QUANTITATIVE RADIOMIC ANALYSIS FOR PROGNOSTIC MEDICAL APPLICATIONS

Chongrui Xu

A thesis submitted in fulfilment of the requirements for
the degree of Master of Philosophy
in the Faculty of Engineering & Information Technologies

UNIVERSITY OF SYDNEY

AUGUST 2019

Abstract

With the global aging population and the deterioration of people's living habits, the morbidity and mortality of cancer continues to increase. Developing methods to prevent cancer, predict cancer, cure cancer and increase life span are the main goals of modern precision medicine. However, traditional medical techniques are not available for some patients as intra-tumour heterogeneity promotes tumour evolution to form diverse diseases. This limits the use of effectiveness of invasive biopsies based on molecular assays. The evolution of this trend has led to breakthroughs in precision medicine to prevent, diagnose and treat cancer more effectively.

Radiomics, a non-invasive and quantitative mining medical imaging information method, could extract molecular biological features and enormous feature combinations to customise individualised treatment and solve the problem of heterogeneity, satisfying the standards of precision medicine. However, it faces many challenges in the feature selection process, including redundant features, irrelevant features and the overfitting risk. More important, people know little about radiomics biological background and its connection to radiology, so it is difficult to apply radiology directly to medicine as it lacks interpretability.

The core of this thesis is radiomic biology analysis that connects radiomic imaging information with molecular biology information to achieve a medical “gold standard” for cancer management. Developing methods to succeed in the feature selection process of data of varying dimensions is the main goal of this paper. Our major contributions in this thesis can be summarised as below:

1. We firstly proposed an unsupervised learning framework to guide supervised learning in the reduction of feature dimensions from large cohorts Non-Small Cell Lung Cancer data (NSCLC) on both clinical data and radiomic data for survival prediction. The novelties include 1) we use a data-driven clustering method; we compare this with outcome-driven classical machine learning models (Cox proportion hazard and Random survival forest). 2) Optimising the clustering algorithm automatically reveals the hidden structure of feature information inherent in unlabelled data sets. 3) Each part has specific selection criteria to overcome common feature selection challenges.
2. An interpretable machine learning approach measures the contribution of features for each case and the connection of radiomics to its underlying biological features to make clinical decisions in leukemia and breast cancer cases. The weight of the feature can be estimated by measuring the distance of the approximate perturbation centre. Related experiments found that the error in the segmentation of bone marrow lesions has no significant impact on the experimental results.
3. Based on the framework of feature selection that we proposed, to ensure the fairness and stability of the data split when processing classification results, cross-validation is embedded in the training process. We further propose a traversal selection method, optimising the computational complexity of the selection process to obtain the most robust feature set.

The proposed selection framework has been verified by multiple evaluation methods on NSCLC patients and compared to state-of-the-art well-known survival algorithms. In this thesis, all experimental datasets include clinical data, anatomical imaging data and molecular imaging data from three hospitals. We focus our research on several common or specific cancer types and assess the value of the radiomic method for cancer prognosis.

Publications

Journal Publications

1. H. Li, **C. Xu**, B. Xin, C. Zheng, Y. Zhao, K. Hao, Q. Wang, R. Wahl, X. Wang, Y. Zhou, " ^{18}F-FDG PET/CT Radiomic Analysis with Machine Learning for Identifying Bone Marrow Involvement in the Patients with Suspected Relapsed Acute Leukemia" *THERANOSTICS,* vol. 9, no. 16, pp.4730-4739, 2019.
2. L. Wang, T. Dong, B. Xin, **C. Xu**, M. Guo, H. Zhang, D. Feng, X. Wang, J. Yu, "Integrative nomogram of CT imaging, clinical, and hematological features for survival prediction of patients with locally advanced non-small cell lung cancer," *European radiology,* pp. 1-10, 2019.
3. H. Li, **C. Xu**, B. Xin, C. Zheng, Y. Zhao, K. Hao, Q. Wang, R. Wahl, X. Wang, Y. Zhou, "Automated PET/CT radiomic analysis with machine learning for identifying bone marrow involvement in the patients with suspected relapsed acute leukemia," *Journal of Nuclear Medicine,* vol. 60, no. supplement 1, pp. 619-619, 2019.

Journal Papers under Review

4. P. Li, X. Wang, **C. Xu**, C. Liu, C. Zheng, M. Fulham, D. Feng, L. Wang, S. Song, G. Huang " Exploration of ^{18}F-FDG PET/CT Radiomic Features to Predict Pathologic Complete Response of Breast Cancer to Neoadjuvant Chemotherapy" submitted to *European Journal of Nuclear Medicine and Molecular Imaging,* (under Review).

Conference Paper

5. B. Xin, **C. Xu**, L. Wang, T. Dong, C. Zheng and X. Wang, " Integrative Clustering and Supervised Feature Selection for Clinical Applications," in *2018 15th International Conference on Control, Automation, Robotics and Vision (ICARCV)*, 2018, pp. 1316-1320: IEEE.

Attribution statements

Chapter 3 and 4 in this thesis was published as [1, 2]. In Chapter 5, the status of the paper is under review. I conducted all of the experiments, integrate analysis and wrote the manuscripts. Therefore, the corresponding author positions are co-first authors.

Acknowledgements

Here, I would like to express my sincere appreciation to all of those who have helped me during my Mphil study.

I would like to firstly thank my supervisor, Associate Prof. Xiuying Wang for providing me the chance to join in the research group. She is the guide on my research road and always gives me professional advice. Her passion for research and her keen judgment often motivates me to face and overcome research challenges. As she often said to me, "When you feel tired, maybe you are in the uphill path of life". I am honoured to be her student.

As my supervisor, she has provided many research opportunities for us so that I can cooperate with many hospitals, and obtain medical technical support and data. I am glad to work with these diligent doctors, and I thank them for their support and cooperation.

I would like to thank Dr. Bowen Xin and Dr. Chaojie Zheng for supporting me both in my research and life. Their research capabilities and enthusiasm always inspire me to better myself.

Last but not least, I would like to thank my parents for unconditionally supporting and accommodating all of my shortcomings. They set a good sample to teach me how to be a confident, optimistic, hardworking person. Thanks to Shanshan Wang for her encouragement and companionship when I face difficulties during my MPhil study.

Contents

List of Figures

List of Tables

Chapter 1. Introduction

1.1. Background and Motivations

Because of the growth of the global population and aging, and people's sub-healthy living conditions such as smoking, obesity, and lack of exercise, the incidence and death rate of cancer continues to increase. According to worldwide cancer statistics in 2018, there are 18.1 million new cases of cancer and 9.6 million deaths in the year, an increase of 4.1 million new cases and an additional 1.4 million deaths compared to 2012 [3, 4]. The development of modern medicine has contributed to the improvement of overall life expectancy. For example, vaccines against Human Papillomavirus (HPV) created can effectively reduce the likelihood of anogenital diseases [5], and the Hepatitis vaccine has become a major factor in the reduction of the morbidity and mortality of Viral hepatitis [6]. However, lower case because the lack of comprehensive population-based cancer registries for understanding the causes and development of cancer vaccines in many low- and middle-income countries, approximately 70 % of total cancer mortality comes from these regions [7]. The main morbidity and mortality is also derived from carcinogenic infections, such as HPV, Hepatitis virus, and Helicobacter pylori. It is estimated that 15% or more of the global cancer burden can be attributed to carcinogenic infections [8]. The evolution of this trend has led to breakthroughs in the new implementation of targeted cancer management to prevent, diagnose and treat cancer more effectively.

Although 30-50% of cancers can be avoided by improving lifestyle habits and implementing preventive measures, effective clinical treatment decision is critical for those who have mid- or late-stage cancers [3]. There are many strategies of cancer treatment including surgery, radiation therapy (RT), chemotherapy, immunotherapy, targeted therapy (TT), hormone therapy (HT) and stem cell transplant (SCT). Because of different cancer types, tumoral sites and progressions, the treatment methods are also discrepant. However, often the special tumor location, large tumor size and worsening body condition greatly increase the difficulty of operation. Hence, other treatment methods are needed to support the limitations of a "one-size-fits-all" approach to cancer treatment [9]. How to determine the most appropriate treatment is a major challenge for cancer management.

1.2. Precision Medicine Challenges

Precision medicine is a treatment that uses the genetic information of the disease to select the treatment that is most likely to be suitable for the patient. Precision medicine is also called personalized treatment as it considers different layers of patient information including demographics, proteomics, blood biomarkers, genomics and imaging, to form different treatment decisions [10-12]. Among them, the genetic tests of tumors have the potential to analyze the applicability of targeted drugs on mutant genes. Genetic tests will help determine the most likely progression, direction and treatment response of tumors, thereby fundamentally assessing the effectiveness of the treatment regimen and eliminating ineffective treatments [13]. However, because of its extremely cost, genetic tests are difficult to be widely used. This series of personalized treatments have three common goals: 1) Understanding the treatment response mechanism. 2) Developing a road against drug resistance. 3) Improving the life expectancy of cancer patients as much as possible [9]. One of the main challenges to achieving these goals is the heterogeneity of cancer, which confines diagnosis and therapeutics.

1.2.1.Tumor Heterogeneity

Patients with the same type, stage and symptoms are likely to have different responses even when they receive the same treatment [13]. Therefore, apparent different symptoms of cancer treatment failure, i.e. the frequent recurrence and metastasis of cancer, reduction of drug sensitivity and prognostic discrepancy, are most likely due to tumor heterogeneity.

Tumor heterogeneity [14] refers to the growth of cancer cells through multiple divisions to show different morphological and phenotypic characteristics, including cell proliferation ability, genetic alteration, signaling pathways, drug response, immune response, cellular metabolism, motility and metastasis potential. These differences interfere with the effectiveness of cancer treatment in all aspects and become one of the main obstacles of precision medicine. From broadly speaking, tumor heterogeneity can be divided into two types: intra-tumor heterogeneity and inter-tumor heterogeneity (inter-patient and intra-patient tumor heterogeneity) (as shown in Figure 1-1).

Tumoral heterogeneity is also characterized by spatial heterogeneity and temporal heterogeneity. The gene and cell subtype spatial heterogeneity within tumors, and the temporal heterogeneity of primary and secondary tumors limit the utility of invasive biopsies based on molecular assays [11, 15]. Similarly, genetic analysis of tumors may deliver incorrect information on cell mass mutations, expression, and metabolism as the spatial and temporal information cannot easily lock on same tissue [16]. This means that even if part of the cancer cells may be inhibited and killed by the targeted drugs, it is fundamentally impossible to completely kill all the cancer cells, and the surviving cancer cells will develop resistance to the drug, commonly known as drug resistance [17] (Figure 1-2). Hence, it is necessary to customize individualized treatment solving the obstacle of heterogeneity and achieving the goals of precision medicine [18].

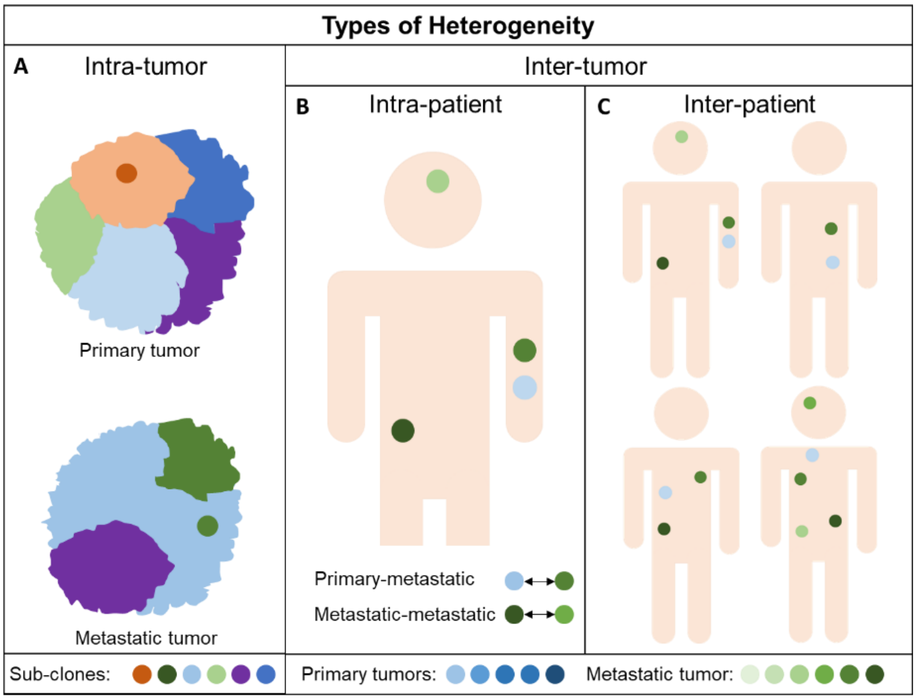


Figure 1-1. The types of tumor heterogeneity. (A) Intratumor or intercellular heterogeneity from primary tumor (top) and metastatic tumor (bottom). (B) Intra-patient heterogeneity comes from the discrepancies among the primary tumor and metastatic tumors or among metastatic tumors. (C) The inter-patient heterogeneity is the differences from different tumors of patients.

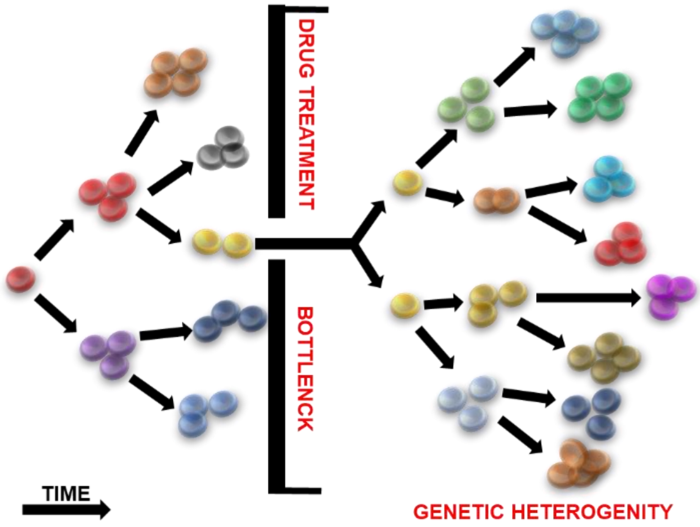


Figure 1-2. The procedure of drug resistance for subclones-cells re-form heterogeneous tumors.

1.2.2. Integrate Multi-level Biomarkers Information

Biomarkers [19, 20] are a repeatable and objective way to record the biological state. It can detect what activities will be changed in the body or as target drug interfered and monitored disease progressions or curative effect. However, different biological indicators may appear in different patients even if they receive the same treatment [21]. Therefore, patients with the same type of cancer can be separated into more sub-groups by consolidating a large variety of molecular marker data to improve the therapeutic effect [22]. Hence, determining how to integrate, extract, select and analyze these biological results is a big challenge.

these important biomarkers can be obtained from multiple omics or multiple layers' molecules, such as genomics, metabolomics, microbiomics and proteomics, and then Integrating these biomarkers analysis to explain pathological mechanisms [22] (Figure 1-3). These biomarkers are widely utilized in many clinical applications, including diagnostic markers, monitoring disease progression markers, disease prognosis markers and curative effect markers [23].

Currently, there are many biomarkers for diagnosing diseases, the commonly used of which are blood tests, urine tests and radio-imaging based parameters in metabolomics for preliminary disease inspections [24]. This does not mean that the rest of three other omics information cannot be applied to the diagnosis of the disease. On the contrary, they fundamentally understand, explore and track disease. Taking the diagnosis of the Non-Small Cells Lung Cancer (NSCLC) subtype as an example, the expression of two genes driven, Epidermal Growth Factor Receptor (EGFR) and KRAS, were used to identify 2 new adenocarcinoma subtypes in 2004 [25, 26]. In the following decade, the researchers found that the three enzyme proteins, PPP3CA, DOT1L and FtsJ

Methyltransferase Domain Containing 1 (FTSJD1) identified new several lung adenocarcinomas subtypes [27]. Some proteomic biomarkers are very effective in the prognosis of the disease. Researchers validated 17 types of circulating inflammatory proteins as biomarkers due to were prognostic ability in lung cancer [28]. Microbiome biomarkers also have the potential to influence cancer progression and prognosis. Because of the pathogenic mechanism of Fusobacterium nucleatum on colorectal cancer and intestinal inflammation, it provides strong support for functional microbial markers [29-34]. If there is no effective biomarker to detect the activities of the tumor, then the targeted treatment will be limited. Therefore, there is an urgent requirement for more mature biomarkers to capture the molecular spectrum of tumors. Medical imaging has huge potential to capture the tumor phenotypes in a non-invasive way [11, 35].

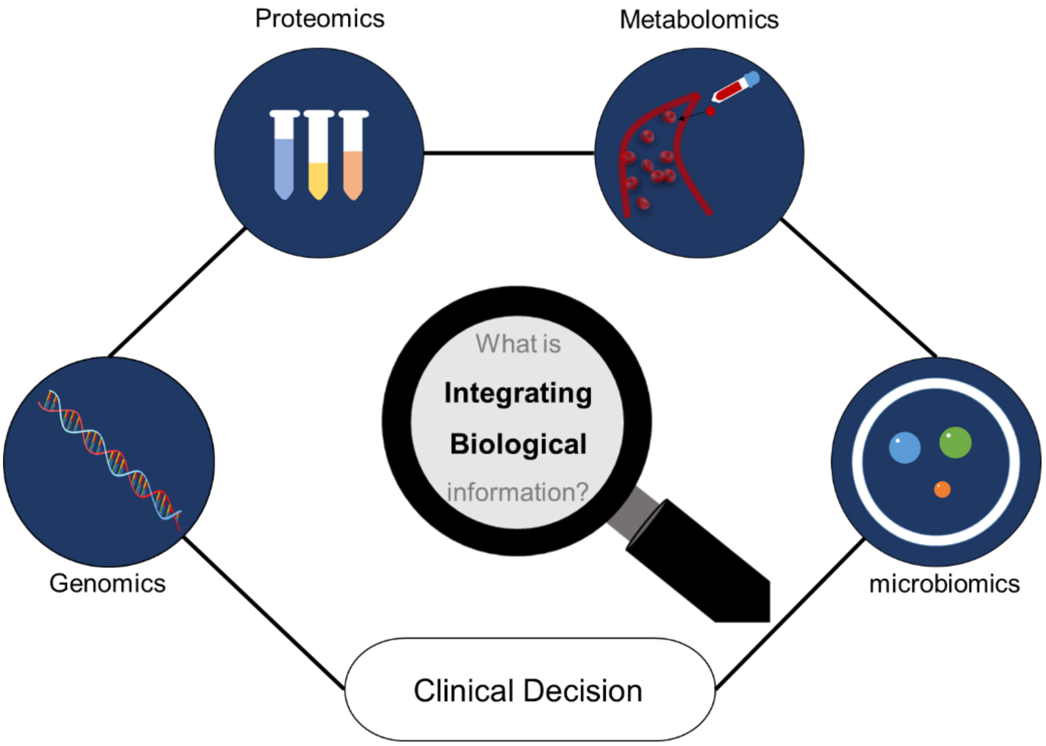


Figure 1-3. The types of biomarkers can be classified by multilayer molecule-omics information, including genomics, proteomics, metabolomics and microbiomics.

1.3. Medical Imaging

The technologies of medical imaging play a core role in the whole process of cancer management [36, 37]. It accurately determines tumor location, size, cancer metastasis, and whether treatment involves critical anatomical structures. Especially, with the rise of genomics and proteomics technologies, it integrates the unique molecular and physiological information of each patient with the anatomical information obtained by conventional imaging methods to detect molecular or physiological changes, assess and modify therapeutic schedules in real-time, simplify the development process of cancer drugs [36]. More importantly, this imaging technology is visualized tissue in non-invasive ways to avoid injury from invasive biopsy [11].

There are many mature imaging technologies, such as Computed Tomography (CT) imaging, Positron-emission Tomography (PET) imaging, Magnetic Resonance Imaging (MRI) and Medical ultrasound imaging. Different modalities of molecular imaging technology can observe different information. CT images can assess the cancer structural features, especially organs composed of soft tissues such as the spinal cord, lung, liver, pancreas, etc., but it cannot describe the functional details of solid tumors and it is difficult to find initial symptoms of cancer. PET images can detect the presence of early cancer cells and their molecule metabolic activities but have a poorer ability to describing structural tissue information [11]. MRI images are superior to CT in soft tissue such as nerves, blood vessels, muscles, etc. However, the detection of lung, liver, pancreas, adrenal gland and prostate is worse than CT and more expensive [38].

Although these imaging techniques are widely implemented in hospitals for treating patients, the expression of information from imaging is limited due to the oversimplification of internal diagnostic criteria, For example, from a clinical perspective, PET-CT imaging is difficult to evaluate the Bone Marrow Involvement (BMI) in suspected relapsed leukemia patients, especially in patients with diffuse leukemia [39]. This is because the available information from these clinical criteria is definitely limited and the assessment is only qualitative rather than quantitative [40, 41]. Therefore, there is an urgent need for an approach for quantitatively mining more valuable information from imaging to diagnose, treat and monitor disease. With the rapid development of hardware devices and imaging agents in medical imaging technology, a computer-assisted standard quantitative extracted features method, known as radiomics, as the true transformative power of medical imaging analysis.

1.4. Contributions

The core of this thesis is radiomic biology analysis that connects radiomic imaging information with molecular biology to achieve a medical “gold standard” for cancer management. Our major contributions in this thesis can be summarized as below:

1) We firstly proposed a framework based on unsupervised learning to guide supervised learning in the reduction of feature dimensions from NSCLC for survival prediction. Different from most feature selection methods, we use the combination of based on data-driven clustering method with outcome-driven classical machine learning models (Cox proportion hazard and Random survival forest) as comparison. Optimizing the clustering algorithm reveals the hidden structure of feature information inherent in unlabeled data sets to ensure the stability, robustness and scalability of the clustering results. Each part has specific selection criteria to overcome these challenges to select the most informative, representative, and non-redundant feature combinations.

2) An interpretable machine learning approach measures the contribution of features for each case and the connections of radiomics with its underlying biological features to seek innovate clinical decision making in related studies. The weight of the feature can be estimated by measuring the distance of the approximate perturbation center. In this thesis, apply on predicted bone marrow biopsy (BMB) conditions of leukemia patients and the status of pathologic complete response (pCR) in patients with locally advanced breast cancer.

3) Based on the framework of feature selection we proposed, in order to ensure the fairness and stability of the data splits when processing the classification results, the cross-validation is embedded in the part of the outcome-driven to select the most representative and reasonable features set. We further propose a traversal selection method optimizing the computational complexity of the selection process. Under the comparison of the mean and variance of 30 times cross-validation, the most robust feature set can be selected. In addition, this approach has played an important role in the evaluation of tumor heterogeneity in breast cancer subtypes and the conclusion is consistent with clinical experiences.

1.5. Thesis Organization

The MPhil thesis is divided into 6 chapters. Chapter 1 provides a general background about cancer management challenges, and giving a literature review for precision medicine in oncologic heterogeneity, biomarkers and medical imaging. Chapter 2 is a literature review summaries about the statistical analysis methods and applications of the quantitative radiomic field including, radiomic workflow, opportunities, biological foundation and understanding interpretation for common data analytics. Chapter 3 is a proposed framework to solve the common problems of feature redundancy, irrelevance and overfittings in conventional feature selection process. The main journal contents of thesis are distributed in Chapter 4 and Chapter 5 that present a published and an under review study on the radiomic basic biology, its application on predicting differential diagnosis of pathologic tissue and prognostic treatment response using machine learning methods. Finally, Chapter 6 will combine these research results with clinical experiences to discuss contributions and future works on prognostic medicine. The brief chapters were summarized as below.

Chapter 1: Introduction

Chapter 1 is describing a general introduction about cancer management, including its origin, process and challenges. Focus on its potentially great value on multi-sources of information, e.g. imaging, biomarkers and genomics, to achieve precision medicine of oncological diagnosis and treatment decision.

Chapter 2: Radiomic Methodology and Applications

Chapter 2 through summaries of current literature research describes the details of radiomics process, pitfalls, chances, analysis methods and its connections with other domain information. Analysis methods are mainly based on feature selection methods for classification, comparison and interpretation, then some of those methods will be adopted and optimized then form a reliable selection algorism in following chapters.

Chapter 3: Integrative Clustering and Supervised Feature Selection for Clinical Applications

Chapter 3 proposes a framework based on unsupervised clustering method to guide supervised learning methods in the reduction of feature dimensions from large cohorts non-small cell lung cancer data for survival prediction to select most informative, representative and non-redundant feature set.

Chapter 4: A Differential Diagnosis Ability of Radiomic Analysis on Detecting the Bone Marrow Involvement Patients of Suspected Relapsed Acute Leukemia

Chapter 4 provides opportunities on how bone marrow biopsy for high incidence of leukemia patients can be superseded by incorporating unsupervised machine learning methods base on radiomics to evaluate the BMI.

Chapter 5: A Prognostic ability of Radiomic Analysis on predicting Pathologic Complete Response of Breast Cancer to Neoadjuvant Chemotherapy

Chapter 5 aimed at defining the neoadjuvant chemotherapy (NAC) postoperative pathologic outcomes by the radiomic original phenotype on machine learning methods. This study assessed whether radiomics can be a useful tool to adjuvantly predict the status of Pathologic Complete Response (pCR) in patients with locally advanced breast cancer. Here as well, the proportion of target variables were strictly controlled to conform to statistic needs.

Chapter 6: Discussion and Future Works

Chapter 6 concludes the MPhil thesis by tiding all radiomic research findings which have positive impacts on the advance of prognostic medicine and describes the further research of medical image analyses and applications.

Chapter 2. Radiomics Methodology and Applications

2.1. Radiomics and Applications

Radiomics is a novel field that automatically extracts numerous and underlying imaging characteristics from standard medical images by a high-throughput quantitative method to be applied on clinical decision support systems [40, 42]. Radiomics can be applied to any cancer-related imaging, because it contains a large number of potential information from medical images. Some medical images of requiring to delineate the Volume of Interest (VOI) include CT imaging, PET imaging and MRI imaging, and others molecular imaging need not delineation, such as pathological imaging, protein imaging and gene imaging. Such related cancer imaging can be explored using radiomics automatically and be repeatedly converted into high dimensional quantitative tumor intensity, shape and texture features [11, 43]. Then the optimal feature selection methods and machine learning models are required objectively characterize data to construct the predictive models of disease progression. Radiomics not only provide a quantitative way to understanding tumor phenotype but also discover the great potential for application to precision medicine. For example, radiomic features predict prognostic responses, distinguish diagnosed tissues or tumors, and can assess cancer genetic differences in many cancer types.

