Medical Image Analytics (Radiomics) with Machine/Deep Learning for Outcome Modeling in Radiation Oncology

by

Lise Wei

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Applied Physics)
in the University of Michigan
2020

Doctoral Committee:

Professor Issam El Naqa, Chair
Professor James M. Balter
Professor Jeffrey A. Fessler
Clinical Assistant Professor Benjamin Rosen
Professor Clayton Scott
Professor Ji Zhu

Lise Wei

liswei@umich.edu

ORCID iD: 0000-0001-7226-2091

ACKNOWLEDGEMENT

First, I would like to express my sincere appreciation to my advisor, Professor Issam El Naqa. I am very grateful that Issam could have me in his group. In the beginning, I am totally new to the field and lacked the required knowledge for the research. I not only learned tremendously from his expertise in machine learning in radiotherapy, but also influenced deeply by his great passion in doing research. Without his help and guide, there is no way that I can accomplish all these work, including the manuscripts, book chapters, presentations, etc. Besides academic advice, he is always supportive and willing to help with other difficulties the students might have, which I really could not appreciate more.

I would also like to thank my committee members. Doing research in an interdisciplinary field is challenging, it is great to have a committee with expertise from different fields to help. I benefited a lot from their insightful suggestions and knowledge.

I would also like to thank everyone from my research group for the support and suggestions. It is a great team that everyone is friendly and very approachable. I would like to thank my Argus colleagues for any kind help they provided as well. I would like to thank Professor Cagliyan Kurdak and Ms Cynthia McNabb from Applied Physics Program for their support.

Finally, I would like to thank my mother, Yujie Bai, and my father, Hu Tian for their unconditional love and support.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ABSLYMPH	absolute lymphocytes
ABSNEUT	absolute neutrophils
ADASYN	adaptive synthetic sampling methods
AFP	alpha-fetoprotein

ALBI	Albumin-Bilirubin Grade
ALT	alanine transaminase
AST	aspartate aminotransferase
AUC	area under the ROC Curve
CAD	computer-aided diagnosis
CBCT	cone beam computed tomography
CCC	Concordance correlation coefficient
CECT	contrast-enhanced computed tomography
CHUM	Centre hospitalier de l'Université de Montréal
CHUS	Centre hospitalier universitaire de Sherbrooke
CI	confidence interval
CNN	convolutional neural network
CPH	Cox proportional hazard model
CT	computed tomography
CTCAE	common terminology criteria for adverse events
CV	cross validation

DCE	Dynamic contrast enhancement
DFS	disease-free survival
ECOG	Eastern Cooperative Oncology Group
EQD	equivalent dose
FCN	fully CNN
FDG	fluorodeoxyglucose
FFDM	full-field digital mammography
GDD	gradient direction distribution
GLCM	gray-level co-occurrence matrix
GLN	grey level nonuniformity
GLRLM	grey-level run length matrix
GLSZM	grey-level size zone matrix
GLV	grey level variance
GPU	graphics processing unit
GRU	gated recurrent unit
GSHT	grey-scale Hough transform

GTV	gross tumor volume
HCC	hepatocellular carcinoma
HGJ	Hôpital général juif
HGRE	high grey level run emphasis
HGZE	high grey level zone emphasis
HIPAA	health insurance portability and accountability act
HMR	Hôpital Maisonneuve-Rosemont
HNC	head and neck cancer
HNN	head and neck
HNSCC	head and neck squamous cell carcinoma
HPV	human papilloma virus
IBSI	image biomarker standardization initiative
ICGR	indocyanine green retention rate
IDM	inverse difference moment
IMRT	intensity-modulated radiation therapy
INR	international normalized ratio

IRB	institutional review board
IRS	intrahepatic recurrence survival
KL	Kullback–Leibler
KM	Kaplan-Meier
LASSO	least absolute shrinkage and selection operator
LBP	local binary pattern
LC	local control
LGRE	large grey level run emphasis
LGZE	large grey level zone emphasis
LRE	long run emphasis
LRHGE	long run high grey level emphasis
LRLGE	long run low grey level emphasis
LSTM	long short term memory
LZE	large zone emphasis
LZHGE	large zone high grey level emphasis
LZLGE	large zone low grey level emphasis

MELD	model for end-stage liver disease
MLP	multilayer perceptron
MPS	myocardial-perfusion SPECT
MRI	magnetic resonance imaging
MSE	mean squared error
NGTDM	neighborhood grey tone difference matrix
NSCLC	non-small cell lung cancer
OPSCC	oropharyngeal squamous cell carcinoma
OS	overall survival
PCA	principal component analysis
PET	positron emission tomography
PFS	progression-free survival
PML	patch-/pixel-based machine learning
PTV	planning target volume
RBM	restricted Boltzmann machine
RECIST	response evaluation criteria in solid tumors

RF	random forests
RFA	radiofrequency ablation
RFE	recursive feature elimination
RLN	run length nonuniformity
RLV	run length variance
mRMR	minimum redundancy maximum relevance
RNN	recurrent neural network
ROC	receiver operating characteristic curve
RT	radiotherapy
SBRT	stereotactic body radiotherapy
SNR	Signal-to-noise ratio
SPECT	single-photon emission computerized tomography
SRE	short run emphasis
SRHGE	short run high grey level emphasis
SRLGE	short run low grey level emphasis
SUV	standardized uptake value

SVM	support vector machine
SZE	small zone emphasis
SZHGE	small zone high grey level emphasis
SZLGE	small zone low grey level emphasis
TACE	trans-arterial chemoembolization
TCGA	the cancer genome atlas
TCIA	the cancer imaging archive
TRIPOD	transparent reporting of a multivariable prediction model for individual prognosis or diagnosis
TSVM	transductive SVM
VAE	variational autoencoder
VMAT	volumetric modulated arc therapy
ZP	zone percentage
ZSN	zone size nonuniformity
ZSV	zone size variance

ABSTRACT

Image-based quantitative analysis (radiomics) has gained great attention recently. Radiomics possesses promising potentials to be applied in the clinical practice of radiotherapy and to provide personalized healthcare for cancer patients. However, there are several challenges along the way that this thesis will attempt to address. Specifically, this thesis focuses on the investigation of repeatability and reproducibility of radiomics features, the development of new machine/deep learning models, and combining these for robust outcomes modeling and their applications in radiotherapy.

Radiomics features suffer from robustness issues when applied to outcome modeling problems, especially in head and neck computed tomography (CT) images. These images tend to contain streak artifacts due to patients' dental implants. To investigate the influence of artifacts for radiomics modeling performance, we firstly developed an automatic artifact detection algorithm using gradient-based hand-crafted features. Then, comparing the radiomics models trained on 'clean' and 'contaminated' datasets.

The second project focused on using hand-crafted radiomics features and conventional machine learning methods for the prediction of overall response and progression-free survival for Y90 treated liver cancer patients. By identifying robust features and embedding prior knowledge in the engineered radiomics features and using bootstrapped LASSO to select robust features, we trained imaging and dose based models for the desired clinical endpoints, highlighting the complementary nature of this information in Y90 outcomes prediction.

Combining hand-crafted and machine learnt features can take advantage of both expert domain knowledge and advanced data-driven approaches (e.g., deep learning). Thus, we proposed a new variational autoencoder network framework that modeled radiomics features, clinical factors, and raw CT images for the prediction of intrahepatic recurrence-free and overall survival for hepatocellular carcinoma (HCC) patients in this third project. The proposed approach was compared with widely used Cox proportional hazard model for survival analysis. Our proposed methods achieved significant improvement in terms of the prediction using the c-index metric

highlighting the value of advanced modeling techniques in learning from limited and heterogeneous information in actuarial prediction of outcomes.

Advances in stereotactic radiation therapy (SBRT) has led to excellent local tumor control with limited toxicities for HCC patients, but intrahepatic recurrence still remains prevalent. As an extension of the third project, we not only hope to predict the time to intrahepatic recurrence, but also the location where the tumor might recur. This will be clinically beneficial for better intervention and optimizing decision making during the process of radiotherapy treatment planning. To address this challenging task, firstly, we proposed an unsupervised registration neural network to register atlas CT to patient simulation CT and obtain the liver's Couinaud segments for the entire patient cohort. Secondly, a new attention convolutional neural network has been applied to utilize multimodality images (CT, MR and 3D dose distribution) for the prediction of high-risk segments. The results showed much improved efficiency for obtaining segments compared with conventional registration methods and the prediction performance showed promising accuracy for anticipating the recurrence location as well.

Overall, this thesis contributed new methods and techniques to improve the utilization of radiomics for personalized radiotherapy. These contributions included new algorithm for detecting artifacts, a joint model of dose with image heterogeneity, combining hand-crafted features with machine learnt features for actuarial radiomics modeling, and a novel approach for predicting location of treatment failure.

CHAPTER 1

Introduction

1.1 Imaging-based Modeling in Radiotherapy

The exponential growth in the use of various imaging modalities for diagnostic, therapeutic and prognostic purposes in noninvasive and quantitative cancer studies has the potential to provide individualized treatments for these patients. Radiomics is an emerging area that enables quantitative image analysis that aims to relate large-scale extracted imaging information to clinical and biological endpoints. The development of quantitative imaging methods along with machine learning techniques has enabled the opportunity to advance data science research towards clinical translation and provide more personalized cancer treatments. Accumulating evidence has indeed demonstrated that non-invasive advanced imaging analytics, i.e., radiomics, can reveal key components of tumor phenotype for multiple three-dimensional lesions at multiple time points over and beyond the course of treatment. These developments in the use of CT, PET, US and MR imaging could augment patient stratification and prognostication buttressing emerging targeted therapeutic approaches. In recent years, deep learning architectures have demonstrated their tremendous potential for image segmentation, reconstruction, recognition, and classification. Many powerful open-source and commercial platforms are currently available to embark in new research areas of radiomics.

In this work, we focused on the therapeutic utilization of imaging information. Radiotherapy relies heavily and often exclusively on medical imaging to determine the extent of disease and spatial location of cancer target and the surrounding healthy tissues [1]. There are two main types of radiotherapy, external beam radiation and internal radiation therapy. External beam radiation contains 3D conformal radiation therapy, intensity modulated radiation therapy (IMRT), stereotactic body radiation therapy (SBRT). SBRT uses focused, high-energy beams to treat small tumors with well-defined edges outside the brain and spinal cord, often in the liver or lung. Internal radiotherapy involves placing radiation sources as close as possible to the tumor site, including interstitial brachytherapy, intracavitary brachytherapy, which use solid sources, or inject the sources into blood, such as radioembolization with yttrium 90 (Y90). Radioembolization combines embolization and radiation therapy to treat liver cancers. Tiny glass or resin beads filled with the radioactive isotope Y90 are placed inside the blood vessels that feed the tumor. This can not only block the supply of blood to the tumor cells but also treat the tumor with a high dose of radiation. Both these radiotherapy methods involve tremendous use of digital images, before, during and post treatment. However, most imaging routines assess relative image signals merely to determine the location and size of tumors. Radiomic features can describe histology [2] and genetic footprint of the tumor [3-5], which are correlated with tumor aggressiveness. Thus, outcomes model building is vital to take full advantage of these potentials of multimodality imaging for cancer care.

Currently, most of the radiomics modeling in radiotherapy tries to answer the question of *what* will happen to the tumor, in terms of local/regional control (classification), or *when* an event (local/regional recurrence of tumors or overall survival) might occur (survival). In our work, we

comprehensively present a new framework to answer not only the questions of *what* and *when*, but also the challenging question *where* the regional recurrent tumor(s) might occur. Regional recurrence is a severe issue for hepatocellular carcinoma (HCC) patients, which is the most common type of primary liver cancer. Being able to answer these questions whether the tumor will recur, when it might occur and even where within the liver is the riskiest region are of great value to design a comprehensive personalized treatment framework and improve the prognosis for these patients.

Quantitative imaging research, however, is complex and key statistical principles should be followed to achieve reproducibility and subsequently realize its full potential. The field of radiomics, in particular, requires a renewed focus on optimal study design/reporting practices and standardization of imaging acquisition, feature selection and rigorous statistical analyses for the field to move forward. In this work, machine and deep learning as major computational vehicles for advanced model building of radiomics-based signatures were investigated with several applications primarily for liver cancer. We also address issues related to common practice challenges in medical physics, such as feature robustness and artifacts in head and neck CT images.

1.2 Conventional and Modern Radiomics

Conventional radiomics-based approach involves extracting a large amount of handcrafted features (e.g., intensity, shape and texture parameters), capturing different characteristics of the regions of interest, some of which may be difficult or even impossible to discern by human vision,

even by an expert trained one [6]. With the objective of outcome modeling (e.g., classification of toxicity, response to treatment, overall or disease-free survival), radiomic features have been related to clinical and biological endpoints by feature selection and subsequent machine learning models construction. Recently, the rise of computational power due to advanced hardware (e.g., GPU, cluster or cloud computing) combined with better algorithms for training neural networks, have facilitated the breakthrough of successful application of deep learning algorithms in many computer vision and analysis tasks, for both natural and medical images [7-10]. However, machine learning, either radiomics or deep learning based methods in medical imaging data, are still in their developmental stage in terms of investigation from a systematic perspective; about how we can tackle some challenges such as scarcity of available data, noisiness in this data, lack or noisy labeling, generalizability and interpretability of the generated models before any patient can benefit directly from this research. This thesis will touch most of these challenges and present new methodology and experiments that hopefully would provide some ideas for relevant studies.

To learn the important features from images, two approaches will be introduced: (1) based on handcrafted features (a.k.a., conventional radiomics) and automatically or (2) machine learnt features from data (a.k.a., modern radiomics with deep learning). The details will be introduced in Chapter 2. In addition, other sources of information were investigated and combined with imaging features as well, including clinical factors and genomics data.

1.3 Outline and Contributions

Medical image quantitative analyses possess the potential to reveal underlying biological mechanisms and to enable a more precise and personalized patient's healthcare. However, the prerequisite is to not only to obtain these images, but also to apply proper algorithms to extract the clinically-relevant information. Thus, in this thesis, we give a brief introduction to radiomics and deep learning based methods by first explaining the radiomic feature extraction, conventional machine learning algorithms (i.e., feature selection and model construction), followed by an introduction to deep learning algorithms. Basic theories of survival analysis, segmentation and registration, multitask learning and evaluation frameworks typically used for outcome modeling (ROC, AUC, c-index, nested cross-validation) will be introduced in Chapter 2 as well.

Subsequently, this thesis explores conventional machine learning methods using handcrafted features (Chapters 3, 4), deep learning methods (Chapter 5) for survival analysis combining imaging features (handcrafted and deep learning based), clinical variables and genetic features, and deep learning methods (Chapter 6) for deformable registration, tumor location prediction and survival net joint training.

In the first project (Chapter 3), the task is to detect streak artifacts in CT head and neck cancer (HNC) images. Streak artifacts are very common in HNC CT images due to dental implants. Though various radiomics signatures have been built using these CT images, the contamination of these images was seldom considered when these models were developed. Thus, we conducted experiments to check how much the model performance will be affected using contaminated

dataset and clean dataset. We then developed a method using handcrafted features to automatically detect the slices that contains artifacts to prepare a clean dataset for further outcome modeling. The second project (Chapter 4) is about Y90, a special type of nuclear medicine treatment, that was applied to liver cancer patients. Due to the limited sample size, we extracted handcrafted features (prior knowledge) from the PET images and used modified LASSO to build robust models. The goal of this work is to make use of the Y90 PET images that patients routinely take during treatment and to predict the tumor response to treatment so that we can closely monitor patients at higher risk. The third project (Chapter 5) is deep learning based actuarial analysis combing machine learnt features, handcrafted features and clinical variables. The tumor is HCC and patients were treated with SBRT. We are interested in the intrahepatic recurrence and overall survival risk for individual patients based on their liver phenotypes and other characteristics. The fourth project (Chapter 6) is an expansion of the third project. Since intrahepatic recurrence is common for post SBRT patients, then the question whether it would be possible to predict the location of the recurred tumor within liver, it will be beneficial to modify the dose delivered to the target and boost the dose to high-risk regions while avoiding functional liver to obtain better prognosis. The uniqueness of this study is three folds: multimodality images were available for this study, with MRI, CT and dose distribution for each patient; acquisition of the Couinaud segment using neural networks; joint training of survival, and location prediction tasks (i.e., multi-task learning).

In the medical field, data scarcity and impurity are serious challenges. Especially in a radiation oncology department, patients that received certain type of treatment, having certain type of

disease are only a subset of the whole patient group, thus, the outcome modeling process would suffer immensely due to inadequate training, noise signals, etc. However, the advancement of machine learning/deep learning methods show great potential for individualized treatment planning, even under these difficulties we affronted. This thesis presents new methodologies and experiments that I have tried to tackle these challenges, and they will hopefully help with the clinical practice someday. Chapters 2-6 are based on four radiomics projects and Chapter 7 discusses some current challenges and future directions.

1.4 Contributions

In this section, we will introduce each projects' brief background, contribution, limitations, and potential applications.

It is a known issue that CT head and neck cancer images usually contain streak artifacts due to dental implant. There are existing algorithms that aim at correcting these artifacts. However, this might introduce artificial noise to the original image and affect the radiomics models built upon. Thus, in this work, we aim at detecting the slices containing artifacts and remove these images before the radiomics modeling, which has not been studied by others based on our knowledge. Hand-crafted features were designed and showed high detection accuracy, which are novel and could be used as a preprocessing step before radiomics modeling in CT head and neck cancer based analysis.

Y90 treatment is a new type of internal radiation therapy that is given to patients who are not suitable for surgery or external beam radiation. Since this is relatively new, there are few

publications in this area. As far as we know, our work represents the first study that combines radiomics features and dose metrics to predict the overall response and progression. This is clinically relevant since it provides the physicians with better information to select the lesions that are not going to respond and could be candidates for other types of treatments. Technically, we proposed a modified LASSO approach that copes with very small sample size and build a more robust model than directly implementing LASSO. This technique is novel and can be applied to other small sample size problems as well.

Though advance stereotactic body radiation therapy (SBRT) has been shown to improve the overall survival rate for hepatocellular carcinoma (HCC) patients, the death rate for HCC patients is still increasing since 2000. The prediction of risk for overall survival is of important clinical interest. Current existing research for overall survival in HCC mostly uses Cox regression or random survival forests models and hand-crafted radiomics features. Our method proposes a comprehensive variational autoencoder survival model that incorporates heterogeneous inputs (image, radiomics features and clinical variables), which outperformed the Cox regression model. This model has the potential application of assisting the physicians in selecting "good" responding patients and adapting treatment for risky patients to achieve better prognosis.

