn RL, Rutman AM, Siripongsakun S, Lu D, Imanbayev G: **A computed tomography radiogenomic biomarker predicts microvascular invasion and clinical outcomes in hepatocellular carcinoma**. *Hepatology* 2015, **62**(3):792-800.

19. Kuo MD, Gollub J, Sirlin CB, Ooi C, Chen X: **Radiogenomic analysis to identify imaging phenotypes associated with drug response gene expression programs in hepatocellular carcinoma**. *Journal of Vascular and Interventional Radiology* 2007, **18**(7):821-830.

20. Kooi T, Litjens G, Van Ginneken B, Gubern-Mérida A, Sánchez CI, Mann R, den Heeten A, Karssemeijer N: **Large scale deep learning for computer aided detection of mammographic lesions**. *Medical image analysis* 2017, **35**:303-312.

21. Gillies RJ, Kinahan PE, Hricak H: **Radiomics: images are more than pictures, they are data**. *Radiology* 2015, **278**(2):563-577.

22. Hanif F, Muzaffar K, Perveen K, Malhi SM, Simjee SU: **Glioblastoma multiforme: a review of its epidemiology and pathogenesis through clinical presentation and treatment**. *Asian Pacific journal of cancer prevention: APJCP* 2017, **18**(1):3.

23. Parker NR, Khong P, Parkinson JF, Howell VM, Wheeler HR: **Molecular heterogeneity in glioblastoma: potential clinical implications**. *Frontiers in oncology* 2015, **5**:55.

24. Inda M-d-M, Bonavia R, Seoane J: **Glioblastoma multiforme: a look inside its heterogeneous nature**. *Cancers* 2014, **6**(1):226-239.

25. Kim HS, Lee KS, Ohno Y, Van Beek EJ, Biederer J: **PET/CT versus MRI for diagnosis, staging, and follow-up of lung cancer**. *Journal of Magnetic Resonance Imaging* 2015, **42**(2):247-260.

26. Takhar AS, Palaniappan P, Dhingsa R, Lobo DN: **Recent developments in diagnosis of pancreatic cancer**. *BMJ: British Medical Journal* 2004, **329**(7467):668.

27. Horsthuis K, Stokkers PC, Stoker J: **Detection of inflammatory bowel disease: diagnostic performance of cross-sectional imaging modalities**. *Abdominal imaging* 2008, **33**(4):407-416.

28. Townsend DW, Cherry SR: **Combining anatomy and function: the path to true image fusion**. *European radiology* 2001, **11**(10):1968-1974.

29. Pascual-Marqui RD, Esslen M, Kochi K, Lehmann D: **Functional imaging with low-resolution brain electromagnetic tomography (LORETA): a review**. *Methods and findings in experimental and clinical pharmacology* 2002, **24**(Suppl C):91-95.

30. Cherry SR: **Multimodality imaging: Beyond pet/ct and spect/ct**. In: *Seminars in nuclear medicine: 2009*. Elsevier: 348-353.

31. Sone S, Takashima S, Li F, Yang Z, Honda T, Maruyama Y, Hasegawa M, Yamanda T, Kubo K, Hanamura K: **Mass screening for lung cancer with mobile spiral computed tomography scanner**. *The Lancet* 1998, **351**(9111):1242-1245.

32. Investigators IELCAP: **Survival of patients with stage I lung cancer detected on CT screening**. *New England Journal of Medicine* 2006, **355**(17):1763-1771.

33. Cercignani M, Bouyagoub S: **Brain microstructure by multi-modal MRI: Is the whole greater than the sum of its parts?** *Neuroimage* 2018, **182**:117-127.

34. LeCun Y, Bengio Y, Hinton G: **Deep learning**. *nature* 2015, **521**(7553):436.

35. Bengio Y, Courville A, Hinton G: **Representation learning: a review and new perspectives," IEEE Transactions on Pattern Analysis and Machine Intelligence**. 2013.

36. Brunese L, Martinelli F, Mercaldo F, Santone A: **Machine learning for coronavirus covid-19 detection from chest x-rays**. *Procedia computer science* 2020, **176**:2212-2221.

37. Brodley C, Kak A, Shyu C, Dy J, Broderick L, Aisen AM: **Content-based retrieval from medical image databases: A synergy of human interaction, machine learning and computer vision**. In: *AAAI/IAAI: 1999*. 760-767.

38. Zhang J, Ma K-K, Er M-H, Chong V: **Tumor segmentation from magnetic resonance imaging by learning via one-class support vector machine**. In: *International Workshop on Advanced Image Technology (IWAIT'04): 2004*. 207--211.

39. Krizhevsky A, Sutskever I, Hinton GE: **Imagenet classification with deep convolutional neural networks**. In: *Advances in neural information processing systems: 2012*. 1097-1105.

40. Kotsiantis SB, Zaharakis I, Pintelas P: **Supervised machine learning: A review of classification techniques**. *Emerging artificial intelligence applications in computer engineering* 2007, **160**:3-24.