Predict prognostic responses: Multiple studies have recorded and verified the prognostic value of radiomics at multiple clinical endpoints. Aerts et al. stated that radiomics can be an effective tool for predicting the response phenotype of patients with NSCLC using gefitinib (an EGFR inhibitor) [44]. Grossmann et al. investigated and confirmed radiomics excellent predictive values based on 291 MRI patients with recurrent glioblastoma treated using bevacizumab [45]. The predictive value of radiomics for postoperative response has also been confirmed by many studies. The extraction of massive radiomic features from pre-treatment MRI or PET-CT images has shown the feasibility to predict pCR status of breast cancer patients after receiving NAC [46-48]. Braman et al. combined the structural ability of radiomic extracted from in the intratumoral and surrounding microenvironment to predict pCR, and the predicted performance was better than that of purely tumor features. the differential prediction based on breast cancer subtypes also reflects the potential applications of tumor heterogeneity [46]. Although several conventional features in PET images, such as SUVmax, SUVmean, and Metabolically active Tumor Volume (MTV), can reflect tumor metabolic behaviour and prognosis-related responses, the over simple Standard Uptake Value (SUV) measurements are not sufficient to reflect the distribution of heterogeneity [49-52]. While radiomics, as a unique tumor phenotype is driven by genetic and biological variation, was confirmed to be superior to traditional PET values in some studies to evaluate its prognostic ability. For example, radiomics predicts the Postoperative response of 20 patients with pre-treatment esophageal cancer after RT. The results show that the performance of the selected radiomic features fitted model is clearly superior to models constructed only using conventional PET features [53]. In addition, radiomic features can be utilized to assess tumor metastasis. Coroller et al. extracted 635 radiomic features from CT images of 182 lung adenocarcinoma patients to identify 35 important features that can predict distant metastases, especially effective for early treatment of patients [54].

Distinguish diagnosed tissues or tumors: Radiomics is very helpful in distinguishing cancer staging. The utility of radiomic features extracted from MRI, pathologic imaging or other medical imaging can be as a multi-parameter effective prediction of glioma stage [55, 56]. In a study for bladder cancer staging prediction, the texture features extracted from MRI based on Diffusion-Weighted Image (DWI) and Apparent Diffusion Coefficient (ADC) can well reflect the difference between early and late-stage bladder cancer; the GLCM features calculated from the ADC are strongly correlated especially with the bladder cancer stage [57]. The use of radiomics to distinguish tumor stages can avoid the risk of invasive surgery and enact accurate individualized treatment for patients.

Tumor tissues in many diseases can also be identified by radiomics. The extracted Gray Level Co-occurrence Matrix (GLCM) texture information from T1 and T2-weighted MRI images can distinguish between brain tumors, edema, necrosis and other tissues [58-60]; In particular, the GLCM features in Three-Dimensional (3D) images are significantly better than Two-Dimensional (2D) [61]. Effectively distinguishing between benign and malignant lesions of the tumor is also one of the capabilities of radiomics. Based on PET-CT texture information diagnoses, the uptake of malignant tumors is more uneven than that of benign tumors [62]. More importantly, the radiomic features (intensity, shape, and texture) predict benign and malignant tumors more accurately than visual analysis by several experienced radiologists [63].

Assess cancer genetic differences: Many studies have shown that imaging features are the expression of tumor genetics and biology, hence because of this causal relationship there are also close connections between them [40]. The linkage between MRI imaging features and three genes (MmmaPrint, Oncotype DX, and PAM50) was validated in a study for predicting breast cancer recurrence. The experimental results indicated that the phenotype of effective diameter has strong correlations with those three genes [64]. In MRI and 3D ultrasound fusion images of prostate cancer, Stoyanova et al. found strong correlations among radiomic features with the genes for immunity, metabolism, and biological adhesion [65]. To this end, various biological mechanisms can be described by different radiomic features.

What the previously mentioned studies have in common is that they discover the connection of quantitative radiomic features with clinical biological information to construct predictive models. The core of radiomics is providing potentially up-to-date tumor phenotypic data and then complementing it with existing clinical and molecular data, such as patient history, tissue biopsy and gene test, to support the clinical decision system [43].

Compared to traditional clinical methods, radiomics has four main advantages that may complement the challenges and shortcomings mentioned in Chapter 1. First of all, radiomics is a low-risk and non-invasive way to detect disease and avoid unnecessary invasive treatments. Secondly, radiomics is a repeatable and objective measurement approach that can be used to assist professional diagnosis by different feature extraction algorithms. Thirdly, radiomics is a comprehensive and quantitative measurement method that may overcome the obstacle of tumors’ heterogeneity as it considers the feature differences across the whole region. Finally, radiomics is not essentially creating new data, but rather the translation and interpretation of existing and untapped data for its biological foundation.

2.2. Radiomics Procedure

The process of radiomics can be categorized in five steps as follows: image acquisition, image segmentation, feature extraction, feature selection and integrated analysis.

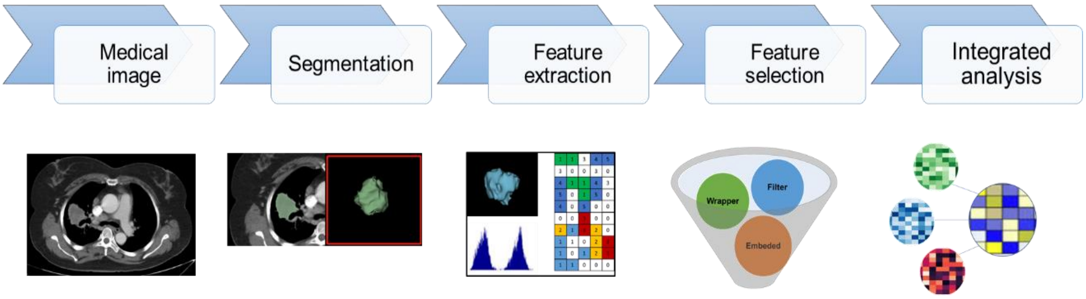


Figure 2-1. The process of radiomics. The First step is the acquisition of radiographic imaging modalities from medical images, such as PET, CT and MRI. Then using segmentation algorisms, we delineate the region of interest in 3D. On the basis of the VOI masks we quantitatively extract radiomic features from images automatically. Next, adopting suitable feature selection methods we select most representative features from different modalities. Finally, we integrate multiple sources of data further analyze and explain radiomic biological foundation.

2.3. Image Acquisition

In conventional clinical image acquisition, differences in various imaging parameters, such as pixel size, matrix size, slice thickness, tumor position, and reconstruction algorithms, need to be considered. When meaningful data is extracted digitally to analyze the image phenotypes, discrepancies in these imaging parameters may lead to some changes that are not due to potential biological effects [43]. It is therefore necessary to use standardized imaging protocols to eliminate the variability of unnecessary confounding factors [44].

2.4. Segmentation

How to accurately segment the VOI is a vital step in radiomic analysis. Experts can flexibly and reproducibly delineate different anatomical regions in various imaging by manual segmentation, but this process is very time-consuming and labor-consuming, and not suitable for radiomic analysis of huge dataset. Hence, many automatic and semi-automatic segmentation algorithms are proposed as much as possible to reduce the operator interaction of the segmentation process, improve time efficiency, and determine the exact boundaries [66].

2.4.1. Image Segmentation Fundamentals

Segmentation algorithms divide the image according to features such as grayscale, color, texture, and shape, making the inter-regions distinct and the intra-regions are similar. The main six main categories for segmentation methods: 1) threshold-based segmentation [67]; 2) edge-based segmentation [68]; 3) region-based segmentation [69]; 4) graph theory-based segmentation [70, 71]; 5) energy functional-based segmentation [72, 73]; 6) neural network-based segmentation [74, 75].

Threshold-based segmentation method calculates one or more grayscale thresholds based on the grayscale values of the image and solves the optimal grayscale threshold according to a certain criterion function. However, this type of method has poor connectivity to the target area and is susceptible to noise [67]. Edge-based segmentation is edge detection based on greyscale. The reflection of the local feature discontinuity of the image reflects the mutation of image characteristics such as grayscale, color, and texture. This method can suppress noise interference but the accuracy of positioning edge is not high, especially for images with uniform gray-scale changes [76]. Region-based segmentation is an important reference for the similarity between adjacent pixels, so that pixel points that are close in spatial position and have similar gray values are connected to each other to form a closed contour. However, although such algorithms are highly efficient, they are highly susceptible to noise and create unavailable areas. Graph-based segmentation removes a particular edge and divides the graph into several subgraphs to achieve segmentation. Commonly used methods are GraphCut, GrabCut and Random Walk. Such methods may result in edge penalty due to lack of color distinction and insufficient contrast at the edges [77, 78]. The energy functional-based segmentation method uses a continuous curve to represent the target edge and defines an energy function such that its curve forms the contour of the target. A commonly used method is parametric active contour model (snake) [79]. This algorithm requires significant human input, so it takes longer to calculate time [66]. Neural network-based segmentation is to train a multi-layer perceptron to obtain a linear decision function, and then utilize the decision function to distinguish the pixels to achieve the goal of segmentation [74, 75]. Although this method is very efficient, it requires a lot of data, constructing the model is very slow and complex, and the segmentation accuracy is strongly related to the amount of data. In addition, for some specific tumor locations, such as the pleural wall or mediastinum, the segmentation performance is poor [66].

There is no universal segmentation algorithm that can be applied to all medical images. The setting of the parameters, the choice of the segmentation algorithm, the imaging method of the region of interest, etc., the results of the segmentation will be completely different. Therefore, it is necessary to evaluate the performance of the segmentation algorithm.

2.4.2.Segmentation Evaluation

Evaluating the accuracy, usability, and consistency of the segmentation method is to compare the segmentation results with the results of professional manual delineation as a Ground Truth (GT). The evaluation methods can be divided into spatial overlap-based and shape-based dissimilarity. Spatial overlap-based methods are Dice's Similarity Coefficient (DSC) and Jaccard index.

DSC is computed the spatial overlap ratio between the segmented VOI and GT as

$$DSC(V_1, V_2) = \frac{2|V_1 \cap V_2|}{|V_1|+|V_2|} \tag{2.1}$$

where V_1 means the volume for using the segmentation algorithm, V_2 is manual delineation volume. The closer the DSC value is to 1, the more accurate the segmentation algorithm.

There is a certain relationship between Jaccard index and DSC, which is JAC = DSC (2 − DSC).

$$JAC(V_1, V_2) = \frac{|V_1 \cap V_2|}{|V_1 \cup V_2|} \tag{2.2}$$

The larger the JAC coefficient value means the higher the sample similarity.

Shape-based dissimilarity method is Hausdorff Distance (HD) defined as

$$HD(V_1, V_2) = max\left\{\underset{i\in s_1}{sup}\ \underset{j\in s_2}{inf}\, d(i,j), \underset{i\in s_2}{sup}\ \underset{j\in s_1}{inf}\, d(i,j)\right\} \tag{2.3}$$

where s_1 and s_2 indicate the boundaries of segmentation volume and manual delineation volume. sup means the largest subset element and inf the smallest subset element. d is the Euclidean distance from i to j. A high HD value indicates low accuracy for segmentation.

2.5. Feature Extraction

Once the defined segmented images have been determined, various imaging features can be prepared to be extraction. In extraction, it is necessary to discretize or quantify the intensity of the ROI to suppress the noise and conveniently calculate the texture features. There are two commonly used discretization methods as below:

Fixed bin number

$$X_{d,k} = \begin{cases} \left\lfloor N_b \frac{x_{gl,k} - min(x_{gl})}{max(x_{gl}) - min(x_{gl})} \right\rfloor + 1 & X_{gl,k} < X_{gl,k,max} \\ N_b & X_{gl,k} = X_{gl,k,max} \end{cases} \tag{2.4}$$

Fixed bin size

$$X_{d,k} = \left\lfloor \frac{x_{gl,k} - min(x_{gl})}{W_b} \right\rfloor + 1 \tag{2.5}$$

where $x_{gl,k}$ and $X_{d,k}$ is the discretised grey-level intensity before and after, respectively. $min(x_{gl})$ and $max(x_{gl})$ is respectively the maximum and minimum discretized gray-level intensity. N_b is bin number and W_b is bin width.

After re-segmentation and discretization of imaging intensity, the radiomic features can be extracted and broadly classified into first-order statistics, shape features, texture information and the features extracted by using different filters features [40].

2.5.1. First-order Statistics

The first-order statistical feature, also called Intensity, is the distribution of voxel intensities within the Region of Interest (ROI) by various basic features, such as the maximum, minimum, mean, range, kurtosis and skewness of gray level intensity. The homogeneity of the ROI can also be reflected from changes in some intensity features. For example, the uniformity is measuring the sum of squares for each gray value. A larger value means a stronger homogeneity or a smaller range of discrete intensity value [80]. The function is defined as below:

$$Uniformity = \sum_{i=1}^{N_g} p(i)^2 \tag{2.6}$$

where N_g is the number of discretized grey level intensity in the imaging mask of VOI. $p(i)$ is the normalized intensity histogram.

2.5.2. Shape Features

The shape feature is based on the edge, angle, ridge, etc. of the ROI to calculate the extracted 2D or 3D geometric features on the non-derivatized image [80]. The calculation of the surface area, compactness, spherical ratio, etc. of the tumor area can reflect the burr, volume and density of the tumor. For example, the value for compactness be related to tumor density, a higher value implies the greater density of ROI, the value is calculated as:

$$Compactness = \frac{V}{\sqrt{\pi A^3}} \tag{2.7}$$

where V is the volume of ROI and A is the surface of ROI.

2.5.3. Texture Features

The texture feature, also called the second-order statistical feature, is a statistic for correlation property of the gray space in the ROI. Commonly used matrices include GLCM, Gray Level Size Zone Matrix (GLSZM), Gray Level Run Length Matrix (GLRLM). Neighbouring Gray Tone Difference Matrix (NGTDM) and Grey Level Distance Zone Matrix (GLDZM) (equal with Gray Level Dependence Matrix (GLDM)) [80, 81]. These kinds of matrices are methods for describing spatial pixel differences by studying the spatial correlation properties of gray scales and thus are the most capable of expressing the inhibition between different parts of the tumor. For example, when the GLCM cluster prominence value is large, the symmetry of the image after decomposing the original image at low frequency can be described. When it is small, it can detect the difference of the region where the gradation value does not change obviously, such as a tumor with unclear boundaries [82]. The function as:

$$Cluster\ prominence = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g}\left(i + j - u_x - u_y\right)^4 p(i,j) \qquad (2.8)$$

where N_g is the number of discretized grey level intensities in the imaging mask of VOI. u_x and u_y be the average intensity of p_x and p_y. $p(i,j)$ is the normalized co-occurrence matrix.

2.5.4. Other filters features

Different filters are used to pre-process the image to extract more radiomic features that may reflect more valuable information than the original features. These filters include wavelet and Laplacian of Gaussian (LoG) filters, and some simple filters such as square, square root, logarithmic and exponential filters [81]. LoG transformation describing metrics, capturing texture features at different coarse levels, especially for areas with unclear boundaries have exceptional performance. Wavelets reveals more valuable subtle characteristics on the basis of the original features by eight different degradation domains.

2.6. Feature selection

From the extracted massive high-dimensional feature set, not all features should be preserved. There are a large number of irrelevant features and redundant features. In the process of dimensionality reduction, there is often the possibility of losing important features. To this end, it is necessary to consider not only the correlation of features in a single data category but also the relevance of features between different data sets. Methodologies to reduce the computation time, improve the prediction performance and select most representative, stable, repeatable pattern is the purpose of feature selection [83-85].

Generally, feature selection can be summed up into four categories: filter method, wrapper method, embedded method and other selection method. Methods that to eliminate redundant and irrelevant features rely primarily on filter and wrappers.

2.6.1. Filter Method

The filter method is to discard the lower-ranked features by pre-processing the ranking features. The criterion of the filter method is to sort the variables according to certain rules. The rules may be based on correlation or mutual information between variables， And determine appropriate threshold criteria to retain variables above the threshold. Hence， filter methods could suppress the least interesting variables. Since filter methods only consider general proxy metrics rather than complex error rates for evaluating features, it is particularly effective in computation time and robust to overfitting (Figure 2-2) [86]. Common filter methods are Mutual Information (MI), Correlation matrix, Relief (RELF) and Significance tests for multivariable data.

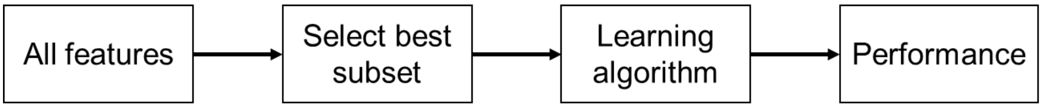


Figure 2-2. Filter method for feature selection

Correlation methods as threshold criteria play an important role in determining independent features, and correlation coefficients can be used to determine which features and target variables are irrelevant. The common types of correlation coefficients are Pearson, Spearman and Intra-class.

Pearson correlation [87] is an assessment for linear correlation between two variables. The calculated method as below:

$$R_{pearson} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} \tag{2.9}$$

Where n is the size of sample, $\bar{x}$ and $\bar{y}$ is the mean of variable X, Y, x_i and y_i are the individual points for X, Y, the range of $R_{pearson}$ from -1 to 1. Where the further the value is away from 0, the more independent the two variables are.

Spearman correlation [87] is a rank variable from Pearson correlation measuring the monotonic relationship between variables. The function was obtained as:

$$R_{spearman} = 1 - \frac{6\sum_{i=1}^{n} d_i}{n(n^2-1)} \tag{2.10}$$

where d_i is the difference for the ranks of variable X and Y, n represents the number of observations.

Intra-Class Correlation (ICC) [88] describes the degree of similarity of units in the same group. The function is as follows:

$$R_{intraclass} = \frac{K}{K-1} \cdot \frac{N^{-1}\sum_{n=1}^{N}(\bar{x}_n - \bar{x})^2}{s^2} - \frac{1}{K-1} \tag{2.11}$$

where K is the number of variable values in each group, and N is the number of paired data values, $\bar{x}_n$ is the mean of sample in a specific group.

Another threshold criterion is a MI sorting criterion based on the measure of dependencies between variables [89]. The MI between the two discrete variables is given by:

$$I(X;Y) = \sum_{x \epsilon X} \sum_{y \epsilon Y} p(X,Y)(x,y) \log\left(\frac{p(X,Y)(x,y)}{p_X(x) p_Y(y)}\right) \tag{2.12}$$

where $p(X,Y)$ is a joint probability density function for two variables. p_X, p_Y are respectively the marginal probability mass function for X, Y. If the variables are continuous, the MI can be obtained by integral replacement summation.

The filter methods need to modify its own data structure to match the appropriate learning algorithm. Since the important features themselves have less information and are easily discarded when combined with other modal features [83, 90], it is necessary to rely on classifiers (wrapper method) to obtain a feature subset.

2.6.2. Wrapper Method

The reference standard for feature selection in the wrapper method is mainly based on the predictive results. According to the multiple evaluations of the internal algorithm, find the subset with the highest prediction performance. It uses a predictive model to score feature subsets. Each new subset is used to train an independent model, which is tested on a hold-out set and then using the error rate of each model to evaluate subsets (Figure 2-3). Wrapper methods need consider the possible interactions between each variable and construct probable feature space. Hence it can be computationally expensive and lead to the overfitting risks to the model.

The wrapper method can be broadly classified into sequential selection algorithms and heuristic search algorithms [85].

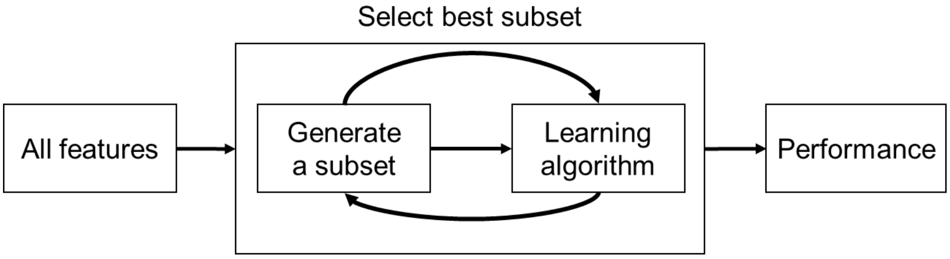


Figure 2-3. Wrapper method for feature selection

The frequently used sequential selection algorithms are Sequential Backward Selection (SBS) and Sequential Feature Selection (SFS). Common heuristic search algorithms are Bayesian (BY), Random Forest (RF) and Support Vector Machines (SVM). We choose a representative method from each type of method to explain the principle.

SFS starts with the feature set being empty, and iteratively selects features with the best univariate prediction score under some objective function. It then adds the remaining features one by one to the current subset and selects the feature with the best classification effect as a new subset. This process is repeated several times and the best feature set can be selected (Figure 2-4). Although the prediction effect of this method is better, the selected feature set may have redundancy problems [91].

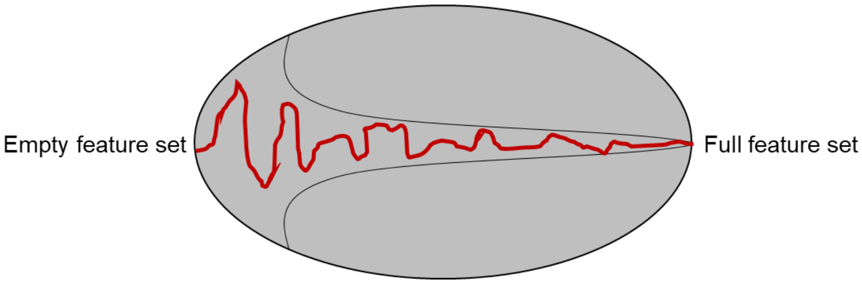


Figure 2-4. The SFS for wrapper method

RF [92] as a heuristic classifier is essentially composed of multiple decision trees. Its decision category is determined by the output mode vote of the subset trees. The original datasets can be randomly separated into a training set and test set by a pre-set ratio. Then the pre-set training set will be replaced by random bootstrap multiple times as the new training sets of decision tree. Then using the extracted cases as predictions evaluate the error values. Different training set will be observed separately for binary decisions and then discover the best split method. The final prediction comes from the vote of all of the decision trees (Figure 2-5). The main advantages of RF are: 1) rapidly distributed processing of high-dimensional data; 2) providing internal importance for interpretation; 3) strong model generalization; 4) balance error capability on unbalanced data set. Therefore, the discriminative ability for RF is often superior to other machine learning methods, especially when applied to high-dimensional albeit small data sets [93].

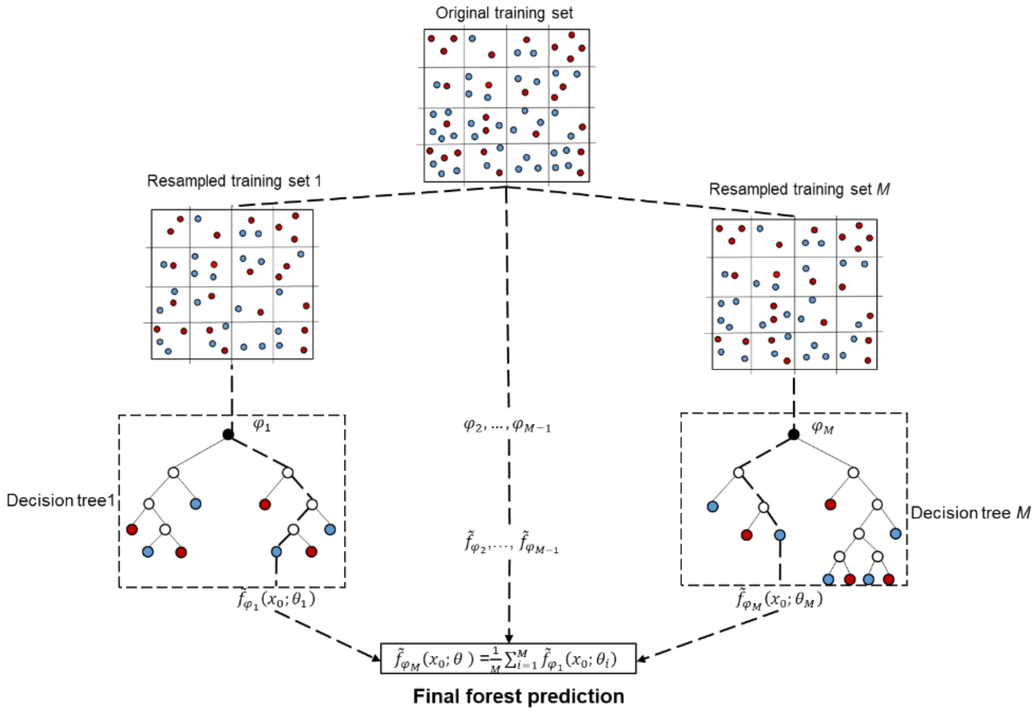


Figure 2-5. The decision process of RF model

As the sample size increases, the number of calculations of the model of the wrapper method also increases exponentially. In addition, the model with poor generalization ability also has the problem of feature redundancy [94]. Therefore, the filter method is often used as a pre-processing of wrapper method to overcome the shortcomings of the other model methods [95]. This is also the main philosophy of the embedded model.

2.6.3. Embedded Method

The main aim of the embedded method is the reduction of the calculation time it takes to classify unnecessary feature sets in wrapper methods and to avoid feature redundancy [96, 97]. Variable selection as part of the training process in embedded method, this can be seen as a combination of a filter and wrapper (Figure 2-6). One of the most well-known embedded methods is Least Absolute Shrinkage and Selection (LASSO).