Recurrence of treated HCC tumor is prevalent (>50%), which contributes to patient death. There are studies that predict whether a patient will have recurrence using radiomics features. We propose a new question – where the recurred tumor might occur in the liver, which by itself is novel. This has not been studied but is very meaningful for HCC patient personalized healthcare.

The method we proposed is a two-step method that firstly obtains the Couinaud segments, and then uses an attention-based CNN to perform recurrence region prediction. This method is first proposed by us shows promising predictive results. The potential application of this includes assisting the physicians getting Couinaud segments for liver either for surgery or radiation therapy purposes, providing more information for physicians about the risky regions, and combining with other modalities to improve the treatment planning to achieve personalized dose coverage.

1.5 Summary of Accomplishments

1.5.1 Abstracts and Presentations

Oral Presentations

"A Multimodality Approach Using Deep Attention Convolutional Neural Networks for Localization of Intrahepatic Liver Cancer Recurrence Post-SBRT", American Association of Physicists in Medicine (AAPM), online, the U.S., July, 2020.

"A Multimodality Approach Using Deep Convolutional Neural Networks for Localization of Intrahepatic Liver Cancer Recurrence Post-SBRT", Great Lakes Chapter American Association of Physicists in Medicine (GLCAAPM), online, the U.S., May, 2020.

"Variational autoencoder and graph-based radiomics modeling of intrahepatic progression risk and overall survival for HCC post-SBRT patients," American Society for Radiation Oncology (ASTRO) Annual Meeting, Chicago, Illinois, the U.S., Sep. 2019.

"CT-based radiomic analysis for prediction of early intrahepatic progression risk in hepatocellular carcinoma patients treated with stereotactic body radiation therapy," American Association of Physicists in Medicine (AAPM), Nashville, TN, the U.S., Aug. 2018.

"Automatic recognition of streak artifacts in CT regions of interest using gradient direction distribution method for radiomics analysis," American Association of Physicists in Medicine (AAPM), Denver CO, the U.S., Aug. 2017.

E-Posters

"Multitask-based supervised deep learning using contrast-enhanced CT (CECT) images for hepatocellular carcinoma (HCC) intrahepatic progression risk analysis," American Association of Physicists in Medicine (AAPM), San Antonio TX, the U.S., Aug. 2019.

"Variational Autoencoder Graph-based Radiomics Modeling of Intrahepatic Progression Risk and Overall Survival for HCC Post-SBRT Patients," Rogel Cancer Center Spring Symposium, Ann Arbor, MI, the U.S., Jun. 2019.

"Y-90 PET radiomics modeling analysis for response prediction within hours of radioembolization in liver cancer patients," Annual Congress of the European Association of Nuclear Medicine, Oct. 2018.

1.5.2 Book chapters

Wei, L., and El Naqa, I. "Feature extraction and qualification." Li, Ruijiang, Lei Xing, Sandy Napel, and Daniel L. Rubin, eds. Radiomics and Radiogenomics: Technical Basis and Clinical Applications. CRC Press, (2019).

Wei, L., and El Naqa, I. "Fundamentals of radiomics in Nuclear Medicine and Hybrid Imaging," Basic Sciences of Nuclear Medicine. Springer Nature, (2019).

1.5.3 Peer-reviewed Journals

Wei, L., Owen, D., Mendiratta-Lala, M., Rosen, B., Cuneo, K., Lawrence, T. S., Ten Haken, R. K., El Naqa, I., "Variational Autoencoder SurvivalNet Radiomics Modeling of Overall Survival for Hepatocellular Carcinoma Patients." Physica Medica (2020), submitted.

Pfaehler, E., **Wei, L.,** Zhovannik, I., Boellaard, R., Dekker, A., Monshouwer, R., El Naqa, I., Gillies, R., Wee, L., Traverso, A., "Repeatability and Reproducibility of Radiomic Features: Review and quality of reporting score." International Journal of Radiation Oncology • Biology • Physics (2020), submitted.

Wei, L., Xu, J., Cui, C., El Naqa I., Dewaraja, Y. K., "Prediction of tumor control in 90Y radioembolization by bootstapped LASSO using PET radiomics features." Journal of Nuclear Medicine, (2019). Under review.

Wei, L. Rosen, B., Vallières, M., Chotchutipan, T., Mierzwa, M., Eisbruch, A., and El Naqa, I. "Automatic recognition and analysis of metal streak artifacts in head and neck computed tomography for radiomics modeling." Physics and Imaging in Radiation Oncology (2019).

Wei, L., Osman, S., Hatt, M., and El Naqa, I. "Machine learning for radiomics-based multi-modality and multi-parametric modeling." The Quarterly Journal of Nuclear Medicine and Molecular Imaging (2019).

Luo, Y., Tseng, H., Cui, S., **Wei, L.,** Ten Haken, R. K., and El Naqa, I. "Balancing accuracy and interpretability of machine learning approaches for radiation treatment outcomes modeling." The British Institute of Radiology, (2019).

Avanzo, M., **Wei, L**., Stancanello, J., Vallières, M., Rao, A., Morin, O., Mattonen, S., El Naqa, I. "Machine and deep learning methods for radiomics." Medical Physics (2019).

Tseng, H., **Wei, L.**, Cui, S., Luo, Y., Ten Haken, R. K., and El Naqa, I. "Machine learning and imaging informatics in oncology." Oncology (2018).

Constanzo, J., **Wei, L.**, Tseng, H., and El Naqa, I. "Radiomics in precision medicine for lung cancer." Translational Lung Cancer Research, (2017).

1.6 References

1. Van den Berge DL, De Ridder M, Storme GA. Imaging in radiotherapy. Eur J Radiol. 2000;36:41-8.
2. Wu W, Parmar C, Grossmann P, Quackenbush J, Lambin P, Bussink J, et al. Exploratory study to identify radiomics classifiers for lung cancer histology. Front Oncol. 2016;6:71.
3. Zhu Y, Li H, Guo W, Drukker K, Lan L, Giger ML, et al. Deciphering genomic underpinnings of quantitative MRI-based radiomic phenotypes of invasive breast carcinoma. Sci Rep. 2015;5:17787.
4. Liu Y, Kim J, Balagurunathan Y, Li Q, Garcia AL, Stringfield O, et al. Radiomic features are associated with EGFR mutation status in lung adenocarcinomas. Clin Lung Cancer. 2016;17:441-8. e6.
5. Yoon HJ, Sohn I, Cho JH, Lee HY, Kim J-H, Choi Y-L, et al. Decoding tumor phenotypes for ALK, ROS1, and RET fusions in lung adenocarcinoma using a radiomics approach. Medicine. 2015;94.
6. Hatt M, Tixier F, Visvikis D, Le Rest CC. Radiomics in PET/CT: more than meets the eye? J Nucl Med. 2017;58:365-6.
7. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition; 2016. p. 770-8.
8. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. Imagenet large scale visual recognition challenge. International journal of computer vision. 2015;115:211-52.
9. Cheng J-Z, Ni D, Chou Y-H, Qin J, Tiu C-M, Chang Y-C, et al. Computer-aided diagnosis with deep learning architecture: applications to breast lesions in US images and pulmonary nodules in CT scans. Sci Rep. 2016;6:24454.
10. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017;542:115.

CHAPTER 2

Background

This chapter is based on the following book chapters and review articles: **Wei, L.,** Osman, S., Hatt, M., and El Naqa, I. "Machine learning for radiomics-based multi-modality and multi-parametric modeling." The Quarterly Journal of Nuclear Medicine and Molecular Imaging (2019); Avanzo, M., **Wei, L**., Stancanello, J., Vallières, M., Rao, A., Morin, O., Mattonen, S., El Naqa, I. "Machine and deep learning methods for radiomics." Medical Physics (2019); and book chapters: "Feature extraction and qualification." Li, Ruijiang, Lei Xing, Sandy Napel, and Daniel L. Rubin, eds. Radiomics and Radiogenomics: Technical Basis and Clinical Applications. CRC Press, (2019); Wei, L., and El Naqa, I. "Fundamentals of radiomics in Nuclear Medicine and Hybrid Imaging," Basic Sciences of Nuclear Medicine. Springer Nature, (2019).

The fundamental idea of radiomics is that medical images are much richer in information than what the human eye can discern. Quantitative imaging features, also called “radiomic features” can provide richer information about intensity, shape, size or volume, and texture of tumor phenotypes using different imaging modalities (e.g., MRI, CT, PET, ultrasound, etc.) [1]. Tumor biopsy-based assays provide limited tumor characterization as the extracted sample may not

always represent the whole heterogeneity spectrum of the patient's tumor, while radiomics can comprehensively assess the three-dimensional tumor landscape by means of extracting relevant imaging information and combining these features into mathematical models for associating these features with clinical and biological endpoints [2]. It implies that applying well-known machine learning methods to radiomic features extracted from medical images, it will make it possible to macroscopically decode the phenotype of many physio-pathological structures and, in theory, solve the inverse problem of inferring the genotype from the phenotype, providing valuable diagnostic, prognostic or predictive information [3, 4].

2.1 Radiomics Features

2.1.1 Preprocessing

Prior to radiomics analysis, preprocessing steps need to be applied to the medical images, which aim at reducing image noise, enhancing image quality, enabling the reproducible and comparable radiomic analysis. For some imaging modalities, such as PET, the images should be converted into a more meaningful representation (standardized uptake value, SUV). Image smoothing can be achieved by averaging or Gaussian filters [5]. Voxel size resampling is important for datasets that have variable voxel sizes [6]. Specifically, an isotropic voxel size is required for some texture feature extraction. There are two main categories of interpolation algorithms: Polynomial and spline interpolation. Nearest neighbor is a zero-order polynomial method that assigns grey-level values of the nearest neighbor to the interpolated point. Bilinear or trilinear interpolation and

bicubic or tricubic interpolation are often used for 2D in-plane interpolation or 3D cases. Cubic spline and convolution interpolation are third order polynomial method that interpolates smoother surface than linear method, while being slower in implementation. Linear interpolation is a rather commonly used algorithm, since it neither leads to the rough blocking artifacts images that are generated by nearest neighbors, nor will it cause out-of-range grey levels that might be produced by higher order interpolation [7].

In the context of feature-based radiomics analysis, as discussed below, computing of textures requires discretization of the grey levels (intensity values). There are two ways to do the discretization: fixed bin number *N* and fixed bin width *B*. For fixed bin number, we first decide a fixed number of *N* bins, and the grey levels that will be discretized into these bins using the formula below:

$$X_{d,k} = \begin{cases} \left\lfloor N_g \frac{X_{gl,k} - X_{gl,min}}{X_{gl,max} - X_{gl,min}} \right\rfloor + 1 & X_{gl,k} < X_{gl,max} \\ N_g & X_{gl,k} = X_{gl,max} \end{cases}, \quad (2.1)$$

where $X_{gl,k}$ is the intensity of kth voxel.

For fixed bin width, starting at a minimum $X_{gl,min}$, a new bin will be assigned for every intensity interval of w_b. Discretized grey levels are calculated as follow:

$$X_{d,k} = \left\lfloor \frac{X_{gl,k} - X_{gl,min}}{w_b} \right\rfloor + 1. \quad (2.2)$$

The fixed bin number method is better when the modality used is not well calibrated. It maintains the contrast and makes the images of different patients comparable, but loses the relationship between image intensity, while fixed bin size method keeps the direct relationship with the original scale. Some investigations about the effect of both methods have shown that fixed bin size method gave better repeatability and thus may be suitable for intra- and inter- patient studies, however, this remains the subject of ongoing research [8, 9]. In CT-radiomics, the image pixel intensity is mapped into the HU values and thus is much more directly comparable and interpretable. MRI-related modalities are more challenging since the pixel intensities are not directly interpretable, rather need to be normalized relative to some standard reference (e.g., contralateral brain, or normal appearing white matter in neuroimaging, psoas muscle in abdominal imaging, etc.). The art of appropriate normalization to ensure high fidelity for the remainder of the analysis pipeline remains a subject of ongoing research.

2.1.2 Static Features

Static features are based on intensity, shape, size (volume), texture and wavelet, describing the geometric property and the distribution of intensities of the ROIs in relation to their spatial distribution.

2.1.2.1 Morphological Features

These are geometrical shape characteristics of ROIs, such as compactness (representing how compact the region is), eccentricity (a measure of non-circularity, describing tumor growth directionality); Euler number (the number of connected objects in a region minus the solidity (this is a measurement of convexity), which may be a characteristics of benign lesions [10, 11].

2.1.2.2 First Order Intensity Features

First-order features are based on first-order histograms that shows the distribution of the voxel intensities in the ROIs. These features summarize the large number of voxel values in ROIs into single values, such as mean, minimum, skewness, etc.

2.1.2.3 Texture features

Broadly speaking, there are three categories for texture analysis: statistical (e.g., second and higher order features), model-based (e.g., Gaussian Markov random fields, Gabor filter and wavelet) and structural methods (e.g., Topological texture descriptors, Invariant histogram). The performance of texture approaches is not affected by tumor position, orientation, size, and brightness. It takes into account the local intensity-spatial distribution, and is also invariant to translation, rotation, affine and perspective transform [12-15]. Among these features, statistical methods have been widely used in the field of radiomics for cancer outcome modeling. Second-order features provide statistical interrelationships between voxels and capture special patterns in the ROIs, which make up for the loss of information associated with the first-order features. Haralick *et al.* introduced the idea of using textural features for image classification [16]. Several texture matrices formed the basis of statistical approaches: the grey level co-occurrence matrix (GLCM) [16],

neighborhood gray tone difference matrix (NGTDM) [17], run-length matrix (RLM) [18], and grey level size-zone matrix (GLSZM) [19]. GLCM illustrates the distribution of the combinations of grey levels of neighboring voxels (pixels) along certain direction, entropy, angular second moment, correlation, contrast, inverse difference moment are commonly used [13]. GLRLM was first proposed by Galloway. It is defined as the frequency of occurrence of contiguous voxels with some run length along certain direction that have the same grey level. It characterizes the distribution of combination of grey levels in different directions. Example features includes short/long run emphasis, low/high grey level run emphasis, etc. GLSZM gives the statistics of groups of voxels that are connected and have a specific grey level, with features such as small/large size zone emphasis. The NGTDM is thought to provide more human-like perception of texture such as: coarseness, contrast, busyness, and complexity.

2.1.2.4 Dynamic features

For dynamic imaging protocols, such as 4D MRI, CT and PET, features based on kinetic analysis using tissue compartment models and parameters related to transport and binding rates can be extracted [20]. Compartmental modeling is used to describe systems varying in time but not in space. In the case of FDG, a 3-compartment model could be used to depict the trapping of FDG-6-Phosphate (FDG6P) in tumor [21-23]. Glucose metabolic uptake rate could be evaluated from compartmental modeling. Values of influx rate constant from compartment modeling was shown to be able to offer an assessment for inflammations at different locations of the body for non-small cell lung carcinoma [24]. For dynamic contrast enhanced MRI, the Toft and Kermode (TK) model is one of the most popular compartment models, providing information about the influx forward

volume transfer constant from plasma into the extravascular-extracellular space [25-28]. Lee *et al.* found that three dynamic parameters were correlated with the dose of radiation delivered to the parotid gland and the degree of radiation-induced parotid atrophy during the treatment of head and neck cancer [29].

The radiomic features are in general defined independently of the image modality. Yet, there are some variations in the nomenclature used depending on the different imaging techniques. For instance, the use of SUV in PET image quantitative analysis, which are used instead of the raw counts intensity values. Therefore, basic features such as maximum, minimum, mean, standard deviation (SD), and coefficient of variation (CV) are usually expressed as SUV_{max}, SUV_{mean}, etc. Total lesion glycolysis (TLG) is defined as the product of volume and mean SUV [30-32].

2.2 Machine and Deep Learning Algorithms for Radiomics

Machine and deep learning algorithms provide us with powerful modeling tools to mine the huge amount of image information available, reveal underlying complex biological mechanisms, and make personalized precision cancer diagnosis and treatment planning possible. Here, we will briefly introduce two main types - feature-engineered (conventional radiomics) and machine learnt (deep learning-based) radiomics modeling methods. Generally speaking, machine learning methods can also be divided into supervised, unsupervised and semi-supervised for both feature-based and featureless methods. We will discuss briefly each of these categories in the following

sections. A workflow diagram illustrating the radiomics analysis process after image acquisition is shown in Fig. 2.1.

2.2.1 Feature-Engineered Radiomics Methods

Traditionally, the radiomic features being extracted are hand-crafted features that capture characteristic patterns in the imaging data, including shape-based, first-, second-, and higher order statistical determinants, model-based (e.g., fractal) and dynamic features as briefly discussed above. Feature-based methods require a segmentation of the region of interest (ROI), either through a manual, semi-automated, or automatic methods.