41. Chan HP, Hadjiiski LM, Samala RK: **Computer-aided diagnosis in the era of deep learning**. *Medical physics* 2020, **47**(5):e218-e227.

42. McCulloch WS, Pitts W: **A logical calculus of the ideas immanent in nervous activity**. *The bulletin of mathematical biophysics* 1943, **5**(4):115-133.

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

44. Schmidhuber J: **Deep learning in neural networks: An overview**. *Neural networks* 2015, **61**:85-117.

45. Hinton G, Deng L, Yu D, Dahl GE, Mohamed A-r, Jaitly N, Senior A, Vanhoucke V, Nguyen P, Sainath TN: **Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups**. *IEEE Signal Processing Magazine* 2012, **29**(6):82-97.

46. Litjens G, Sánchez CI, Timofeeva N, Hermsen M, Nagtegaal I, Kovacs I, Hulsbergen-Van De Kaa C, Bult P, Van Ginneken B, Van Der Laak J: **Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis**. *Scientific reports* 2016, **6**:26286.

47. Simonyan K, Zisserman A: **Very deep convolutional networks for large-scale image recognition**. *arXiv preprint arXiv:14091556* 2014.

48. He K, Zhang X, Ren S, Sun J: **Delving deep into rectifiers: Surpassing human-level performance on imagenet classification**. In: *Proceedings of the IEEE international conference on computer vision: 2015*. 1026-1034.

49. Ioffe S, Szegedy C: **Batch normalization: Accelerating deep network training by reducing internal covariate shift**. In: *International conference on machine learning: 2015*. 448-456.

50. Tan M, Le Q: **Efficientnet: Rethinking model scaling for convolutional neural networks**. In: *International conference on machine learning: 2019*. PMLR: 6105-6114.

51. He K, Zhang X, Ren S, Sun J: **Deep residual learning for image recognition**. In: *Proceedings of the IEEE conference on computer vision and pattern recognition: 2016*. 770-778.

52. Jegou H, Perronnin F, Douze M, Sánchez J, Perez P, Schmid C: **Aggregating local image descriptors into compact codes**. *IEEE transactions on pattern analysis and machine intelligence* 2012, **34**(9):1704-1716.

53. Perronnin F, Dance C: **Fisher kernels on visual vocabularies for image categorization**. In: *Computer Vision and Pattern Recognition, 2007 CVPR'07 IEEE Conference on: 2007*. IEEE: 1-8.

54. Bishop CM: **Neural networks for pattern recognition**: Oxford university press; 1995.

55. Hochreiter S, Schmidhuber J: **Long short-term memory**. *Neural computation* 1997, **9**(8):1735-1780.

56. Srivastava RK, Greff K, Schmidhuber J: **Highway networks**. *arXiv preprint arXiv:150500387* 2015.

57. Kumar A, Kim J, Lyndon D, Fulham M, Feng D: **An ensemble of fine-tuned convolutional neural networks for medical image classification**. *IEEE journal of biomedical and health informatics* 2017, **21**(1):31-40.

58. Tran PV: **A fully convolutional neural network for cardiac segmentation in short-axis MRI**. *arXiv preprint arXiv:160400494* 2016.

59. Pan SJ, Yang Q: **A survey on transfer learning**. *IEEE Transactions on knowledge and data engineering* 2010, **22**(10):1345-1359.

60. Pratt LY: **Discriminability-based transfer between neural networks**. In: *Advances in neural information processing systems: 1993*. 204-211.

61. Ben-David S, Blitzer J, Crammer K, Pereira F: **Analysis of representations for domain adaptation**. In: *Advances in neural information processing systems: 2007*. 137-144.

62. Glorot X, Bordes A, Bengio Y: **Domain adaptation for large-scale sentiment classification: A deep learning approach**. In: *Proceedings of the 28th international conference on machine learning (ICML-11): 2011*. 513-520.

63. Shin H-C, Roth HR, Gao M, Lu L, Xu Z, Nogues I, Yao J, Mollura D, Summers RM: **Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning**. *IEEE transactions on medical imaging* 2016, **35**(5):1285-1298.

64. van Ginneken B, Setio AA, Jacobs C, Ciompi F: **Off-the-shelf convolutional neural network features for pulmonary nodule detection in computed tomography scans**. In: *Biomedical Imaging (ISBI), 2015 IEEE 12th International Symposium on: 2015*. IEEE: 286-289.

65. Bar Y, Diamant I, Wolf L, Lieberman S, Konen E, Greenspan H: **Chest pathology detection using deep learning with non-medical training**. In: *Biomedical Imaging (ISBI), 2015 IEEE 12th International Symposium on: 2015*. IEEE: 294-297.