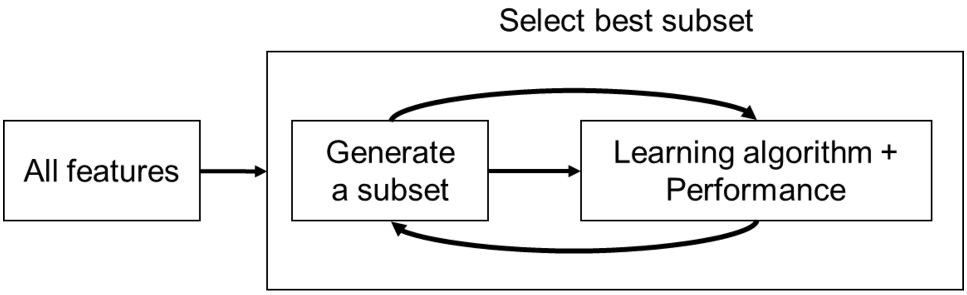


Figure 2-6. Embedded method for feature selection

LASSO [98] is a compression estimation that refines the model by constructing a penalty function and compresses mandatorily specific regression coefficients within a certain region and then selects a variable with a coefficient of zero. This improves the prediction and interpretability of the regression model. Mathematically, it has very simple and elegant expression as follows：

$$\hat{\beta}^{lasso} = \underset{\beta}{arg\,min} \sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_{j=i}^{p} x_{ij}\beta_j\right)^2 \; subject\ to \;\; \sum_{j=1}^{p}\left|\beta_j\right| \le t \quad (2.13)$$

Where x_{ij} is the i^{th} row and the j^{th} column of the covariate matrix X. $x_i \epsilon R^d$ is d dimensional vector. $\beta \epsilon R^d, \beta_0 \epsilon R$ are required optimized parameters. $y_i \epsilon R$ is the outcome. t is a predetermined free parameter used to assess the degree of regularization. The smaller the t value, the stronger the compression of the estimated parameters and the better the selection effect.

2.6.4. Other Selection Methods

The above three types of methods are all supervised learning feature selection techniques. In some applications, some unsupervised learning methods assist feature selection to improve the reliability and fairness of the selection methods assisted feature selection [99]. Clustering algorithm is one of the main unsupervised learning methods. It groups the data by some hidden structure of unmarked data, and divides these samples into several disjoint subsets. There are many clustering algorithms, such as k-mean, hierarchical and spectral clustering. These methods can be density-based, partition-based, grid-based or model-based. The different clustering methods have different grouping effectiveness, so the method of consensus clustering is needed. Consensus clustering refers to repeatedly voting clusters on existing data sets that classify high-dimensional feature sets into highly correlated classes, and finally finding the most stable and reliable clustering set [100]. Define the consensus clustering C^* as:

$$C^* = \underset{C}{arg\,min} \sum_{i=1}^{n} d(C, C_i) \tag{2.14}$$

where $C_1, C_2, \ldots, C_i$ are each clustering results, d is the distance measure to control the median partition problem.

The principle to automatically compute the number of clusters is based on the consensus cumulative function (CDF) and a specific clustering algorithm [101]. The criterion for the optimal number of clusters is decided by the Proportion of Ambiguous Clustering (PAC), defined as measure that the maximum value of the ratio of the area difference between the CDF curve and the horizontal axis of three consecutive clusters (Figure 2-7).

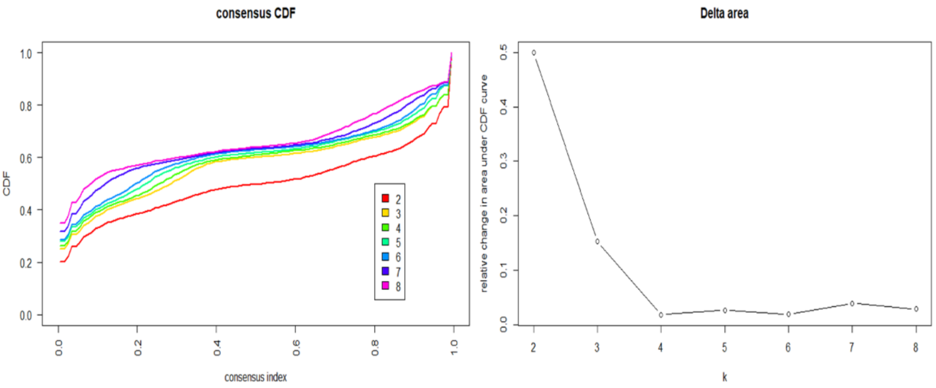


Figure 2-7. The Proportion of ambiguous clustering

The selection criteria for clustering method is similar to the filter method as both consider the correlation between variables. Therefore, the clustering method is also often used as a pre-processing method for other feature selection methods to quickly determine the redundant features.

2.7. Integrated Analysis

After extracting and selecting representative quantitative features, a crucial step is to establish and analyze predictive models to find the link between radiomics and its underlying biological information. Such models can be divided into two categories for different clinical purposes. One type is the prediction model for classification. The output variables of the model are discrete variables. These models can be used to predict tumors staging, tissue identification and treatment plan, etc. the other type is the regression prediction model. The output variables of such models are usually continuous variables. It is often used to predict the clinical endpoints of patients. However, many of the current machine learning model frameworks are black-box models, and the internal specific calculation prediction criteria are not clear, which makes the clinical trustworthiness of the model greatly reduced. To this end, a trusted machine learning training and analysis framework is urgent.

For the classification prediction, the model is interpreted using the Local Interpretable Model-agnostic Explanations (LIME) algorithm, which can show the details of the patient predictions and calculate the contribution of different image features [102]. For the patient's survival prediction, the nomogram is included in the survival analysis system, which can score and visualize the importance of different features, greatly enhancing the interpretability and trust of survival analysis [103]. In addition, correlation analysis can be used to mine the connection between radiomics and clinical biological information. After the model training is completed, the model needs to be verified to prevent overfitting. Commonly used verification forms are cross-validation and external validation [104, 105]. In the result analysis, the discriminability of the selected radiomic features was evaluated using several parameters, which comes from a confusion matrix for a comprehensive assessment of feature effects [106]. Hence, the goal is to propose a comprehensive analysis framework that integrates radiomic imaging, molecular assays and clinical information to accurately support biological therapy.

Chapter 3. Integrative Clustering and Supervised Feature Selection for Clinical Applications

Widely used biomedical data offers new chances to discover knowledge for quality healthcare. However, how optimal selection of the most informative, representative and non-redundant features from the complicated datasets need to be better explored. To address this challenge, in this chapter we propose a framework for reducing the feature dimension based on unsupervised learning to guide supervised learning. This is a feasible method to use the clustering method as a pre-processing procedure for the other three types of feature selection methods (filter, wrapper, embedded), which can maximize the elimination of redundant features and the most effective dimension reduction. Here we performed survival predictions on pre-treatment data for non-small cell lung cancer, using two typical wrapper methods for survival analysis for comparison. Identify suitable selection criteria to ensure that important variables are not lost (as determined by specific data and score changes).

This chapter is organized into four sections: in 3.1 section, the description for solving feature selection challenges. In 3.2 section, we introduce two wrapper methods (univariate Cox and univariate Random Survival Forest, referred as uni-cox and uni-rsf) explanation in the part of informative feature identification. In 3.3 section is representative feature identification that a novel consensus clustering method combined with informative feature identification to select the representative feature set. In 3.4 section is adopting pairwise correlation analysis to reduce redundancy. In 3.5 and 3.6 section are the experiment results and discussion for each selection part.

3.1. Research Motivation

With big data rapidly development, a large variety of valuable digital data are widely used in the areas of trade, energy, media [107], and in biomedical applications [108]. A mature machine learning model can be successfully used to improve the performance of diagnosis, prognosis and prediction in clinical medicine. However, there are often facing many problems of feature redundancy, feature irrelevance and noisy information. How to select the features that are highly informative (relevant to the output), non-redundant (not correlated with other features) and representative (efficiently representing the original data) [83] is critically vital for improving performance of machine learning models [109, 110]. To address these embarrassments , in this chapter, we propose an integrative unsupervised learning (consensus clustering) and supervised learning feature selection framework. In this framework, the outcome-driven of supervised parts aims to keep the most informative features, while the data-driven of unsupervised clustering part supplemented the supervised method with automatically grouping information based on hidden data structure correlation. By integrating both information from the supervised and unsupervised method, our method could select the optimal feature set, where features are highly relevant to the outcome and independent from each other.

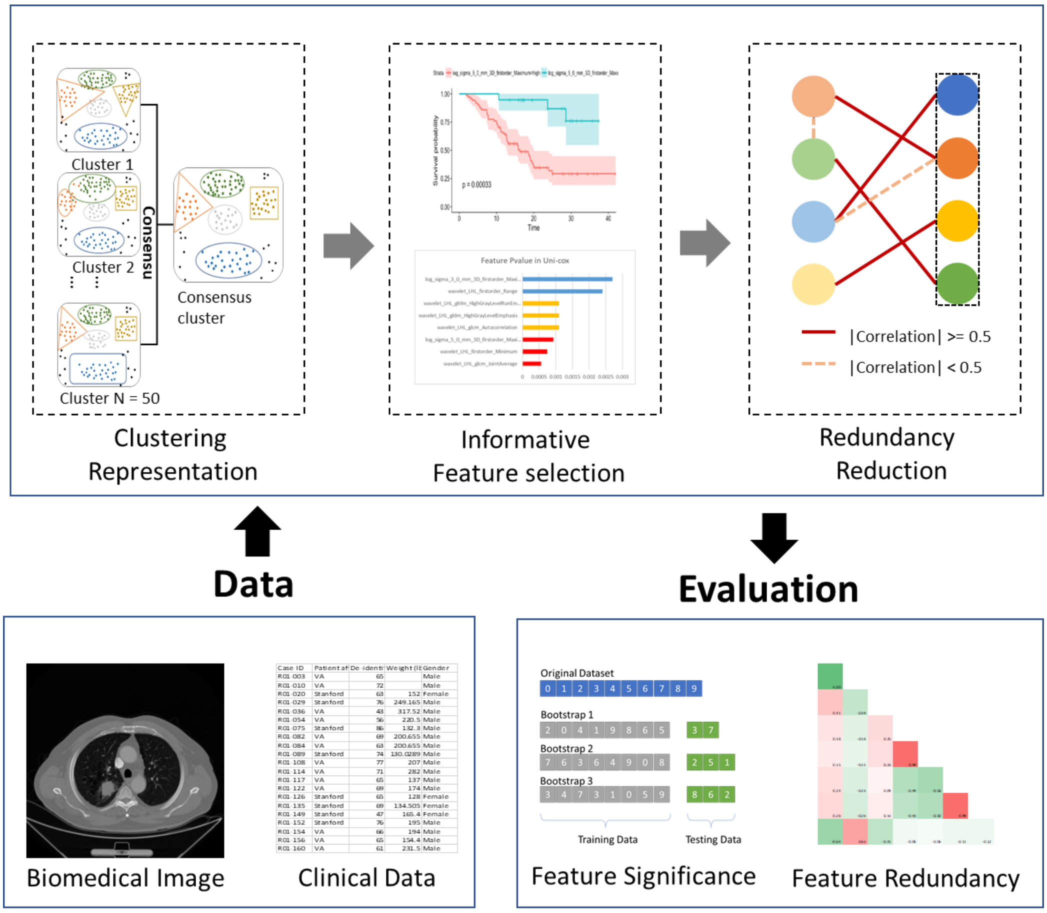


Figure 3-1. The workflow of the proposed framework

3.2. Informative Feature Identification

Informative features refer to the features with the excellent ability to predict the clinical outcome, i.e. predict time-to-event survival label. To identify and sort features, supervised feature selection approaches can be utilized due to its outstanding performance to find the feature relevance to the output. In this chapter, we adopt uni-cox [111] as our supervised feature selection approach, a classical wrapper classifier specialized for time-to-event prediction. The rank feature criteria of in Cox could have three ways including, P-value, hazard ratio value and C-index. We have already compared more detailed experiments for these three and decided P-value as rank criteria from each Cox model (CPH) to evaluate features statistical significance with a score test. The uni-rsf selection [112] method as a comparison to identify informative features based on the classical machine learning model RF. Although RF has the fame of high performance, the selection method often suffers overfitting problem and lack the capability of generalization.

The CPH model is one of the most general regression models as it does not base on any hypothesis of behind survival distribution. The CPH utilizes hazard function as a response, which can be shown as:

$$h(t, x) = h_0(t) exp(\beta' x) \tag{3.1}$$

where h0 means an arbitrary baseline hazard and $\beta = (\beta_1, \ldots, \beta_p)'$ is an array of unknown regression coefficients. As h_0 is independent of covariates, the regression coefficients β can be gained by maximizing the partial log-likelihood as:

$$L(\beta) = \prod_{i=1}^{k} \frac{exp(\beta' x_i)}{\sum_{l \in R(t_i)} exp(\beta' x_l)} \tag{3.2}$$

where k denotes distinct ordered survival times while x_i denotes the covariate value with distinct ordered survival times.

In our method, we applied univariate CPH model on univariate features to sort the importance for each feature. The importance value is P-values gained by Wald test, describing the features statistical significance. The univariate Random Survival Forest (RSF) model uses Bootstrap sampling to split nodes and recursively generate a total classifier as a CPH comparison method. Then, informative feature rank ascendingly by sorting the P-values for CPH and feature importance for RSF, where features with low P-values and high importance are selected as informative features.

3.3. Representative Feature Identification

Representative features are defined with the capability to describe its hidden structure grouping and predict the label, as well as with low redundancy. Unsupervised clustering is used to reveal dataset hidden structure based on inherent feature information without labels. The clustering method votes the whole feature space into several highly correlated feature clusters and gives each feature with a fixed number. Then, the representative features can be preserved from each cluster according to the informative feature rank gained by supervised feature selection. Hence, we significantly reduce the intra-cluster redundancy in this part.

Consensus clustering [101] is adopted in our framework as it can automatically decide the number of clusters within the PAC algorithm and utilize multiple times resampling to ensure a stable, robust and scalable clustering result. This clustering approach ascertains the final clusters utilizing the consensus across multiple runs of one specific clustering method, similar to ensemble learning in supervised learning. To adapt the consensus clustering algorithm, the data need pre-processing, such as feature omitting and standardization. Subsequently, the number of clusters can be acquired with the criteria which give the highest median cluster consensus on all clusters. Cluster consensus is described as the average consensus between all pairs of features from a single cluster.

3.4. Redundancy Reduction

Redundant reduction refers to as fewer feature sets as possible with minimum internal correlation within the subset. Although redundant features have been dramatically reduced by highly correlated feature clusters in representatives feature selection part, there still persists highly correlation among feature representatives for inter-cluster. Therefore, we further utilized pairwise correlation analysis to avoid redundancy.

The pairwise correlation analysis evaluates the redundancy by building a Pearson Correlation Matrix (PCM) for each pair of features in the representative subsets. Pearson correlation is a strict linear correlation measurement, specific calculated correlation function in formula 2.9 before we mentioned.

Because of the symmetrical matrix structure of Pearson correlation for every two variables,

The PCM can purely compute the symmetrical upper or downer triangle from the matrix. We remove the redundancy following below criteria: 1) Determine the feature pairs with correlation coefficients larger than a threshold (the value depends on the dataset structure and label differences). 2) Keeping the one with larger feature informativeness (smaller P-value) in each pair.

3.5. Evaluation Methods

The proposed feature selection approach was assessed with feature significance and redundancy. The feature significance was evaluated by feeding the selected features into the Cox survival model and assessing its performance by Concordance index (C-index). C-index is a professional measure for assessing survival model efficiency, the generalization of the Area Under the Curve (AUC) for continuous survival data. The range of C-index is from 0 to 1 where 1 means the best accuracy, vice versa. In this research, apparent C-index and cross-validated C-index are both computed. The former was training and testing features capabilities on the same data while the latter is training and validating features on random resampled data with bootstrapping strategy to overcome the overfitting risks [113].

The feature redundancy is evaluated with the average correlation coefficients, defined as the mean coefficients in upper or downer PCM:

$$CM = \frac{1}{N}\sum_{1}^{N} r \tag{3.3}$$

where r means the Pearson correlation computed from formula 2.9 and N is the count of feature pairs in the simplified PCM.

3.6. Experimental Results

3.6.1. Data Description

We assessed the proposed feature selection framework on two distinctive dimensional datasets. The first one is a complicated CT image feature dataset while the other one is a lower-dimensional clinical tabular dataset.

- The CT image features come from 97 NSCLC late-stage patients including 1045 radiomic features extracted from pretreatment CT images, including 13 morphological features, 18 first-order features, 68 textural features, 258 LoG features and 688 wavelet features [40]. Because of its high-dimensional data structure, the feature selection will be more difficult as it requires higher time complexity and space complexity.
- The clinical tabular dataset includes 14 demographic and 20 immune features. The dataset has pre-processed with missing value handling because the amount number of none values list in the dataset. If the data missed over 40% of the total data, the whole columns will be discarded.

3.6.2. Assessment of Selected Representative Features

The first experiment was used to evaluate the feature significance and redundancy of preserved representative features. These features were gained by keeping different number of informative features from each cluster. Next, we compared the C-index of each feature set with that chosen with state-of-the-art survival prediction methods when feature number was same. Figure 3-2 compares the discriminative ability of selected representative sets and features selected with state-of-the-art methods. The graph shows the highest performance is achieved with the feature sets selected by our proposed method, achieving 0.64. When the selected number of features increase too large, the further inter-cluster redundancy reduction is needed.

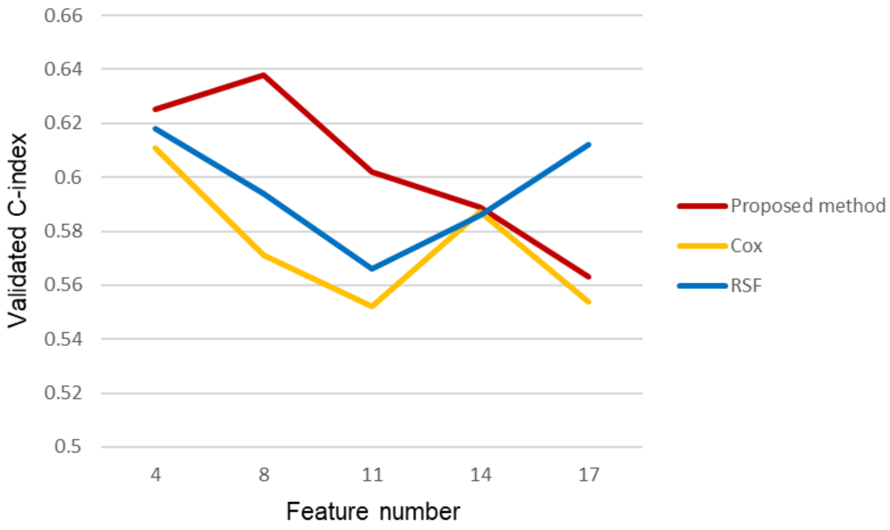


Figure 3-2. Feature significance of representative features

Figure 3-3 illustrates the feature redundancy of selected feature sets with average correlation coefficients. The result indicates that persevered representative feature sets have overall lower internal correlation compared with feature subsets selected by other two methods, averagely lower than 0.4.

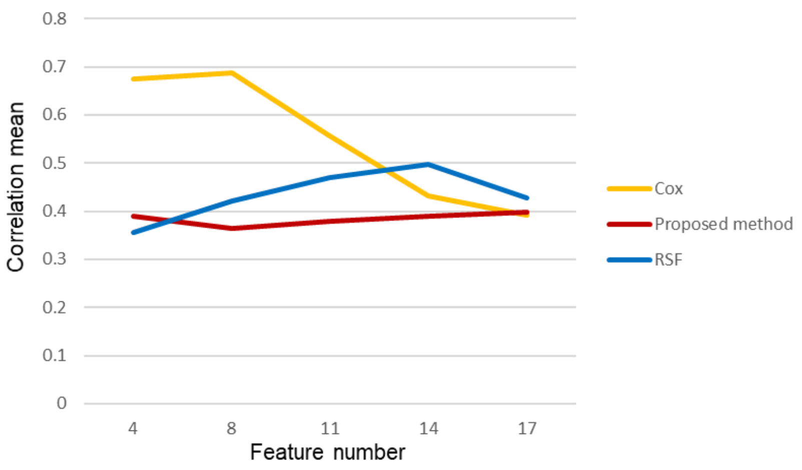


Figure 3-3. Feature redundancy of representative features

3.6.3. Assessment of Non-Redundant Representative Features

The second experiment was utilized to assess the feature significance and redundancy of kept non-redundant representative feature sets after redundancy further reduced. The non-redundant representative feature set can be gained by further reducing inter-cluster redundancy. The coefficient threshold to select high correlated features was set as 0.5. Figure 3-4 compares the performance and feature number of selected subset before and after inter-cluster redundancy reduction when different numbers of features are selected. The result represents the sets of redundancy reduction have a smaller number of features but even larger discriminative ability. Furthermore, the performance of selected non-redundant representative features has gradually better stability no matter how many features selected from a cluster.

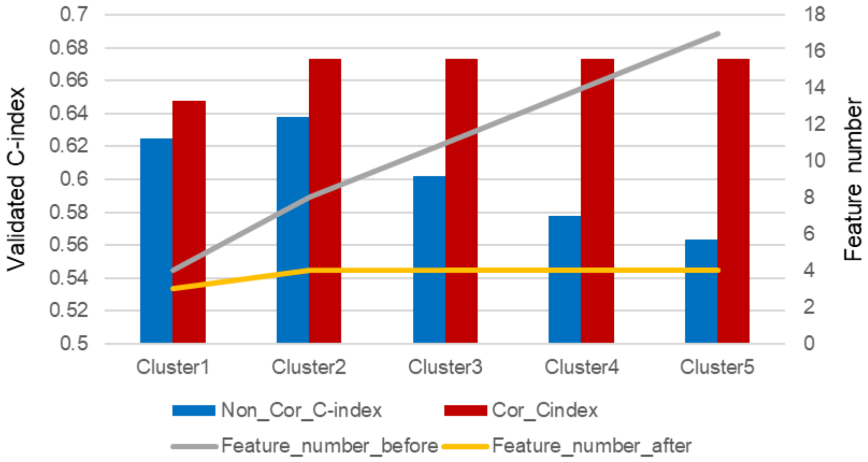


Figure 3-4. Feature significance of non-redundant representative features

Figure 3-5 compares the average correlation coefficients and feature number of selected subset before and after redundant reduction. The result illustrates inter-cluster redundancy reduced sets have dramatically less redundancy and feature number. To conclude, the final result by our proposed framework selected four non-redundant representative features achieving 0.673 while the kept four selected features are completely independent.

Figure 3-5. Feature redundancy of non-redundant representative features

3.6.4. Assessment of Proposed Method on Data with Different Dimension

The third experiment was designed to verify that our framework worked well on both CT image feature set and clinical dataset. The result in Figure 3-6 shows that proposed approach obtained superior results when compared with RSF and Cox selection on the two datasets, reaching 0.685 and 0.673 respectively for two datasets. Moreover, the experiment result also illustrates the proposed framework has a greater advantage on the complicated imaging feature datasets.

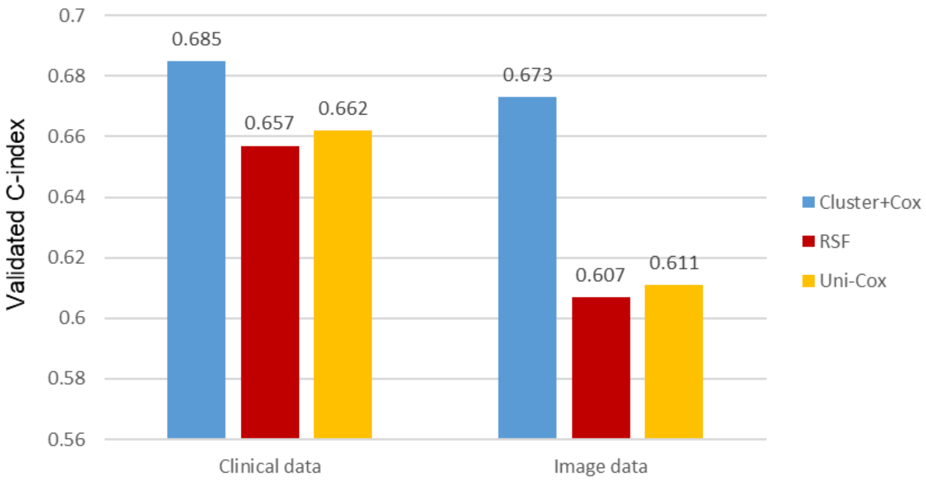


Figure 3-6. Assessment of the whole framework on different datasets

3.6.5. Assessment of Overfitting

The last experiment was intended to show that further inter-cluster redundancy reduction successfully avoided the overfitting risk. We compared the apparent C-index and validated C-index of the representative feature subsets before and after further redundancy reduction, as illustrated in Figure 3-7. Before redundancy reduction, the apparent C-index increases continually with the feature number; however, the reduced validated C-index curve demonstrate that the high C-index obtained with a large number of features may be because of overfitting. After redundancy reduction, the gap between apparent C-index and validated C-index narrows, with C-index value stabilized at 0.699 and 0.673 respectively, regardless of the increase of the feature numbers, representing is successfully handled.

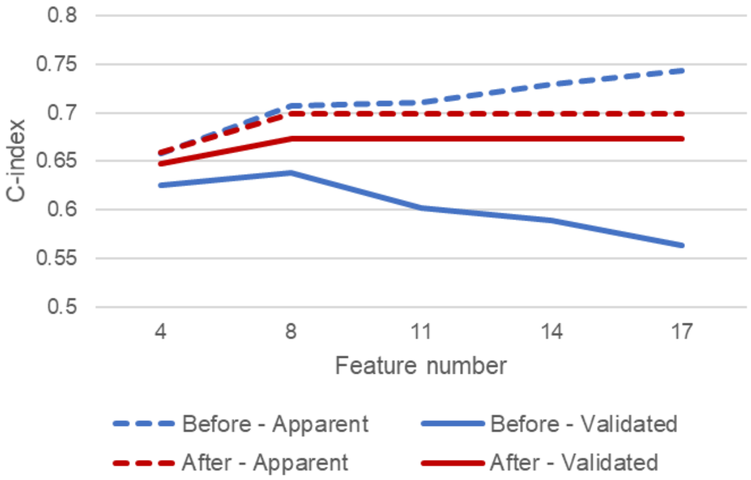


Figure 3-7. Assessment of overfitting

3.7. Chapter Summary

In this chapter, we proposed an integrative clustering and supervised feature selection framework to select informative, representative, and non-redundant features from both high-dimensional imaging feature set and low-dimensional clinical dataset. By comparing the selection of each part from we proposed framework with the conventional survival models, we prove that our model is completely superior to other methods. Our selected features had both strongly discriminative ability and lower internal correlation as well as suppression the overfitting problem.