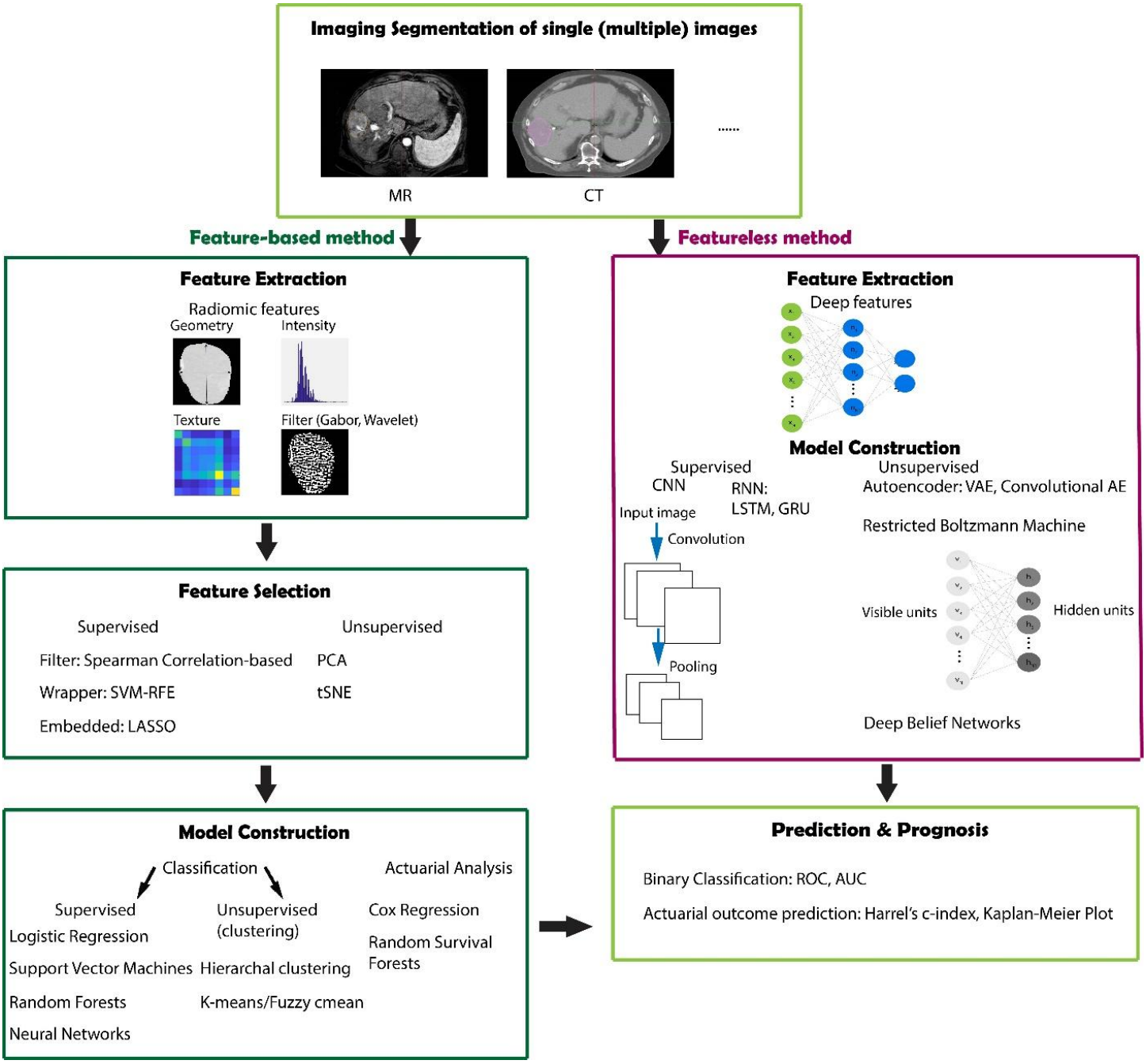


Fig. 2.1 Workflow for radiomics analysis with feature-based (conventional machine learning) and featureless (deep learning) approaches.

Hundreds or even thousands of radiomic features are not uncommon when we deal with outcome modeling. Feature selection and/or extraction thus is a crucial step that aims at obtaining the optimal feature subset or feature representation that correlates most with the endpoint and meanwhile correlates least between each other. After the feature subset is obtained, various machine learning algorithms can be applied based on them. Sometimes, the feature selection and model construction can be implemented together, called the *embedded method*, such as LASSO [33]. In contrast, *wrapper methods* select the features based on the models' performance for different subsets of features, for which we need to rebuild the model again after features are selected, for instance, recursive feature elimination SVM (SVM-RFE). *Filter method*s also separate the feature selection and the model construction processes, the uniqueness of it is their independence of the classifier being used for the subsequent model building, such as Pearson correlation-based feature ranking. In any feature selection method, it is essential to ensure that there is no "double dipping" into the training data for both feature selection, hyperparameter optimization and model selection. Rather the methods of "nested cross-validation" should be used in order to prevent overfitting or incorrect estimates of generalization.

According to whether or not the labels (ground truths) are used, feature selection and extraction can be divided into supervised, unsupervised and semi-supervised ways. The three feature selection methods discussed above are mostly supervised. Examples of unsupervised methods are principal component analysis (PCA) [34], clustering and t-Distributed Stochastic Neighbor Embedding (t-SNE) [35]. PCA uses an orthogonal linear transformation to convert the data into a new coordinate system so that large variances are projected to orthogonal coordinates. Clustering

is another feature extraction algorithm, which aims at finding relevant features and combining them by their cluster centroids based on some similarity measure, such as K-means and hierarchical clustering [36]. tSNE is a dimension reduction method that is capable of retaining the local structure (pairwise similarity) of data, while revealing some important global structure.

In the medical field, two types of questions are mainly investigated, binary problems (classification), such as whether or not a disease has recurred, the patient is alive beyond certain time threshold, etc.; and survival analysis, that is able to show if a risk factor or treatment affects time-to-event. For the classification problem logistic regression fits the coefficients of the variables to predict a logit transformation of the probability of the presence of the event. SVM, frequently used in CAD [37] and radiomics [36, 38-40], learns an optimal hyperplane that separates the classes as wide as possible, while trying to balance with misclassified cases. SVM can also perform nonlinear classification using the "kernel trick" -- different basis functions (e.g., radial basis function), mapping to higher dimensional feature space. The hyperplane maximizes the margin between the two classes in a nonlinear feature space. SVM also tolerates some points on the wrong side of the boundary, thus improving model robustness and generalization [41]. Random forests (RF) is based on decision trees, a popular concept in machine learning especially in the field of medicine, because their representation of hypotheses as sequential "if-then" resembles human reasoning [42]. RF applies bootstrap aggregating to decision trees and improve the performance by lowering the high variance of the trees [43].

Neural networks, though usually used in the one step context, can also be used in conventional feature selection and modeling [40, 44, 45]. These algorithms are mainly for supervised learning, while in particular in the medical field, there are a lot of data without labeling, in these cases, semi-supervised learning can be applied to make use of the unlabeled data combined with the small amount of labeled data. The self-training is bootstrapped with additional labelled data obtained from its predictions [46]. The transductive SVM (TSVM) tries to keep the unlabeled data as far away from the margin as possible [47]. Graph-based methods construct a graph connecting similar observations and enable the class information being transported through the graph [48].

For the survival analysis, Cox regression [49], random survival forests [50] and support vector survival [51] methods are also available to investigate the presence of a set of variables that may affect survival time.

2.2.2 Machine Learnt Radiomics Methods

Though hand-crafted features introduced above provide prior knowledge, they also suffer from the tedious designing process and may not faithfully capture the underlying imaging feature information. Alternatively, with the development of deep learning technologies based on multi-layer neural networks, especially convolutional neural networks (CNN), the extraction of machine learnt features is becoming widely applicable recently. In deep learning, the processes of data representation and prediction (e.g., classification or regression) are performed jointly [52]. In such a case, multi-stack neural layers of varying modules (e.g., convolution or pooling) with linear/non-linear activation functions perform the task of learning the representations of data with multiple

levels of abstraction and subsequent fully connected layers are tasked with classification, for instance. A typical scenario to get such features is to use the data representation CNN layers as feature extractor. Each hidden layer module within the network transforms the representation at one level. For example, the first level may represent edges in an image oriented in a particular direction, the second may detect motifs in the observed edges, the third could recognize objects from ensembles of motifs [52]. Patch-/pixel-based machine learning (PML) methods use pixel/voxel values in images directly instead of features calculated from segmented objects as in other approaches [52, 53]. Thus, PML removes the need for segmentation, one of the major sources of variability of radiomic features. Moreover, the data representation removes the feature selection portion eliminating associated statistical bias in the process. For the CNN network, either self-designed (from scratch) or existing structures, e.g., VGG [54], Resnet [55], can be used. Depending on the data size, we can choose to fix the parameters or fine tune the network using our data, also called *transfer learning*. Instead of using deep networks as feature extractors, we can use them directly for the whole modeling process. Similarly, to the conventional machine learning methods, there are also supervised, unsupervised and semi-supervised methods. CNN are similar to regular neural networks, but the architecture is modified to fit to the specific input of large-scale images. Inspired by the Hubel and Wiesel's work on the animal visual cortex [56], local filters are used to slide over the input space in CNNs, which not only exploit the strong local correlation in natural images, but also reduce the number of weights significantly by sharing weights for each filter. Recurrent neural networks (RNN) can use their internal memory to process sequence inputs and take the previous output as inputs. There are two popular types of RNNs – Long short-term

memory (LSTM) [57] and Gated recurrent units (GRUs) [58]. They were invented to solve the problem of vanishing gradient for long sequences by internal gates that are able to learn which data in the sequence is important to keep or discard. Deep autoencoders (AE), which are unsupervised learning algorithms, have been applied to medical imaging for latent representative feature extraction. There are variations to the AEs, such as variational autoencoders that resemble the original AE and variational Bayesian methods to learn a probability distribution that represents the data [58], convolutional autoencoders that preserve spatial locality [59], etc. Another unsupervised method is the restricted Boltzmann machine (RBM), which consists of visible and hidden layers [60]. The forward pass learns the probability of activations given the inputs, while the backward pass tries to estimate the probability of inputs given activations. Thus, the RBMs lead to the joint probability distribution of inputs and activations. Deep belief networks (DBNs) can be regarded as a stack of RBMs, where each RBM communicates with previous and subsequent layers. RBMs are quite similar with AEs, however, instead of using deterministic units, like RELU, RBMs use stochastic units with certain distribution. As mentioned above, labeled data is limited, especially in the medical field. Neural network based semi-supervised approaches combine unsupervised and supervised learning by training the supervised network with an additional loss component from the unsupervised generative models (e.g. AEs, RBMs) [61].

Machine learning methods are highly effective with large number of samples; however, they suffer from overfitting pitfalls with limited training samples. For deep learning, data augmentation (affine transformation of the images) during training is commonly implemented. Transfer learning is another way to reduce the difficulty in training. Using deep models trained on other dataset (natural

images) and then fine-tune on the target dataset. The structures of the networks can also be modified to reduce overfitting, such as, by adding dropout and batch normalization layers. Dropout randomly deactivates a fraction of the units during training and can be viewed as a regularization technique that adds noise to the hidden units [62]. Batch normalization reduces the internal covariate shift by normalizing for each training mini-batch [63].

Comparing with feature-based methods, deep learning methods are more flexible and can be used with some modifications in various tasks. In addition to classification, segmentation, registration, and lesion detection are widely explored by deep learning techniques. Fully CNN (FCN), trained end-to-end, merge features learnt from different stages in the encoders and then upsampling low resolution feature maps by deconvolutions [64]. Unet, built upon FCN, with the pooling layers being replaced by upsampling layers, resulted in a nearly symmetric U-shaped network [65]. Skipping structures combines the context information with the unsampled feature maps to achieve higher resolution.

Selecting the proper methods for each task is important. Basically, it is the art of balancing the information obtained from data and prior knowledge, and balancing the complexity of the models (capacity), which determines the bias and the variance. The more complex a model is, the less bias there will be, but the more variance as well. Thus, in practice, we need to deal with these trade-offs and there is no onetime answer. Conventional machine learning and deep learning methods both have their advantages and disadvantages. If the sample size is too small, using complex models will easily lead to overfitting and poor performance on new data. On the other hand, with

adequate data, complex models can represent the underlying relations of the data comprehensively and give better results. In terms of data-driven analysis and using prior knowledge, when the data size is small, it is usually beneficial to take advantage of the prior knowledge (radiomics features), if the prior knowledge is not misleading. With more data, we can rely more on the data and learn the information we need, which will avoid the risk that the prior knowledge is incorrect or irrelevant.

2.3 Software Tools for Radiomics

In most published research studies in radiomics, in-house developed methods are used. However, some research groups developed image analysis/radiomic software tools, both commercial or open source, available to the scientific community. The main goals of these tools are: 1) to speed up the development of competences based on more recent skills on radiomics; 2) to allow reproducibility and comparability of results from different research groups, and 3) to standardize both feature definitions and computation methods to guarantee the reliability of radiomic results [66, 67].

Table 2.1 shows a list of the software, web platforms, and toolkits available free of charge for the extraction of radiomics features, along with some of their main functionalities and relevant information. Given the high pace of radiomic developments, the list is not exhaustive and does not intend to cover all possible solutions. Furthermore, considering recent and increased interest in the radiomic field, many other dedicated tools are under development. All the open source solutions shown in this overview have been implemented by research teams and are capable of analyzing

CT, MRI, and PET, some of them can process also other medical images, such as mammography, radiography, or ultrasound.

Four software programs (MaZda [68], LifeX [69], ePAD [70], IBEX [71]) offer the possibility of manually or automatically segmenting medical images. Three toolkits (HeterogeneityCAD [2], PyRadiomics/Radiomics [72], QIFE [73]) are designed exclusively for the extraction of features. They can be embedded in more complete solutions (e.g. 3D Slicer [74]). Morphological, first, second and third order statistical features can be extracted by all software solutions, except for ePAD. Four of them (TexRAD, MaZda, PyRadiomics/Radiomics, IBEX) offer also the possibility of extracting features from filtered images. Of note, MEDomicsLab is an open-source software currently being developed by a consortium of research institutions, which will be available in the second half of 2019.

Table 2.1 Open access software programs for radiomics analysis.

Software/ Toolbox	MaZda	lifeX	ePAD	QIFE	HeterogeneityCAD	PyRadiomics / Radiomics	QuantImage	Texture Analysis Toolbox	IBEX	MEDomicsLab
Research group	Institute of Electronics, Technical University of Lodz, Poland	IMIV, CEA, Inserm, CNRS, Univ. Paris-Sud, Université	Rubin Lab, Stanford University	Sandy Napel, Stanford University	V.Narayan, J. Jagadeesan	Dana-Farber Cancer Institute, Brigham Women's Hospital Harvard Medical	University of Applied Science and Arts, Western Switzerland	M. Vallières	The University of Texas MD Anderson Cancer Center,	MEDomics consortium

		Paris Saclay				School, Boston			Houston, Texas	
Image modalities	CT, MRI, PET	CT, MRI, PET, ultrasound	CT, MRI, radiography	CT, MRI, PET	CT, MRI, PET	CT, MRI, PET		CT, MRI, PET	CT, MRI, PET	CT, MRI, PET
Segmentation	YES	YES	YES	NO	NO	NO	NO	NO	YES	NO
Segmentation methods	manual, automatic (threshold, flood-filling)	manual, automatic (threshold, snake)	Manual	/	/	/	/	/	manual, automatic (threshold)	/
Radiomic features: morphology	YES	YES	NO	YES	YES	YES	YES	YES	YES	YES
statistical 1° order	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
statistical 2° order	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
statistical 3° order	YES	YES	NO	YES	YES	YES	YES	YES	YES	YES
Filtering	YES	NO	NO	NO	NO	YES	NO	NO	YES	YES
Feature selection	YES	NO	NO	NO	NO	NO	NO	YES	NO	YES
Feature selection methods	Fisher score, classification error, corr.	/	/	/	/	/	/	Maximal information coefficient	/	False discovery avoidance, Elastic Net,

	coeff, **m**utual informat., minimal classification error									minimum Redundancy Maximum Relevance
Stratification	NO	NO	NO	NO	NO	NO	NO	NO	NO	YES

2.4 Deep Learning Based Survival Analysis

Marrying deep learning with survival analysis has become a trend in the actuarial analysis. There are mainly two ways of conducting deep learning based survival analysis, one is the proportional hazard based continuous method, the other is discrete time survival analysis.

Faraggi and Simon first extended Cox regression with neural networks in 1995 by replacing the linear predictor of the Cox regression model with formula by a one hidden layer multilayer perceptron (MLP), as shown in Eqn. (2.3) [75]. Katzman *et al.* in 2018 then modified previous work using novel neural network structure (DeepSurv), which outperformed the traditional Cox model:

$$h(t|x) = h_0(t)\exp[g(x)], \qquad g(x) = \beta^T x, \tag{2.3}$$

Zhu *et al.* extended the work to pathology images by replacing the MLP to convolutional layers [76]. This neural network-based Cox model still has the proportionality assumption. Kvamme *et al*. introduced non-proportional Cox model by making the relative risk function depend on time:

$$h(t|x) = h_0(t)\exp[g(t,x)], \tag{2.4}$$

The $g(t,x)$ handles time as a regular covariate, so that the loss function is still the same partial likelihood as classical Cox model [77].

An alternative approach is to discretize the time and compute the survival function on the time grid. In this way, no assumptions are made about the underlying stochastic processes and the distribution of survival times are directly learned. Lee *et al.* proposed a method called DeepHit that uses a deep neural network to learn the probability mass function by log-likelihood and a ranking loss [78]. Fotso *et al.* used a multi-task logistic regression with a neural network to calculate the survival probabilities [79].

2.5 Validation and Benchmarking of Radiomics Models

Once models are developed using the selected predictors, quantifying the predictive ability of the models (validation) is necessary. Based on the TRIPOD criteria, there are 4 types of validation: 1a. Developing and validating on the same data, which gives apparent performance. This evaluation is usually optimistic estimation of the true performance. 1b. Developing the models using all the data, then using resampling techniques to evaluate the performance. 2a. Randomly split the data into 2 groups for development and validation separately. 2b. Split the data non-

randomly (e.g., by location or time), which is stronger than 2a. 3 & 4. Develop the model using one data set and validate on separate data [80]. It is ideal if there is a separate dataset for external validation, however, in the frequent case that only a single data set is available, internal validation (1b) is required. Two popular resampling methods are bootstrapping and cross-validation. Feature selection, which is required before machine learning, should precede cross-validation, or it will lead to a selection bias due to the leak of information by the pre-filtering of the features [81]. In practice, nested cross-validation is also commonly used, which is a stronger version of TRIPOD type 1b.

Radiomic classifiers output a score that indicates the likelihood of one event to happen, and a threshold, to generate positive or negative predictions according to the task at hand. For example, fewer FPs would be required if we are implementing a conservative experiment, thus larger threshold will be preferred. Classifiers are evaluated using either a numeric metric (e.g., accuracy), or the so-called confusion matrix, or a graphical representation of performance, such as a receiver operating characteristic curve (ROC), a two-dimensional graph with TP rate being the Y axis, and FP rate the X axis. It has the advantage that they show classifier performance without regard to threshold and class distribution, thus widely used in model evaluation. The area under an ROC curve (AUC) is more convenient when comparing, and is equivalent to the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance [82]. For survival analysis, Harrell's c-index [83] is commonly used to measure discrimination ability of the model, which is motivated by Kendall's tau correlation. Harrell defines the overall c-index as the proportion of all usable pairs in which the predicted risk

probabilities and outcomes are concordant (Usable pairs are two cases that at least one of them is event) [84].