66. Ciompi F, de Hoop B, van Riel SJ, Chung K, Scholten ET, Oudkerk M, de Jong PA, Prokop M, van Ginneken B: **Automatic classification of pulmonary peri-fissural nodules in computed tomography using an ensemble of 2D views and a convolutional neural network out-of-the-box**. *Medical image analysis* 2015, **26**(1):195-202.

67. Tajbakhsh N, Shin JY, Gurudu SR, Hurst RT, Kendall CB, Gotway MB, Liang J: **Convolutional neural networks for medical image analysis: Full training or fine tuning?** *IEEE transactions on medical imaging* 2016, **35**(5):1299-1312.

68. Pham DL, Xu C, Prince JL: **Current methods in medical image segmentation**. *Annual review of biomedical engineering* 2000, **2**(1):315-337.

69. Poldrack RA: **Region of interest analysis for fMRI**. *Social Cognitive and Affective Neuroscience* 2007, **2**(1):67-70.

70. Raab SS, Grzybicki DM, Janosky JE, Zarbo RJ, Meier FA, Jensen C, Geyer SJ: **Clinical impact and frequency of anatomic pathology errors in cancer diagnoses**. *Cancer* 2005, **104**(10):2205-2213.

71. Fan J, Yau DK, Elmagarmid AK, Aref WG: **Automatic image segmentation by integrating color-edge extraction and seeded region growing**. *IEEE transactions on image processing* 2001, **10**(10):1454-1466.

72. Hu S, Hoffman EA, Reinhardt JM: **Automatic lung segmentation for accurate quantitation of volumetric X-ray CT images**. *IEEE transactions on medical imaging* 2001, **20**(6):490-498.

73. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S: **Dermatologist-level classification of skin cancer with deep neural networks**. *Nature* 2017, **542**(7639):115-118.

74. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, Venugopalan S, Widner K, Madams T, Cuadros J: **Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs**. *Jama* 2016, **316**(22):2402-2410.

75. Miranda E, Aryuni M, Irwansyah E: **A survey of medical image classification techniques**. In: *2016 international conference on information management and technology (ICIMTech): 2016*. IEEE: 56-61.

76. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, Van Der Laak JA, Van Ginneken B, Sánchez CI: **A survey on deep learning in medical image analysis**. *Medical image analysis* 2017, **42**:60-88.

77. Pan Y, Huang W, Lin Z, Zhu W, Zhou J, Wong J, Ding Z: **Brain tumor grading based on neural networks and convolutional neural networks**. In: *2015 37th annual international conference of the IEEE engineering in medicine and biology society (EMBC): 2015*. IEEE: 699-702.

78. Kriegsmann M, Haag C, Weis C-A, Steinbuss G, Warth A, Zgorzelski C, Muley T, Winter H, Eichhorn ME, Eichhorn F: **Deep learning for the classification of small-cell and non-small-cell lung cancer**. *Cancers* 2020, **12**(6):1604.

79. Yang L, Yang J, Zhou X, Huang L, Zhao W, Wang T, Zhuang J, Tian J: **Development of a radiomics nomogram based on the 2D and 3D CT features to predict the survival of non-small cell lung cancer patients**. *European Radiology* 2019, **29**:2196-2206.

80. An J, Oh M, Kim S-Y, Oh Y-J, Oh B, Oh J-H, Kim W, Jung JH, Kim HI, Kim J-S: **PET-Based Radiogenomics Supports mTOR Pathway Targeting for Hepatocellular Carcinoma**. *Clinical Cancer Research* 2022, **28**(9):1821-1831.

81. Lee HW, Cho H-h, Joung J-G, Jeon HG, Jeong BC, Jeon SS, Lee HM, Nam D-H, Park W-Y, Kim CK: **Integrative radiogenomics approach for risk assessment of post-operative metastasis in pathological T1 renal cell carcinoma: a pilot retrospective cohort study**. *Cancers* 2020, **12**(4):866.

82. Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P, Moore S, Phillips S, Maffitt D, Pringle M: **The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository**. *Journal of digital imaging* 2013, **26**(6):1045-1057.

83. Network CGAR: **Comprehensive genomic characterization defines human glioblastoma genes and core pathways**. *Nature* 2008, **455**(7216):1061.

84. Armitage JO: **A clinical evaluation of the International Lymphoma Study Group classification of non-Hodgkin's lymphoma**. *Blood* 1997, **89**(11):3909-3918.

85. McGuire WL: **Breast cancer prognostic factors: evaluation guidelines**. In.: Oxford University Press; 1991.

86. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES: **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles**. *Proceedings of the National Academy of Sciences* 2005, **102**(43):15545-15550.

87. Khan J, Wei JS, Ringner M, Saal LH, Ladanyi M, Westermann F, Berthold F, Schwab M, Antonescu CR, Peterson C: **Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks**. *Nature medicine* 2001, **7**(6):673.

88. Dunning AM, Healey CS, Pharoah PD, Teare MD, Ponder BA, Easton DF: **A systematic review of genetic polymorphisms and breast cancer risk**. *Cancer Epidemiology and Prevention Biomarkers* 1999, **8**(10):843-854.