Chapter 4. A Differential Diagnosis Values of Radiomic Analysis on Detecting the Bone Marrow Involvement Patients of Suspected Relapsed Acute Leukemia

This chapter focuses on how to use machine learning to assess the radiomics ability of differential diagnosis. 18F-FDG PET-CT imaging is utilized clinically for the diagnosis of extramedullary lesions in patients with relapsed Acute Leukemia (AL). However, the visual analysis of 18F-FDG diffuse bone marrow uptake in the detection of BMI in routine clinical practice remains challenging. How to use machine learning-based radiomics to assess whether invasive biopsy can be replaced for tissue identification, which plays a vital role in the diagnosis of diseases and avoiding harm to body. This chapter aims to improve and evaluate the diagnostic performance of 18F-FDG PET-CT in the detection of BMI for patients with suspected relapsed AL by analysis composed of three major stages, including feature extraction, important feature selection with model construction and model validation.

This chapter can be divided into five parts: the overview of introduction and motivation, data description, PET-CT radiomic analysis with machine learning, results, discussion and chapter summary.

4.1. Overview

AL is a hematological malignancy characterized by a dramatic increase in the number of immature blood cells. Despite the high rates of initial complete remission, relapse keeps a powerful clinical challenge and has become the main cause of failure in medical treatment [114]. Leukemia relapse can happen intramedullary or extramedullary. Patients usually suffer multiple BMB in the follow-up to records the intramedullary relapse [115]. However, BMB is an invasive way and only assess a small part of the whole bone marrow. ^{18}F-fluorodesoxyglucose positron emission tomography/computed tomography (^{18}F-FDG PET/CT) has been certified to monitor more missed extramedullary lesions than routine examinations [116-121].

The differentiate diagnosis of ^{18}F-FDG PET/CT-based BMI has not been completely assessed because of the imperfect systematic and immature related studies. From the accessible studies, most are case reports, we could conjecture that diffuse uptake is the main pattern [122-125], and its incidence is much higher than that in the lymphoma studies [126]. It is really difficult to decide whether diffuse uptake is BMI by visual analysis, because the judgment relies on the physician's experience, and both malignant and benign phenotypes may be similar [124, 127, 128]. About some lymphomatous bone marrow researches, diffuse uptake was evaluated as BMI negative [129, 130], while in other studies it was evaluated as BMI positive [131, 132]. Because of the high incidence of diffuse uptake in AL patients, it is not suitable to take diffuse uptake as positive or negative for BMI in patients with suspected relapsed AL. To conclude, the clinical ^{18}F-FDG PET/CT-based diagnosis of BMI in relapsed AL still face challenges.

Radiomics quantitatively extracted and mined considerable medical imaging features to reveal features that are impenetrable to the naked eye. It has been utilized in many solid tumors [133-135], while seldom utilized in bone marrow evaluation. A recently published study found that ^{18}F-FDG PET-based radiomic analysis was effective in detecting BMI [136]. We surmise that high-dimensional, high-throughput radiomic features from both PET and CT images would provide a comprehensive strategy for extracting the pattern of BMI, and thereby would be effective in improving the diagnostic capability of ^{18}F-FDG PET/CT in patients with suspected relapsed AL.

4.2. Dataset Description

4.2.1. Patients

The study was approved by the institution review board, and no written informed consent is required. This study retrospectively analyzed images of AL patients who received ^{18}F-FDG PET/CT at Peking University People's Hospital from January 2012 to February 2019. The collection criteria were: 1) acute myeloid leukemia or acute lymphoblastic leukemia patients who reached complete remission after induction chemotherapy, 2) the patient’s age are all great than 16, 3) clinically suspected recurrence, but not yet received treatment, 4) no chemotherapy or granulocyte stimulation-factor less than 30 days, 5) BMB has been accomplished within seven days. The simple statistics of collected patients are listed in Table 3-1. The patients were separated into two groups, one is 35 patients between January 2012 and February 2018 as training set and 6 patients between March 2018 and February 2019 as independent validation set.

Table 4-1. Demographic and clinical characteristics of patients.

Characteristics	Total number (*n*=41)	BMB positive (*n* =18)	BMB negative (*n* =23)	P-value
Age (years), median (range)	35.2 (17~75)	38.1 (18~75)	32.9 (17~49)	0.276
Gender (female/ male)	15/ 26	4/ 14	11/ 12	0.089
Leukemia subtype (ALL/ AML)	17/ 24	5/ 13	12/ 11	0.116
With extramedullary relapse/ without	24/ 17	11/ 7	13/ 10	0.767
Laboratory parameters				
WBC (G/L), mean (SD)	6.62 (4.70)	8.19 (6.27)	5.35 (2.38)	0.092
Hb (g/dL), mean (SD)	114.45 (23.07)	111.10 (21.53)	117.17(24.42)	0.427
ESR (mm/h), mean (SD)	38.33 (26.45)	37.17 (22.16)	39.50 (32.32)	0.887
CRP (mg/L), mean (SD)	13.013 (21.92)	18.32 (28.59)	7.08 (7.21)	0.199

ALL: acute lymphoblastic leukemia, AML: acute myeloid leukemia, WBC: white blood cell, Hb: hemoglobin, ESR: erythrocyte sedimentation rate, CRP: C reaction protein

4.2.2. Image Acquisition and Reconstruction Parameters

All patients fasted at least 6 h before scan, and the blood glucose level was controlled less than 8.3 mM (range 4.7~8.0 mM). 18F-FDG (provided by Atom high-tech Co., Ltd., Beijing, China) was injected intravenously with a weight-base dose of 5.55 MBq/kg (0.15 mCi/kg). After one hour (60 ± 5 min, range 54~63 min) 18F-FDG injection, the PET scan between the base of skull and the middle of the thigh was performed on a Discovery VCT (GE Healthcare, Milwaukee, Wisconsin, USA) with a 64-slice spiral CT. CT scan was firstly performed with a tube voltage of 140 Kev and a tube current of 80 mAs. The matrix size of CT was 512 × 512 and the voxel size $1.0 \times 1.0 \times 3.3$ mm^3. The PET images were collected in 3D mode for 2.5 min/bed and were revised for attenuation with a CT-based attenuation revise method. The PET images were reconstructed using an iterative algorithm (ordered-subset expectation maximization with 2 iterations, 28 subsets) and 6-mm full width at half maximum (FWHM) of Gaussian filter. The matrix size of PET was 128 × 128 and the voxel size $5.5 \times 5.5 \times 3.3$ mm^3.

4.3. PET/CT radiomic analysis with machine learning

As shown in Figure 4-1, the radiomic analysis constituted of three main stages. Firstly, based on the manual delineation of the VOIs from CT and then ascertained on PET, our model automatically extracted high-dimensional imaging features from both PET and CT VOIs; then important and discriminative features for pattern extraction were selected using harnessed correlation analysis and machine learning models; and finally, a machine learning-based prediction model was validated for the classification of BMB cases.

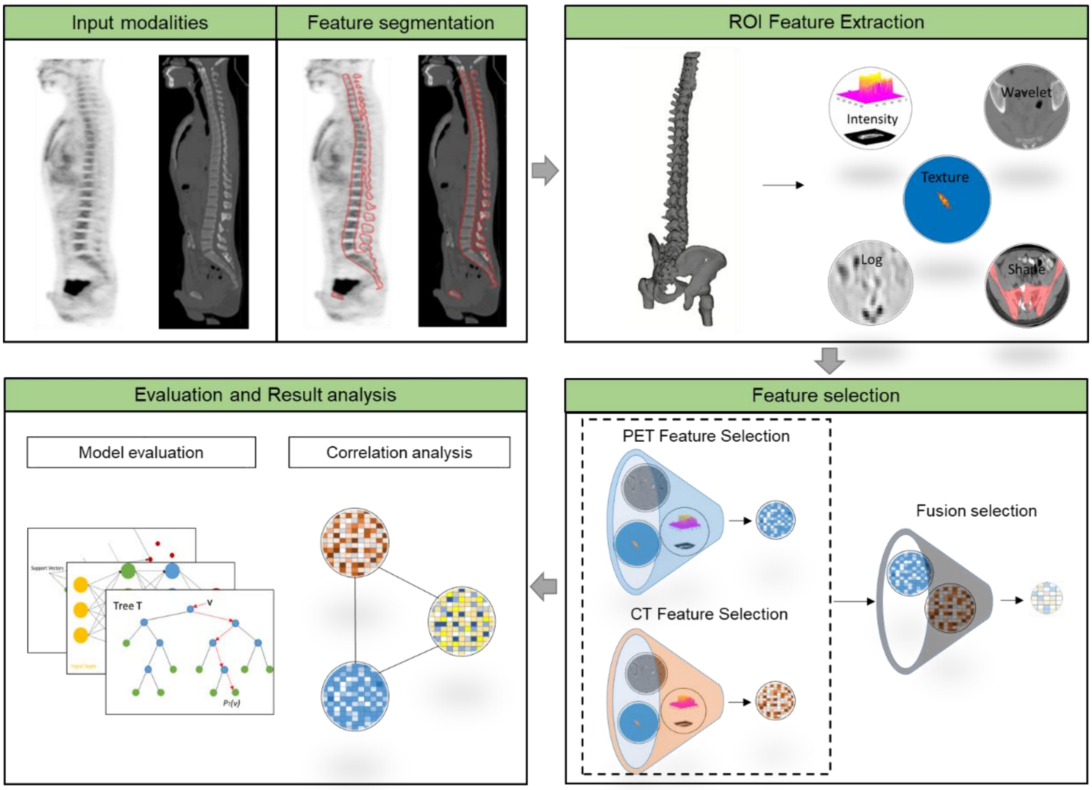


Figure 4-1. The flowchart of radiomics features extraction, selection and evaluation.

4.3.1. Radiomic Feature Extraction

The first stage was radiomic feature extraction. A semi-automatic procedure for axial skeleton VOI definition is described in a previous study which shows high reproducibility [136]. A software XD3 (Mirada Medical) was used for PET-CT image display and processing. The VOI including the spine and the pelvis was firstly determined by CT densities of Hounsfield units >130, and then all irrelevant bone areas were manually excluded. The final CT VOIs were then displayed on fused PET images to check if there were possible regions of increased 18F-FDG uptake near the skeleton, including extramedullary lesions and bladder. Areas of contiguous bone involvement and bone hyperplasia and sclerosis were also manually excluded.

From PET/CT VOIs, in total 1826 quantitative features including 781 features from PET and 1045 from CT were extracted. We extracted the radiomics features with the PyRadiomics package [81] (https://github.com/Radiomics/pyradiomics) which is compliant with the Imaging Biomarker Standardization Initiative [80]. From this package, we extracted the radiomics features from the original PET and CT images, filtered images with coiflet wavelet and LoG respectively. The images were discretized with a fixed bin size of 25 HU, which was quite commonly used in radiomics literature [137-139]. The extracted features reflected the disease characteristics including intensity distribution, texture pattern, morphological information, and spatial locations, as well as wavelet features [81]. The detailed list of extracted features was provided in the Supplementary Materials. Conventional PET metrics were also considered with equivalent features included in the features list. Specifically, the maximum and mean of the SUV were represented by the “Intensity Histogram” features “Maximum” and “Mean” from the original PET image, and the MTV could be represented by

“Morphology” feature “Volume”. Texture patterns were represented statistically by some common matrix, such as GLCM, GLSZM, and GLRLM. In addition, features from LoG and wavelet images were able to depict subtle texture features at different coarseness levels and frequency domains.

4.3.2. Feature Selection and Model Construction

The second stage was important feature selection with model construction. To reduce the high dimensionality of features, our selection strategy incorporated both intrinsic and statistical feature relationship as well as an outcome-driven machine learning model. To ensure that the feature-set was accurately clustered, we first repeated consensus cluster sampling for n=50 times to achieve the most stable groups. And then, to select the most important features, our selection process included (Figure 4-2) : 1) from each cluster, the most representative features were selected based on RF [92] tree importance (importance ≥ 0.01), 2) key features were selected from the representative features by univariate RF using the area under the curve (AUC ≥ 0.7), 3) to further eliminate the remaining redundant features, we then utilized the pairwise PCM, 4) Recursive Feature Elimination (RFE) [140] was adopted to select the most important features to form radiomic pattern. Thereby, the machine learning prediction model could be constructed only with the selected important features using RF algorithm.

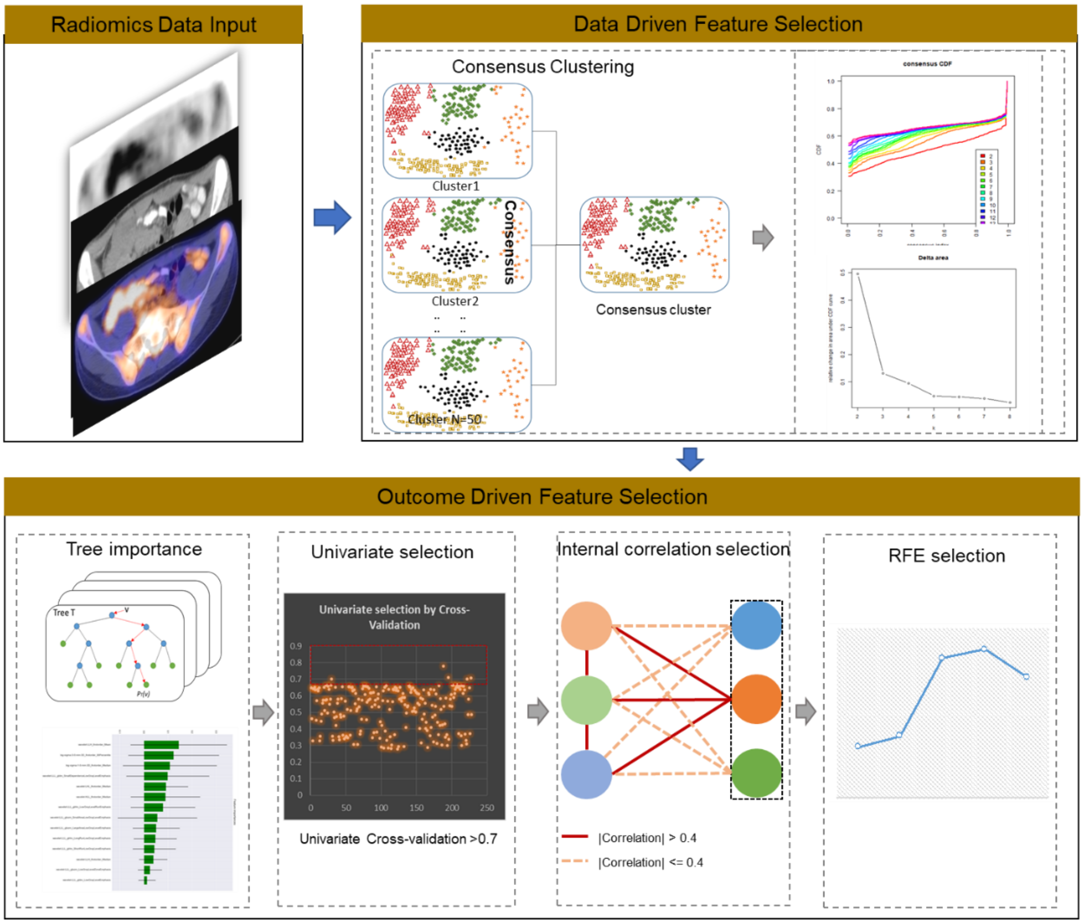


Figure 4-2. The process of feature selection.

4.3.3. Model Validation

The last stage was model validation. The machine learning model was trained by a Stratified ten-fold cross-validation on the training dataset, and the proportion of the positive-negative sample ratio in training and testing sets were approximately the same as in the original data set. To validate the robustness and stability of the machine learning model, we utilized both cross-validations and independent validations to assess the performance of the model. Ten-fold cross-validations were performed within the training group. As to the independent validations, the model was trained with the entire training group and then evaluated on the independent validation group.

Feature importance ranking was adopted in the RF model to evaluate the representative value of selected features. The feature-set was continuously and randomly permuted and scored, and the importance scores of the variable were obtained.

The performance of the pattern in this model was evaluated using receiver Operating Characteristic Curve (ROC). Wilcoxon test was utilized for feature P-values ($P \leq 0.05$) selection for both continuous and classification variables. The sensitivity, specificity, accuracy, Positive Predictive Value (PPV) and Negative Predictive Value (NPV) were also computed by Confusion matrix-derived metrics.

4.4. Experimental Results

4.4.1. Clinical visual analysis

The visual analysis was performed on the patients of training group with 16 BMB positive and 19 BMB negative patients. According to the visual analysis, 3 patients were classified as focal uptake, 8 as normal uptake and 24 patients were classified as diffuse uptake. Visual analysis correctly diagnosed all focal uptake patients and 7 out of 8 normal uptake patients. However, as to the diffuse uptake cases, visual analysis correctly diagnosed 14 cases, with 7 TP and 7 TN, failed in 10 cases with 5 FP and 5 FN. In summary, visual analysis achieved a successful diagnosis in 68.6% (24/35) of patients. The AUC of the visual analysis was 0.681 (95% confidence interval was 0.502-0.828). Its sensitivity, specificity, accuracy, PPV and NPV was 62.5%, 73.7%, 68.6%, 66.7% and 70.0%, respectively.

4.4.2. Feature selection and machine learning model

Feature selection procedure and results are illustrated as Figure 4-3. It could be observed that although Morphology features were extracted from images, these features were eliminated due to their statistical insignificance by statistical analysis. The texture features from original CT image were all eliminated due to their less importance determined by the RF algorithm. The following feature univariate RF selection showed that the features from original PET and CT images were less predictive in comparison with the features from LoG filtered and Wavelet decomposed images. Finally, after RFE process, the machine learning model consisted of two PET and one CT features (Table 4-2). It could be observed that the three selected features were all from the wavelet decomposed images capturing the textural information with low pass filters applied to the first two dimensions and high pass filter applied to the last dimension. The feature values extracted from the experimental dataset are normalized and summarized in Table 4-3. These values were assigned different weights when performing the model prediction.

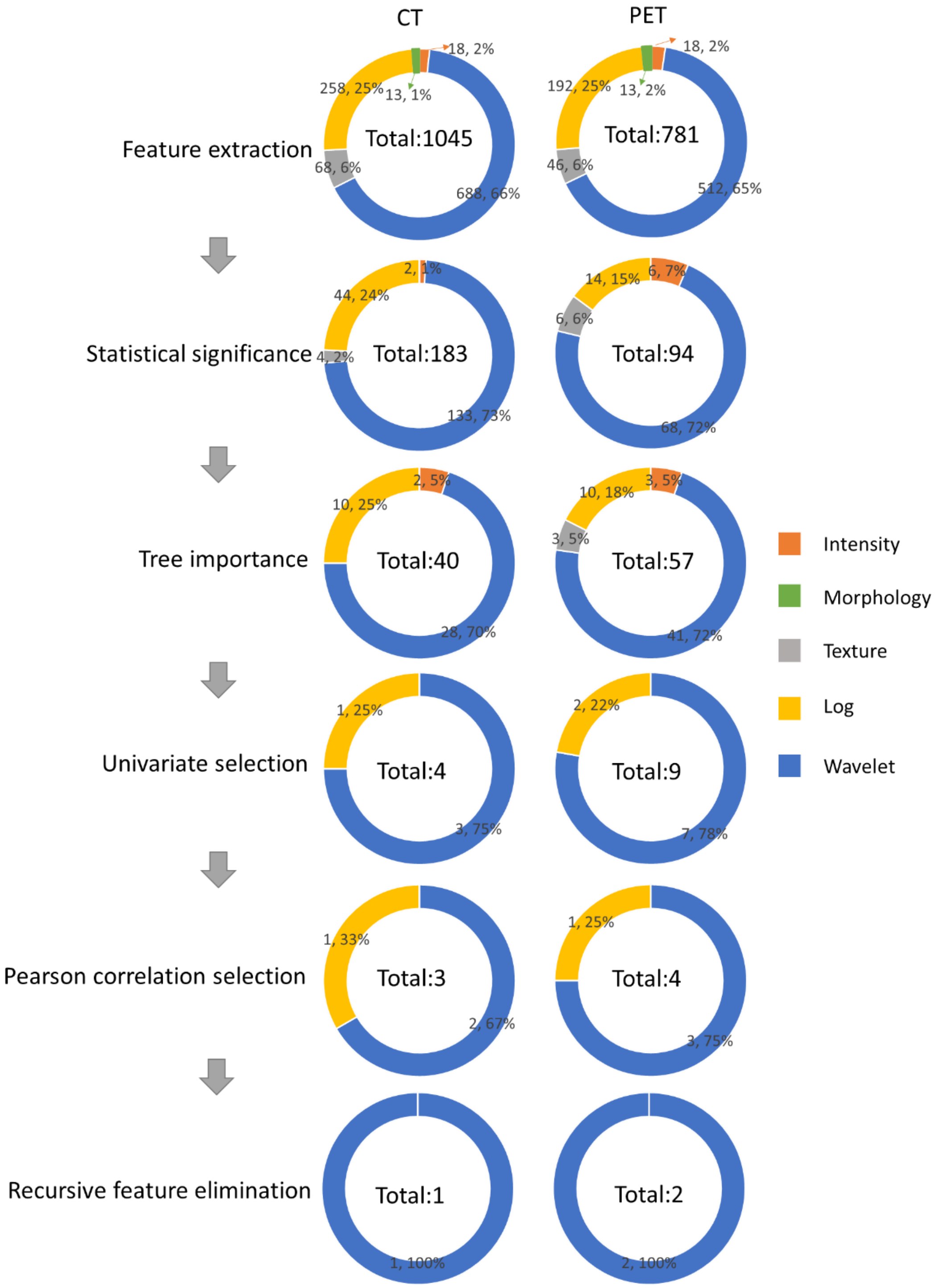


Figure 4-3. The results of feature reductions.

Table 4-2. The features selected from the trained machine learning model and their meanings.

Feature name	Feature definition and meaning
Wavelet-LLH_GLRLM_Run Entropy_PET	Formula： $F_{rlm.rl.entr} = \sum_{i=1}^{N_g}\sum_{j=1}^{N_r} p_{ij}log_2p_{ij}$ (4.1) where N_g is the number of discretized grey level intensity in the mask of VOI, N_r is the maximal possible run length in the mage. p_{ij} is normalized the run length matrix. i is a discretized grey level and j is occurrences of runs with length in matrix. Measuring the distribution of gray levels randomness from an image filter by a mid-frequency wavelet. The higher the value, the stronger the heterogeneity in the texture patterns.
Wavelet-LLH_firstorder _kurtosis _PET	Formula： $F_{kurt} = \frac{\frac{1}{N_v}\Sigma_{k=1}^{N_v}(X_{d,k}-\mu)^4}{(\frac{1}{N_v}\Sigma_{k=1}^{N_v}(X_{d,k}-\mu)^2)^2}$ (4.2) where N_v is the intensities set included in the ROI intensity mask denoted as $\{X_{d,1}, X_{d,2}, \dots, X_{d,N_v}\}$. μ is average gray-level intensity within the VOI. Measuring the peak of image VOI pixel value distribution in a mid-decomposition domain by using wavelet filter. The lower the value implies the mass of distribution concentrated towards a peak close to the mean value, vice versa.
Wavelet-LLH_GLRLM_SRHGLE _CT	Formula: $F_{rlm.srlge} = \frac{1}{\mathrm{N_S}}\sum_{i=1}^{N_g}\sum_{j=1}^{N_r}\frac{i^2r_{ij(\theta)}}{j^2}$ (4.3) where N_g is the number of discretized grey level intensity in the mask of VOI, N_r is the maximal possible run length in the image. r_{ij} be the run length matrix for an arbitrary direction θ. $\mathrm{N_S} = \sum_{i=1}^{N_g}\sum_{j=1}^{N_r} r_{ij}$ be the number of runs in the image along angle θ. i is a discretized grey level and j is occurrences of runs with length in matrix. Measuring the distribution of homogeneity by measuring the short run length distribution of higher gray values after mid-pass wavelet filter.

GLRLM： gray level run length matrix, LLH： low, low, and high frequency, SRHGLE： short run high gray-level emphasis.

Table 4-3. The mean± standard (SD), range and P-value of the features of the BMB positive and negative patients.

Features	**BMB positive**		**BMB negative**		**P-value**
	Mean± SD	range	Mean± SD	Range	
Wavelet-LLH _GLRLM_RunEntropy _PET	0.453±1.190	-1.02~3.09	-0.381±0.528	-1.26 ~0.81	0.022
Wavelet-LLH_firstorder_kurtosis _PET	-0.443±0.320	-0.62~0.72	0.375±1.183	-0.61~3.21	0.008
Wavelet-LLH_ GLRLM_SRHGLE_ CT	-0.404±0.273	-0.93~-0.04	0.34 ±1.24	-0.69~5.21	0.001

GLRLM： gray level run length matrix, LLH： low, low, and high frequency, SRHGLE： short run high gray-level emphasis

The model was evaluated with both cross-validation and independent validation. In the cross-validation, the model correctly predicted 31 patients with 14 TP and 17 TN, incorrectly predicted 4 (2 FP and 2 FN) patients of ^{18}F-FDG diffuse uptake. The machine learning model achieved a successful diagnosis in 88.6% (31/35) of patients, which was significantly higher than that of visual analysis by using Pearson Chi-square test (P=0.041). The AUC of the model was 0.885 (95% confidence interval was 0.732-0.968), which was significantly higher than that of visual analysis (P=0.046). Its sensitivity, specificity, accuracy, PPV and NPV was 87.5%, 89.5%, 88.6%, 87.5% and 89.5%, respectively. As to the independent validations, the prediction model could achieve 83.3% (5/6) accuracy on the independent validation dataset. Among the six patients, one (out of two) focal uptake patient was incorrectly predicted as FN, while all the diffuse uptake and normal uptake patients were correctly predicted.

4.4.3. Results analysis and interpretation

Results from the study show that the differences between the two methods mainly existed in the diagnosis of the patients with diffuse uptake. The machine learning model achieved 83.3% (20/24) prediction accuracy, in comparison with 58.3% (14/24) accuracy from visual analysis. Among the 10 visually failed diffuse uptake cases, the machine learning model correctly predicted 9 of them. Visual analysis correctly diagnosed the other three cases in which the machine learning model failed.