Kaplan-Meier (KM) curves is used to estimate the survival function from lifetime data, and also used to compare different risk groups. The risk groups can be patients that are treated with certain plan and the control group, or they can be the outputs from a survival model (e.g., Cox model) that divides the patients into high and low risk groups. It is highly recommended to visualize confidence intervals of the curves. The log rank test gives a quantitative evaluation of the statistical significance of the difference for different curves, which is also widely provided for KM curves [85].

Prediction accuracy is important, however, in the medical field, due to limited sample size, noisiness in the data, and intrinsic uncertainty rooted in the complex system, understanding the correlation and contribution of different components (imaging features, clinical factors, genomic data, etc.) might be more important clinically, which unfortunately, is not adequately addressed in current studies in our field. In this thesis, we will present some preliminary results in terms of the interpretation of the developed models in Chapters 5 & 6 in addition to the widely used evaluation metrics described previously.

2.6 Repeatability and Reproducibility of Radiomic Features

In radiomics, *repeatability* is measured by extraction of features from repeated acquisition of images under identical or near-identical conditions and acquisition parameters [86], whereas *reproducibility*, is assessed when measuring system or parameters differ. These can be assessed by use of digital or physical phantoms.

Radiomic features are more and more used in clinical research, but before a clinical implementation is possible, a standardization of image pre-processing, image discretization, and feature aggregation across centers is necessary. Moreover, a consensus about which features are repeatable and reproducible needs to be drawn as only these features should be used in the clinic. It is difficult to draw a general conclusion about which features are repeatable and reproducible. This is due to the large variability of settings and tumor types which were analyzed. Moreover, the variability in metrics used for assessing repeatability/reproducibility makes it almost impossible to compare studies between each other. However, most studies reported on the robustness of first-order and local textural features such as GLCM and GLRLM features, while global textural features (such as GLSZM features) were found to be less robust.

In general, reconstruction settings and image noise have a high impact on radiomic feature values for all imaging modalities. This implies that multi-center radiomic studies require harmonized images in terms of image reconstruction setting and signal-to-noise ratio. This harmonization can be achieved by, e.g., harmonizing image reconstruction methods as well as image post-processing across centers.

Moreover, there was a consensus that the segmentation method has an impact on radiomic feature values. Therefore, it is important to use an accurate as well as repeatable segmentation in the radiomics workflow. At the moment, most segmentations are performed manually what comes with a high inter-observer variability as well as with a low repeatability. The identification of an automatized segmentation algorithm most suitable for tumor segmentation is important.

Due to the sensitivity of radiomic features to various factors, it is crucial to report in detail each step performed during radiomic analysis. Only a detailed report makes a study reproducible itself and gives the opportunity to compare different studies.

2.7 References

1. Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, Van Stiphout RG, Granton P, et al. Radiomics: extracting more information from medical images using advanced feature analysis. Eur J Cancer. 2012;48:441-6.
2. Aerts HJ, Velazquez ER, Leijenaar RT, Parmar C, Grossmann P, Carvalho S, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nature communications. 2014;5:1-9.
3. Panth KM, Leijenaar RT, Carvalho S, Lieuwes NG, Yaromina A, Dubois L, et al. Is there a causal relationship between genetic changes and radiomics-based image features? An in vivo preclinical experiment with doxycycline inducible GADD34 tumor cells. Radiother Oncol. 2015;116:462-6.
4. Kumar V, Gu Y, Basu S, Berglund A, Eschrich SA, Schabath MB, et al. Radiomics: the process and the challenges. Magn Reson Imaging. 2012;30:1234-48.
5. Bagher‐Ebadian H, Siddiqui F, Liu C, Movsas B, Chetty IJ. On the impact of smoothing and noise on robustness of CT and CBCT radiomics features for patients with head and neck cancers. Med Phys. 2017;44:1755-70.
6. Shafiq‐ul‐Hassan M, Zhang GG, Latifi K, Ullah G, Hunt DC, Balagurunathan Y, et al. Intrinsic dependencies of CT radiomic features on voxel size and number of gray levels. Med Phys. 2017;44:1050-62.

7. Zwanenburg A, Leger S, Vallières M, Löck S. Image biomarker standardisation initiative. arXiv preprint arXiv:161207003. 2016.
8. van Velden FH, Kramer GM, Frings V, Nissen IA, Mulder ER, de Langen AJ, et al. Repeatability of radiomic features in non-small-cell lung cancer [18 F] FDG-PET/CT studies: impact of reconstruction and delineation. Mol Imaging Biol. 2016;18:788-95.
9. Leijenaar RT, Nalbantov G, Carvalho S, Van Elmpt WJ, Troost EG, Boellaard R, et al. The effect of SUV discretization in quantitative FDG-PET Radiomics: the need for standardized methodology in tumor texture analysis. Sci Rep. 2015;5:11075.
10. Jain AK. Fundamentals of digital image processing. Englewood Cliffs, NJ: Prentice Hall; 1989.
11. O'Sullivan F, Roy S, O'Sullivan J, Vernon C, Eary J. Incorporation of tumor shape into an assessment of spatial heterogeneity for human sarcomas imaged with FDG-PET. Biostat. 2005;6:293-301. doi:10.1093/biostatistics/kxi010.
12. Castleman KR. Digital image processing. Englewood Cliffs, N.J.: Prentice Hall; 1996.
13. Haralick R, Shanmugam K, Dinstein I. Texture Features for Image Classification. IEEE Trans on Sys Man and Cyb SMC. 1973;3:610-21.
14. Zhang J, Tan T. Brief review of invariant texture analysis methods. Pattern Recognition. 2002;35:735-47. doi:http://dx.doi.org/10.1016/S0031-3203(01)00074-7.
15. Castellano G, Bonilha L, Li LM, Cendes F. Texture analysis of medical images. Clinical Radiology. 2004;59:1061-9. doi:http://dx.doi.org/10.1016/j.crad.2004.07.008.
16. Haralick RM, Shanmugam K. Textural features for image classification. IEEE Transactions on systems, man, and cybernetics. 1973:610-21.
17. Amadasun M, King R. Textural features corresponding to textural properties. IEEE Transactions on systems, man, and Cybernetics. 1989;19:1264-74.
18. Galloway MM. Texture analysis using grey level run lengths. NASA STI/Recon Technical Report N. 1974;75.
19. Thibault G, Fertil B, Navarro C, Pereira S, Cau P, Levy N, et al. Shape and texture indexes application to cell nuclei classification. International Journal of Pattern Recognition and Artificial Intelligence. 2013;27:1357002.
20. Watabe H, Ikoma Y, Kimura Y, Naganawa M, Shidahara M. PET kinetic analysis--compartmental model. Ann Nucl Med. 2006;20:583-8.
21. Graham MM, Peterson LM, Hayward RM. Comparison of simplified quantitative analyses of FDG uptake. Nuclear Medicine and Biology. 2000;27:647-55.
22. Patlak CS, Blasberg RG. Graphical evaluation of blood-to-brain transfer constants from multiple-time uptake data. Generalizations. J Cereb Blood Flow Metab. 1985;5:584-90.
23. Morris ED, Endres CJ, Schmidt KC, Christian BT, Muzic RF, Fisher RE. Kinetic modeling in positron emission tomography. Emission Tomography: The Fundamentals of PET and SPECT Academic, San Diego. 2004.
24. Yang Z, Zan Y, Zheng X, Hai W, Chen K, Huang Q, et al. Dynamic FDG-PET imaging to differentiate malignancies from inflammation in subcutaneous and in situ mouse model for non-small cell lung carcinoma (NSCLC). PLoS One. 2015;10:e0139089.

25. Chikui T, Obara M, Simonetti AW, Ohga M, Koga S, Kawano S, et al. The principal of dynamic contrast enhanced MRI, the method of pharmacokinetic analysis, and its application in the head and neck region. International journal of dentistry. 2012;2012.
26. Tofts PS. Modeling tracer kinetics in dynamic Gd ‐ DTPA MR imaging. J Magn Reson Imaging. 1997;7:91-101.
27. Tofts PS, Brix G, Buckley DL, Evelhoch JL, Henderson E, Knopp MV, et al. Estimating kinetic parameters from dynamic contrast ‐ enhanced T1 ‐ weighted MRI of a diffusable tracer: standardized quantities and symbols. Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine. 1999;10:223-32.
28. Leach M, Morgan B, Tofts P, Buckley D, Huang W, Horsfield M, et al. Imaging vascular function for early stage clinical trials using dynamic contrast-enhanced magnetic resonance imaging. Eur Radiol. 2012;22:1451-64.
29. Lee FK-h, King AD, Ma BB-Y, Yeung DK-w. Dynamic contrast enhancement magnetic resonance imaging (DCE-MRI) for differential diagnosis in head and neck cancers. Eur J Radiol. 2012;81:784-8.
30. Benz MR, Allen-Auerbach MS, Eilber FC, Chen HJJ, Dry S, Phelps ME, et al. Combined Assessment of Metabolic and Volumetric Changes for Assessment of Tumor Response in Patients with Soft-Tissue Sarcomas. J Nucl Med. 2008;49:1579-84. doi:10.2967/jnumed.108.053694.
31. Erdi YE, Macapinlac H, Rosenzweig KE, Humm JL, Larson SM, Erdi AK, et al. Use of PET to monitor the response of lung cancer to radiation treatment. Eur J Nucl Med. 2000;27:861-6.
32. Larson SM, Erdi Y, Akhurst T, Mazumdar M, Macapinlac HA, Finn RD, et al. Tumor Treatment Response Based on Visual and Quantitative Changes in Global Tumor Glycolysis Using PET-FDG Imaging. The Visual Response Score and the Change in Total Lesion Glycolysis. Clin Positron Imaging. 1999;2:159-71. doi:S1095039799000163 [pii].
33. Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological). 1996;58:267-88.
34. Wold S, Esbensen K, Geladi P. Principal component analysis. Chemometrics and intelligent laboratory systems. 1987;2:37-52.
35. Maaten Lvd, Hinton G. Visualizing data using t-SNE. Journal of machine learning research. 2008;9:2579-605.
36. Parmar C, Leijenaar RT, Grossmann P, Velazquez ER, Bussink J, Rietveld D, et al. Radiomic feature clusters and prognostic signatures specific for lung and head & neck cancer. Sci Rep. 2015;5:11044.
37. Tang J, Rangayyan RM, Xu J, El Naqa I, Yang Y. Computer-aided detection and diagnosis of breast cancer with mammography: recent advances. IEEE Trans Inf Technol Biomed. 2009;13:236-51.
38. Wang J, Kato F, Oyama-Manabe N, Li R, Cui Y, Tha KK, et al. Identifying triple-negative breast cancer using background parenchymal enhancement heterogeneity on dynamic contrast-enhanced MRI: a pilot radiomics study. PLoS One. 2015;10.

39. Mattonen SA, Palma DA, Johnson C, Louie AV, Landis M, Rodrigues G, et al. Detection of local cancer recurrence after stereotactic ablative radiation therapy for lung cancer: physician performance versus radiomic assessment. International Journal of Radiation Oncology* Biology* Physics. 2016;94:1121-8.
40. Ypsilantis P-P, Siddique M, Sohn H-M, Davies A, Cook G, Goh V, et al. Predicting response to neoadjuvant chemotherapy with PET imaging using convolutional neural networks. PLoS One. 2015;10.
41. Chen S, Zhou S, Yin FF, Marks LB, Das SK. Investigation of the support vector machine algorithm to predict lung radiation‐induced pneumonitis. Med Phys. 2007;34:3808-14.
42. El Naqa I, Li R, Murphy MJ. Machine learning in radiation oncology. Theory Appl. 2015:57-70.
43. Yang D, Rao G, Martinez J, Veeraraghavan A, Rao A. Evaluation of tumor‐derived MRI‐texture features for discrimination of molecular subtypes and prediction of 12‐month survival status in glioblastoma. Med Phys. 2015;42:6725-35.
44. Wu W, Parmar C, Grossmann P, Quackenbush J, Lambin P, Bussink J, et al. Exploratory study to identify radiomics classifiers for lung cancer histology. Front Oncol. 2016;6:71.
45. Nie K, Shi L, Chen Q, Hu X, Jabbour SK, Yue N, et al. Rectal cancer: assessment of neoadjuvant chemoradiation outcome based on radiomics of multiparametric MRI. Clin Cancer Res. 2016;22:5256-64.
46. Rosenberg C, Hebert M, Schneiderman H. Semi-supervised self-training of object detection models. WACV/MOTION. 2005;2.
47. Joachims T. Transductive inference for text classification using support vector machines. Icml; 1999. p. 200-9.
48. Blum A, Lafferty J, Rwebangira MR, Reddy R. Semi-supervised learning using randomized mincuts. Proceedings of the twenty-first international conference on Machine learning; 2004. p. 13.
49. Cox DR. Regression models and life‐tables. Journal of the Royal Statistical Society: Series B (Methodological). 1972;34:187-202.
50. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. The annals of applied statistics. 2008;2:841-60.
51. Van Belle V, Pelckmans K, Van Huffel S, Suykens JA. Support vector methods for survival analysis: a comparison between ranking and regression approaches. Artif Intell Med. 2011;53:107-18.
52. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521:436-44.
53. Suzuki K. Pixel-based machine learning in medical imaging. International journal of biomedical imaging. 2012;2012.
54. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:14091556. 2014.
55. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition; 2016. p. 770-8.

56. Hubel DH, Wiesel TN. Receptive fields and functional architecture of monkey striate cortex. The Journal of physiology. 1968;195:215-43.
57. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput. 1997;9:1735-80.
58. Cho K, Van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:14061078. 2014.
59. Li F, Qiao H, Zhang B. Discriminatively boosted image clustering with fully convolutional auto-encoders. Pattern Recognition. 2018;83:161-73.
60. Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. Science. 2006;313:504-7.
61. Kingma DP, Mohamed S, Rezende DJ, Welling M. Semi-supervised learning with deep generative models. Adv Neural Inf Process Syst; 2014. p. 3581-9.
62. Hinton GE, Srivastava N, Krizhevsky A, Sutskever I, Salakhutdinov RR. Improving neural networks by preventing co-adaptation of feature detectors. arXiv preprint arXiv:12070580. 2012.
63. Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:150203167. 2015.
64. Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. Proceedings of the IEEE conference on computer vision and pattern recognition; 2015. p. 3431-40.
65. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. International Conference on Medical image computing and computer-assisted intervention: Springer; 2015. p. 234-41.
66. Boellaard R. Standards for PET image acquisition and quantitative data analysis. J Nucl Med. 2009.
67. Mackin D. Measuring Computed Tomography Scanner Variability of Radiomics Features. Investig. radiology 50, 757–65. 2015.
68. Szczypiński PM, Strzelecki M, Materka A, Klepaczko A. MaZda—a software package for image texture analysis. Comput Methods Programs Biomed. 2009;94:66-76.
69. Nioche C, Orlhac F, Boughdad S, Reuzé S, Goya-Outi J, Robert C, et al. LIFEx: a freeware for radiomic feature calculation in multimodality imaging to accelerate advances in the characterization of tumor heterogeneity. Cancer Res. 2018;78:4786-9.
70. Rubin DL, Willrett D, O'Connor MJ, Hage C, Kurtz C, Moreira DA. Automated tracking of quantitative assessments of tumor burden in clinical trials. Transl Oncol. 2014;7:23-35.
71. Zhang L, Fried DV, Fave XJ, Hunter LA, Yang J, Court LE. IBEX: an open infrastructure software platform to facilitate collaborative work in radiomics. Med Phys. 2015;42:1341-53.
72. Van Griethuysen JJ, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V, et al. Computational radiomics system to decode the radiographic phenotype. Cancer Res. 2017;77:e104-e7.
73. Echegaray S, Bakr S, Rubin DL, Napel S. Quantitative Image Feature Engine (QIFE): an open-source, modular engine for 3D quantitative feature extraction from volumetric medical images. J Digit Imaging. 2018;31:403-14.

74. Fedorov A, Beichel R, Kalpathy-Cramer J, Finet J, Fillion-Robin J-C, Pujol S, et al. 3D Slicer as an image computing platform for the Quantitative Imaging Network. Magn Reson Imaging. 2012;30:1323-41.
75. Faraggi D, Simon R. A neural network model for survival data. Stat Med. 1995;14:73-82.
76. Zhu X, Yao J, Huang J. Deep convolutional neural network for survival analysis with pathological images. 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM): IEEE; 2016. p. 544-7.
77. Kvamme H, Borgan Ø, Scheel I. Time-to-event prediction with neural networks and Cox regression. Journal of Machine Learning Research. 2019;20:1-30.
78. Lee C, Zame WR, Yoon J, van der Schaar M. Deephit: A deep learning approach to survival analysis with competing risks. Thirty-Second AAAI Conference on Artificial Intelligence; 2018.
79. Fotso S. Deep neural networks for survival analysis based on a multi-task framework. arXiv preprint arXiv:180105512. 2018.
80. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. Br J Surg. 2015;102:148-58.
81. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference, and prediction: Springer Science & Business Media; 2009.
82. Ren J, Zhou J, Ding W, Zhong B. Clinicopathological characteristics and imaging features of pulmonary adenocarcinoma with micropapillary pattern. Zhonghua zhong liu za zhi [Chinese journal of oncology]. 2014;36:282-6.
83. Harrell Jr FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med. 1996;15:361-87.
84. Pencina MJ, D'Agostino RB. Overall C as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. Stat Med. 2004;23:2109-23.
85. Royston P, Altman DG. External validation of a Cox prognostic model: principles and methods. BMC Med Res Methodol. 2013;13:33.
86. Hatt M, Tixier F, Pierce L, Kinahan PE, Le Rest CC, Visvikis D. Characterization of PET/CT images using texture analysis: the past, the present… any future? Eur J Nucl Med Mol Imaging. 2017;44:151-65.

CHAPTER 3

Automatic Recognition and Analysis of Metal Streak Artifacts in Head and Neck Computed Tomography for Radiomics Modeling

This chapter developed a new algorithm for automatic recognition and analysis of artifacts for radiomics modeling and is based on the paper: **Wei, L**. Rosen, B., Vallières, M., Chotchutipan, T., Mierzwa, M., Eisbruch, A., and El Naqa, I. "Automatic recognition and analysis of metal streak artifacts in head and neck computed tomography for radiomics modeling." Physics and Imaging in Radiation Oncology (2019).