89. Knudson AG: **Mutation and cancer: statistical study of retinoblastoma**. *Proceedings of the National Academy of Sciences* 1971, **68**(4):820-823.

90. Ford D, Easton DF, Bishop DT, Narod SA, Goldgar DE: **Risks of cancer in BRCA1-mutation carriers**. *The Lancet* 1994, **343**(8899):692-695.

91. Kobayashi S, Boggon TJ, Dayaram T, Jänne PA, Kocher O, Meyerson M, Johnson BE, Eck MJ, Tenen DG, Halmos B: **EGFR mutation and resistance of non–small-cell lung cancer to gefitinib**. *New England Journal of Medicine* 2005, **352**(8):786-792.

92. Van't Veer LJ, Dai H, Van De Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, Van Der Kooy K, Marton MJ, Witteveen AT: **Gene expression profiling predicts clinical outcome of breast cancer**. *nature* 2002, **415**(6871):530.

93. Celis JE, Kruhøffer M, Gromova I, Frederiksen C, Østergaard M, Thykjaer T, Gromov P, Yu J, Pálsdóttir H, Magnusson N: **Gene expression profiling: monitoring**

**transcription and translation products using DNA microarrays and proteomics**. *FEBS letters* 2000, **480**(1):2-16.

94. Choras RS: **Image feature extraction techniques and their applications for CBIR and biometrics systems**. *International journal of biology and biomedical engineering* 2007, **1**(1):6-16.

95. Hyvarinen A, Oja E, Hoyer P, Hurri J: **Image feature extraction by sparse coding and independent component analysis**. In: *Pattern Recognition, 1998 Proceedings Fourteenth International Conference on: 1998*. IEEE: 1268-1273.

96. Liu J, Tang YY, Cao Y: **An evolutionary autonomous agents approach to image feature extraction**. *IEEE Transactions on Evolutionary Computation* 1997, **1**(2):141-158.

97. Haralick RM, Shanmugam K: **Textural features for image classification**. *IEEE Transactions on systems, man, and cybernetics* 1973(6):610-621.

98. Zhang X, Cui J, Wang W, Lin C: **A study for texture feature extraction of high-resolution satellite images based on a direction measure and gray level co-occurrence matrix fusion algorithm**. *Sensors* 2017, **17**(7):1474.

99. Malik J, Belongie S, Leung T, Shi J: **Contour and texture analysis for image segmentation**. *International journal of computer vision* 2001, **43**(1):7-27.

100. Wang D, Srihari SN: **Classification of newspaper image blocks using texture analysis**. *Computer Vision, Graphics, and Image Processing* 1989, **47**(3):327-352.

101. Leone A, Distante C: **Shadow detection for moving objects based on texture analysis**. *Pattern Recognition* 2007, **40**(4):1222-1233.

102. Liu Y, Zhang D, Lu G, Ma W-Y: **A survey of content-based image retrieval with high-level semantics**. *Pattern recognition* 2007, **40**(1):262-282.

103. du Buf JH, Kardan M, Spann M: **Texture feature performance for image segmentation**. *Pattern recognition* 1990, **23**(3-4):291-309.

104. Zhang D, Wong A, Indrawan M, Lu G: **Content-based image retrieval using Gabor texture features**. *IEEE Transactions PAMI* 2000:13-15.

105. Jain AK, Farrokhnia F: **Unsupervised texture segmentation using Gabor filters**. *Pattern recognition* 1991, **24**(12):1167-1186.

106. Laine A, Fan J: **Texture classification by wavelet packet signatures**. *IEEE Transactions on pattern analysis and machine intelligence* 1993, **15**(11):1186-1191.

107. Mallat SG: **A theory for multiresolution signal decomposition: the wavelet representation**. *IEEE transactions on pattern analysis and machine intelligence* 1989, **11**(7):674-693.

108. Rui Y, Huang TS, Chang S-F: **Image retrieval: Current techniques, promising directions, and open issues**. *Journal of visual communication and image representation* 1999, **10**(1):39-62.

109. Deng Y, Manjunath B, Kenney C, Moore MS, Shin H: **An efficient color representation for image retrieval**. *IEEE Transactions on image processing* 2001, **10**(1):140-147.

110. Stricker MA, Orengo M: **Similarity of color images**. In: *Storage and Retrieval for Image and Video Databases III: 1995*. International Society for Optics and Photonics: 381-393.

111. Swain MJ, Ballard DH: **Color indexing**. *International journal of computer vision* 1991, **7**(1):11-32.

112. Van De Sande K, Gevers T, Snoek C: **Evaluating color descriptors for object and scene recognition**. *IEEE transactions on pattern analysis and machine intelligence* 2010, **32**(9):1582-1596.