Two representative cases from visual analysis and machine learning model are illustrated by Figure 4-4. As shown in Figure 4-5 for the distribution histogram of the three normalized features among all experimental data, there existed BMB positive and BMB negative patients sharing same feature value ranges. Therefore, BMB positive and negative patients could not be discriminated from an individual feature (with mean accuracy of 70.8%, 72.7% and 76.7% respectively for Kurtosis, RunEntropy and SRHGLE features). As to the case 3A, according to the first and third features, since there were more BMB negative patients than positive ones exhibiting the same feature value, these two features would suggest that the patient was more probably to be BMI negative. However, the distribution of the second feature was against this negative suggestion. As to the case 3B, although all three features were suggesting that the patient was more likely to be negative, the possibility of a positive case could not be eliminated, given that a few positive cases were exhibiting the same feature values.

The machine learning model quantitatively combined these features for the final prediction while considering their diverse contributions. The contributions of features could be explained by the weighting coefficients derived from LIME model which is a local linear approximation of the trained prediction model [31]. The LIME model perturbed the feature values and observed the resulted changes in prediction. The features, which the prediction was more sensitive to, would be assigned higher weight values. Positive weights indicated that the increase in the corresponding features would be more supporting a positive prediction, while negative weights would indicate the changes supporting a negative prediction. The right column of Figure 4-3 shows the weights of the features employed in the prediction of the two representative cases, and the predictions were derived from the linear combinations of the features weights and features values.

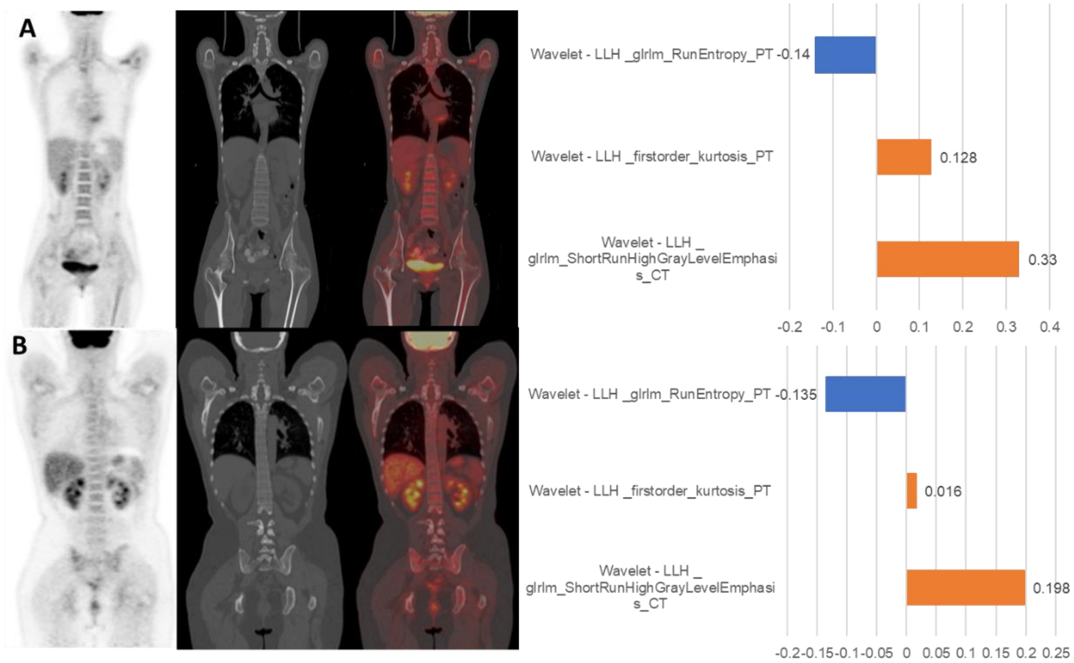


Figure 4-4. Two representative cases from visual analysis and machine learning model. The patient displayed on panel (A) was BMB negative. The machine learning model correctly predicted it. The value of three features were -0.03, -0.57 and -0.23, respectively. The visual diagnosis was false positive. The patient displayed on panel (B) was BMB positive. The machine learning model correctly predicted it. The value of the three features were -0.48, -0.34 and -0.28, respectively. The visual diagnosis was false negative. From left to right, coronal PET, CT, fusion image and the approximated features weights from LIME interpretation.

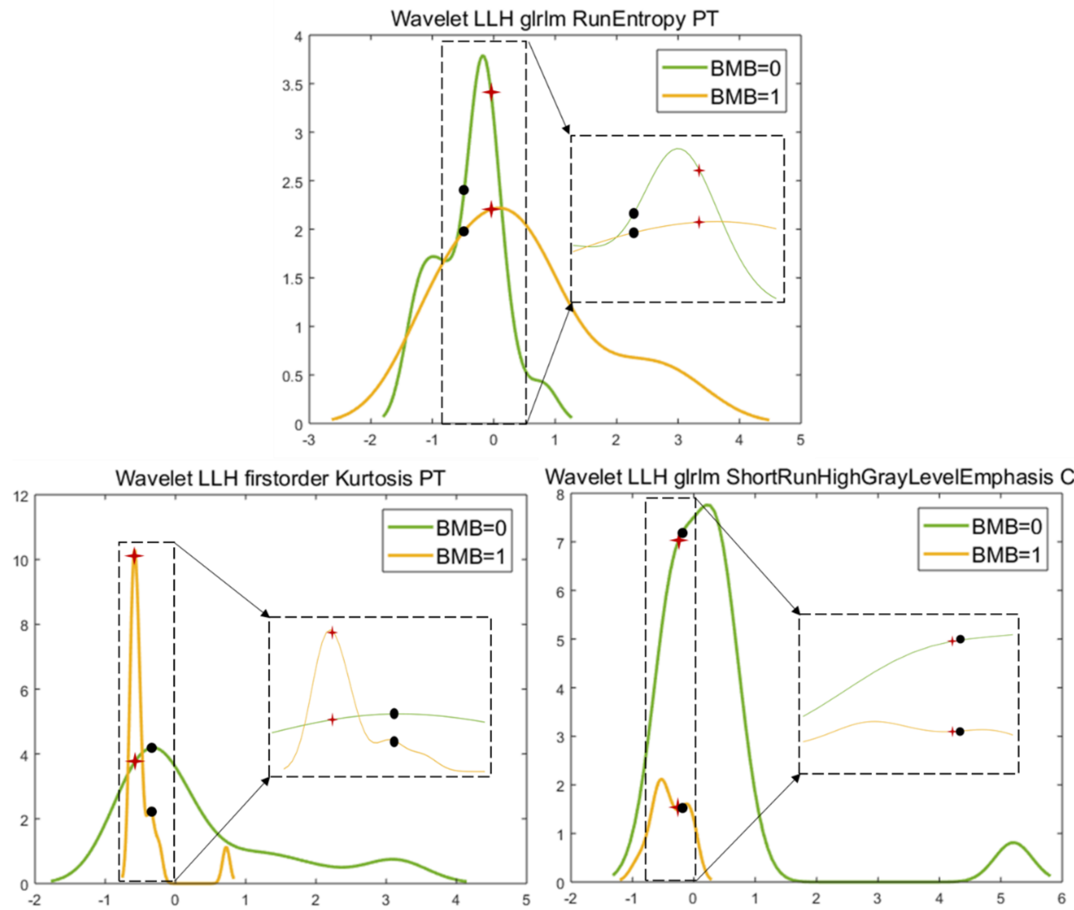


Figure 4-5. Distribution histograms (feature values as x-axis, and value frequency in the dataset as y-axis) of the 3 radiomic features selected by the trained machine learning model with the corresponding feature values of the 2 representative cases (red crosses for case 3A, and black spots for case 3B). The zoomed and scaled views of the distributions are indicated with dashed frames (BMB=0 for bone marrow biopsy negative, and BMB=1 for bone marrow biopsy positive).

4.4.4. Intra- and Inter-observer Variability

We have evaluated the influence of both intra-observer and inter-observer variabilities on the performance of our prediction model. The volume overlap error rate (VOE) was used to measure both intra-observer and inter-observer agreement rate where 1 indicating two identical VOIs and 0 indicating no overlap between the VOIs. From our experimental validations (as detailed below), the performance of the prediction model was not affected by the intra-observer and inter-observer variabilities. The predictions would be slightly affected by the faulty VOIs of large variations, but the intra- and inter-observer variability would not as large as the faulty VOIs to influence the predictions.

To evaluate the influence of intra-observer variability, 16 (out of 35) patients were randomly picked for a second review and adjustment. As shown in Figure 4-6, the mean intra-observer agreement rate with standard deviation was 0.949±0.044, which led to feature value differences as recorded in Table 4-4. We then performed ten-fold cross-validations on the adjusted patients VOIs. The validation demonstrated that such a minor variability did not pose alteration on prediction decision, justified by 87.5% (14/16) accuracy with two (out of 10) diffuse uptake patient predicted as FPs, which were the identical decision as that in the previous cross-validations.

To evaluate the influence of inter-observer variabilities, we further collected six new patient datasets for independent validation. The six datasets included two diffuse uptake patients, two focal uptake patients and two normal uptake patients. For each new patient, two sets of VOIs were delineated by (1) the same experienced senior operator for the initial 35 patients, and (2) a team of two junior operators. As shown in Figure 4-5, the inter-observer agreement rate was 0.921±0.074, leading to feature value differences as recorded in Table 4-4. Then, we respectively performed independent validations on the two sets of VOIs from the six new patients. The comparison of independent validations results demonstrated that there was no alteration on prediction decision, justified by the same accuracy of 83.3% (5/6) with one (out of two) focal uptake patients incorrectly predicted as FN, while all the diffuse uptake and normal uptake patients were correctly predicted.

In particular, for the two intra- and inter-observer cases with 82.74% and 79.23% as the lowest VOI agreement rates, the model produced consistent predictions.

To further investigate the influence of higher variations, we randomly picked 13 patients from the initial 35 patients and intentionally exclude various portions of the VOI of the spinal cord to create a new set of VOIs. The agreement rate between these two sets of VOIs was 0.667±0.064 (range 0.509~0.754). The experimental validation showed that the majority of predictions (11 out of 13 patients) remained consistent while two predictions were inversed.

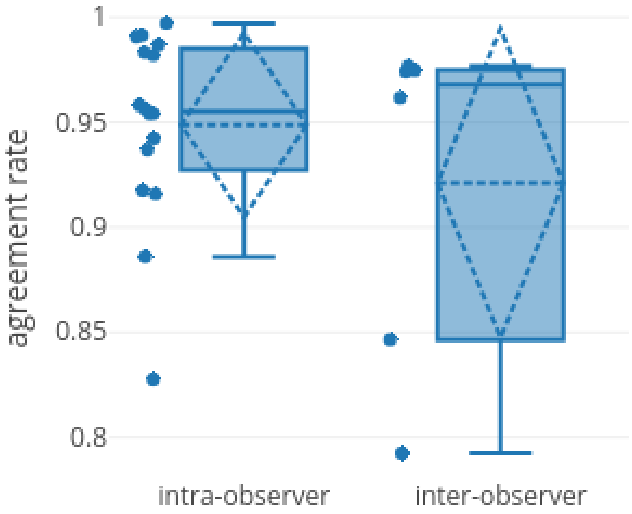


Figure 4-6. Intra-observer and inter-observer agreement rate based on VOE.

Table 4-4. Differences on normalized feature values caused by intra-observer and inter-observer variabilities.

	Intra-observer Cases		Inter-observer Cases	
	Mean± SD	Range	Mean± SD	Range
Wavelet-LLH_GLRLM_RunEntropy_PET	0.018±0.036	-0.066~0.092	-0.001±0.069	-0.11~0.1
Wavelet-LLH_firstorder_kurtosis_PET	-0.012±0.475	-0.787~1.624	0.03±0.136	-0.1~0.28
Wavelet-LLH_GLRLM_SRHGLE_CT	-0.003±0.012	-0.007~0.048	-0.006±0.008	-0.02~0

4.5. Discussion

To solve the well-recognized difficulties of visual analysis of BMI, we constructed a ^{18}F-FDG PET/CT radiomic analysis in the patients with suspected relapsed AL. To the best of our knowledge, there have been no previous studies using radiomic features with machine learning methods to assess leukemic bone marrow uptake, and it is a relatively large-scale bone marrow ^{18}F-FDG PET/CT study.

Considering the sample size, we employed the RF prediction model in our study. As evaluated by Gunduz et al [93], the RF model substantially outperformed other techniques on both real-life and simulated data regarding the task of robust classification in the high dimension low sample size context. Floares et al [141] further justified that the RF method would derive accurate and robust model from omics data of small sample size. Such characteristic made RF model more suitable to our study where radiomic pattern would be derived from high dimensional data (a total of 1826 features for each patient) of limited number of sample studies. Additionally, according to the theory of Chalkidou et al [142], 10 to 15 patients are minimally required to test one radiomic feature, our model reduced the number of features to 3 features and would be valid to minimize false detection rates regarding the sample size in our study.

The first finding of this study is that the machine learning model achieved a high accuracy for detecting the BMI, outperforming that of visual analysis, and was particularly excellent in analyzing diffuse uptake patterns. The diagnostic value of machine learning model statistically outperformed visual analysis in terms of AUC (0.885 vs. 0.681, P=0.046), and the successful diagnosis rate of machine learning model was significantly higher than that of visual analysis (88.6% vs. 68.6%, P=0.041). For the diffuse uptake patients, the machine learning model achieved 83.3% (20/24) prediction accuracy, in comparison with 58.3% (14/24) accuracy from visual analysis. The independent validation further justified the excellence of the machine learning model for diffuse uptake pattern. To the best of our knowledge, this is the first study to apply artificial intelligence technology to improve the ^{18}F-FDG PET/CT-based clinical diagnosis of BMI in the patients with suspected relapsed AL. A comparable radiomic analysis result was reported in the patients with diffuse large B cell lymphoma, where the AUC of a first-order Skewness feature in detecting BMI was 0.821, and its sensitivity and specificity was 81.8% and 81.7%, respectively [136]. The Skewness feature and its variants were also extracted in our experiments, and their performances (mean accuracy of 52%, range 34.7%~67.2%) were all lower than that of the individual three features we selected, and thereby also lower than the performance of our radiomic pattern (provided in supplementary materials).

Another finding is that this study provided an interpretable insight into the output of BMI from the machine learning model. Because of the complexity and opacity of algorithms, machine learning methods are often criticized as black boxes. We attempted to interpret the results of model predictions based on the LIME model. LIME approximated the machine learning model as a local linear model which is a linear combination of the feature values and the corresponding relative weighting coefficients. With the derived weights of features, the driving factors of the machine learning model prediction could be extracted. A more detailed explanation is in the results section.

Interestingly, a CT feature became an integral part of the model in the present study. Although the value of features extracted from unenhanced low-dose CT has been demonstrated in the studies of non-small cell lung cancer [143], lymphoma [144] and esophageal cancer [145], there are no such published studies on bone marrow. Based on the experience of visual analysis, CT is suitable to visualize cortical and trabecular bone, while not a routine method for bone marrow assessment [146, 147]. In the present study, the CT feature contributed with a relatively high weight in some patients. However, the value of CT features on BMI requires a larger number of research samples for further confirmation.

In addition, in comparison to the PET conventional metrics (SUVmax, SUVmean, MTV and TLG), our selected radiomics features possessed much stronger correlations with BMB (as detailed in the supplementary materials). The equivalent features to the three conventional metrics, i.e. SUVmax, SUVmean and MTV, were initially included in the extracted radiomics set. However, these three equivalent features were excluded automatically by our feature selection procedure on the basis of their discriminative contributions. We calculated another conventional metric, TLG=MTV*SUVmean [148]. The prediction accuracy for these four individual conventional metrics were 53.9%, 44.2%, 50.5% and 51.5% respectively. Further comparison analysis on the correlations with BMB was performed between PET conventional metrics and our three selected radiomics features (Table 4-5). The comparison showed that the BMB correlation values of our selected radiomics features were 0.42, -0.41 and -0.38 while the correlation values of the four PET conventional metrics were 0.29, 0.19, 0.22 and -2.33E-01.

Table 4-5. The correlation matrix of selected features, PET conventional metrics and BMB.

Features	**TLG**	**SUVmean**	**MTV**	**SUVmax**	**BMB**
Wavelet-LLH_GLRLM_RunEntropy	0.89	0.85	0.24	-6.63E-02	**0.42**
Wavelet-LLH_firstorder_Kurtosis	-0.22	-0.24	-0.16	6.00E-01	**-0.41**
Wavelet-LLH_GLRLM_SRHGLE	0.1	0.04	-0.45	-2.70E-03	**-0.38**
BMB	**0.29**	**0.19**	**0.22**	**-2.33E-01**	1

The last finding is that our automated radiomic analysis method could serve as a non-invasive test option complementing the visual analysis for the diagnosis of suspected relapsed AL. For the 11 failed cases in visual analysis, our machine learning model correctly predicted 10 of them by analyzing the radiomic features purely based on the PET/CT scans. And that would suggest our model being an eligible non-invasive test option complementing the visual analysis for a more comprehensive and accurate diagnosis.

For the next stage, we will be performing translational research by (1) harnessing automated bone segmentation software with machine learning-based prediction model for automated processing and analysis platform, and (2) installing the software platform in our collaborative hospitals for multi-center study for the standardization of the imaging biomarkers for BMB.

4.6. Chapter Summary

We propose an interpretable machine learning approach measures the contribution of features for each case and the connections of radiomics with its underlying biological features to seek innovate clinical decision making in related studies. ^{18}F-FDG PET/CT radiomic analysis with machine learning model presents an objective and efficient mechanism for monitoring the BMI in suspected relapsed AL, and could serve as a non-invasive way selection supplementing the visual evaluation to obtain a more comprehensive, precise and reasonable diagnosis. It is recommended in particular for the diagnostic BMI in the patients with diffuse uptake.

Chapter 5. A Prognostic Values of Radiomic Analysis on Predicting Pathologic Complete Response of Breast Cancer to Neoadjuvant Chemotherapy

This chapter are mainly evaluating radiomic prognostic ability base on machine learning. Because of the non-invasive advantage of radiomics, a prognostic radiomic feature set is very meaningful for making treatment programs.

The status of pCR is considered as a gold standard assess NAC postoperative pathologic outcomes of locally advanced breast cancer patients. However, even patients have some tumor staging or subtypes, the efficacy of pCR may be completely different. 18F-FDG PET-CT imaging has special value in tumor staging, prognosis prediction and therapeutic assessment. This chapter aims to mine radiomic predictors from PET/CT imaging features for non-invasively prognostic therapeutic efficacy in breast cancer patients prior to NAC.

5.1. Overview

Breast cancer is a malignant tumor that is one of the leading causes of cancer-related death among women worldwide, and morbidity and mortality have increased in recent years [149]. According to the National Central Cancer Registry of China (NCCR), the number of new cases in China in 2015 was similar with the number of morbidity in America in 2019 about 268,600 [150]. NAC is commonly utilized in locally advanced, inflammatory breast cancer with a primary lesion greater than 3 cm [151]. pCR is often used as a substitute for Disease-Free Survival (DFS) and Overall Survival (OS) and is the primary endpoint of neoadjuvant chemotherapy [152]. Therefore, it is vital to detect those patients who will have a pCR after NAC so the toxicity of ineffective chemotherapy is avoided and other treatment options are considered. Moreover, a high proportion about 10-35 % of patients fail to respond to NAC and these patients have a higher risk of relapse and lower survival [153, 154]. Therefore, it would be valuable to be able to identify the women who will respond to NAC.

The postoperative pathology result is the golden standard of pCR, however, some assessment of NAC treatment response is done prior to any surgical intervention and imaging plays an important role in this assessment. Conventional anatomical imaging techniques, such as X-ray CT and MRI, provide volumetric and morphometric changes of tumor size using RECIST criteria but these are insufficient to evaluate early treatment response [155, 156]. PET/CT with 18F-FDG is a more sensitive and specific tool for the noninvasive assessment of tumor metabolism and can provide an earlier indication of response and so reduce ineffective treatment-related toxicity [157]. More recently, investigators have suggested that an early change in ^{18}F-FDG uptake in the primary breast tumor is a potential predictive biomarker of response to NAC but these changes can take at least two courses of NAC before tumor metabolic changes become apparent [158-160].

Radiomics is an emerging field where various techniques are used to extract/interpret information in quantitative medical imaging and it has a great potential to assist in patient management [161, 162]. Radiomics has been applied to oncologic detection, diagnosis, tumor monitoring response to treatment and prognosis [163-165]. A recent report showed that pre-treatment ^{18}F-FDG PET/CT metabolic radiomics had a close relationship with tumor response to NAC and risk of recurrence in locally advanced breast cancer, but this research did not establish a specific therapy response prediction model [166]. Our aim was to identify radiomic predictors of treatment response and pCR from a single ^{18}F-FDG PET/CT scan and construct an automated model.

5.2. Dataset Description

5.2.1. Patients

The study was retrospective and done in a single center. It was approved by the Institutional Review Board; the need for informed consent was waived. We reviewed records and images of patients with breast cancer (n = 163) who underwent NAC at Fudan University Shanghai Cancer Center (Shanghai, China) between January 2012 and July 2018. Tumor staging was performed on the basis of the American Joint Committee on Cancer TNM Staging Manual, 8th Edition [167]. Inclusion criteria were female sex, biopsy-confirmed breast cancer, unilateral tumor, ^{18}F-FDG PET/CT scan done < 7 days before NAC, completion of the full course of NAC (the details of the NAC can be found in the supplementary material). Exclusion criteria were: multifocal or multicentric breast cancer (n = 12); non-standard NAC regimen (n = 15). There were 138 patients, with 50 pCR patient, who met the inclusion criteria. 50 non-pCR patients were randomly selected with the 50 pCR patients for finally analysis to ensure uniform data distribution. The clinical-pathologic characteristics of the 100 patients are listed in Table 5-1. All the patients included in the study, the median (interquartile range, IQR) age of all patients was 47.37 (26-76) years. Immunohistochemical (IHC) parameters included Ki-67, Estrogen Receptor (ER) / Progesterone Receptor (PR) status, and Human Epidermal Growth Factor Receptor (HER2). The breast cancer molecular subtypes were classified as: ER/PR-positive and HER2-negative (Luminal A); ER/PR-positive and HER2-negative with high Ki-67 (Luminal B1); ER/PR-positive and HER2-positive (Luminal B2); ER/PR negative and HER2-positive; or Triple-Negative (TN). pCR was defined as no histological evidence of infiltration of tumor cells in breast and axillary tissues, including ductal carcinoma in situ (yp $T_{0/is}$).

Table 5-1. The baseline characteristics of the study population.

	Training set		Testing set	
	Number	%	Number	%
Number	70		30	
Age median(range)	49.5(26-76)		47(28-68)	
Molecular subtype				
Luminal	31	44.29	17	56.67
Her 2+	17	24.28	10	33.33
TN	22	31.43	3	10
Response to NAC				
pCR	35		15	
non-pCR	35		15	

5.2.2.Phantom PET-CT Studies

^{18}F-FDG was produced in our cyclotron (Siemens CTI RDS Eclips ST, Knoxville, Tennessee, USA) using Explora FDG_4 synthesis module; radiochemical purity was > 95%. All the patients were requested to fast at least 6 h. At the time of the tracer injection (7.4 MBq/kg), the patient's blood glucose level was < 10 mmol/L. Before and after radioisotope injection, patients were kept lying comfortably in a quiet, dimly lit room. Scanning began 60 mins after radioisotope injection; data were acquired on a Siemens Biograph 16HR PET/CT scanner (Siemens Healthineers, Hoffman Estates, IL, USA) with a transaxial intrinsic spatial resolution was 4.1 mm (full-width at half-maximum) in the center of the field of view. Data acquisition procedure was as follows: CT scanning was first performed, from the proximal thighs to head, with 120 kV, 80 ~ 250 mA, pitch 3.6, rotation time 0.5. Immediately after CT scanning, a PET emission scan that covered the identical transverse field of view was obtained.

Acquisition time was 2 ~ 3 min per bed position. PET image data sets were reconstructed iteratively by applying CT data for attenuation correction, and co-registered images were displayed on a workstation.

5.3. Statistical Analysis

5.3.1. Image Segmentation and Feature Extraction

The overall workflow of this study is shown in Figure 5-1. The three-dimensional Gross Tumor Volume (GTV) was performed by two expert physicians using *3D Slicer* software [168, 169] on all PET/CT images to minimize operator bias. The registered PET-CT images and the delineated VOIs, were used to extract 1105 image features. The radiomic features can be divided into Shape, Intensity and Texture features and features based on LoG and Wavelet filter. The Shape feature includes edges, angles, and ridges so that 3D geometric features such as surface area, compactness, and spherical ratio can be extracted. Intensity features included maximum, mean and the average absolute deviation of the gray level intensity values in the VOIs. The texture feature, is a 2nd order feature, and is used to express internal heterogeneity difference between different parts of the lesion from common matrices (Table 5-2). LoG and Wavelet features included extracted the first order, shape, and second-order features from the different frequencies and resolutions for subtle features. These features enable a comprehensive comparison of the spatial position voxel with its surrounding voxels. Overfitting is a concern because of the high dimensionality of the radiomic features available, therefore, the dataset was randomly divided into training (70 patients with 35 pCR) and validation (30 patients with 15pCR) groups to reduce this risk. Feature selection and model construction were based on the training group.