3.1 Introduction

With recent advances in medical imaging technologies, contrast enhanced computed tomography (CT), magnetic resonance (MR), and positron emission tomography (PET) imaging are being routinely acquired during the diagnosis, staging and radiotherapy treatment planning of head and neck cancers (HNC). The most widely used imaging modality for diagnosis and therapy is CT, which can assess the tissue/ lesion density, shape and texture, and has been a good image-based data resource for patient's outcome modeling (e.g., radiomics) [1]. A large number of imaging features can be extracted from CT images for such radiomics analysis. These features are widely

explored in HNC CT image analysis (e.g., segmentation, predictive and prognostic biomarkers, etc.). Aerts *et al.* found that the CT radiomic signature constructed from non-small cell lung cancer (NSCLC) patients preserved significant prognostic performance for head and neck squamous cell carcinoma (HNSCC). Significant associations were discovered between the radiomics features and gene-expression patterns [2]. Zhang *et al.* found that CT texture features such as primary mass entropy and histogram skewness were independent predictors of overall survival in a dataset of 72 HNSCC patients [3]. However, it is sometimes overlooked that the existence of metal artifacts in CT HNC images, due to dental implants, may corrupt the reliability and the precision of such radiomics analysis and may cause misleading results. Metal artifacts in the original images can lead to changes in underlying texture features, which form the basis of radiomics analysis [4, 5]. Large amounts of promising studies were carried out for CT-based HNC radiomic analysis. Artifacts influence was not taken into account or at least wasn't mentioned in these articles. Until recently, less attention was paid into this critical issue in CT HNC image analyses. Bogowicz *et al.* conducted studies aimed to predict tumor local control (LC) after radiochemotherapy of HNSCC and human papilloma virus (HPV) status using CT radiomics. In their study, contours were manually removed from artifact-affected slices. Scans with more than half of the contoured slices affected by metal artifacts were excluded in the analysis [6]. Elhalawani *et al.* also excluded slices with metal artifacts in their HPV prediction model [7]. For these studies that excluded the artifact-affected slices or patients, manual filtering was applied, which is a very time-consuming process. There are also some approaches proposed for the metal artifacts reduction (MAR) [8-12]. Yet, these methods are likely to introduce new artifacts to images, degrade their resolution, and

influence the statistical distribution of the original images, rendering them detrimental to any subsequent radiomic analysis [13, 14]. To overcome these challenges, we proposed a novel method that enabled the classification of artifact-affected slices/ROIs using extracted features automatically and efficiently, which has the potential to simplify the preprocessing and make the radiomic signatures more reliable. We have applied our algorithm on an external dataset to investigate the impact of artifacts on radiomics modeling as well. Our current approach aims to flag images with artifacts so as to build more robust radiomic models with artifact-free images.

3.2 Methods and Materials

A total of 131 oropharyngeal squamous cell carcinoma patients (3513 slices, among which 360 slices had visually identified metal artifacts in the regions of interest [ROIs]), treated at the University of Michigan Department of Radiation Oncology, and a set of 220 head and neck squamous cell carcinoma patients (17956 slices) from a previously published dataset, treated at four hospitals in Canada were included in this study [15]. The two datasets will be referred to as UM data and the Canadian data, respectively. We determined the ground truth non-artifact slices by visually inspecting all slices in the UM data set and only looking at the tumor ROI. This means that if the tumor ROI on a given slice did not contain metal artifacts, it would be considered as a negative sample even if other parts of that same ROI contained artifacts, which will help save valuable data. For the Canadian data, due to the very large amount of slices, we visually determined if any ROI contained artifacts, as opposed to slice-by-slice. For this case, after we obtained predicted slice label, if for a ROI, there was at least one slice that was labeled to contain artifact,

the whole ROI would be labeled as artifact-present. Fig. 3.1 shows an example slice of the artifact-affected ROI. The UM data was randomly and equally split into training and test sets. Training set was used to train a random forests artifacts detection model (all hyper-parameters and parameters), then applied to the holdout test set. The Canadian data was split by hospital: 148 patients from Hôpital général juif (HGJ) and Centre hospitalier universitaire de Sherbrooke (CHUS) were used as training set for the proposed radiomics model for distant metastases in Vallières *et al.* study

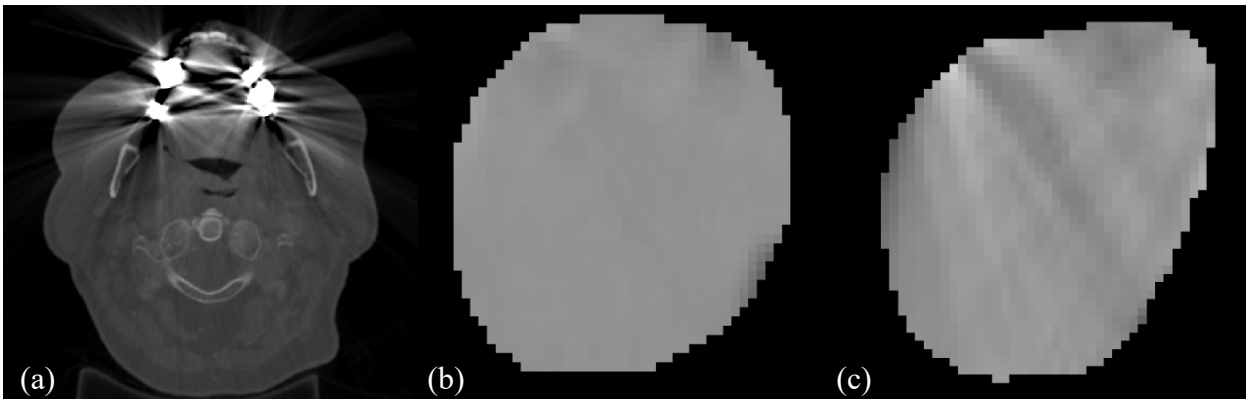


Fig. 3.1 (a) whole field CT image containing metal artifacts; (b) ROI (tumor) without artifacts; (c) ROI (tumor) with artifacts.

[15]. Seventy-two patients from Hôpital Maisonneuve-Rosemont (HMR) and Centre hospitalier de l'Université de Montréal (CHUM) combined were used as test set for evaluation. Fig. 3.2 shows a brief workflow. Table 3.1 gives more details about the datasets.

Table 3.1 Datasets information.

Data sets	Patients	ROIs	ROIs with artifacts	ROIs without artifacts	Slices	Slices with artifacts	Slices without artifacts
UM	131	131	63	68	3515	360	3155
Canada	220	344	105	239	17956	NA	NA

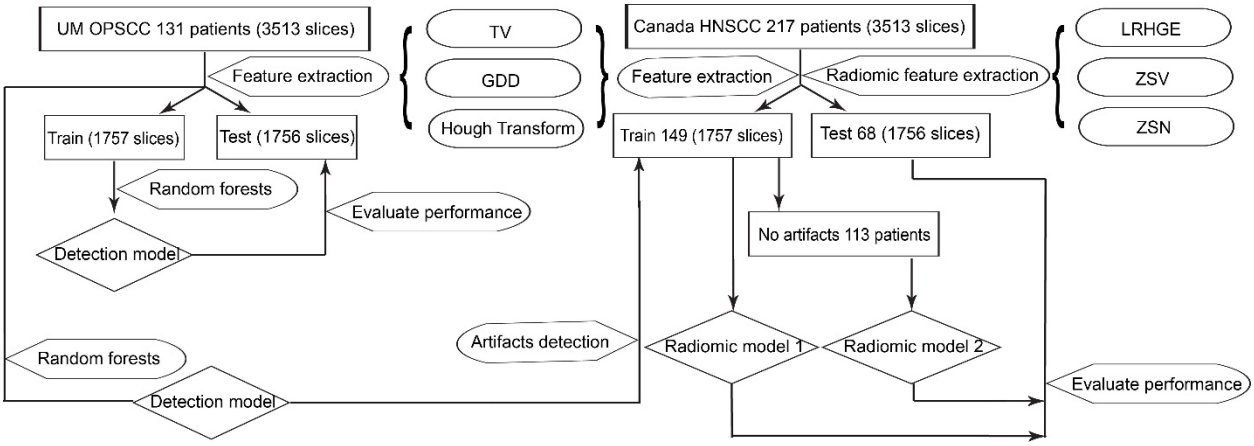


Fig. 3.2 Brief workflow for artifacts detection and impact on radiomic model performance.

3.2.1 Features design and extraction

3.2.1.1 Total variation based-feature

The concept of total variation (TV) was introduced first by Rudin *et al.* [16] for noise removal in image processing using first-order norms, since noisy images tended to have a high TV value compared with noise-free images. Similarly, metal artifacts led to an increased TV value relative to that of the regions of interest (ROIs) without artifacts, so the TV score could be taken as a measure of the artifacts. Below is the formula for calculating the TV (feature 1):

$$TV(I) = (\sum_{y=1}^{N_y}\sum_{x=1}^{N_x}|I(x,y) - I(x-1,y)| + |I(x,y) - I(x,y-1)|)/N_{\mathrm{ROI}} \quad (3.1)$$

where $I(x,y)$ is the intensity for pixel (x,y), N_x, N_y are number of pixels along the two directions, N_{ROI} is the number of pixels in the ROI. TV sums the absolute values of two-dimensional gradients for each pixel point of an image $I(x,y)$, here we are referring to in-plane directions. Additionally, TV values were normalized by dividing by the number of pixels in ROIs to exclude the influence of image size.

3.2.1.2 Gradient direction distribution (GDD) based features

Compared with TV gradient magnitude information, GDD features extracted gradient direction information. Some image preprocessing procedures were necessary prior to the extraction. The aim of pre-processing was to improve the image data quality by suppressing unwanted distortions and enhance the metal artifacts for preparation of the feature extraction. As shown in Fig. 3.3(a), the raw ROI images were noisy and the artifacts were hard to detect directly. We cropped and resized the original ROIs so that they had comparable size to provide a good estimate of the distribution. First, we resized ROIs to 25 × 25 pixels (median size of ROI slices in UM dataset), then cropped outer pixels to remove any edge effects, such that all images had the same size of 16 × 16. With size-modified ROIs, the gradient direction of each pixel point in the ROIs was approximated using the Sobel operator [17]. The direction ranged from −180° to 180° counterclockwise from the positive x-axis. In Fig. 3.3(b) and (e), the gradient direction map of the

ROIs was plotted. In these plots, the streak artifacts were more pronounced than in the original images. Histograms provided useful information about image statistics, from which we could extract discriminant features to help with artifacts detection. For ROIs with artifacts there should be a dominant direction of the gradient orientation distribution, while the ROIs without artifacts would tend to have more uniform distribution. We extracted the maximum gradient direction percentage (feature 2) from the histogram of angles for the gradient $H(\theta)$ with 36 bins (bin width of 10°):

$$\text{Max gradient direction} = \frac{\max\left(H(\theta_i)\right)}{\sum_i H(\theta_i)} \tag{3.2}$$

where i is the bin index. Due to varying tumor shapes, even though the bounding ROI box was modified, there were still some non-tumor parts of the original images included in the analyzed region. To remove shape effects, another complementary feature was calculated (feature 3):

$$\text{ratio of pixels in tumor ROI} = \frac{N_{\text{ROI}}}{16 \times 16} \tag{3.3}$$

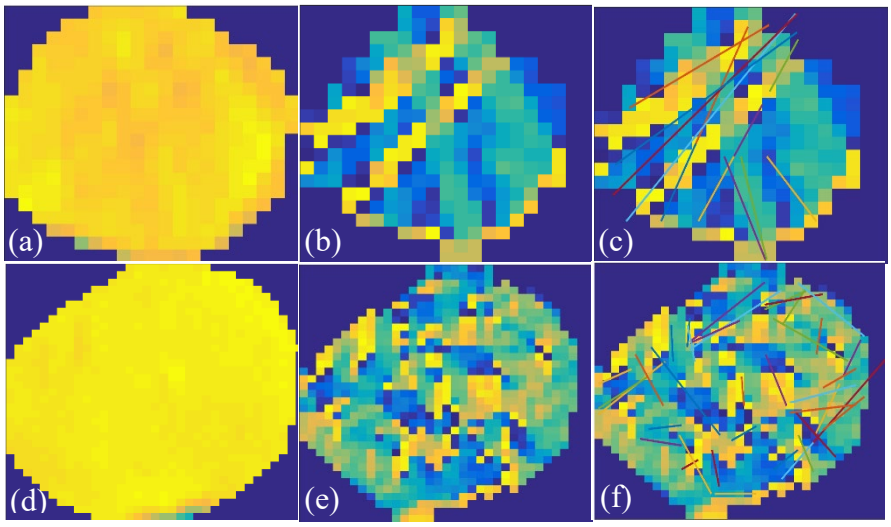


Fig. 3.3 (a) Original ROI with artifacts; (b) corresponding Gradient direction map of the ROI; (c) detected lines by modified Hough transform; (d)–(f) are similar with (a), (b), (c), while without artifacts.

3.2.1.3 Grey-scale Hough transform based features

Conventional Hough transform is a well-known method for line detection [18-20]. However, conventional Hough transform (CHT) requires input images to be edge-enhanced binary images, which are obtained by edge detection algorithms followed by thresholding or thinning. This will lead to loss of information and it requires selecting a threshold (harder for rich contrast, blurry, and band structure images). In our case, the artifacts were not obvious, and they were dispersed band structures. We used an extended Hough transform algorithm that dealt specifically with grey-scale images and avoided the thresholding process encountered in the conventional Hough transform [21]. Keck *et al.* proposed the use of direct output from the edge operator [22]. Thus, there was no threshold used to suppress the edges. Instead, the intensities in the edge image was considered to be the weighting coefficient for the Hough transform – grey-scale Hough transform

(GSHT). However, the traditional edge operators performed poorly in the ROIs in our application, since the lines in our case were dispersed with relatively gradually changing intensities. Instead of applying the edge operators, we input the gradient direction map for GSHT. Subsequently, we applied a local maxima filter to the obtained Hough map. The modified GSHT algorithm was summarized in Table 3.2. We used the GSHT as feature extractor instead of directly line detection because of dispersive characteristic of the artifacts. As shown in Fig. 3.3(f), if there were too many lines detected, it indicated high noisiness of the image and also absence of artifacts. Hence, the number of lines detected was a distinctive feature (feature 4). Since most of the artifact lines extended through the whole tumor, the ratio of the line length detected over the length of the tumor along that same direction should be close to one for artifact lines (feature 5).

$$\text{maximum ratio of detected lines} = \max\left(\frac{L_{\text{detected lines}}}{L_{\text{tumor along the same direction}}}\right) \tag{3.4}$$

For tumors with artifacts, the length of detected lines should be relatively large. Thus, the number of lines with ratios larger than a priori threshold was another distinct feature:

$$\text{feature 6} = \text{number of lines, if } \left(\frac{L_{\text{detected lines}}}{L_{\text{tumor along the same direction}}}\right) > 0.6 \tag{3.5}$$

The threshold for large lines we used here was empirically found to be 0.6 using our training data. For ROIs with artifacts, the lines detected were expected to have similar orientations, while those false lines had directions without any regular pattern. We counted the number of lines that had

similar orientations with the maximum ratio line in one ROI. Here, we defined similar orientation as angle difference smaller than δ (20°) using prior knowledge.

$$\text{feature } 7 = \text{number of lines, if } \left| D_{\text{line}} - D_{\text{argmax}\left(\frac{L_{\text{detected lines}}}{L_{\text{tumor along the same direction}}}\right)} \right| < \delta \qquad (3.5)$$

where D represents the direction of a line. After designed these seven features, we did principal component analysis (PCA) for these features to explain the variance in the data.

Table 3.2 Modified Hough transform for artifact detection.

(1) Obtain GDD map as input;

(2) Initiate the accumulators $H(\rho, \theta)$ to all zeros;

(3) For each point (x, y) in GDD map:

$$\theta = \text{gradient direction at } (x, y)$$

$$\rho = x\cos\theta + y\sin\theta$$

$$H(\rho, \theta) = H(\rho, \theta) + I(x, y)$$

(4) Non-maxima suppression to get the local maxima, then thresholding (m fraction of the maximum, *m*=0.1) to eliminate non-candidates, followed by selection of connected regions (pixels in the group larger than *n*, *n*=4 in our case).

(5) Each selected region corresponds to a line candidate. Several features can be extracted for these candidates: ρ the distance to the image origin; θ the orientation of the line; number of line candidates; the ratio of detected line length and the length of tumor at the same position (ρ, θ).

3.2.2 Random forests artifacts detection classifier construction

In summary, we have devised 7 features for the automatic detection of artifacts which were summarized in Table 3.3. Tree-based methods are commonly used in machine learning to build predictive models by partitioning the feature space into a set of rectangles. We randomly split UM data into training and testing sets (with equal samples). Random forests were implemented on the training set to construct the detection model. A Bayesian optimizer was used to optimize the 5-fold cross-validated loss objective function to tune the hyper-parameters (minimum leaf size and number of trees used) to control the tree depth. Then, we fixed the hyper-parameters to re-train the model. The trained model was applied to the testing set, with 10 times 5-fold cross-validation to provide the confidence interval for the training results. Feature importance was computed as well. Slice level artifacts detection model was trained and tested on the training and testing sets of UM data, while ROI level detection was trained on the UM data and tested both on the testing sets of UM data and externally on the Canadian data. Furthermore, based on the variance explained using principal component analysis (PCA), we tried different models using all the 7 features and less features to examine whether we could simplify the features we used and still obtain a generalizable model.

Table 3.3 Extracted features.