113. Fang J, Qiu G: **A colour histogram based approach to human face detection**. 2003.

114. Mackiewicz MW, Fisher M, Jamieson C: **Bleeding detection in wireless capsule endoscopy using adaptive colour histogram model and support vector classification**. In: *Medical Imaging 2008: Image Processing: 2008*. International Society for Optics and Photonics: 69140R.

115. Smeulders AW, Worring M, Santini S, Gupta A, Jain R: **Content-based image retrieval at the end of the early years**. *IEEE Transactions on pattern analysis and machine intelligence* 2000, **22**(12):1349-1380.

116. Yang M, Kpalma K, Ronsin J: **A survey of shape feature extraction techniques**. In.: In-Tech; 2008.

117. Belongie S, Malik J, Puzicha J: **Shape matching and object recognition using shape contexts**. *IEEE transactions on pattern analysis and machine intelligence* 2002, **24**(4):509-522.

118. Basri R, Costa L, Geiger D, Jacobs D: **Determining the similarity of deformable shapes**. *Vision Research* 1998, **38**(15-16):2365-2385.

119. Sebastian TB, Klein PN, Kimia BB: **Recognition of shapes by editing their shock graphs**. *IEEE Transactions on pattern analysis and machine intelligence* 2004, **26**(5):550-571.

120. Soltanian-Zadeh H, Rafiee-Rad F: **Comparison of multiwavelet, wavelet, Haralick, and shape features for microcalcification classification in mammograms**. *Pattern recognition* 2004, **37**(10):1973-1986.

121. Gao XW, Podladchikova L, Shaposhnikov D, Hong K, Shevtsova N: **Recognition of traffic signs based on their colour and shape features extracted using human vision models**. *Journal of Visual Communication and Image Representation* 2006, **17**(4):675-685.

122. Brandt S, Laaksonen J, Oja E: **Statistical shape features for content-based image retrieval**. *Journal of Mathematical Imaging and Vision* 2002, **17**(2):187-198.

123. Funkhouser T, Min P, Kazhdan M, Chen J, Halderman A, Dobkin D, Jacobs D: **A search engine for 3D models**. *ACM Transactions on Graphics (TOG)* 2003, **22**(1):83-105.

124. Berretti S, Del Bimbo A, Pala P: **Description, matching and retrieval by content of 3D objects**. In: *Digital Libraries: Research and Development.* Springer; 2007: 288-297.

125. Ankerst M, Kastenmüller G, Kriegel H-P, Seidl T: **3D shape histograms for similarity search and classification in spatial databases**. In: *International Symposium on Spatial Databases: 1999*. Springer: 207-226.

126. Tangelder JW, Veltkamp RC: **A survey of content based 3D shape retrieval methods**. In: *Shape Modeling Applications, 2004 Proceedings: 2004*. IEEE: 145-156.

127. Gillies RJ, Kinahan PE, Hricak H: **Radiomics: Images Are More than Pictures, They Are Data**. *Radiology* 2016, **278**(2):563-577.

128. Segal E, Sirlin CB, Ooi C, Adler AS, Gollub J, Chen X, Chan BK, Matcuk GR, Barry CT, Chang HY: **Decoding global gene expression programs in liver cancer by noninvasive imaging**. *Nature biotechnology* 2007, **25**(6):675.

129. Engle Jr RL: **Attempts to use computers as diagnostic aids in medical decision making: a thirty-year experience**. *Perspectives in biology and medicine* 1992, **35**(2):207-219.

130. Chan HP, Galhotra S, Vyborny CJ, MacMahon H, Jokich PM: **Image feature analysis and computer-aided diagnosis in digital radiography. I. Automated detection of microcalcifications in mammography**. *Medical physics* 1987, **14**(4):538-548.

131. Way TW, Hadjiiski LM, Sahiner B, Chan HP, Cascade PN, Kazerooni EA, Bogot N, Zhou C: **Computer-aided diagnosis of pulmonary nodules on CT scans: Segmentation and classification using 3D active contours**. *Medical physics* 2006, **33**(7Part1):2323-2337.

132. Jiang Y, Nishikawa RM, Schmidt RA, Metz CE, Giger ML, Doi K: **Improving breast cancer diagnosis with computer-aided diagnosis**. *Academic radiology* 1999, **6**(1):22-33.

133. Sahiner B, Petrick N, Chan H-P, Hadjiiski LM, Paramagul C, Helvie MA, Gurcan MN: **Computer-aided characterization of mammographic masses: accuracy of mass segmentation and its effects on characterization**. *IEEE transactions on medical imaging* 2001, **20**(12):1275-1284.

134. Hu B, Guo H, Zhou P, Shi Z-L: **Characteristics of SARS-CoV-2 and COVID-19**. *Nature Reviews Microbiology* 2021, **19**(3):141-154.

135. Li YC, Bai WZ, Hashikawa T: **The neuroinvasive potential of SARS-CoV2 may play a role in the respiratory failure of COVID-19 patients**. *Journal of medical virology* 2020, **92**(6):552-555.