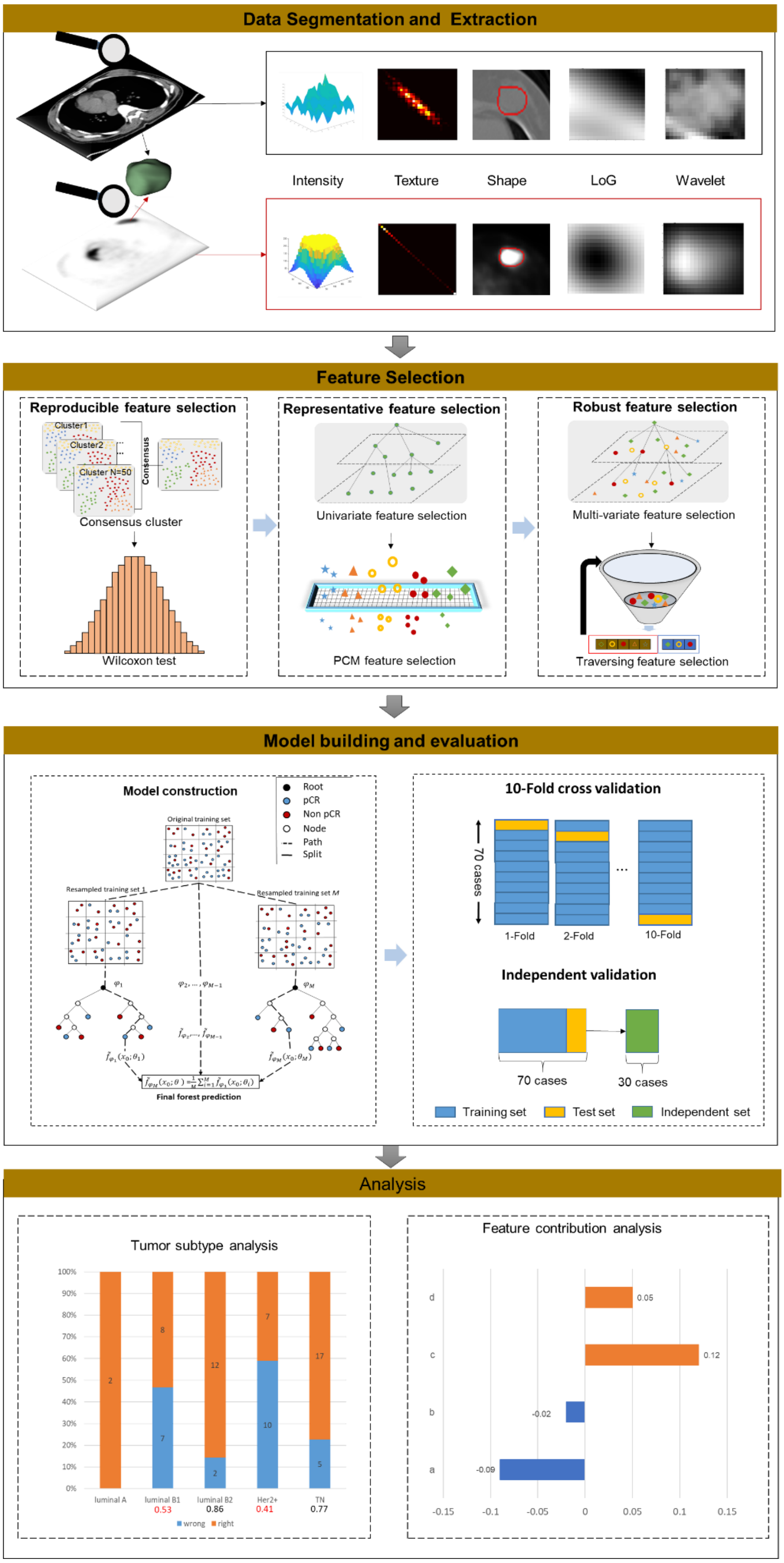


Figure 5-1. The workflow of radiomic features extraction, selection and model construction.

5.3.2. Feature Selection

For feature selection, we first identified reproducible features using an unsupervised learning method, i.e. consensus clustering repeated multiple times, in combination with a Wilcoxon test. We then used a Pairwise PCM to identify pairs of highly correlated features ($|r| \geq 0.76$ for PET and $|r| \geq 0.4$ for CT) from inner clusters for non-redundant features. Each pair was represented by one with the highest discriminative ability to preserve most representative features, that is, keep larger one and average accuracy higher than 0.5. Efficacy prediction by using univariate RF. Then, the most robust features for NAC prediction could then be derived based on these inter-cluster representative features. Traversal Feature Selection (TFS) based on multi-variate RF was applied on the inter-cluster representative features to select the radiomic predictor with best prediction performance. Selection details and results are shown in Figure 5-2 and Table 5-2.

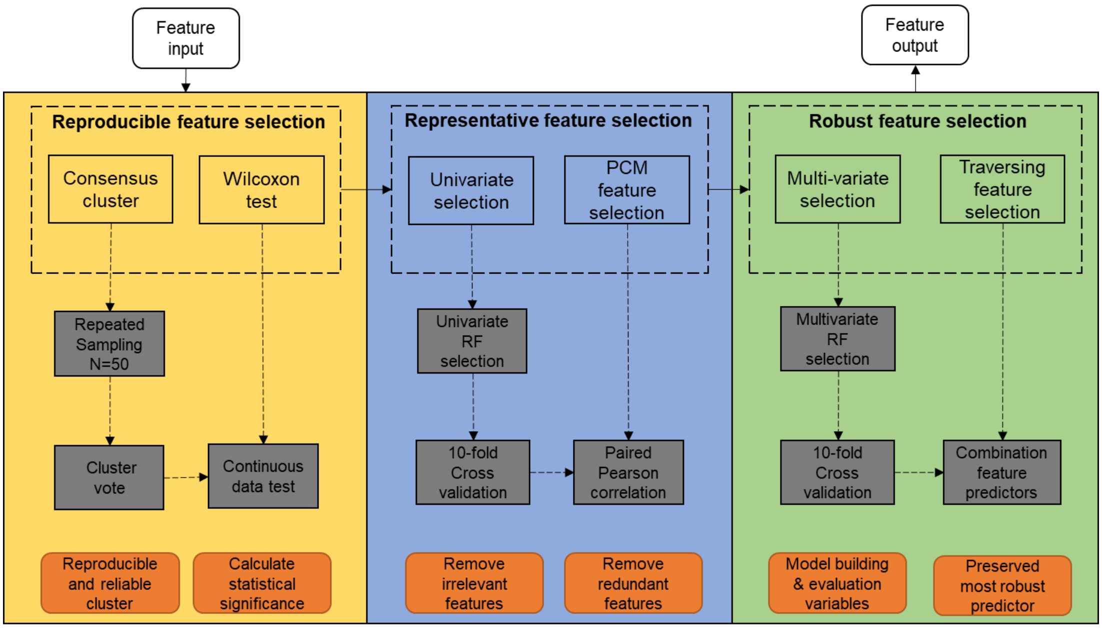


Figure 5-2. The chart of feature selection procedure

Table 5-2. The selected features details in the study

Feature name	Observation	Feature Equation
wavelet-LHH_glcm_DifferenceAverage.CT	GLCM_DA_LHH_CT	Formula: $F_{cm.dif.avg} = \sum_{k=0}^{N_g-1} kp_{x-y}(k)$ (5.1) where $\mathrm{N_g}$ is the number of discretized grey level intensity in the anatomical imaging mask of VOI. $\mathrm{p_{x-y}(k) = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i,j)}$, where $\mathrm{\|i-j\| = k}$, and $\mathrm{k = 0,1,\dots,N_g-1}$. Measuring the local intensity variation through evaluating the connections the pairs of similar intensity values with different intensity values by after mid-pass wavelet filter. A higher value means a greater disparity from among neighboring voxels.
wavelet-HLH_glrlm_LowGrayLevelRunEmphasis.CT	GLRLM_LGRE_HLH_CT	Formula: $F_{rlm.lgre} = \frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_r} \frac{P(i,j\|\theta)}{i^2}}{N_r(\theta)}$ (5.2) where $\mathrm{N_g}$ is the number of discretized grey level intensity in the anatomical imaging mask of VOI. $\mathrm{N_r}$ is the maximal possible run-length in the image. $\mathrm{P(i,j\|\theta)}$ be an arbitrary direction θ for run length matrix. $\mathrm{N_r(\theta)}$ is the number of runs in the image along angle θ. Measuring the distribution of low gray level values in a mid-decomposition domain by using a wavelet filter. A higher value indicates a higher density of lower gray values in the image.
wavelet-HLH_glrlm_RunVariance.PT(RunLenthVariance)	GLRLM_RV_HLH_PT	Formula: $F_{rlm.rl.var} = \sum_{i=1}^{N_g} \sum_{j=1}^{N_r} p(i,j\|\theta)\,(j-\mu)^2$ (5.3) where $\mathrm{\mu = \sum_{i=1}^{N_g} \sum_{j=1}^{N_r} p(i,j\|\theta)j}$. $\mathrm{N_g}$ is the number of discretized grey level intensity in the molecular imaging mask of VOI. $\mathrm{N_r}$ is the maximal possible run-length in the image. $\mathrm{p(i,j\|\theta)}$ be an arbitrary direction θ for run-length matrix. Measuring the variance in runs over the grey levels for run lengths from an image filter by a mid-frequency wavelet.
wavelet-HHH_glszm_ZoneVariance.PT(ZoneSizeVariance)	GLSZM_ZV_HHH_PT	Formula: $F_{szm.zs.var} = \sum_{i=1}^{N_g} \sum_{j=1}^{N_s} p(i,j)(j-\mu)^2$ (5.4) where $\mathrm{\mu = \sum_{i=1}^{N_g} \sum_{j=1}^{N_s} p(i,j)j}$. $\mathrm{N_g}$ is the number of discretized grey level intensity in the

		molecular imaging mask of VOI. N_s is the number of discreet zone sizes in the image. $p(i, j)$ is a normalized size zone matrix. Measuring the variance in zone counts over different zone sizes grey levels in a high-decomposition domain by wavelet filter. The variance can reflect the volumes of size zones for the VOI.

5.3.3. Radiomic Predictor and Model Construction

Both unsupervised learning and supervised learning methods contributed to the derivation of the prognostic predictor. Consensus clustering is implemented to ensure that the RF derives the fairness of representative features from each cluster. The RF method was capable of evaluating the feature discriminativeness. It was effective in two aspects of radiomic feature analysis: selection of representative features and derivation of the robust radiomic predictor.

In the training process, using the multivariate RF model to construct the radiomic predictor. RF is essentially composed of 100 decision trees in our model. The original training set will be replaced by random selection multiple times as the new training sets of decision tree. Then different structures of training set will be observed separately binary decisions, and the final prediction comes from the vote of all of it. The AUC, sensitivity, specificity, positive predictive value, and negative predictive value were then computed to establish model performance.

5.3.4. Analysis and Interpretation

10-fold Stratified cross-validation was adopted on training set and mimic the positive-negative sample ratio of training set to split-folds as similar proportion. 30 times cross-validation were settled in this study guaranteeing the data split fairness of selected predictor to evaluate the robustness of selected radiomic predictor. On the basis of build model on training group, independent validation group was applied to evaluate the stability and performance of the predictive model. Confusion matrix-derived metrics computed sensitivity, specificity, accuracy, PPV and NPV from the machine learning model.

We used the LIME model to measure the approximate perturbation range and hence the contribution of each feature to the prediction model. The weight of the feature was estimated by measuring the distance from the range center. Positive weights indicated that an increase in the corresponding feature would support a positive prediction, while negative weights indicated the change supported a negative prediction. The LIME explanation of function was obtained as follows:

$$\varepsilon(\mathrm{x}) = \underset{\mathrm{g} \in \mathrm{G}}{\operatorname{argmin}} \ \mathcal{L}(f, g, \pi_x) + \Omega(g) \tag{5.5}$$

where gϵG as an explanation model, G is a potentially interpretable model, $\Omega(g)$ is a measure for complexity of model G, here, it represents the depth of the tree. $f(x)$ is predictive probability that the case belongs to a certain class1. π_x as a proximity measure to define locality around x. $\mathcal{L}(f, g, \pi_x)$ measures the faithfulness of g in approximating f defined by π_x.

5.4. Experimental Results

5.4.1. Clinical features analysis and signature construction

Table 5-3 illustrates the performance for univariate clinical features and clinical predictor: 1) each clinical feature performance was evaluated using 30 times of cross-validation in second columns, 2) with TFS selected and assessed, three immunohistochemical parameters and a clinical feature age, as clinical predictor applied on training set and independent validation set and the performances show in third and fourth column.

Table 5-3. The performance of the selected clinical features

Features	Uni_mean_accuracy	TFS accuracy	Independent set
ER	0.7	0.757	0.5
PR	0.619		
HER2	0.324		
Age	0.529		
Ki-67	0.610		
Molecular Subtype	0.676		

5.4.2. Clinical and Radiomic Features Analysis and Results

In our experiments, we compared the four different radiomic predictors: 1) CT features only; 2) PET features only; 3) PET and CT features; 4) clinical features and PET and CT features. The most prognostic predictor, based on the average prediction performance of various feature combinations (Table 5-4) had 5 PET, 3 CT and 4 clinical features (30 patients with 15pCR).

Table 5-4. Selected features in the prediction model

Category	Selected features
PET radiomic features	wavelet-LHH_glrlm_RunVariance wavelet-LHH_gldm_DependenceVariance wavelet-HLH_glrlm_RunVariance log-sigma-3-0-mm-3D_firstorder_Maximum wavelet-HHH_glszm_ZoneVariance
CT radiomic features	wavelet-HLH_glrlm_LowGrayLevelRunEmphasis.CT wavelet-LHH_glcm_DifferenceAverage.CT wavelet-LLH_glszm_SizeZoneNonUniformity.CT
Clinical features	ERP PRP HER2 Age
PET-CT radiomic features	wavelet-HLH_glrlm_LowGrayLevelRunEmphasis.CT wavelet-LHH_glcm_DifferenceAverage.CT wavelet-HLH_glrlm_RunVariance.PT wavelet-HHH_glszm_ZoneVariance.PT
Predictive predictor	Age wavelet-HLH_glrlm_LowGrayLevelRunEmphasis.CT wavelet-LHH_glcm_DifferenceAverage.CT wavelet-HLH_glrlm_RunVariance.PT

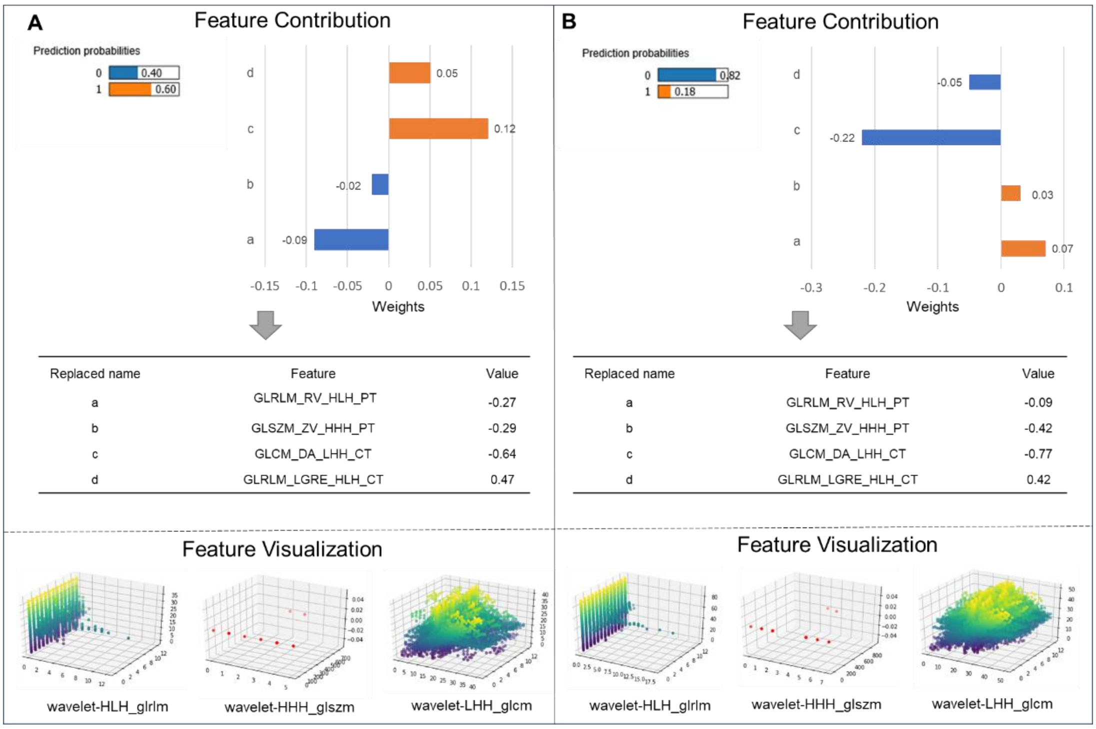
	wavelet-HHH_glszm_ZoneVariance.PT

5.4.3. The Performance of pCR Predictions

Table 5-5 shows the prognostic performance comparison of the three different types of radiomic predictors applied on the training group and on the independent validation group as well. The combination of 2 PET radiomic features (GLRLM_RV_HLH_PT and GLSZM_ZV_HHH_PT) with 2 CT radiomic features (GLRLM_LGRE_HLH_CT and GLCM_DA_LHH_CT) and add the Age information of patients could achieve the accuracy of 0.857 and 0.8 in the training group and the independent validation group respectively. The prediction model showed that the independent validation had a sensitivity of 0.733 and a specificity of 0.867. It could be observed that the overall performance of the PET-CT adds age information radiomic predictor consistently outperformed the radiomic predictors based on single modality in both ten-fold validation and independent validation. Figure 5-3 provides images of two patients as representative examples.

Table 5-5. The prediction performance of selected radiomic features.

		AUC	Accuracy	Sensitivity	Specificity	PPV	NPV
PET	training set	0.656	0.714	0.833	0.625	0.625	0.833
	testing set	0.6	0.7	0.6	0.8	0.75	0.667
CT	training set	0.755	0.785	0.714	0.857	0.833	0.75
	testing set	0.698	0.733	0.733	0.733	0.733	0.733
PET-CT	training set	0.844	0.857	0.875	0.833	0.875	0.833
	testing set	0.722	0.767	0.733	0.8	0.786	0.75
Predictor	training set	0.948	0.857	0.75	1	1	0.75
	testing set	0.73	0.8	0.733	0.867	0.846	0.765



Replaced name	Feature	Value
a	GLRLM_RV_HLH_PT	-0.27
b	GLSZM_ZV_HHH_PT	-0.29
c	GLCM_DA_LHH_CT	-0.64
d	GLRLM_LGRE_HLH_CT	0.47

Replaced name	Feature	Value
a	GLRLM_RV_HLH_PT	-0.09
b	GLSZM_ZV_HHH_PT	-0.42
c	GLCM_DA_LHH_CT	-0.77
d	GLRLM_LGRE_HLH_CT	0.42

Figure 5-3. The patient displayed on panel (A) was a pCR patient after NAC. It was a 55 years old Her 2 positive subtype breast cancer patient. The model correctly predicted it. The value of four radiomic features were -0.27, -0.29, -0.64 and 0.47, respectively. The patient displayed on panel (B) was a non-pCR patient after NAC. It was a 59 years old lunimal B1 subtype breast cancer patient. The model also correctly predicted it. The value of four radiomic features were -0.09, -0.42, -0.77 and 0.42, respectively.

5.4.4. Association of Radiomic features with Clinical Variables

We investigated possible correlations of immunohistochemical information with the radiomic features. We found that GLCM_DA_LHH_CT had a significant negative correlation with molecular subtype while positive correlation with ER (p = 0.02 and p = 0.04 respectively). In addition, GLRLM_RV_HLH_PT and Ki-67 indicators also showed a significant negative correlation (p = 0.03) (Figure 5-4).

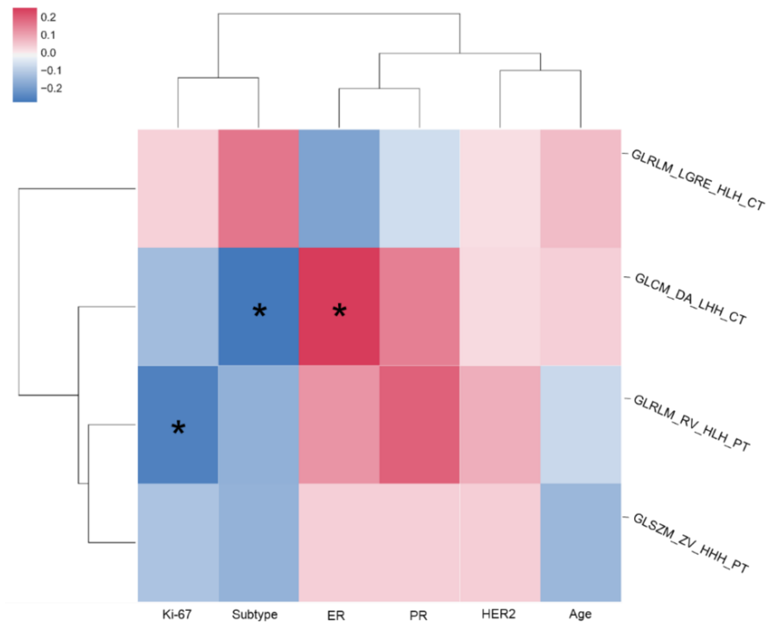


Figure 5-4. Correlation analyses of selected radiomic features with clinical features. The "* "means strong correlation, $p \leq 0.05$. The red color denotes positive correlation, the blue color denotes negative correlation, and the shade of color indicates the correlation intensity.

5.4.5. Failed Analysis with Its Heterogeneity

To examine the cases where PET-CT radiomic pattern failed, we retrospectively analysed the experimental results of ten-fold validation based on the training group of 70 patients. To reduce the bias introduced into the prediction model due to the insufficient training, we performed 30 times random ten-fold validation, and summarized the wrong prediction frequency of these 30 times validations for each case. Cases with consistent prediction error would be marked as wrong prediction and Figure 5-5 summarizes the results of these examinations. It could be observed that the PET-CT radiomic pattern was more capable of predicting the treatment prognosis for luminal B2 and TN. In addition, the performance of radiomic patterns based on pure PET or CT was also examined. Among the 24 cases which could not be dealt with by PET-CT radiomic pattern, there were 13 cases, as shown in Table 5-6, exhibiting better prediction results using pure PET radiomic pattern, and there were 8 cases, as shown in Table 5-7, exhibiting better prediction results using pure CT radiomic pattern. There were also 4 Her2+ cases which could be predicted solely based on the image features, as shown in Table 5-8.

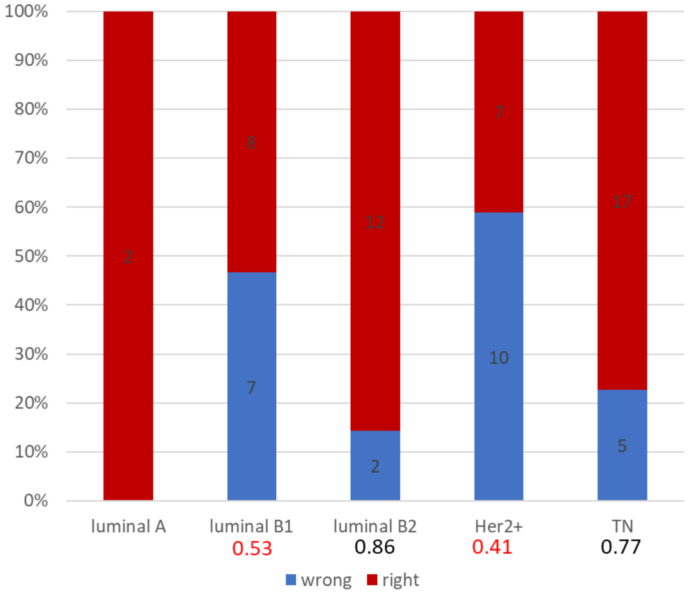


Figure 5-5. Retrospective examinations of cases using PET-CT radiomic pattern in ten-fold validations.

Table 5-6 Cases with better prediction results using PET radiomic pattern.

Pid	Prognostic	Subtype	CT	PET	PET-CT
10367735	0	luminal B1	30	0	9
10619045	0	luminal B1	14	0	12
10788498	0	luminal B1	0	5	18
117575	0	luminal B1	30	2	30
10551997	0	luminal B2	19	4	20
10783695	0	Her2+	30	0	13
10857910	1	Her2+	0	1	14
10554442	1	Her2+	0	0	16
10782145	0	Her2+	30	2	30
10480212	0	Her2+	29	0	30
10356750	0	TN	30	0	13
10627797	0	TN	0	0	30
10689877	1	TN	5	0	30

Table 1-7. Cases with better prediction results using CT radiomic pattern.

Pid	Prognostic	Subtype	CT	PET	PET-CT
10788498	0	luminal B1	0	5	18
10350033	0	luminal B1	8	30	30
10779233	0	luminal B2	0	30	30
10857910	1	Her2+	0	1	14
10554442	1	Her2+	0	0	16
10628341	1	TN	0	12	14
10627797	0	TN	0	0	30
10689877	1	TN	5	0	30

Table 5-8. Cases could not be predicted solely based on radiomic patterns.

Pid	Prognostic	Subtype	CT	PET	PET-CT
10326389	1	Her2+	30	18	22
10354311	1	Her2+	29	23	29
10293278	1	Her2+	30	29	30
10333807	1	Her2+	30	30	30

5.5. Discussion

In this study, our main findings are: 1) we constructed a model using radiomic features to identify those patients who would have a pCR following NAC from a single pre-treatment PET-CT scan; 2) the radiomic features were associated the immunohistochemistry parameters. To the best of our knowledge this is the first pCR prediction model for NAC using a pretreatment ^{18}F-FDG PET/CT and immunohistochemistry; 3) building a pCR prediction model was able to accurately distinguish between pCR and non-pCR in a highly heterogeneous cohort of ^{18}F-FDG PET/CT textures.

The first finding of this study was that we constructed a model with high accuracy using radiomic features to identify those patients who would have a pCR following NAC from a single pre-treatment PET-CT scan. In these experiments, we analyzed the prediction ability of clinical information and immunohistochemistry parameters, more details in Table 5-3. Some of univariate performance of these clinical features conform to our univariate criteria, and TFS evaluated the clinical predictors with the best predictive performance (accuracy = 0.757) in training set which is nearby the performance of radiomics predictor (accuracy = 0.767). However, it showed a poor prediction performance in the independent validation set (accuracy = 0.5). When feeding with the radiomic features and the clinical data into the prediction model, a more stable predictive capability was achieved in the independent set. The experimental comparisons validated that the non-invasive PET/CT radiomic predictors outperformed clinical factors obtained from invasive biopsy pathological results. The work reported by Antunovic et al included 79 patients with breast cancer, the study built four different models for pCR prediction in two methods. One model only used the clinical data and another used high-order imaging features, in all models the area under the curve values ranged from 0.70 to 0.73. Our study, however, combined the clinical data and the radiomics features in the one model [170]. Besides, the pre NAC ^{18}F-FDG PET/CT standard parameters such as the SUVmax, TLG [171, 172] or the ^{18}F-FDG uptake changes between pre and post two cycles of NAC [159, 173] have been widely studied the correlation with breast cancer NAC response. In this study, we also analyzed the PET/CT standard parameters, but these features were excluded during the feature selection.

In combination with age, the compound radiomic predictors achieved an accuracy of 0.8, a sensitivity of 0.733 and a specificity of 0.867 for pCR as the end point（Table 5-4）. This finding is consistent with previous reports where younger patients had a higher rate of pCR when compared to older patients [174-176]. But in overall study population, the recurrence and death were worse in young age breast cancer patients who more likely to have an aggressive biological subtype. So more attention should be paid for the young breast cancer to achieve a better long-term outcomes.