Extraction Method	Total variation	GDD		Modified grey-scale Hough transform			
Feature index	1	2	3	4	5	6	7
Name	Total variation	Maximum gradient direction	Ratio of pixels inside ROIs	Number of lines detected	Maximum ratio of detected lines	Number of lines larger than a certain threshold	Number of lines with similar orientation with the longest line detected

3.2.3 Evaluation of impact of artifacts in tumor ROIs on radiomic prediction performance

For all the 148 train and 72 test ROIs, we implemented the feature extraction method described above. A random forests classifier (that classifies the presence of metal artifacts in each slice) using all the samples of UM data (131 patients, 3513 slices) was constructed, and applied to the Canadian data (220 patients, 344 ROIs, 17,956 slices) to obtain the predicted labels for whether or not one slice has artifacts and then determine if the ROI contains artifacts. The ground truth for these data are visually determined. Radiomic models for distant metastases were built on three sets of data: (1). all 148 train samples; (2). samples without artifacts based on the algorithm; (3) samples without artifacts based on visual detection. The three models were further tested on test set (72 patients containing no metal artifacts). For the model construction details, please refer to the paper [15]. The prediction results were presented by plotting the receiver operating characteristic (ROC)

curve and calculating the corresponding area under the curve (AUC). The clinical patient characteristics were evaluated for patients without artifacts and all the patients in Canadian data to make sure the subgroup (without artifacts) clinical characteristics were not biased. For categorical variables, Pearson's Chi-squared test was carried out, and for continuous variables, pairwise t-test was used to check if there was significant bias or deviance for the subgroup compared to the whole set.

3.3 Results

Fig. 3.3(a) showed the training of objective function in terms of number of trees and minimum leaf size. The optimal values for these two parameters were 42 trees and 9 minimum number of leaf node observations. Fig. 3.3(b) showed that the training AUC achieved 0.91 (95% CI: 0.89–0.94), testing 0.89. The out-of-bag feature importance, measured by bootstrapping technique in random forests algorithm, was also calculated and presented in Fig. 3.3(c) [23]. The ranking showed that first four important features were total variation, max GDD, number of lines detected by Hough transform and ratio of valid pixels in the images and contributed to 99% of the unexplained variation using PCA. Since the first 4 features could explain most of the variance, we examined the models using less features. The slice level AUC on UM data saturated after 4 features (~0.90), details are summarized in Table 3.4. The confusion matrix for ROI level performance on UM and Canadian data using different features (4–7) was shown in Table 3.5. The results for 5–7 features were the same for UM (combined in the table), with an accuracy of 0.70/0.77, specificity of 0.66/0.83, sensitivity of 0.74/0.71, F1 score 0.73/0.77 for 4 feature and 5–7 feature models,

respectively, as shown in Table 3.6. Since a 4-feature model didn't perform well for ROI level classification, we tested only 5 and 7 feature models on Canadian data.

Table 3.4 AUC vs. feature number being used in UM data of slice level artifacts detection.

# of features	3	4	5	6	7
Train AUC (95% CI)	0.86 (0.83-0.89)	0.90 (0.88-0.93)	0.92 (0.88-0.93)	0.91 (0.88-0.93)	0.91 (0.90-0.93)
Test AUC	0.87	0.91	0.92	0.91	0.92

Table 3.5 Confusion matrices for UM and Canadian data of ROI artifacts.

UM (4 features/5~7 features)	Positive	Negative
Predicted positive	26/25	10/5
Predicted negative	9/10	19/24
Canada (5 features/7 features)	Positive	Negative
Predicted positive	81/73	47/29
Predicted negative	24/32	192/210

Table 3.6 Performance for UM and Canadian data of ROI artifacts.

	Feature #	Accuracy	Specificity	Sensitivity	F1 score
UM	4 features	0.70	0.66	0.74	0.73
	5-7 features	0.77	0.83	0.71	0.77
Canada	5 features	0.79	0.80	0.77	0.69
	7 features	0.82	0.88	0.70	0.71

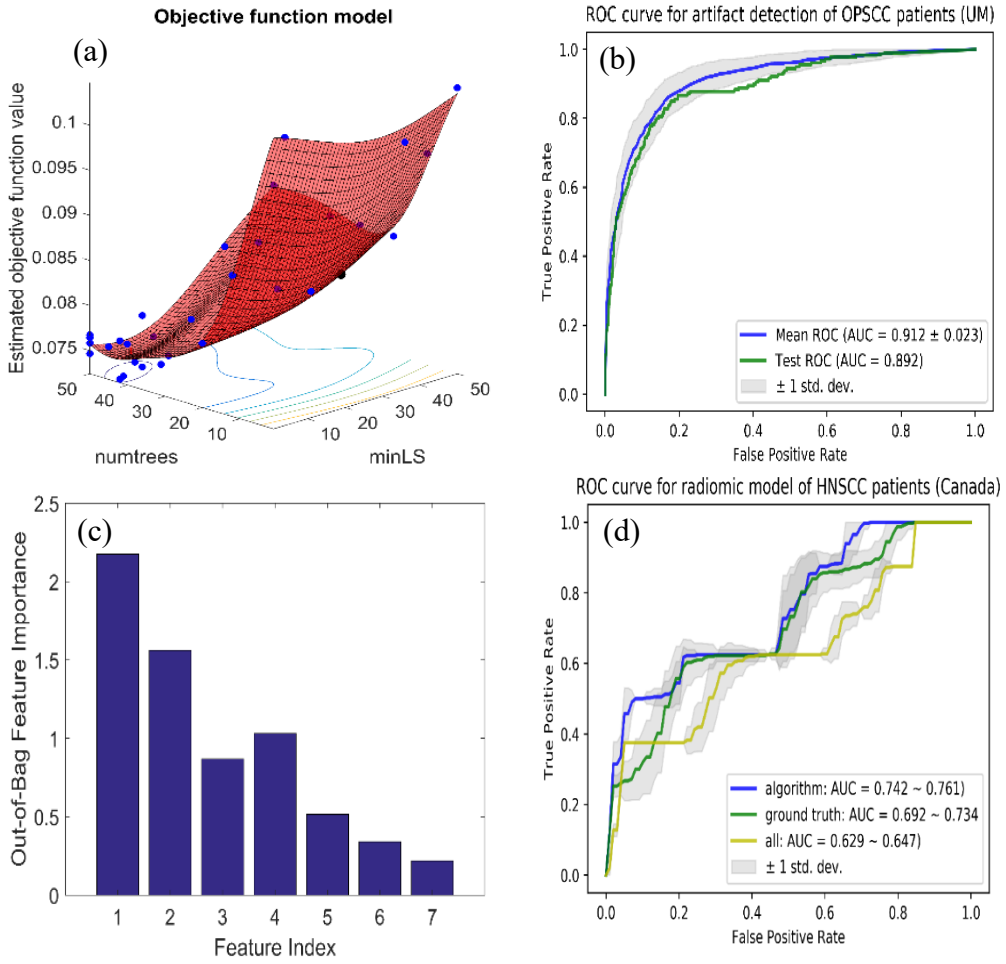


Fig. 3.4 (a) Optimization of hyper-parameters for random forests: number of trees (41) and minimum leaf size (17); (b) ROC curve for test data, with AUC of 0.89; (c) Out-of-bag feature importance; (c) Radiomic model test results for distant metastases using: all train samples (148 patients, yellow); samples filtered by our artifacts detection algorithm (107 patients, blue) and samples filtered visually (100 patients, green).

The confusion matrix and corresponding metric results were shown in Tables 3.5 and 3.6 as well, with an accuracy of 0.79/0.82, specificity of 0.80/0.88, sensitivity of 0.77/0.70, F1 score 0.69/0.71 for 5 feature and 7 feature models, respectively. After checking the clinical characteristics, we found that none of the characteristics showed significant deviance from the original dataset. Fig. 3.4(d) showed the results for the radiomic models. Radiomic model constructed using samples

without artifacts, either filtered by our algorithm or visually, yielded a substantially better performance than using the original training set, which included 32% (48/148) artifacts patients. The AUC were 0.64 (95% CI: 0.63–0.65), 0.71 (95% CI: 0.69–0.73), and 0.75 (95% CI: 0.74–0.76) for the radiomic models trained on all train samples, samples excluding artifacts affected ones by our algorithm and by visual detection, respectively.

3.4 Discussion

In this study, a set of features extracted from total variation, gradient direction distribution and grey-scale Hough transform algorithm was designed. UM testing AUC of 0.89 showed that the proposed approach was able to accurately classify slices with metal artifacts. The robustness of these features was further validated by relatively good performance on external Canadian data. Confusion matrix for external validation was used since the slice-by-slice labels in Canadian data were not obtained due to the large sample size (17,596 slices). A ROI would be labeled positive, if one or more slices contained artifacts. Though first 4 features explained most of the variance and slice level AUC saturated after 4 features, adding feature 5 increased the ROI level performance (accuracy, specificity, and F1 score) for UM data, due to less false positive, and similar true positive cases. This was reasonable since the last few features were mainly designed to regularize or reduce the false positive classification. Similar trend was captured in the Canadian data as well, the specificity increased from 0.80 to 0.88 with more features. While, the decrease of sensitivity (0.77–0.70) for 7 features-model was due to less true positives. In all, 5 features-model was comparable for slice-by-slice detection to the 7 features-model. For the ROI level detection,

5 features-model resulted in less artifacts dataset, while 7 features-model tended to reserve more samples but with more artifacts cases as well. The slice level model generalizability was not harmed by adding more features, probably because random forests algorithm is able to select the most robust features for the task. AUC around 0.90 suggested the probability of correctly ranking a positive – negative pair was 0.90. In general, it is pretty good performance for a classification task. To the best of our knowledge, we did not find literature implementing this kind of metal artifacts detection, thus it was hard to compare how good the accuracy of 0.82 was. However, the UM data ROI level accuracy was 0.77, with the slice level AUC 0.90. Thus, we could infer that the slice level AUC for external data was probably comparable. In addition, the artifact-free subset filtered by our algorithm showed improvement of performance, which also proved the goodness of this level of classification accuracy for radiomics modeling. Based on the context that researchers usually remove the slices affected when building models not the ROIs, our technique should be applicable and meaningful. Another thing to notice was the lower UM ROI accuracy, which could be due to the different artifacts proportion, with UM having 55% and Canada having 31% artifact-affected ROIs. Hence, it made sense that the specificity as well as the accuracy would be lower, with comparable sensitivity for UM data. Leijenaar *et al.* tested a radiomics signature derived from non-small cell lung cancer (NSCLC) patients on an external dataset of oropharyngeal squamous cell carcinoma (OPSCC) patients (n = 542) [24]. They visually identified ROIs with artifacts, and resulted in a subset of 275 patients with artifacts. Their radiomics signature was validated on all the data, subset of patients with and without artifacts within the delineated tumor regions. They found that the features preserved discriminative value on both with and without

artifacts subsets, however, they still suggested that there was an influence of CT artifacts on the model fit, which indicated a need for remodeling excluding samples with artifacts. This was consistent with our finding. Their research focused on validating the radiomics signature on head and neck tumor ROIs with and without artifacts to see the robustness of the features. We investigated the influence of presence of artifacts for the model construction and corresponding test performance on artifact-free data. Another study related to ours is the one by Ger *et al.* [5]. They investigated metal artifacts caused by dental fillings and beam-hardening artifacts caused by bone. They found at least 73% of feature values were affected by the streak artifacts. And almost all features were robust with removal of up to 50% of the original GTV. In summary, they showed that metal artifacts affect radiomic feature values, suggesting that regions containing such artifacts should not be included in radiomics data set. Their research provided further support for the necessity of removing artifact-affected images before radiomics modeling. We were also interested in understanding the nature of the misclassified cases. Some examples of both false negatives and false positives were shown in Fig. 3.5. The main challenge we met with in this detection task was the subtleness of the metal artifacts or small signal to-noise ratio (SNR) of the ROIs. A lot of the misses were the cases with artifacts that were subtle and hard to detect. The false positive cases were some slices with line-like structures inside while not being a true artifact. Finally, one thing to point out is that if the radiomic features are from 3D ROIs, we might have to remove the artifact-affected patients. Given the fact that around 30–50% of patients have metal artifacts, the radiomic models developed in this way might be suitable for not affected patients only. However, if we extract features from 2D slices, then we can remove the affected slices without excluding the

patient. While, we do acknowledge that the artifact classification can be more beneficial to develop radiomics models which are more robust against the streak artifacts, which is out of our scope for this study.

3.5 Conclusion

In conclusion, we have developed a new method for CT artifacts detection in tumor regions for head and neck patients; achieved UM test dataset prediction AUC of 0.89 using random forests algorithm and investigated the impact of presence of artifacts for head and neck CT images using internal and external datasets. We recommend using the proposed automatic algorithm to filter samples before CT head and neck radiomics analysis.

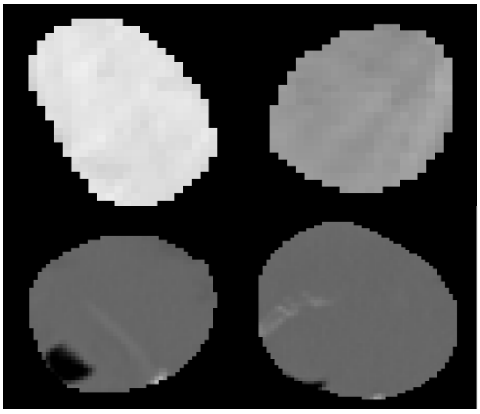


Fig. 3.5 Misclassified images: Top row shows misses and bottom row shows false positive cases.

3.6 References

1. Coroller TP, Grossmann P, Hou Y, Velazquez ER, Leijenaar RTH, Hermann G, et al. CT-based radiomic signature predicts distant metastasis in lung adenocarcinoma. Radiother Oncol. 2015;114:345-50. doi:10.1016/j.radonc.2015.02.015.
2. Aerts HJWL, Velazquez ER, Leijenaar RTH, Parmar C, Grossmann P, Cavalho S, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nature Communications. 2014;5:4006. doi:10.1038/ncomms5006.
3. Zhang H, Graham CM, Elci O, Griswold ME, Zhang X, Khan MA, et al. Locally advanced squamous cell carcinoma of the head and neck: CT texture and histogram analysis allow independent prediction of overall survival in patients treated with induction chemotherapy. Radiology. 2013;269:801-9.
4. Yu H, Scalera J, Khalid M, Touret A-S, Bloch N, Li B, et al. Texture analysis as a radiomic marker for differentiating renal tumors. Abdominal Radiology. 2017;42:2470-8.
5. Ger RB, Craft DF, Mackin DS, Zhou S, Layman RR, Jones AK, et al. Practical guidelines for handling head and neck computed tomography artifacts for quantitative image analysis. Comput Med Imaging Graph. 2018;69:134-9.
6. Bogowicz M, Riesterer O, Ikenberg K, Stieb S, Moch H, Studer G, et al. Computed tomography radiomics predicts HPV status and local tumor control after definitive radiochemotherapy in head and neck squamous cell carcinoma. International Journal of Radiation Oncology* Biology* Physics. 2017;99:921-8.
7. Head MACC, Group NQIW. Investigation of radiomic signatures for local recurrence using primary tumor texture analysis in oropharyngeal head and neck cancer patients. Sci Rep. 2018;8.
8. Abdoli M, Dierckx RA, Zaidi H. Metal artifact reduction strategies for improved attenuation correction in hybrid PET/CT imaging. Med Phys. 2012;39:3343-60. doi:10.1118/1.4709599.
9. Boas FE, Fleischmann D. Evaluation of two iterative techniques for reducing metal artifacts in computed tomography. Radiology. 2011;259:894-902. doi:10.1148/radiol.11101782.
10. Joemai RM, de Bruin PW, Veldkamp WJ, Geleijns J. Metal artifact reduction for CT: development, implementation, and clinical comparison of a generic and a scanner-specific technique. Med Phys. 2012;39:1125-32. doi:10.1118/1.3679863.
11. Xu C, Verhaegen F, Laurendeau D, Enger SA, Beaulieu L. An algorithm for efficient metal artifact reductions in permanent seed implants. Med Phys. 2011;38:47-56. doi:10.1118/1.3519988.
12. Yazdi M, Lari MA, Bernier G, Beaulieu L. An opposite view data replacement approach for reducing artifacts due to metallic dental objects. Med Phys. 2011;38:2275-81. doi:10.1118/1.3566016.
13. Katsura M, Sato J, Akahane M, Kunimatsu A, Abe O. Current and Novel Techniques for Metal Artifact Reduction at CT: Practical Guide for Radiologists. Radiographics. 2018;38:450-61.
14. Gjesteby L, De Man B, Jin Y, Paganetti H, Verburg J, Giantsoudi D, et al. Metal artifact reduction in CT: where are we after four decades? IEEE Access. 2016;4:5826-49.

15. Vallières M, Kay-Rivest E, Perrin LJ, Liem X, Furstoss C, Aerts HJ, et al. Radiomics strategies for risk assessment of tumour failure in head-and-neck cancer. Sci Rep. 2017;7:10117.
16. Rudin LI, Osher S, Fatemi E. Nonlinear total variation based noise removal algorithms. Phys D. 1992;60:259-68. doi:10.1016/0167-2789(92)90242-f.
17. Sobel I. An Isotropic 3 3 Image Gradient Operator; 2014.
18. VC HP. Method and means for recognizing complex patterns. Google Patents; 1962.
19. Duda R, Hart P. Use ofthe Hough transform to detectlinesand curvesin pictures. Commun ACM. 1972;15.
20. Ballard DH. Generalizing the Hough transform to detect arbitrary shapes. Pattern recognition. 1981;13:111-22.
21. Peng T. Detect lines in grayscale image using Hough Transform. version 1.0 ed: mathworks; 2005.
22. Ruwwe C, Zölzer U, Duprat O. Hough transform with weighting edge-maps. Visualization Imaging and Image Processing; 2005.
23. James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning: Springer; 2013.
24. Leijenaar RT, Carvalho S, Hoebers FJ, Aerts HJ, Van Elmpt WJ, Huang SH, et al. External validation of a prognostic CT-based radiomic signature in oropharyngeal squamous cell carcinoma. Acta Oncol. 2015;54:1423-9.

CHAPTER 4

Tumor Response Prediction in Y90 Radioembolization with PET-based Radiomics Features and Absorbed Dose Metrics

This chapter is developed a radiomics algorithm for tumor response prediction in ^{90}Y PET images and is based on the paper: **Wei, L**., Xu, J., Cui, C., El Naqa I., Dewaraja, Y. K., "Prediction of tumor control in ^{90}Y radioembolization by bootstrapped LASSO using PET radiomics features." European Journal of Nuclear Medicine and Molecular Imaging, Physics (2019). *Under review.*

4.1 Introduction

Delivering external radiation to multifocal/large liver tumors is a challenging task due to the damage of surrounding normal liver parenchyma. Hence, when disease burden is high, selective internal radiation delivery is preferred. Transarterial radioembolization (RE), with preferential delivery of glass or resin microspheres embedded with beta-emitting ^{90}Y to hepatic tumors is an established treatment for unresectable hepatocellular carcinoma (uHCC) and liver metastases [1, 2]. Ability to predict lesion-level response immediately after therapy can facilitate adaptive

therapies following RE by selecting lesion(s) predicted to be non-responding to the initial treatment for subsequent highly focal external stereotactic radiation.