136. Faghani S, Khosravi B, Moassefi M, Conte GM, Erickson BJ: **A Comparison of Three Different Deep Learning-Based Models to Predict the MGMT Promoter Methylation Status in Glioblastoma Using Brain MRI**. *Journal of Digital Imaging* 2023:1-10.

137. Hagerty JR, Stanley RJ, Almubarak HA, Lama N, Kasmi R, Guo P, Drugge RJ, Rabinovitz HS, Oliviero M, Stoecker WV: **Deep learning and handcrafted method fusion: higher diagnostic accuracy for melanoma dermoscopy images**. *IEEE journal of biomedical and health informatics* 2019, **23**(4):1385-1391.

138. Nair JKR, Saeed UA, McDougall CC, Sabri A, Kovacina B, Raidu B, Khokhar RA, Probst S, Hirsh V, Chankowsky J: **Radiogenomic Models Using Machine Learning Techniques to Predict EGFR Mutations in Non-Small Cell Lung Cancer**. *Canadian Association of Radiologists Journal* 2020:0846537119899526.

139. Li Z-C, Bai H, Sun Q, Li Q, Liu L, Zou Y, Chen Y, Liang C, Zheng H: **Multiregional radiomics features from multiparametric MRI for prediction of MGMT methylation status in glioblastoma multiforme: a multicentre study**. *European radiology* 2018, **28**:3640-3650.

140. Aerts HJWL RVE, Leijenaar RTH, Parmar C, Grossmann P, Carvalho S, Bussink J, Monshouwer R, Haibe-Kains B, Rietveld D, Hoebers F, Rietbergen MM, Leemans CR, Dekker A, Quackenbush J, Gillies RJ, Lambin P.: **Data From NSCLC-Radiomics-Genomics**. In. The Cancer Imaging Archive; 2015.

141. Edge SB, Compton CC: **The American Joint Committee on Cancer: the 7th edition of the AJCC cancer staging manual and the future of TNM**. *Annals of surgical oncology* 2010, **17**(6):1471-1474.

142. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M: **NCBI GEO: archive for functional genomics data sets—update**. *Nucleic acids research* 2012, **41**(D1):D991-D995.

143. Wolf I, Vetter M, Wegner I, Böttger T, Nolden M, Schöbinger M, Hastenteufel M, Kunert T, Meinzer H-P: **The medical imaging interaction toolkit**. *Medical image analysis* 2005, **9**(6):594-604.

144. Bakr S, Gevaert O, Echegaray S, Ayers K, Zhou M, Shafiq M, Zheng H, Benson JA, Zhang W, Leung AN: **A radiogenomic dataset of non-small cell lung cancer**. *Scientific data* 2018, **5**(1):1-9.

145. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR: **STAR: ultrafast universal RNA-seq aligner**. *Bioinformatics* 2013, **29**(1):15-21.

146. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L: **Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks**. *Nature protocols* 2012, **7**(3):562-578.

147. Aerts HJ, Velazquez ER, Leijenaar RT, Parmar C, Grossmann P, Carvalho S, Bussink J, Monshouwer R, Haibe-Kains B, Rietveld D: **Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach**. *Nature communications* 2014, **5**:4006.

148. Van Griethuysen JJ, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V, Beets-Tan RG, Fillion-Robin J-C, Pieper S, Aerts HJ: **Computational radiomics system to decode the radiographic phenotype**. *Cancer research* 2017, **77**(21):e104-e107.

149. Su C, Jiang J, Zhang S, Shi J, Xu K, Shen N, Zhang J, Li L, Zhao L, Zhang J: **Radiomics based on multicontrast MRI can precisely differentiate among glioma subtypes and predict tumour-proliferative behaviour**. *European radiology* 2019, **29**(4):1986-1996.

150. Liu X, Li Y, Qian Z, Sun Z, Xu K, Wang K, Liu S, Fan X, Li S, Zhang Z: **A radiomic signature as a non-invasive predictor of progression-free survival in patients with lower-grade gliomas**. *NeuroImage: Clinical* 2018, **20**:1070-1077.

151. Fu Y, Cao L, Guo G, Huang TS: **Multiple feature fusion by subspace learning**. In: *Proceedings of the 2008 international conference on Content-based image and video retrieval: 2008*. 127-134.

152. Ding C, Peng H: **Minimum redundancy feature selection from microarray gene expression data**. *Journal of bioinformatics and computational biology* 2005, **3**(02):185-205.

153. Hua J, Xiong Z, Lowey J, Suh E, Dougherty ER: **Optimal number of features as a function of sample size for various classification rules**. *Bioinformatics* 2005, **21**(8):1509-1515.

154. Kohane IS, Butte AJ, Kho A: **Microarrays for an integrative genomics**: MIT press; 2002.

155. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP: **Molecular signatures database (MSigDB) 3.0**. *Bioinformatics* 2011, **27**(12):1739-1740.