The second finding was that some higher-order imaging features have close associations with immunohistochemical parameters and importantly these features were included in the prediction model. It was found that GLCM_DA_LHH_CT has a significant negative correlation with molecular subtype ($p = 0.02$) while positive correlation with ER ($p = 0.04$). Previous studies have shown that the pCR rate varied with breast cancer molecular subtypes, specifically, the TN and HER2 positive molecular subtypes showed higher pCR rate of NAC [177, 178]. Some researches also reported that the high-order imaging features have closer associations with molecular phenotypes, risk of recurrence risk and prognosis [179, 180]. The study also found that the GLRLM_RV_HLH_PT and Ki-67 indicators also showed a significant negative correlation ($p = 0.03$) (Figure 5-4). Ki-67, a marker of cell proliferation activity, is a specific nuclear antigen expressed in all phases of the cell-cycle. Studies have demonstrated the clinical validity of Ki-67 as a predictive marker in the neoadjuvant setting and as a prognostic factor in responder/non-responder groups [181, 182]. These reported that the high Ki-67 expression before NAC was associated with higher pCR rate than low Ki-67 in breast cancer patients.

Intratumoral heterogeneity, where there are tumor cells with different genotypes and phenotypes within the same tumor, is considered to be a major factor in tumor evolution/progression and resistance to chemo- and radiotherapy [183]. Radiomics offers a potential means to characterize heterogeneity given that tissue biopsies only sample a very small part of the tumor. Breast cancer is a heterogeneous disease that varies in presentation, morphological features, behavior, and response to therapy [184]. The high BC mortality is directly endowed by the lack of effective markers to estimate the risk of cancer progression and to predict tumor response to chemotherapy and radiotherapy. The PET/CT radiomics might provide a more precise method for tumor molecular and effective marker determination. The subgroup analysis results indicated strong heterogeneity that the PET/CT prediction radiomic pattern was more capable of predicting the luminal B2 among luminal molecular subtype tumors, we could predict pCR with an accuracy of 0.86. Relative to luminal breast cancers, the TN and Her2 + subtypes had more frequently achieving pCR [185, 186], and the subgroup analysis showed that TN patients response to NAC was accurately detected on the basis of pre-NAC PET/CT. This work was aim to validate the radiomic for the noninvasive prediction of pCR in NAC recipients based on pretreatment 18F-FDG PET/CT. Although this study has achieved preliminary success, when applied to clinic, many attentions should be paid. For the same prediction model, different molecular subtypes showed different accuracy, a MRI image radiomic showed the same conclusion [166]. When applying the model to clinical patients, it is necessary to clarify the molecular classification tendencies of the model. Or applying different prediction models for the different molecular subtypes to achieve more individualized and accurate results.

This study represented a preliminary success for the pCR prediction model based on the breast cancer ^{18}F-FDG PET/CT image, which could potentially assist personalized treatment and reduce ineffective treatment. But our research still has some limitations and much more work remains to be done. First of all, the study was a retrospective work so many patients without complete information being kicked out of the research, at last only 100 patients image data were included. For the future, expanding the sample size and multicenter research should be performed in order to obtain more accurate and reliable results. Moreover, the PET/CT scan of post two NAC cycles was also often performed in clinical work, the advanced image analysis should be further evaluated in a large population.

5.6. Chapter Summary

We propose a more reasonable and fairer feature selection framework to select the most reproducible, representative and robust feature set. We further proposed a traversal selection method optimizing the computational complexity of the selection process to make sure robust feature set can be selected. Then, construction a model to predict pCR prior to NAC using four radiomic PET/CT features, derived from a single pre-treatment PET/CT scan and patient age. Further, we found the strong correlation between radiomic features and immunohistochemical receptor expression, then reveal its underlying biological basis and implication.

Chapter 6. Conclusion and Future Work

6.1. Conclusion

In this thesis, we proposed three novel models on machine learning to integrate radiomic imaging, molecular assays and clinical information, and identifying the connections between them.

Firstly, we proposed a mature framework that involves an unsupervised clustering method guiding supervised learning methods, based on late-stage NSCLC for survival prediction. Unlike common feature selection methods, we add consensus clustering as a pre-processing step for selection. We selected PCM as an independent selection to avoid the problem of overfitting as it follows a strictly linear relationship. To validate the effectiveness of our framework, we compare our feature selection process against two classical machine learning models specialising in survival prediction. In addition, our framework can work well not only on imaging feature sets but also on clinical datasets.

Secondly, to evaluate the ability of radiomic differential diagnosis, we established a selection framework based on the BMB conditions of leukemia patients. Unlike previous frameworks: 1) we embedded tree importance to ensure that representative features can be selected; 2) RFE is adopted to form radiomic patterns; 3) an interpretable machine learning approach (LIME) measures the contribution of features for each case to seek innovative clinical decisions; and 4) we evaluated the influence of both intra-observer and inter-observer variabilities on the performance of our prediction model; we determinted that the error in the segmentation variability of bone marrow lesions has no significant impact on the predictions.

Thirdly, to assess the ability of radiomic treatment prognosis, we constructed a framework to predict the status of pCR in patients with locally advanced breast cancer. Stratified cross-validation is embedded in the training process to ensure the fairness and stability of the data splits. Then we proposed a TFS method to optimise the computational complexity of the selection process. Analysing the mean and variance of each fold in 30-fold cross-validation, we selected the most robust feature set. This approach has played an important role in the evaluation of tumor heterogeneity in breast cancer subtypes and the conclusion is consistent with clinical experiments. Most importantly, aside from using LIME to evaluate feature contribution, we also calculated the correlation coefficient between radiomics with clinical data to reveal its biological insight.

6.2. Future Work

This thesis aims to explore the potential of radiomics in prognostic medical applications and to understand the relationships between radiomics, oncology and clinical biology. Because of the high dimensional data structure, we have proposed mature feature selection frameworks in chapters 3, 4 and 5 to select the most suitable radiomic pattern. And we have evaluated and validated radiomics' ability to differentiate diagnostic and prognostic treatments by using advanced machine learning algorithms (chapters 4, 5). This Mphil study can be used to develop predictive imaging biomarkers in order to identify patients who will benefit the most (or least) from a particular treatment.

To this end, work in the future stem from the following:

1) Adhere to the standardisation of image acquisition and processing to increase the reproducibility of radiomic analysis. The different imaging parameters are inevitable to define radiomic features. We conjecture that special image transformation and interpolation produces the best performance; this is followed by making reproducible experimental results and rational interpretations [66].

2) Add deep features to find more important features in the related study. Deep learning, [187] a kind of multi-hidden layer machine learning model, has become a hot spot in medical image analysis. Unlike traditional hand-crafted feature models, although it faces the challenges of big data requirements and high computational complexity, it can implicitly and automatically learn deeper features from data samples and discover complex structures in high-dimensional datasets [188].

3) Establish more fundamental biological connections such as imaging data with genes. As the most basic biomarker, genes define the basic structure and phenotype of organisms. Imaging-genomics [189] combines medical imaging with genomics to better understand the relationship between radiomics and genomes.

4) Compare and assess radiomics' predictive power of other microenvironments in the tumour. In addition to providing insight of tumour characteristics, microenvironment features are also important for the guidance of medical decisions as it can influence carcinogenesis, cancer progression and therapy response [190]. Therefore, this should be appear in the future work.

Bibliography

[1] B. Xin, C. Xu, L. Wang, T. Dong, C. Zheng, and X. Wang, "Integrative Clustering and Supervised Feature Selection for Clinical Applications," in *2018 15th International Conference on Control, Automation, Robotics and Vision (ICARCV)*, 2018, pp. 1316-1320: IEEE.

[2] H. Li *et al.*, "18F-FDG PET/CT Radiomic Analysis with Machine Learning for Identifying Bone Marrow Involvement in the Patients with Suspected Relapsed Acute Leukemia," *Theranostics,* vol. 9, no. 16, p. 4730, 2019.

[3] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: a cancer journal for clinicians,* vol. 68, no. 6, pp. 394-424, 2018.

[4] B. Stewart and C. P. Wild, "World cancer report 2014," 2014.

[5] S. M. Garland *et al.*, "Quadrivalent vaccine against human papillomavirus to prevent anogenital diseases," *New England Journal of Medicine,* vol. 356, no. 19, pp. 1928-1943, 2007.

[6] S. Ogholikhan and K. Schwarz, "Hepatitis vaccines," *Vaccines,* vol. 4, no. 1, p. 6, 2016.

[7] F. Bray *et al.*, *Planning and developing population-based cancer registration in low- and middle-income settings*. IARC (International Agency for Research on Cancer), 2014.

[8] M. Plummer, C. de Martel, J. Vignat, J. Ferlay, F. Bray, and S. Franceschi, "Global burden of cancers attributable to infections in 2012: a synthetic analysis," *The Lancet Global Health,* vol. 4, no. 9, pp. e609-e616, 2016.

[9] F. S. Collins and H. Varmus, "A new initiative on precision medicine," *New England journal of medicine,* vol. 372, no. 9, pp. 793-795, 2015.

[10] S. L. Sangisetty and T. J. Miner, "Malignant ascites: a review of prognostic factors, pathophysiology and therapeutic measures," *World journal of gastrointestinal surgery,* vol. 4, no. 4, p. 87, 2012.

[11] P. Lambin *et al.*, "Radiomics: extracting more information from medical images using advanced feature analysis," *European journal of cancer,* vol. 48, no. 4, pp. 441-446, 2012.

[12] J. L. Jameson and D. L. Longo, "Precision medicine—personalized, problematic, and promising," *Obstetrical & Gynecological Survey,* vol. 70, no. 10, pp. 612-614, 2015.

[13] S. J. Aronson and H. L. Rehm, "Building the foundation for genomics in precision medicine," *Nature,* vol. 526, no. 7573, p. 336, 2015.

[14] A. Marusyk and K. Polyak, "Tumor heterogeneity: causes and consequences," *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer,* vol. 1805, no. 1, pp. 105-117, 2010.

[15] M. Gerlinger *et al.*, "Intratumor heterogeneity and branched evolution revealed by multiregion sequencing," *New England journal of medicine,* vol. 366, no. 10, pp. 883-892, 2012.

[16] M. Gerlinger and C. Swanton, "How Darwinian models inform therapeutic failure initiated by clonal heterogeneity in cancer medicine," *British journal of cancer,* vol. 103, no. 8, p. 1139, 2010.

[17] N. C. Turner and J. S. Reis-Filho, "Genetic heterogeneity and cancer drug resistance," *The lancet oncology,* vol. 13, no. 4, pp. e178-e185, 2012.

[18] J. L. Gevertz, G. T. Gillies, and S. Torquato, "Simulating tumor growth in confined heterogeneous environments," *Physical Biology,* vol. 5, no. 3, p. 036010, 2008.

[19] A. Mordente, E. Meucci, G. E. Martorana, and A. Silvestrini, "Cancer biomarkers discovery and validation: state of the art, problems and future perspectives," in *Advances in Cancer Biomarkers*: Springer, 2015, pp. 9-26.

[20] N. L. Henry and D. F. Hayes, "Cancer biomarkers," *Molecular oncology,* vol. 6, no. 2, pp. 140-146, 2012.

[21] R. A. Mukhtar, O. Nseyo, M. J. Campbell, and L. J. Esserman, "Tumor-associated macrophages in breast cancer as potential biomarkers for new treatments and diagnostics," *Expert review of molecular diagnostics,* vol. 11, no. 1, pp. 91-100, 2011.

[22] A. J. Vargas and C. C. Harris, "Biomarker development in the precision medicine era: lung cancer as a case study," *Nature Reviews Cancer,* vol. 16, no. 8, p. 525, 2016.

[23] J. A. Ludwig and J. N. Weinstein, "Biomarkers in cancer staging, prognosis and treatment selection," *Nature Reviews Cancer,* vol. 5, no. 11, p. 845, 2005.

[24] X. Chen and S. T. Wong, "Cancer theranostics: An introduction," in *Cancer Theranostics*: Elsevier, 2014, pp. 3-8.

[25] T. Kosaka, Y. Yatabe, H. Endoh, H. Kuwano, T. Takahashi, and T. Mitsudomi, "Mutations of the epidermal growth factor receptor gene in lung cancer: biological and clinical implications," *Cancer research,* vol. 64, no. 24, pp. 8919-8923, 2004.

[26] W. Pao and N. Girard, "New driver mutations in non-small-cell lung cancer," *The lancet oncology,* vol. 12, no. 2, pp. 175-180, 2011.

[27] J. D. Campbell *et al.*, "Distinct patterns of somatic genome alterations in lung adenocarcinomas and squamous cell carcinomas," *Nature genetics,* vol. 48, no. 6, p. 607, 2016.

[28] Y. J. Kim *et al.*, "Verification of the biomarker candidates for non-small-cell lung cancer using a targeted proteomics approach," *Journal of proteome research,* vol. 14, no. 3, pp. 1412-1419, 2015.

[29] M. Castellarin *et al.*, "Fusobacterium nucleatum infection is prevalent in human colorectal carcinoma," *Genome research,* vol. 22, no. 2, pp. 299-306, 2012.

[30] K. Mima *et al.*, "Fusobacterium nucleatum in colorectal carcinoma tissue and patient prognosis," *Gut,* vol. 65, no. 12, pp. 1973-1980, 2016.

[31] L. Flanagan *et al.*, "Fusobacterium nucleatum associates with stages of colorectal neoplasia development, colorectal cancer and disease outcome," *European journal of clinical microbiology & infectious diseases,* vol. 33, no. 8, pp. 1381-1390, 2014.

[32] T. Yu *et al.*, "Fusobacterium nucleatum promotes chemoresistance to colorectal cancer by modulating autophagy," *Cell,* vol. 170, no. 3, pp. 548-563. e16, 2017.

[33] K. Nosho *et al.*, "Association of Fusobacterium nucleatum with immunity and molecular alterations in colorectal cancer," *World journal of gastroenterology,* vol. 22, no. 2, p. 557, 2016.

[34] M. Koi, Y. Okita, and J. M. Carethers, "Fusobacterium nucleatum Infection in colorectal cancer: linking inflammation, DNA mismatch repair and genetic and epigenetic alterations," *Journal of the anus, rectum and colon,* vol. 2, no. 2, pp. 37-46, 2018.

[35] S. S. Yip and H. J. Aerts, "Applications and limitations of radiomics," *Physics in Medicine & Biology,* vol. 61, no. 13, p. R150, 2016.

[36] R. Weissleder, "Molecular imaging in cancer," *Science,* vol. 312, no. 5777, pp. 1168-1171, 2006.

[37] L. Fass, "Imaging and cancer: a review," *Molecular oncology,* vol. 2, no. 2, pp. 115-152, 2008.

[38] K. Iniewski, *Medical Imaging*. Wiley Online Library, 2009.

[39] H. Li *et al.*, "Automated PET/CT radiomic analysis with machine learning for identifying bone marrow involvement in the patients with suspected relapsed acute leukemia," *Journal of Nuclear Medicine,* vol. 60, no. supplement 1, pp. 619-619, 2019.

[40] H. J. Aerts *et al.*, "Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach," *Nature communications,* vol. 5, p. 4006, 2014.

[41] S. Narang, M. Lehrer, D. Yang, J. Lee, and A. Rao, "Radiomics in glioblastoma: current status, challenges and potential opportunities," *Translational Cancer Research,* vol. 5, no. 4, pp. 383-397, 2016.

[42] P. Lambin *et al.*, "Radiomics: the bridge between medical imaging and personalized medicine," *Nature Reviews Clinical Oncology,* vol. 14, no. 12, p. 749, 2017.

[43] R. J. Gillies, P. E. Kinahan, and H. Hricak, "Radiomics: images are more than pictures, they are data," *Radiology,* vol. 278, no. 2, pp. 563-577, 2015.

[44] H. J. Aerts *et al.*, "Defining a radiomic response phenotype: a pilot study using targeted therapy in NSCLC," *Scientific reports,* vol. 6, p. 33860, 2016.

[45] P. Grossmann *et al.*, "Quantitative imaging biomarkers for risk stratification of patients with recurrent glioblastoma treated with bevacizumab," *Neuro-oncology,* vol. 19, no. 12, pp. 1688-1697, 2017.

[46] N. M. Braman *et al.*, "Intratumoral and peritumoral radiomics for the pretreatment prediction of pathological complete response to neoadjuvant chemotherapy based on breast DCE-MRI," *Breast Cancer Research,* vol. 19, no. 1, p. 57, 2017.

[47] M. Fan, G. Wu, H. Cheng, J. Zhang, G. Shao, and L. Li, "Radiomic analysis of DCE-MRI for prediction of response to neoadjuvant chemotherapy in breast cancer patients," *European journal of radiology,* vol. 94, pp. 140-147, 2017.

[48] S. Ha, S. Park, J.-I. Bang, E.-K. Kim, and H.-Y. Lee, "Metabolic radiomics for pretreatment 18 F-FDG PET/CT to characterize locally advanced breast cancer: histopathologic characteristics, response to neoadjuvant chemotherapy, and prognosis," *Scientific reports,* vol. 7, no. 1, p. 1556, 2017.

[49] F. H. van Velden *et al.*, "Evaluation of a cumulative SUV-volume histogram method for parameterizing heterogeneous intratumoural FDG uptake in non-small cell lung cancer PET studies," *European journal of nuclear medicine and molecular imaging,* vol. 38, no. 9, pp. 1636-1647, 2011.

[50] N.-M. Cheng, Y.-H. D. Fang, and T.-C. Yen, "The promise and limits of PET texture analysis," *Annals of nuclear medicine,* vol. 27, no. 9, pp. 867-869, 2013.

[51] F. Yang, M. A. Thomas, F. Dehdashti, and P. W. Grigsby, "Temporal analysis of intratumoral metabolic heterogeneity characterized by textural features in cervical cancer," *European journal of nuclear medicine and molecular imaging,* vol. 40, no. 5, pp. 716-727, 2013.

[52] A. Marusyk, V. Almendro, and K. Polyak, "Intra-tumour heterogeneity: a looking glass for cancer?," *Nature Reviews Cancer,* vol. 12, no. 5, p. 323, 2012.

[53] H. Zhang *et al.*, "Modeling pathologic response of esophageal cancer to chemoradiation therapy using spatial-temporal 18F-FDG PET features, clinical parameters, and demographics," *International Journal of Radiation Oncology* Biology* Physics,* vol. 88, no. 1, pp. 195-203, 2014.

[54] T. P. Coroller *et al.*, "CT-based radiomic signature predicts distant metastasis in lung adenocarcinoma," *Radiotherapy and Oncology,* vol. 114, no. 3, pp. 345-350, 2015.

[55] J.-b. Qin *et al.*, "Grading of gliomas by using radiomic features on multiple magnetic resonance imaging (MRI) sequences," *Medical science monitor: international medical journal of experimental and clinical research,* vol. 23, p. 2168, 2017.

[56] X. Wang *et al.*, "Machine learning models for multiparametric glioma grading with quantitative result interpretations," *Frontiers in neuroscience,* vol. 12, 2018.

[57] X. Zhang *et al.*, "Radiomics assessment of bladder cancer grade using texture features from diffusion-weighted imaging," *Journal of Magnetic Resonance Imaging,* vol. 46, no. 5, pp. 1281-1288, 2017.

[58] L. Kjaer, P. Ring, C. Thomsen, and O. Henriksen, "Texture analysis in quantitative MR imaging: tissue characterisation of normal brain and intracranial tumours at 1.5 T," *Acta Radiologica,* vol. 36, no. 2, pp. 127-135, 1995.

[59] S. Herlidou-Meme *et al.*, "MRI texture analysis on texture test objects, normal brain and intracranial tumors," *Magnetic resonance imaging,* vol. 21, no. 9, pp. 989-993, 2003.

[60] E. Sala *et al.*, "Unravelling tumour heterogeneity using next-generation imaging: radiomics, radiogenomics, and habitat imaging," *Clinical radiology,* vol. 72, no. 1, pp. 3-10, 2017.

[61] D. Mahmoud-Ghoneim, G. Toussaint, J.-M. Constans, and D. Jacques, "Three dimensional texture analysis in MRI: a preliminary evaluation in gliomas," *Magnetic resonance imaging,* vol. 21, no. 9, pp. 983-987, 2003.

[62] H. Wang *et al.*, "Comparison of machine learning methods for classifying mediastinal lymph node metastasis of non-small cell lung cancer from 18 F-FDG PET/CT images," *EJNMMI research,* vol. 7, no. 1, p. 11, 2017.

[63] I. Petkovska *et al.*, "Pulmonary nodule characterization: a comparison of conventional with quantitative and visual semi-quantitative analyses using contrast enhancement maps," *European journal of radiology,* vol. 59, no. 2, pp. 244-252, 2006.

[64] H. Li *et al.*, "MR imaging radiomics signatures for predicting the risk of breast cancer recurrence as given by research versions of MammaPrint, Oncotype DX, and PAM50 gene assays," *Radiology,* vol. 281, no. 2, pp. 382-391, 2016.

[65] R. Stoyanova *et al.*, "Association of multiparametric MRI quantitative imaging features with prostate cancer gene expression in MRI-targeted prostate biopsies," *Oncotarget,* vol. 7, no. 33, p. 53362, 2016.

[66] V. Kumar *et al.*, "Radiomics: the process and the challenges," *Magnetic resonance imaging,* vol. 30, no. 9, pp. 1234-1248, 2012.

[67] A. Makandar and B. Halalli, "Threshold based segmentation technique for mass detection in mammography," *J Comput,* vol. 11, no. 6, pp. 472-478, 2016.

[68] A. Pratondo, C.-K. Chui, and S.-H. Ong, "Robust edge-stop functions for edge-based active contour models in medical image segmentation," *IEEE Signal Processing Letters,* vol. 23, no. 2, pp. 222-226, 2015.

[69] R. Achanta, F. Estrada, P. Wils, and S. Süsstrunk, "Salient region detection and segmentation," in *International conference on computer vision systems*, 2008, pp. 66-75: Springer.

[70] M. W. Khan, "A survey: image segmentation techniques," *International Journal of Future Computer and Communication,* vol. 3, no. 2, p. 89, 2014.

[71] F. Ma, M. Bajger, J. P. Slavotinek, and M. J. Bottema, "Two graph theory based methods for identifying the pectoral muscle in mammograms," *Pattern Recognition,* vol. 40, no. 9, pp. 2592-2602, 2007.

[72] A. Yezzi, S. Kichenassamy, A. Kumar, P. Olver, and A. Tannenbaum, "A geometric snake model for segmentation of medical imagery," *IEEE Transactions on medical imaging,* vol. 16, no. 2, pp. 199-209, 1997.

[73] Y. Yang and B. Wu, "Convex image segmentation model based on local and global intensity fitting energy and split Bregman method," *Journal of Applied Mathematics,* vol. 2012, 2012.

[74] M. Yasmin, M. Sharif, and S. Mohsin, "Neural networks in medical imaging applications: A survey," *World Applied Sciences Journal,* vol. 22, no. 1, pp. 85-96, 2013.

[75] F. Visin *et al.*, "Reseg: A recurrent neural network-based model for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2016, pp. 41-48.

[76] C. Liu, W. Liu, and W. Xing, "An improved edge-based level set method combining local regional fitting information for noisy image segmentation," *Signal Processing,* vol. 130, pp. 12-21, 2017.

[77] Y. Y. Boykov and M.-P. Jolly, "Interactive graph cuts for optimal boundary & region segmentation of objects in ND images," in *Proceedings eighth IEEE international conference on computer vision. ICCV 2001*, 2001, vol. 1, pp. 105-112: IEEE.

[78] C. Rother, V. Kolmogorov, and A. Blake, "Grabcut: Interactive foreground extraction using iterated graph cuts," in *ACM transactions on graphics (TOG)*, 2004, vol. 23, no. 3, pp. 309-314: ACM.

[79] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," *International journal of computer vision,* vol. 1, no. 4, pp. 321-331, 1988.

[80] A. Zwanenburg, S. Leger, M. Vallières, and S. Löck, "Image biomarker standardisation initiative," *arXiv preprint arXiv:1612.07003,* 2016.

[81] J. J. van Griethuysen *et al.*, "Computational radiomics system to decode the radiographic phenotype," *Cancer research,* vol. 77, no. 21, pp. e104-e107, 2017.

[82] X. Yang *et al.*, "Ultrasound GLCM texture analysis of radiation-induced parotid-gland injury in head-and-neck cancer radiotherapy: An in vivo study of late toxicity," *Medical physics,* vol. 39, no. 9, pp. 5732-5739, 2012.

[83] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of machine learning research,* vol. 3, no. Mar, pp. 1157-1182, 2003.

[84] M. H. Law, M. A. Figueiredo, and A. K. Jain, "Simultaneous feature selection and clustering using mixture models," *IEEE transactions on pattern analysis and machine intelligence,* vol. 26, no. 9, pp. 1154-1166, 2004.

[85] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Computers & Electrical Engineering,* vol. 40, no. 1, pp. 16-28, 2014.

[86] G. H. John, R. Kohavi, and K. Pfleger, "Irrelevant features and the subset selection problem," in *Machine Learning Proceedings 1994*: Elsevier, 1994, pp. 121-129.

[87] M. M. Mukaka, "A guide to appropriate use of correlation coefficient in medical research," *Malawi Medical Journal,* vol. 24, no. 3, pp. 69-71, 2012.

[88] S. Wu, C. M. Crespi, and W. K. Wong, "Comparison of methods for estimating the intraclass correlation coefficient for binary responses in cancer prevention cluster randomized trials," *Contemporary clinical trials,* vol. 33, no. 5, pp. 869-880, 2012.

[89] J. A. Thomas and T. Cover, "Elements of information theory," *John Wiley & Sons, Inc., New York. Toni, T., Welch, D., Strelkowa, N., Ipsen, A., and Stumpf, MPH (2009),"Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems," Journal of the Royal Society Interface,* vol. 6, pp. 187-202, 1991.

[90] Z. Xu, I. King, M. R.-T. Lyu, and R. Jin, "Discriminative semi-supervised feature selection via manifold regularization," *IEEE Transactions on Neural networks,* vol. 21, no. 7, pp. 1033-1047, 2010.

[91] T. Rückstieß, C. Osendorfer, and P. van der Smagt, "Sequential feature selection for classification," in *Australasian Joint Conference on Artificial Intelligence*, 2011, pp. 132-141: Springer.