Radiomics, a branch of quantitative image analysis, can capture heterogeneity characteristics of regions of interest (ROIs) by extracting relevant features from medical images (CT, MR, PET) has been widely explored in the literature and shown to provide predictive capability of treatment response in different cancers [3-11]. Specifically, in patients undergoing transarterial ^{90}Y radioembolization in uHCC, Blanc-Durand *et al.* showed that pre-treatment FDG-PET derived radiomics features (strength for PFS, variance, strength, low intensity run short emphasis and contrast for OS) for whole liver are independent negative predictors for progression-free survival (PFS) and overall survival (OS) [12]. Gensure *et al.* found tumor contrast-enhanced CT based texton and local binary pattern (LBP) features both achieve high accuracy in discriminating patient response to radioembolization (RE) with ^{90}Y resin microspheres in terms of serologic response and survival status [13]. Recent studies, by our group and others have reported on the association between post-therapy ^{90}Y imaging derived lesion absorbed dose and outcome (response, survival) in patients treated with ^{90}Y radioembolization for primary and metastatic liver cancer [14-18]. However, to our knowledge, our current study is the first investigation to combine lesion radiomics features with absorbed dose metrics to predict outcome. Furthermore, our study relies on radiomics features from post-treatment ^{90}Y PET imaging, unlike prior studies that used conventional FDG PET-derived features, which makes it unique in this respect. Compared with FDG-PET, ^{90}Y PET is considerably more noisy due to the low true coincidence rate associated with a low yield-positron in the presence of high random coincidence rates [19]. However, recent ^{90}Y PET/CT

studies have reported good quantitative accuracy and contrast-to-noise for dosimetry applications, using time-of-flight (TOF), longer acquisitions, optimized reconstruction parameters and partial volume correction [18, 20]. Although ^{90}Y can also be imaged by bremsstrahlung SPECT, the poor spatial resolution and challenges of correcting for bremsstrahlung scatter, makes ^{90}Y PET potentially better suited for radiomics analysis.

A major challenge of radiomics modeling especially with limited data is the robustness of the extracted features, as highlighted in recent review articles [21-23]. Variabilities can result from contouring, reconstruction algorithms, filtering, even different scans with the same setting. Another challenge is the risk of overfitting when dealing with relatively small datasets. Therefore, in this study both issues are addressed by: (1) conducting a phantom study to identify robust features, particularly to assess reconstruction and variability issues; and (2) applying a modified LASSO approach with bootstrap resampling for robust modeling. To mitigate analysis bias, nested cross-validation was used to train (feature selection, model construction) and test the outcome model (evaluation).

4.2 Methods and Materials

4.2.1 Patient cohort

The study included patients with primary and secondary intrahepatic malignancies who had ^{90}Y PET/CT imaging performed after ^{90}Y radioembolization with glass microspheres (Theraspheres) at University of Michigan (UM) Medical Center as part of an ongoing dosimetry research study.

Selection criteria for ^{90}Y PET/CT imaging were: well defined lesions >2 mL, ability to undergo imaging, follow-up at UM and informed consent. The patient and lesion characteristics for the 36 lobar treatments (30 patients, 105 lesions, 6 patients had treatment to right and left lobes at different time points.) are summarized in Table 4.1. The treating physician followed standard guidelines to deliver 80-150 Gy to the treated liver with empirical adjustments within this range based on clinical factors. The ^{90}Y PET/CT imaging was approved by the institutional review board, and all subjects signed an informed consent form.

4.2.2 ^{90}Y PET/CT Imaging and dosimetry

Images were acquired on a Siemens Biograph mCT PET/CT within a couple of hours of the RE procedure (prior to discharge) with an acquisition time of ~30 minutes to cover the entire liver and partial lung. PET reconstruction parameters were selected based on phantom studies considering both activity recovery and noise: 1 iteration, 21 subsets of 3D OS-EM with time-of-flight and resolution recovery and a 5 mm Gaussian post-filter [18]. The PET matrix size was 200×200 with a pixel size 4.07×4.07 mm and a slice thickness of 3 mm. The CT was performed in low dose mode (120 kVp; 80 mAs) during free-breathing. The CT matrix size was 512×512 with a pixel size of 0.97×0.97 mm and a slice thickness of 2 mm.

Table 4.1 Patient/lesion characteristics of the cohort and sub-cohort (HCC and metastasis).

	Cohort
Disease	
Primary HCC	13 (43%)
Liver Metastasis	17 (57%)
Total Patients	30
Total Therapies	36

Number of lesions		
	HCC	35 (33%)
	Liver Metastasis	70 (67%)
	Total Lesions	105
Cirrhotic livers		11 (37%)
Lesion volume (mL)		Median [range]
	Primary HCC	11.5 [2.1-204]
	Liver Metastasis	9.3 [2.2-833]
Number of lesions per patient		3 [1-5]

PET images were transformed to CT-space and the CT-derived density map were input to our DPM Monte Carlo code [18] to generate dose-rate maps that were converted into absorbed dose maps by accounting for ^{90}Y physical decay. Mean absorbed doses to segmented lesions were reported following partial volume correction based on volume-dependent recovery coefficients, determined from a phantom study [18].

4.2.3 Radiomics: lesion segmentation, PET data preprocessing and feature extraction

Lesion segmentation was performed on diagnostic quality contrast enhanced baseline CT or MRI by a radiologist specializing in hepatic malignancies (RK), which is considered a gold standard. Note that variability due to contouring can be a source of error, but has been addressed in several previous studies [6, 24, 25]. The diagnostic scan was then rigidly registered to the CT of the ^{90}Y PET/CT and the contours were transformed with fine tuning when mis-registration was evident on MIM (MIM Software Inc, Cleveland, OH). In some cases, where the lesions were well visualized

on the non-contrast low-dose CT of the PET/CT they were directly defined on this CT in order to minimize mis-registration effects. Up to 5 (largest) lesions > 2 mL were segmented per patient.

Lesion contours and ^{90}Y PET images were input to an in-house developed (Matlab, MathWorks Inc., Natick, MA) radiomics toolbox (benchmarked by image biomarker standardization - ISBI) that run as an extension on MIM. Our radiomics code is shared at https://github.com/mvallieres/radiomics. All subsequent analyses were performed in MATLAB. First, a root-squared transform was applied to the PET images to reduce quantum noise effects [26].

The full intensity range of the tumor region was quantized to a smaller number of gray levels (Ng) before computation of the features. The quantization algorithm used is Lloyd-Max algorithm, which attempts to minimize the mean-squared quantization error of the output. Ng was experimentally chosen as 32 [27]. The features were extracted from 3D ^{90}Y PET images, which were interpolated to isotropic voxel size (0.97 mm). 46 features, including volume, one shape feature (sphericity), 4 global features, and 40 texture features from gray-level co-occurrence matrix (GLCM), gray-level run length matrix (GLRLM), gray-level size zone matrix (GLSZM), and neighborhood gray-tone difference matrix (NGTDM), were extracted. All the feature extraction followed the Image biomarker standardization initiative (IBSI) guidance [28]. These features represent the spectrum of commonly used features, especially in PET imaging [6, 29, 30]. We further opted for extraction parameters following the ISBI guidelines due to the limited sample

size, we didn't explore further parameterization or less commonly used features. Table 4.2 presents the list of radiomics features used in this study.

4.2.4 Lesion-level Study Endpoints

Two endpoints applied at the lesion-level were considered: overall response (OR) classification at first follow-up and time to progression at lesion level. For OR assessment, diagnostic CT or MR at first follow up was used by a radiologist (RK) to measure percentage reduction in lesion diameter relative to baseline according to RECIST criteria [31]. Lesions that met the RECIST criteria of partial or complete response were classified as responding (OR = 1) and others (stable or progressive disease) were classified as non-responding (OR = 0). For time to progression, clinical follow-up images and records were assessed by a radiologist (RK) and was defined as the time in months from the date of ^{90}Y therapy to date of local progression. Lesions without evidence of progression were right censored at the last date of hepatic imaging. This included lesions that had not progressed at the time of death. Any lesion that had additional liver lesion directed therapy after ^{90}Y treatment were also right censored at the date of the additional treatment.

4.2.5 Phantom study to assess radiomics feature repeatability and reproducibility

A ^{90}Y PET/CT study with a liver/lung torso phantom consisting of a 'warm' liver compartment and three 'hot' lesion inserts (29 mL ellipsoid, 16 mL sphere, 8 mL sphere) with an insert-to-liver

activity concentration ratio of 5:1 was performed. The total activity in the phantom was 1.9 GBq and the activity concentrations in the inserts were 6.0-7.3 MBq/mL and liver minus inserts was 1.2 MBq/mL. To assess radiomics feature repeatability 5 consecutive 30 min acquisitions under identical conditions were performed on the same PET/CT system as in the patient studies. To assess sensitivity to reconstruction parameters and filtering, each of the 5 scans were reconstructed with 1 and 2 OS-EM iterations (21 subsets) and with and without Gaussian post-filter. The activity concentrations, acquisition time and parameters used in the phantom study were chosen to reflect conditions for imaging following ^{90}Y RE, hence, the noise-level was clinically relevant.

4.2.6 Statistical Analysis

4.2.6.1 Phantom feature robustness study

Concordance correlation coefficient (CCC) metric assumed each observation was independent as has been commonly reported in repeatability/reproducibility studies [22, 32]. Thus, in the robustness study of our extracted radiomics features, CCCs were computed for the different scans, different iterations, and with/without Gaussian filtering. For each of the 45 radiomics features (without volume), the resulting CCCs were averaged, and features with larger than 0.85 [22, 33-36] average CCC -robust radiomics feature set, were further investigated in the patient radiomics modeling.

4.2.6.2 Lesion overall response and progression modeling studies

Univariate analysis

Univariate association between the features (or absorbed dose) and OR classification was investigated using Spearman's rank correlation. Univariate analysis for the features (or absorbed dose) and progression was investigated by Cox regression.

Multivariate analysis -- modified Be-LASSO

In order to select robust features, build generalized models and evaluate unbiased model performance, a nested cross-validation (CV) framework has been employed (details are shown in Fig. 4.1). In the outer loop, 10 times 5-fold cross validation was used to estimate the model performance. On the training set of each inner loop, *N* times bootstrap was performed. For each resampling training set, optimal lambda λ hyper-parameter for LASSO was tuned by another cross-validation process. Subsequently, features with non-zero coefficients were recorded. With *N* resampled training sets, N sets of features were recorded. The frequency of a certain feature being selected by LASSO was calculated and thus a ranking list of the features was obtained. Then, *M* times bootstrap logistic regression modeling was used to estimate the model order. Specifically, models using top i $(i = 1, \dots, n = \text{number of features})$ ranked features were developed and mean AUC/c-index for each model order with confidence interval was obtained and the model order corresponding to highest AUC/c-index within one standard error was selected [37]. After we obtained the model order and top selected features, final model in each outer loop was obtained by retraining on the training set and applied on the outer test set (Here, *N* and *M* were both 100).

With the developed method, models were constructed using the 15 robust radiomics features set, lesion volume and mean absorbed dose (AD) (15+1+1=17). Since there are two subgroups in this patient cohort, the developed models were applied to both subgroups to assess if the tumor response correlated differently for primary HCC and metastatic lesions. The ROC curve (AUC) and c-index were used to evaluate the lesion OR and progression model performance, respectively. The confidence intervals were calculated by the bootstrap method [38]. The statistical analysis was performed using MATLAB R2019a and RStudio 1.1.463. The Bonferroni correction was applied to account for the family-wise error rate [39]. Overall, 17 features (dose + volume +15 radiomics features) were tested; therefore, p-values $< 0.05/17=0.003$ was considered significant. For the whole set of features (dose + volume + 45 radiomics features), p- values $< 0.05/47=0.001$ were considered significant. Meanwhile, due to the existence of unbalance in the dataset, especially for progression analysis (events 14/103), Adaptive Synthetic Sampling Approach (ADASYN) was applied for the multivariate analysis to see if it can improve the performance [40].

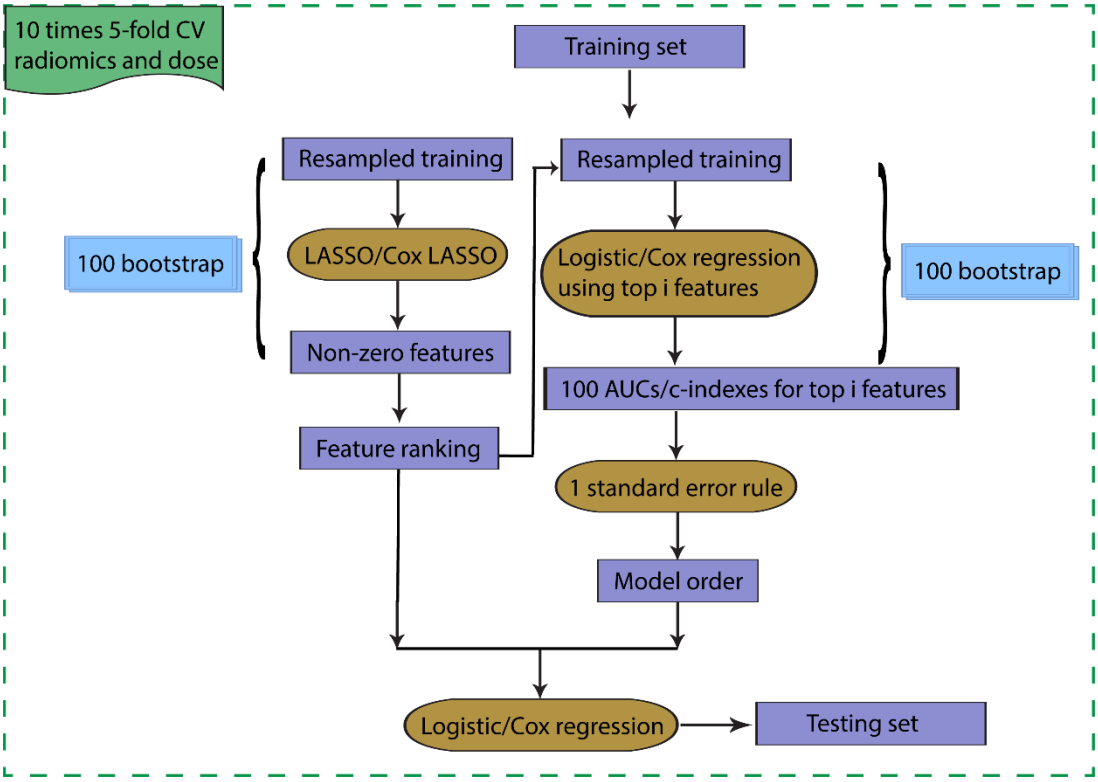


Fig. 4.1 Summary of radiomics model construction and evaluation.

4.3 Results

4.3.1 Phantom based reproducibility and robustness of radiomics features

Table 4.2 shows the mean CCC values from the liver phantom radiomics studies, assessed over the 5 repeat scans, OS-EM iterations 1/2, with/without Gaussian filtering and across all conditions (scans and parameters). CCC for sphericity is always 1 because the shape feature does not depend on the PET scan. There are in total 15 features that have mean CCC > 0.85: 1 global feature sphericity, 1 GLCM feature correlation, 2 GLRLM features grey level nonuniformity (GLN), and run length nonuniformity (RLN), 7 GLSZM features large zone emphasis (LZE), grey level nonuniformity (GLN), zone size nonuniformity (ZSN), zone percentage (ZP), large zone low grey

level emphasis (LZLGE), large zone high grey level emphasis (LZHGE), grey level variance (GLV), 4 NGTDM features coarseness, busyness, complexity and strength. The average CCCs for repeatability (same conditions, different scans) and reproducibility (different iterations and filtering) have similar results as shown in Table 4.2. Comparing with the mean CCC for both repeatability and reproducibility, there is 1 more robust feature for repeatability (dissimilarity), 6 more robust features (variance, contrast, dissimilarity, LGRE, SRLGE, GLV_GLRLM) and 2 less robust features (LZHGE, GLV_GLSZM) for different iterations, 2 less robust features (ZSN, LZHGE) for with/without filtering.

Table 4.2 Mean CCC for 5 repeat scans of the liver phantom, OS-EM iterations 1/2, with/without Gaussian filtering and across all conditions.