156. Ettinger DS, Akerley W, Bepler G, Blum MG, Chang A, Cheney RT, Chirieac LR, D'Amico TA, Demmy TL, Ganti AKP: **Non–small cell lung cancer**. *Journal of the national comprehensive cancer network* 2010, **8**(7):740-801.

157. Jin M, Inoue S, Umemura T, Moriya J, Arakawa M, Nagashima K, Kato H: **Cyclin D1, p16 and retinoblastoma gene product expression as a predictor for prognosis in non-small cell lung cancer at stages I and II**. *Lung cancer* 2001, **34**(2):207-218.

158. Zheng C-Y, Lam S-K, Li Y-Y, Ho JC-M: **Arsenic trioxide-induced cytotoxicity in small cell lung cancer via altered redox homeostasis and mitochondrial integrity**. *International journal of oncology* 2015, **46**(3):1067-1078.

159. Regala RP, Weems C, Jamieson L, Khoor A, Edell ES, Lohse CM, Fields AP: **Atypical protein kinase Cι is an oncogene in human non–small cell lung cancer**. *Cancer research* 2005, **65**(19):8905-8911.

160. Regala RP, Weems C, Jamieson L, Copland JA, Thompson EA, Fields AP: **Atypical protein kinase Cι plays a critical role in human lung cancer cell growth and tumorigenicity**. *Journal of Biological Chemistry* 2005, **280**(35):31109-31115.

161. Cohen AJ, Bunn PA, Franklin W, Magill-Solc C, Hartmann C, Helfrich B, Gilman L, Folkvord J, Helm K, Miller YE: **Neutral endopeptidase: variable expression in human lung, inactivation in lung cancer, and modulation of peptide-induced calcium flux**. *Cancer research* 1996, **56**(4):831-839.

162. Chatterjee A, Vallières M, Dohan A, Levesque IR, Ueno Y, Bist V, Saif S, Reinhold C, Seuntjens J: **An empirical approach for avoiding false discoveries when applying high-dimensional radiomics to small datasets**. *IEEE Transactions on Radiation and Plasma Medical Sciences* 2018, **3**(2):201-209.

163. Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P, Moore S, Phillips S, Maffitt D, Pringle M *et al*: **The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository**. *Journal of Digital Imaging* 2013, **26**(6):1045-1057.

164. Bakr S, Gevaert O, Echegaray S, Ayers K, Zhou M, Shafiq M, Zheng H, Benson JA, Zhang W, Leung ANC *et al*: **A radiogenomic dataset of non-small cell lung cancer**. *Scientific Data* 2018, **5**(1):180202.

165. Aerts HJWL, Velazquez ER, Leijenaar RTH, Parmar C, Grossmann P, Carvalho S, Bussink J, Monshouwer R, Haibe-Kains B, Rietveld D *et al*: **Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach**. *Nature Communications* 2014, **5**(1):4006.

166. Zhang K, Liu X, Shen J, Li Z, Sang Y, Wu X, Zha Y, Liang W, Wang C, Wang K *et al*: **Clinically Applicable AI System for Accurate Diagnosis, Quantitative**

**Measurements, and Prognosis of COVID-19 Pneumonia Using Computed Tomography**. *Cell* 2020, **181**(6):1423-1433.e1411.

167. van Griethuysen JJM, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V, Beets-Tan RGH, Fillion-Robin J-C, Pieper S, Aerts HJWL: **Computational Radiomics System to Decode the Radiographic Phenotype**. *Cancer Res* 2017, **77**(21):e104-e107.

168. Haralick RM, Shanmugam K, Dinstein I: **Textural Features for Image Classification**. *IEEE Transactions on Systems, Man, and Cybernetics* 1973, **SMC-3**(6):610-621.

169. Galloway MM: **Texture analysis using gray level run lengths**. *Computer Graphics and Image Processing* 1975, **4**(2):172-179.

170. Thibault G, Fertil B, Navarro C, Pereira S, Cau P, Levy N, Sequeira J, Mari J-L: **Shape and texture indexes application to cell nuclei classification**. *International Journal of Pattern Recognition and Artificial Intelligence* 2013, **27**(01):1357002.

171. Amadasun M, King R: **Textural features corresponding to textural properties**. *IEEE Transactions on Systems, Man, and Cybernetics* 1989, **19**(5):1264-1274.

172. Weszka JS, Dyer CR, Rosenfeld A: **A Comparative Study of Texture Measures for Terrain Classification**. *IEEE Transactions on Systems, Man, and Cybernetics* 1976, **SMC-6**(4):269-285.

173. Kozachenko LF, Leonenko NN: **Sample estimate of the entropy of a random vector**. *Problemy Peredachi Informatsii* 1987, **23**(2):9-16.

174. Guyon I, Weston J, Barnhill S, Vapnik V: **Gene selection for cancer classification using support vector machines**. *Machine learning* 2002, **46**(1):389-422.