[92] L. Breiman, "Random forests," *Machine learning,* vol. 45, no. 1, pp. 5-32, 2001.

[93] N. Gunduz and E. Fokoué, "Robust classification of high dimension low sample size data," *arXiv preprint arXiv:1501.00592,* 2015.

[94] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial intelligence,* vol. 97, no. 1-2, pp. 273-324, 1997.

[95] Y. Zhang, S. Li, T. Wang, and Z. Zhang, "Divergence-based feature selection for separate classes," *Neurocomputing,* vol. 101, pp. 32-42, 2013.

[96] T. Abeel, T. Helleputte, Y. Van de Peer, P. Dupont, and Y. Saeys, "Robust biomarker identification for cancer diagnosis with ensemble feature selection methods," *Bioinformatics,* vol. 26, no. 3, pp. 392-398, 2009.

[97] V. Bolón-Canedo, N. Sánchez-Maroño, and A. Alonso-Betanzos, "A review of feature selection methods on synthetic data," *Knowledge and information systems,* vol. 34, no. 3, pp. 483-519, 2013.

[98] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological),* pp. 267-288, 1996.

[99] J. G. Dy and C. E. Brodley, "Feature selection for unsupervised learning," *Journal of machine learning research,* vol. 5, no. Aug, pp. 845-889, 2004.

[100] P. Xanthopoulos, "A review on consensus clustering methods," in *Optimization in science and engineering*: Springer, 2014, pp. 553-566.

[101] S. Monti, P. Tamayo, J. Mesirov, and T. Golub, "Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data," *Machine learning,* vol. 52, no. 1-2, pp. 91-118, 2003.

[102] M. T. Ribeiro, S. Singh, and C. Guestrin, "Model-agnostic interpretability of machine learning," *arXiv preprint arXiv:1606.05386,* 2016.

[103] L. Glasser and R. Doerfler, "A brief introduction to nomography: graphical representation of mathematical relationships," *International Journal of Mathematical Education in Science and Technology,* pp. 1-12, 2018.

[104] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Ijcai*, 1995, vol. 14, no. 2, pp. 1137-1145: Montreal, Canada.

[105] P. A. Devijver and J. Kittler, *Pattern recognition: A statistical approach*. Prentice hall, 1982.

[106] S. V. Stehman, "Selecting and interpreting measures of thematic classification accuracy," *Remote sensing of Environment,* vol. 62, no. 1, pp. 77-89, 1997.

[107] S. Kudyba and M. Kwatinetz, "Introduction to the big data era," *Big Data, Mining, and Analytics,* pp. 1-17, 2014.

[108] L. Wang, Y. Wang, and Q. Chang, "Feature selection methods for big data bioinformatics: A survey from the search perspective," *Methods,* vol. 111, pp. 21-31, 2016.

[109] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," *Computational and structural biotechnology journal,* vol. 13, pp. 8-17, 2015.

[110] C. C. Chen, H. Schwender, J. Keith, R. Nunkesser, K. Mengersen, and P. Macrossan, "Methods for identifying SNP interactions: a review on variations of Logic Regression, Random Forest and Bayesian logistic regression," *IEEE/ACM transactions on computational biology and bioinformatics,* vol. 8, no. 6, pp. 1580-1591, 2011.

[111] D. R. Cox, *Analysis of survival data*. Routledge, 2018.

[112] S. Pölsterl, S. Conjeti, N. Navab, and A. Katouzian, "Survival analysis for high-dimensional, heterogeneous medical data: Exploring feature extraction as an alternative to feature selection," *Artificial intelligence in medicine,* vol. 72, pp. 1-11, 2016.

[113] T. A. Gerds, M. W. Kattan, M. Schumacher, and C. Yu, "Estimating a time-dependent concordance index for survival prediction models with covariate dependent censoring," *Statistics in Medicine,* vol. 32, no. 13, pp. 2173-2184, 2013.

[114] N. Kröger, M. Bishop, S. Giralt, and A. Wayne, "Third International workshop on the biology, prevention, and treatment of relapse after stem cell transplantation," ed: Nature Publishing Group, 2018.

[115] M.-E. Percival, C. Lai, E. Estey, and C. S. Hourigan, "Bone marrow evaluation for diagnosis and monitoring of acute myeloid leukemia," *Blood reviews,* vol. 31, no. 4, pp. 185-192, 2017.

[116] A. Cistaro, F. Saglio, S. Asaftei, P. Fania, M. Berger, and F. Fagioli, "The role of 18F-FDG PET/CT in pediatric lymph-node acute lymphoblastic leukemia involvement," *Radiology case reports,* vol. 6, no. 4, p. 503, 2011.

[117] F. Stölzel *et al.*, "18F-FDG-PET/CT for detection of extramedullary acute myeloid leukemia," *Haematologica,* vol. 96, no. 10, pp. 1552-1556, 2011.

[118] A. S. W. I. Cribe, M. Steenhof, C. W. Marcher, H. Petersen, H. Frederiksen, and L. S. Friis, "Extramedullary disease in patients with acute myeloid leukemia assessed by (18) F-FDG PET," *European journal of haematology,* vol. 90, no. 4, pp. 273-278, 2013.

[119] W.-l. Zhou, H.-b. Wu, L.-j. Wang, Y. Tian, Y. Dong, and Q.-s. Wang, "Usefulness and pitfalls of F-18-FDG PET/CT for diagnosing extramedullary acute leukemia," *European journal of radiology,* vol. 85, no. 1, pp. 205-210, 2016.

[120] S. Elojeimy, A. S. Luana, and M. T. Parisi, "Use of 18F-FDG PET-CT for Detection of Active Disease in Acute Myeloid Leukemia," *Clinical nuclear medicine,* vol. 41, no. 3, pp. e137-40, 2016.

[121] I. Cunningham and B. Kohno, "18FDG-PET/CT: 21st century approach to leukemic tumors in 124 cases," *American journal of hematology,* vol. 91, no. 4, pp. 379-384, 2016.

[122] K. Su, Y. Nakamoto, K. Nakatani, K. Kurihara, N. Hayakawa, and K. Togashi, "Diffuse homogeneous bone marrow uptake of FDG in patients with acute lymphoblastic leukemia," *Clinical nuclear medicine,* vol. 38, no. 1, pp. e33-4, 2013.

[123] G. K. Parida, R. Soundararajan, A. Passah, C. Bal, and R. Kumar, "Metabolic Skeletal Superscan on 18F-FDG PET/CT in a Case of Acute Lymphoblastic Leukemia," *Clinical nuclear medicine,* vol. 40, no. 7, pp. 567-568, 2015.

[124] M. K. Arimoto *et al.*, "Increased bone marrow uptake of 18F-FDG in leukemia patients: preliminary findings," *SpringerPlus,* vol. 4, no. 1, p. 521, 2015.

[125] Z. Su *et al.*, "Philadelphia chromosome-positive acute myeloid leukemia with masses and osteolytic lesions: finding of 18F-FDG PET/CT," *Frontiers of medicine,* vol. 11, no. 3, pp. 440-444, 2017.

[126] H. J. Adams, R. A. Nievelstein, and T. C. Kwee, "Opportunities and limitations of bone marrow biopsy and bone marrow FDG-PET in lymphoma," *Blood reviews,* vol. 29, no. 6, pp. 417-425, 2015.

[127] P. Salaun *et al.*, "Analysis of FDG PET diffuse bone marrow uptake (BMU) and splenic uptake (SU) in staging of Hodgkin lymphoma (HL): A reflect of disease infiltration or just inflammation?," *Journal of Nuclear Medicine,* vol. 50, no. supplement 2, pp. 646-646, 2009.

[128] M. V. Knopp, H. Bischoff, A. Rimac, F. Oberdorfer, and G. van Kaick, "Bone marrow uptake of fluorine-18-fluorodeoxyglucose following treatment with hematopoietic growth factors: initial evaluation," *Nuclear medicine and biology,* vol. 23, no. 6, pp. 845-849, 1996.

[129] L. Berthet *et al.*, "In newly diagnosed diffuse large B-cell lymphoma, determination of bone marrow involvement with 18F-FDG PET/CT provides better diagnostic performance and prognostic stratification than does biopsy," *Journal of Nuclear Medicine,* vol. 54, no. 8, pp. 1244-1250, 2013.

[130] H. J. Adams, T. C. Kwee, R. Fijnheer, S. V. Dubois, R. A. Nievelstein, and J. M. de Klerk, "Bone marrow FDG-PET/CT in Hodgkin lymphoma revisited: do imaging and pathology match?," *Annals of nuclear medicine,* vol. 29, no. 2, pp. 132-137, 2015.

[131] H. J. Adams, T. C. Kwee, R. Fijnheer, S. V. Dubois, R. A. Nievelstein, and J. M. de Klerk, "Bone marrow 18F-fluoro-2-deoxy-D-glucose positron emission tomography/computed tomography cannot replace bone marrow biopsy in diffuse large B-cell lymphoma," *American journal of hematology,* vol. 89, no. 7, pp. 726-731, 2014.

[132] C. Soydal *et al.*, "prognostic importance of bone marrow uptake on baseline 18F-FDG positron emission tomography in diffuse large B cell lymphoma," *Cancer Biotherapy and Radiopharmaceuticals,* vol. 31, no. 10, pp. 361-365, 2016.

[133] L. Antunovic *et al.*, "[18 F] FDG PET/CT features for the molecular characterization of primary breast tumors," *European journal of nuclear medicine and molecular imaging,* vol. 44, no. 12, pp. 1945-1954, 2017.

[134] F. Lucia *et al.*, "Prediction of outcome using pretreatment 18 F-FDG PET/CT and MRI radiomics in locally advanced cervical cancer treated with chemoradiotherapy," *European journal of nuclear medicine and molecular imaging,* vol. 45, no. 5, pp. 768-786, 2018.

[135] M. Sollini, L. Cozzi, L. Antunovic, A. Chiti, and M. Kirienko, "PET Radiomics in NSCLC: state of the art and a proposal for harmonization of methodology," *Scientific reports,* vol. 7, no. 1, p. 358, 2017.

[136] N. Aide, M. Talbot, C. Fruchart, G. Damaj, and C. Lasnon, "Diagnostic and prognostic value of baseline FDG PET/CT skeletal textural features in diffuse large B cell lymphoma," *European journal of nuclear medicine and molecular imaging,* vol. 45, no. 5, pp. 699-711, 2018.

[137] M. L. Welch *et al.*, "Vulnerabilities of radiomic signature development: The need for safeguards," *Radiotherapy and Oncology,* vol. 130, pp. 2-9, 2019.

[138] T. H. Dou, T. P. Coroller, J. J. van Griethuysen, R. H. Mak, and H. J. Aerts, "Peritumoral radiomics features predict distant metastasis in locally advanced NSCLC," *PloS one,* vol. 13, no. 11, p. e0206108, 2018.

[139] R. Yuan, S. Shi, J. Chen, and G. Cheng, "Radiomics in RayPlus: a Web-Based Tool for Texture Analysis in Medical Images," *Journal of Digital Imaging,* vol. 32, no. 2, pp. 269-275, 2019.

[140] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine learning,* vol. 46, no. 1-3, pp. 389-422, 2002.

[141] A. G. FLOARES, M. FERISGAN, D. ONITA, A. CIUPARU, G. A. CALIN, and F. B. MANOLACHE, "The Smallest Sample Size for the Desired Diagnosis Accuracy."

[142] A. Chalkidou, M. J. O'Doherty, and P. K. Marsden, "False discovery rates in PET and CT studies with texture features: a systematic review," *PloS one,* vol. 10, no. 5, p. e0124165, 2015.

[143] T. Win *et al.*, "Tumor heterogeneity and permeability as measured on the CT component of PET/CT predict survival in patients with non–small cell lung cancer," *Clinical Cancer Research,* vol. 19, no. 13, pp. 3591-3599, 2013.

[144] B. Ganeshan *et al.*, "CT-based texture analysis potentially provides prognostic information complementary to interim fdg-pet for patients with hodgkin's and aggressive non-hodgkin's lymphomas," *European radiology,* vol. 27, no. 3, pp. 1012-1020, 2017.

[145] B. Ganeshan, K. Skogen, I. Pressney, D. Coutroubis, and K. Miles, "Tumour heterogeneity in oesophageal cancer assessed by CT texture analysis: preliminary evidence of an association with tumour metabolism, stage, and survival," *Clinical radiology,* vol. 67, no. 2, pp. 157-164, 2012.

[146] S. J. Vinnicombe and R. H. Reznek, "Computerised tomography in the staging of Hodgkin's disease and non-Hodgkin's lymphoma," *European journal of nuclear medicine and molecular imaging,* vol. 30, no. 1, pp. S42-S55, 2003.

[147] T. C. Kwee, J. M. de Klerk, and R. A. Nievelstein, "Imaging of bone marrow involvement in lymphoma: state of the art and future directions," *The Scientific World Journal,* vol. 11, pp. 391-402, 2011.

[148] E.-S. Choi, S.-G. Ha, H.-S. Kim, J. H. Ha, J. C. Paeng, and I. Han, "Total lesion glycolysis by 18 F-FDG PET/CT is a reliable predictor of prognosis in soft-tissue sarcoma," *European journal of nuclear medicine and molecular imaging,* vol. 40, no. 12, pp. 1836-1842, 2013.

[149] W. Hong and E. Dong, "The past, present and future of breast cancer research in China," *Cancer Lett,* vol. 351, no. 1, pp. 1-5, Aug 28 2014.

[150] W. Chen *et al.*, "Cancer statistics in China, 2015," (in eng), *CA Cancer J Clin,* vol. 66, no. 2, pp. 115-32, Mar-Apr 2016.

[151] W. J. Gradishar *et al.*, "NCCN Guidelines Insights Breast Cancer, Version 1.2016," (in eng), *J Natl Compr Canc Netw,* vol. 13, no. 12, pp. 1475-85, Dec 2015.

[152] X. Kong, M. S. Moran, N. Zhang, B. Haffty, and Q. Yang, "Meta-analysis confirms achieving pathological complete response after neoadjuvant chemotherapy predicts favourable prognosis for breast cancer patients," *Eur J Cancer,* vol. 47, no. 14, pp. 2084-90, Sep 2011.

[153] A. S. Caudle *et al.*, "Predictors of tumor progression during neoadjuvant chemotherapy in breast cancer," *J Clin Oncol,* vol. 28, no. 11, pp. 1821-8, Apr 10 2010.

[154] W. Chu *et al.*, "Diffusion-weighted imaging in identifying breast cancer pathological response to neoadjuvant chemotherapy: A meta-analysis," *Oncotarget,* vol. 9, no. 6, pp. 7088-7100, Jan 23 2018.

[155] E. A. Eisenhauer *et al.*, "New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1)," *Eur J Cancer,* vol. 45, no. 2, pp. 228-47, Jan 2009.

[156] J. Shang *et al.*, "Comparison of RECIST, EORTC criteria and PERCIST for evaluation of early response to chemotherapy in patients with non-small-cell lung cancer," *Eur J Nucl Med Mol Imaging,* vol. 43, no. 11, pp. 1945-53, Oct 2016.

[157] D. Fuster *et al.*, "Preoperative staging of large primary breast cancer with [18F]fluorodeoxyglucose positron emission tomography/computed tomography compared with conventional imaging procedures," *J Clin Oncol,* vol. 26, no. 29, pp. 4746-51, Oct 10 2008.

[158] D. Groheux *et al.*, "(1)(8)F-FDG PET/CT for the Early Evaluation of Response to Neoadjuvant Treatment in Triple-Negative Breast Cancer: Influence of the Chemotherapy Regimen," *J Nucl Med,* vol. 57, no. 4, pp. 536-43, Apr 2016.

[159] D. Groheux *et al.*, "Baseline Tumor 18F-FDG Uptake and Modifications After 2 Cycles of Neoadjuvant Chemotherapy Are Prognostic of Outcome in ER+/HER2- Breast Cancer," *J Nucl Med,* vol. 56, no. 6, pp. 824-31, Jun 2015.

[160] D. Groheux, D. Mankoff, M. Espie, and E. Hindie, "(1)(8)F-FDG PET/CT in the early prediction of pathological response in aggressive subtypes of breast cancer: review of the literature and recommendations for use in clinical trials," *Eur J Nucl Med Mol Imaging,* vol. 43, no. 5, pp. 983-93, May 2016.

[161] M. Hatt, F. Tixier, D. Visvikis, and C. Cheze Le Rest, "Radiomics in PET/CT: More Than Meets the Eye?," *J Nucl Med,* vol. 58, no. 3, pp. 365-366, Mar 2017.

[162] H. J. Aerts *et al.*, "Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach," (in eng), *Nat Commun,* vol. 5, p. 4006, Jun 3 2014.

[163] P. Lambin *et al.*, "Radiomics: extracting more information from medical images using advanced feature analysis," *Eur J Cancer,* vol. 48, no. 4, pp. 441-6, Mar 2012.

[164] S. S. Yip and H. J. Aerts, "Applications and limitations of radiomics," *Phys Med Biol,* vol. 61, no. 13, pp. R150-66, Jul 7 2016.

[165] M. Kirienko *et al.*, "Prediction of disease-free survival by the PET/CT radiomic signature in non-small cell lung cancer patients undergoing surgery," *Eur J Nucl Med Mol Imaging,* vol. 45, no. 2, pp. 207-217, Feb 2018.

[166] S. Ha, S. Park, J. I. Bang, E. K. Kim, and H. Y. Lee, "Metabolic Radiomics for Pretreatment (18)F-FDG PET/CT to Characterize Locally Advanced Breast Cancer: Histopathologic Characteristics, Response to Neoadjuvant Chemotherapy, and Prognosis," (in eng), *Sci Rep,* vol. 7, no. 1, p. 1556, May 8 2017.

[167] H. I. Kim *et al.*, "Refining the eighth edition AJCC TNM classification and prognostic groups for papillary thyroid cancer with lateral nodal metastasis," *Oral Oncol,* vol. 78, pp. 80-86, Mar 2018.

[168] A. Fedorov *et al.*, "3D Slicer as an image computing platform for the Quantitative Imaging Network," *Magn Reson Imaging,* vol. 30, no. 9, pp. 1323-41, Nov 2012.

[169] Z. Siavashpour, M. R. Aghamiri, R. Jaberi, H. R. Dehghan-Manshadi, M. Sedaghat, and C. Kirisits, "Evaluating the utility of "3D Slicer" as a fast and independent tool to assess intrafractional organ dose variations in gynecological brachytherapy," *Brachytherapy,* vol. 15, no. 4, pp. 514-523, Jul-Aug 2016.

[170] L. Antunovic *et al.*, "PET/CT radiomics in breast cancer: promising tool for prediction of pathological response to neoadjuvant chemotherapy," *Eur J Nucl Med Mol Imaging,* Mar 26 2019.

[171] K. Ogino *et al.*, "Utility of FDG-PET/CT in the evaluation of the response of locally advanced breast cancer to neoadjuvant chemotherapy," *Int Surg,* vol. 99, no. 4, pp. 309-18, Jul-Aug 2014.

[172] M. K. Farrugia, S. Wen, G. M. Jacobson, and M. A. Salkeni, "Prognostic factors in breast cancer patients evaluated by positron-emission tomography/computed tomography before neoadjuvant chemotherapy," *World J Nucl Med,* vol. 17, no. 4, pp. 275-280, Oct-Dec 2018.

[173] A. M. T. Schmitz *et al.*, "Monitoring tumor response to neoadjuvant chemotherapy using MRI and 18F-FDG PET/CT in breast cancer subtypes," *PLoS One,* vol. 12, no. 5, p. e0176782, 2017.

[174] L. Spring *et al.*, "Pathologic Complete Response After Neoadjuvant Chemotherapy and Long-Term Outcomes Among Young Women With Breast Cancer," *J Natl Compr Canc Netw,* vol. 15, no. 10, pp. 1216-1223, Oct 2017.

[175] L. C. Collins *et al.*, "Pathologic features and molecular phenotype by patient age in a large cohort of young women with breast cancer," *Breast Cancer Res Treat,* vol. 131, no. 3, pp. 1061-6, Feb 2012.

[176] F. De Lima Vazquez *et al.*, "Retrospective analysis of breast cancer prognosis among young and older women in a Brazilian cohort of 738 patients, 1985-2002," *Oncol Lett,* vol. 12, no. 6, pp. 4911-4924, Dec 2016.

[177] G. S. Ajmani, T. A. James, O. Kantor, C. H. Wang, and K. A. Yao, "The Impact of Facility Volume on Rates of Pathologic Complete Response to Neoadjuvant Chemotherapy Used in Breast Cancer," *Ann Surg Oncol,* vol. 24, no. 11, pp. 3157-3166, Oct 2017.

[178] W. Haque, V. Verma, S. Hatch, V. Suzanne Klimberg, E. Brian Butler, and B. S. Teh, "Response rates and pathologic complete response by breast cancer molecular subtype following neoadjuvant chemotherapy," *Breast Cancer Res Treat,* vol. 170, no. 3, pp. 559-567, Aug 2018.

[179] S. Y. Huang *et al.*, "Exploration of PET and MRI radiomic features for decoding breast cancer phenotypes and prognosis," *NPJ Breast Cancer,* vol. 4, p. 24, 2018.

[180] S. Henderson *et al.*, "Interim heterogeneity changes measured using entropy texture features on T2-weighted MRI at 3.0 T are associated with pathological response to neoadjuvant chemotherapy in primary breast cancer," *Eur Radiol,* vol. 27, no. 11, pp. 4602-4611, Nov 2017.

[181] R. Chen *et al.*, "Assessment of the predictive role of pretreatment Ki-67 and Ki-67 changes in breast cancer patients receiving neoadjuvant chemotherapy according to the molecular classification: a retrospective study of 1010 patients," *Breast Cancer Res Treat,* vol. 170, no. 1, pp. 35-43, Jul 2018.

[182] C. M. Schlotter, L. Tietze, U. Vogt, C. V. Heinsen, and A. Hahn, "Ki67 and lymphocytes in the pretherapeutic core biopsy of primary invasive breast cancer: positive markers of therapy response prediction and superior survival," *Horm Mol Biol Clin Investig,* vol. 32, no. 2, Sep 22 2017.

[183] R. E. Ellsworth, H. L. Blackburn, C. D. Shriver, P. Soon-Shiong, and D. L. Ellsworth, "Molecular heterogeneity in breast cancer: State of the science and implications for patient care," *Semin Cell Dev Biol,* vol. 64, pp. 65-72, Apr 2017.

[184] N. Eiro, L. O. Gonzalez, M. Fraile, S. Cid, J. Schneider, and F. J. Vizoso, "Breast Cancer Tumor Stroma: Cellular Components, Phenotypic Heterogeneity, Intercellular Communication, Prognostic Implications and Therapeutic Opportunities," *Cancers (Basel),* vol. 11, no. 5, May 13 2019.

[185] Y. Yang *et al.*, "Prognostic impact of AJCC response criteria for neoadjuvant chemotherapy in stage II/III breast cancer patients: breast cancer subtype analyses," *BMC cancer,* vol. 16, no. 1, p. 515, 2016.

[186] K. Kuroi *et al.*, "Comparison of different definitions of pathologic complete response in operable breast cancer: a pooled analysis of three prospective neoadjuvant studies of JBCRG," *Breast Cancer,* vol. 22, no. 6, pp. 586-595, 2015.

[187] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature,* vol. 521, no. 7553, p. 436, 2015.

[188] Z. Lei and Z. Yi, "Big data analysis by infinite deep neural networks," *Journal of computer research and development,* vol. 53, no. 1, pp. 68-79, 2016.

[189] P. M. Thompson, N. G. Martin, and M. J. Wright, "Imaging genomics," *Current opinion in neurology,* vol. 23, no. 4, p. 368, 2010.

[190] A. Nagelkerke, J. Bussink, A. E. Rowan, and P. N. Span, "The mechanical microenvironment in cancer: how physics affects tumours," in *Seminars in cancer biology*, 2015, vol. 35, pp. 62-70: Elsevier.

Appendix

18F-FDG	18F-fluorodesoxyglucose	LoG	Laplacian of Gaussian
2D	Two-Dimensional	MI	Mutual Information
3D	Three-Dimensional	MRI	Magnetic Resonance Imaging
ADC	Apparent Diffusion Coefficient	MTV	Metabolically active Tumor Volume
AL	Acute Leukemia	NAC	Neoadjuvant Chemotherapy
AUC	Area Under the Curve	NCCR	National Central Cancer Registry
BMB	Bone Marrow Biopsy	NGTDM	Neighbouring Gray Tone Difference Matrix
BMI	Bone Marrow Involvement	NPV	Negative Predictive Value
BY	Bayesian	NSCLC	Non-Small Cell Lung Cancer
C-index	Concordance index	OS	Overall Survival
CPH	Cox	PAC	Proportion of Ambiguous Clustering
CT	Computed Tomography	PCM	Pearson Correlation Matrix
DFS	Disease-Free Survival	pCR	pathologic Complete Response
DSC	Dice’s Similarity Coefficient	PET	Positron-Emission Tomography
DWI	Diffusion-Weighted Image	PPV	Positive Predictive Value
EGFR	Epidermal Growth Factor Receptor	PR	Progesterone Receptor
ER	Estrogen Receptor	RELF	Relief
FTSJD1	FtsJ Methyltransferase Domain Containing 1	RF	Random Forest
GLCM	Gray Level Co-occurrence Matrix	RFE	Recursive Feature Elimination
GLDM	Gray Level Dependence Matrix	ROC	Operating Characteristic Curve
GLDZM	Grey Level Distance Zone Matrix	ROI	Region of Interest
GLRLM	Gray Level Run Length Matrix	RSF	Random Survival Forest
GLSZM	Gray Level Size Zone Matrix	RT	Radiation Therapy
GT	Ground Truth	SBS	Sequential Backward Selection
GTV	Gross Tumor Volume	SCT	Stem Cell Transplant
HD	Hausdorff Distance	SFS	Sequential Feature Selection
HER2	Human Epidermal Growth Factor Receptor	SUV	Standard Uptake Value
HPV	Human Papillomavirus	SVM	Support Vector Machines
HT	Hormone Therapy	TFS	Traversal Feature Selection
ICC	Intra-Class Correlation	TN	Triple-Negative
IHC	Immunohistochemical	TT	Targeted Therapy
LASSO	Least Absolute Shrinkage and Selection	VOE	Volume Overlap Error
LIME	Local Interpretable Model-agnostic Explanations	VOI	Volume of Interest