175. Robnik-Šikonja M, Kononenko I: **Theoretical and empirical analysis of ReliefF and RReliefF**. *Machine learning* 2003, **53**(1):23-69.

176. Tibshirani R: **Regression shrinkage and selection via the lasso**. *Journal of the Royal Statistical Society: Series B (Methodological)* 1996, **58**(1):267-288.

177. Zou H, Hastie T: **Regularization and variable selection via the elastic net**. *Journal of the royal statistical society: series B (statistical methodology)* 2005, **67**(2):301-320.

178. Papazian L, Calfee CS, Chiumello D, Luyt C-E, Meyer NJ, Sekiguchi H, Matthay MA, Meduri GU: **Diagnostic workup for ARDS patients**. *Intensive care medicine* 2016, **42**(5):674-685.

179. Kwee TC, Kwee RM: **Chest CT in COVID-19: what the radiologist needs to know**. *RadioGraphics* 2020, **40**(7):1848-1865.

180. Torres R, Judson-Torres RL: **Research techniques made simple: feature selection for biomarker discovery**. *Journal of Investigative Dermatology* 2019, **139**(10):2068-2074. e2061.

181. Dougherty ER, Brun M: **On the number of close-to-optimal feature sets**. *Cancer Informatics* 2006, **2**:117693510600200011.

182. Parasher A: **COVID-19: Current understanding of its pathophysiology, clinical presentation and treatment**. *Postgraduate medical journal* 2021, **97**(1147):312-320.

183. Xiao L, Sakagami H, Miwa N: **ACE2: the key molecule for understanding the pathophysiology of severe and critical conditions of COVID-19: demon or angel?** *Viruses* 2020, **12**(5):491.

184. Baker SA, Kwok S, Berry GJ, Montine TJ: **Angiotensin-converting enzyme 2 (ACE2) expression increases with age in patients requiring mechanical ventilation**. *PLoS One* 2021, **16**(2):e0247060.

185. Ginestra JC, Mitchell OJ, Anesi GL, Christie JD: **COVID-19 Critical Illness: A Data-Driven Review**. *Annual review of medicine* 2022, **73**:95.

186. Bakhshandeh B, Sorboni SG, Javanmard A-R, Mottaghi SS, Mehrabi M-R, Sorouri F, Abbasi A, Jahanafrooz Z: **Variants in ACE2; potential influences on virus infection and COVID-19 severity**. *Infection, Genetics and Evolution* 2021, **90**:104773.

187. Guo X, Chen Z, Xia Y, Lin W, Li H: **Investigation of the genetic variation in ACE2 on the structural recognition by the novel coronavirus (SARS-CoV-2)**. *Journal of Translational Medicine* 2020, **18**(1):1-9.

188. Warfield SK, Zou KH, Wells WM: **Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation**. *IEEE transactions on medical imaging* 2004, **23**(7):903-921.

189. Isensee F, Jaeger PF, Kohl SA, Petersen J, Maier-Hein KH: **nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation**. *Nature methods* 2021, **18**(2):203-211.

190. McKinley R, Meier R, Wiest R: **Ensembles of densely-connected CNNs with label-uncertainty for brain tumor segmentation**. In: *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part II 4: 2019*. Springer: 456-465.

191. Kamnitsas K, Ledig C, Newcombe VF, Simpson JP, Kane AD, Menon DK, Rueckert D, Glocker B: **Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation**. *Medical image analysis* 2017, **36**:61-78.

192. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A: **Going deeper with convolutions**. In: *Proceedings of the IEEE conference on computer vision and pattern recognition: 2015*. 1-9.

193. Gaur L, Bhatia U, Jhanjhi N, Muhammad G, Masud M: **Medical image-based detection of COVID-19 using deep convolution neural networks**. *Multimedia systems* 2021:1-10.

194. Alom MZ, Yakopcic C, Nasrin MS, Taha TM, Asari VK: **Breast cancer classification from histopathological images with inception recurrent residual convolutional neural network**. *Journal of digital imaging* 2019, **32**:605-617.

195. Chen X, Zeng M, Tong Y, Zhang T, Fu Y, Li H, Zhang Z, Cheng Z, Xu X, Yang R: **Automatic prediction of MGMT status in glioblastoma via deep learning-based MR image analysis**. *BioMed Research International* 2020, **2020**.

196. Carrillo J, Lai A, Nghiemphu P, Kim H, Phillips H, Kharbanda S, Moftakhar P, Lalaezari S, Yong W, Ellingson B: **Relationship between tumor enhancement, edema, IDH1 mutational status, MGMT promoter methylation, and survival in glioblastoma**. *American Journal of Neuroradiology* 2012, **33**(7):1349-1355.

197. Cherry SR, Badawi RD, Karp JS, Moses WW, Price P, Jones T: **Total-body imaging: transforming the role of positron emission tomography**. *Science translational medicine* 2017, **9**(381):eaaf6169.