	Radiomics and other metrics	Mean CCC for 5 scans	Mean CCC for iterations 1/2	Mean CCC for with/without Gaussian filtering	Mean CCC across all conditions
	Volume	NA	NA	NA	NA
Global	Sphericity	**1.000**	**1.000**	**1.000**	**1.000**
	Variance	0.465	**0.957**	0.439	0.621
	Coefficient of variation	0.567	0.689	0.499	0.585
	Skewness	0.384	0.724	0.855	0.654
	Kurtosis	0.228	0.415	0.501	0.381
GLCM	Energy	0.183	0.747	0.144	0.358
	Contrast	0.825	**0.931**	0.695	0.817
	Entropy	0.239	0.712	0.336	0.429
	Homogeneity	0.682	0.835	0.732	0.750
	IDM	0.627	0.821	0.717	0.722
	Correlation	**0.897**	**0.967**	**0.922**	**0.929**

	SumMean	0.158	0.686	0.134	0.326
	Variance_GLCM	0.689	0.640	0.485	0.604
	Dissimilarity	**0.869**	**0.912**	0.806	0.845
GLRLM	SRE	0.517	0.787	0.642	0.648
	LRE	0.629	0.788	0.537	0.651
	GLN	**0.998**	**0.998**	**0.996**	**0.997**
	RLN	**0.997**	**0.997**	**0.991**	**0.995**
	RP	0.611	0.814	0.654	0.693
	LGRE	0.056	**0.867**	0.186	0.370
	HGRE	0.145	0.706	0.174	0.342
	SRLGE	0.051	**0.885**	0.152	0.363
	SRHGE	0.279	0.811	0.269	0.453
	LRLGE	0.077	0.814	0.302	0.398
	LRHGE	0.280	0.333	0.184	0.266
	GLV	0.674	**0.934**	0.679	0.762
	RLV	0.633	0.845	0.500	0.659
GLSZM	SZE	0.046	0.370	0.441	0.286
	LZE	**0.948**	**0.910**	**0.879**	**0.912**
	GLN	**0.953**	**0.985**	**0.988**	**0.975**
	ZSN	**0.928**	**0.895**	0.839	**0.887**
	ZP	**0.880**	**0.919**	**0.927**	**0.909**
	LGZE	-0.023	0.703	-0.174	0.169
	HGZE	0.384	0.606	0.556	0.515
	SZLGE	0.091	0.832	-0.161	0.254
	SZHGE	0.042	0.342	-0.067	0.106
	LZLGE	**0.885**	**0.978**	**0.949**	**0.937**
	LZHGE	**0.877**	0.834	0.719	**0.850**
	GLV	**0.880**	0.782	**0.870**	**0.862**
	ZSV	0.541	0.724	0..659	0.641

NGTDM	Coarseness	**0.920**	**0.982**	**0.922**	**0.941**
	Contrast	0.733	0.841	0.424	0.666
	Busyness	**0.940**	**0.958**	**0.909**	**0.936**
	Complexity	**0.893**	**0.988**	**0.924**	**0.935**
	Strength	**0.958**	**0.983**	**0.921**	**0.954**
Dose	Mean absorbed dose	NA	NA	NA	NA

4.3.2 Lesion dosimetry and outcome data

A total of 105 lesions > 2 mL were segmented. The average lesion volume was 45 mL (median:10 ml, range:2 - 833). The average lesion absorbed dose was 336 Gy (median: 265, range:1-1271). The response rate according to RECIST applied at the lesion level was 31% (32/105). The number of metastasis and primary HCC lesions are 70 and 35, respectively, with lesion specific response rate being 26% (9/35) and 33% (23/70) for the 2 groups. There are 103 lesions that have progression data, two metastatic lesions were excluded due to lack of follow-up. The number of progression events for all the lesions was 14 (4 HCC, 10 metastatic). The mean time-to-event are 322 days (median: 229 days, range: 44-1174 days). The mean time-to-event was 342 days (median: 309, range: 50-1174) for metastatic lesions and 284 days (median: 199 days, range: 44-860 days) for HCC. Kaplan Meier analysis showed that the time to progression for HCC and metastasis was not statistically significantly different (p-value=0.49)

4.3.2 Outcome models: Radiomics, absorbed dose, and combined models

4.3.2.1 Univariate analysis

The univariate results for volume, radiomics features and absorbed dose are shown in Table 4.3 and Table 4.4 (with Table 4.3 showing all the features and Table 4.4 showing only the 15 robust radiomics features). These are the Spearman correlation between specific features (or absorbed dose) and OR, and the univariate Cox regression results for progression. Volume has been shown to correlate with patient prognosis for different cancer types [41]. In our study, the Spearman coefficients of volume in terms of OR is -0.215 (p-value = 0.028). Among the 46 radiomics features (including volume), 10 features are significant (p-value < 0.001) for OR: 2/9 GLCM features, 3/13 GLRLM features, 4/13 GLSZM features and 1/5 NGTDM features. Among the 15 robust radiomics features, 8 features are significant for OR: LZE (p-value= 0.0005), ZP (p-value= 0.0004), LZLGE (p-value= 0.001), LZHGE (p-value= 0.002), GLV (p-value= 0.0009), Coarseness (p-value= 0.003), Busyness (p-value= 0.001), and Strength (p-value= 0.003). Absorbed dose is a significant predictor of the OR (p-value= 0.0003). In comparison, among the 46 radiomics features (including volume), no features are significant for progression. ZSN, a robust feature, is the most significant one (p-value= 0.063) for progression. Absorbed dose is a marginally significant predictor for progression (p-value= 0.005).

Table 4.3 Spearman correlation coefficients between lesion-level overall response and all radiomics features/absorbed dose with corresponding p-values (with Bonferroni correction). Univariate Cox regression with c-index, hazard ratio and corresponding p-values for progression are also indicated.

	Radiomics and other metrics	Spearman correlation for OR	P value for OR	C-index for progression	Hazard Ratio for progression	P value for progression
	Volume	-0.215	0.028	0.565	0.282	0.417
Global	Sphericity	0.142	0.148	0.590	0.728	0.313
	Variance	-0.128	0.193	0.789	0.193	0.002
	Coefficient of variation	-0.241	0.013	0.657	1.129	0.298
	Skewness	-0.161	0.100	0.741	1.298	0.116
	Kurtosis	0.133	0.177	0.657	1.147	0.226
GLCM	Energy	-0.137	0.164	0.537	1.103	0.333
	Contrast	0.247	0.011	0.484	0.885	0.711
	Entropy	0.212	0.030	0.484	0.837	0.158
	Homogeneity	-0.348	**0.0003**	0.489	1.275	0.233
	IDM	-0.351	**0.0002**	0.488	1.280	0.240
	Correlation	-0.216	0.027	0.438	1.019	0.950
	SumMean	0.206	0.035	0.747	0.651	0.023
	Variance_GLCM	-0.069	0.487	0.704	0.588	0.037
	Dissimilarity	0.277	0.004	0.484	0.858	0.609
GLRLM	SRE	0.368	**0.0001**	0.532	0.744	0.169
	LRE	-0.374	**8.321e-05**	0.482	1.109	0.361
	GLN	-0.269	0.006	0.600	0.297	0.323
	RLN	-0.236	0.015	0.639	0.213	0.201
	RP	0.366	**0.0001**	0.507	0.791	0.197
	LGRE	-0.121	0.218	0.702	1.140	0.213
	HGRE	0.183	0.061	0.775	0.538	0.011
	SRLGE	-0.055	0.578	0.737	1.299	0.073
	SRHGE	0.213	0.029	0.754	0.532	0.015

	LRLGE	-0.226	0.021	0.644	1.077	0.481
	LRHGE	-0.031	0.753	0.820	0.674	0.136
	GLV	0.269	0.006	0.633	0.507	0.097
	RLV	0.295	0.002	0.563	0.577	0.169
GLSZM	SZE	0.072	0.465	0.745	0.558	0.006
	LZE	-0.333	**0.0005**	0.562	0.415	0.629
	GLN	-0.121	0.218	0.734	0.326	0.088
	ZSN	-0.081	0.412	0.752	0.358	0.063
	ZP	0.341	**0.0004**	0.491	0.804	0.502
	LGZE	-0.010	0.920	0.663	1.438	0.156
	HGZE	-0.034	0.732	0.629	0.919	0.733
	SZLGE	0.039	0.691	0.613	1.194	0.453
	SZHGE	-0.004	0.970	0.764	0.586	0.058
	LZLGE	-0.317	**0.001**	0.460	0.872	0.760
	LZHGE	-0.300	0.002	0.676	0.006	0.348
	GLV	0.320	**0.0009**	0.549	0.491	0.136
	ZSV	-0.233	0.017	0.601	1.383	0.007
NGTDM	Coarseness	0.285	0.003	0.601	1.027	0.930
	Contrast	0.228	0.020	0.518	0.725	0.448
	Busyness	-0.307	**0.001**	0.482	0.522	0.585
	Complexity	0.244	0.012	0.609	1.124	0.657
	Strength	0.284	0.003	0.669	1.110	0.321
Dose	Mean absorbed dose	0.345	**0.0003**	0.819	0.121	0.005

Inter-feature correlation is shown in the correlation heat map of Fig. 4.2. GLN, RLN, LZE, LZHGE are highly correlated with volume (Spearman coefficients > 0.85). In general, the radiomics features are highly correlated with each other (except sphericity). Though most of the radiomics

features are still significantly correlated with dose (except sphericity, GLN, and ZSN), the correlation of radiomics features with dose is generally lower than radiomics features amongst them, as shown in Table 4.4.

Table 4.4 Summary of statistical analysis for volume, the 15 robust radiomics features and absorbed dose with Bonferroni correction.

	Features	Spearman correlation with absorbed dose	P value for dose correlation	Spearman correlation with OR	P value for OR	C-index for progression	Hazard Ratio for progression	P value for progression
	Volume	-0.262	0.007	-0.215	0.028	0.565	0.282	0.417
Global	Sphericity	0.061	0.539	0.142	0.148	0.590	0.728	0.313
GLCM	Correlation	-0.340	3.882e-4	-0.216	0.027	0.438	1.019	0.950
GLRLM	GLN	-0.362	1.45e-4	-0.269	0.006	0.600	0.297	0.323
	RLN	-0.252	0.010	-0.236	0.015	0.639	0.213	0.201
GLSZM	LZE	-0.482	1.989e-7	-0.333	0.0005	0.562	0.415	0.629
	GLN	-0.078	0.427	-0.121	0.218	0.734	0.326	0.088
	ZSN	-0.057	0.565	-0.081	0.412	0.752	0.358	0.063
	ZP	0.483	1.828e-7	0.341	0.0004	0.491	0.804	0.502
	LZLGE	-0.548	1.485e-9	-0.317	0.001	0.460	0.872	0.760
	LZHGE	-0.293	0.002	-0.300	0.002	0.676	0.006	0.348
	GLV	0.467	5.104e-7	0.320	0.0009	0.549	0.491	0.136
NGTDM	Coarseness	0.379	6.789e-5	0.285	0.003	0.601	1.027	0.930
	Busyness	-0.509	2.862e-8	-0.307	0.001	0.482	0.522	0.585
	Complexity	0.324	7.596e-4	0.244	0.012	0.609	1.124	0.657

	Strength	0.245	0.012	0.284	0.003	0.669	1.110	0.321
DOSE	Mean absorbed dose	NA	NA	0.345	0.0003	0.819	0.121	0.005

4.3.2.2 Multivariate analysis

Given the limited sample size, we included both primary and metastasis cases in the modeling. For the subset of robust features, the model order is 2 for both OR and progression endpoints, with top 2 features for OR being absorbed dose and zone percentage (ZP), and for progression being absorbed dose and ZSN. Fig. 4.3 shows the model order determination for the robust features and absorbed dose. The top 5 features are shown in Table 4.2 for OR and progression models. (Model

order determination and the top 5 features using all the radiomics features and absorbed dose are presented in the supplemental materials Fig. 4.4 and table 4.3).

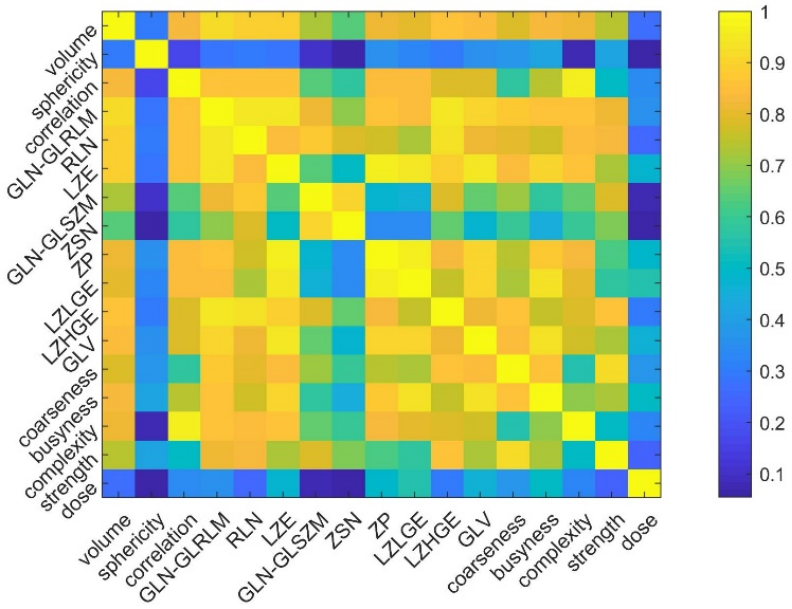


Fig. 4.2 Spearman correlation heat map for radiomics features, volume and absorbed dose.

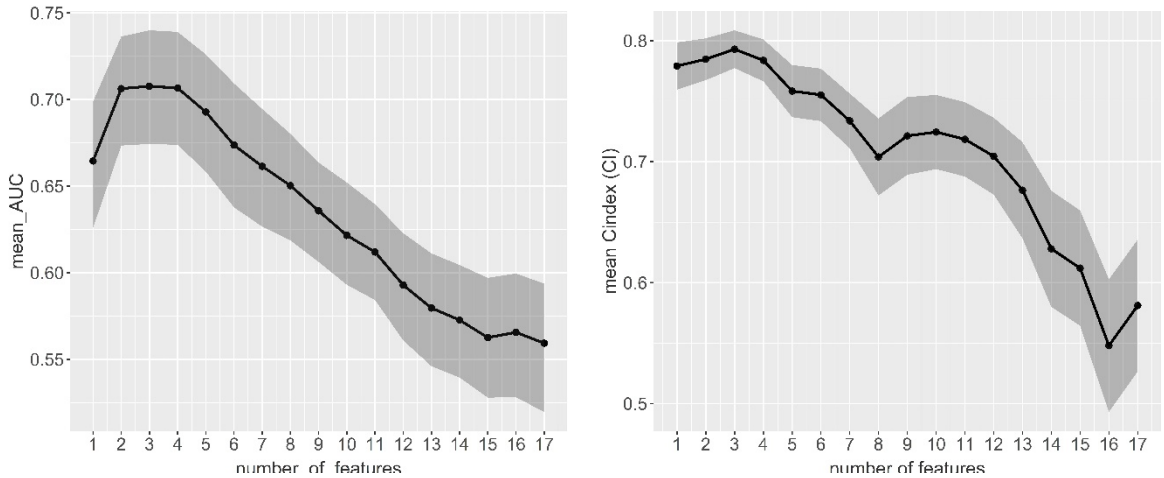


Fig. 4.4 Radiomics_robust+dose model order determination for OR (left) and progression (right): average AUC/c-index vs. number of top features included.

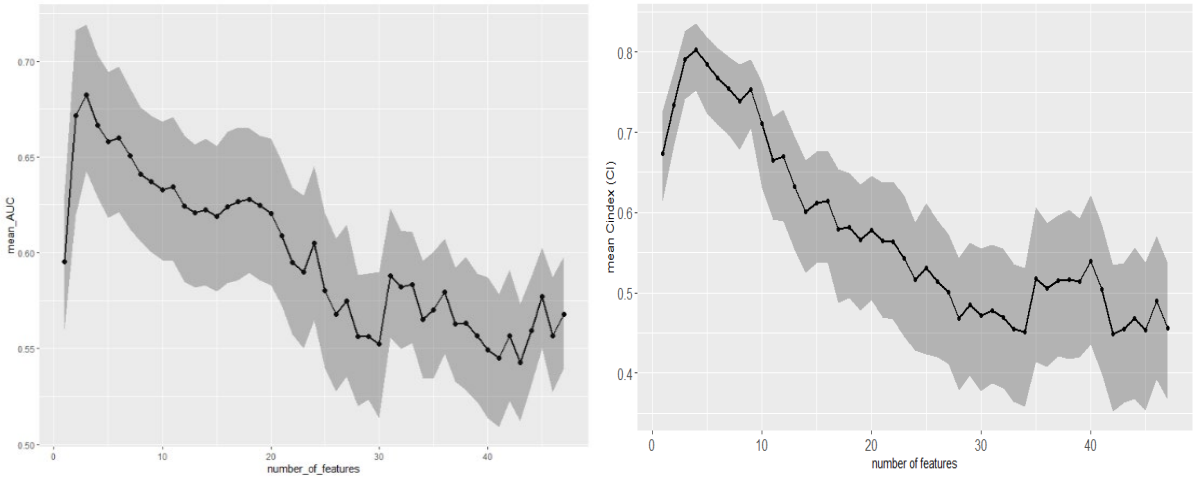


Fig. 4.3 Radiomics_all+dose model order determination for OR (left) and PFS (right). Average AUC/c-index vs. number of top features included. When using all the radiomics features, the average model order calculated using nested cross validation is 2 for OR classification and 3 for PFS, with top 2 features being variance and absorbed dose and top 3 features being variance, absorbed dose and LRHGE, respectively. The nested CV AUCs for radiomics_all+dose is 0.672 (0.620-0.716) for OR and 0.791 (95%CI: 0.740-0.825) for progression.

Table 4.5 Top 5 features for the combined models with robust radiomics features, volume and absorbed dose.

OR	Progression
Mean absorbed dose	Mean absorbed dose
ZP	ZSN
Sphericity	Strength
GLV	Complexity
Coarseness	Sphericity

Table 4.6 Top 5 features for the combined models with **all** radiomics features, volume and absorbed dose.

OR	Progression
Variance	Variance
Mean absorbed dose	Mean absorbed dose
Sphericity	LRHGE
SZE	SZHGE
LRHGE	Kurtosis

After the model order and top features were decided, nested cross-validation was applied to estimate the performance of the final model. Table 4.5 lists the results for models with ZP only, ZSN only, absorbed dose only and the combined models (radiomics robust + dose). When considering the entire cohort, for the combined models the average AUCs for OR (0.729 (95% CI: 0.702-0.758)), and the average c-indexes for progression (0.803 (95% CI: 0.790-0.815) are superior to the corresponding values for the absorbed dose only and ZP/ZSN only models. The results for the subgroups of primary and metastasis cases are shown in Table 4.5 as well. For the OR model in the subgroup of HCC, the radiomics only model shows the best performance with average AUC of 0.762 (95% CI: 0.680-0.834), and in the subgroup of metastasis, the absorbed dose only model shows the best performance with average AUC of 0.696 (95% CI: 0.654-0.737). However, for the progression analysis, in both subgroups the combined model outperforms the

individual models. Fig. 4.5 shows the ROC curve for OR using radiomics alone, dose alone and combined models, and Fig. 4.6 shows the Kaplan-Meier plot for progression for the combined models, respectively. Log-rank test was used for the comparison of high and low risk groups for progression. The cutoff was median value of the predicted Cox survival probability. The weights of OR model and progression models are shown below. Artificially increasing the number of cases using ADASYN was evaluated as well, but found no substantial difference.

OR model (Generalized linear regression model):

$$\text{logit}(y) \sim -0.892 + 0.520\ ZP + 0.488\ Dose$$

Distribution = Binomial

Progression Cox model:

$$h(t) \sim h0(t) * \exp(-0.530\ ZSN + -1.707\ Dose)$$

Table 4.7 Average AUC/c-index for individual and combined models with all the lesions, HCC lesions and metastasis lesions

	Average AUC (95 % confidence intervals)		
OR Model	All (105)	Primary HCC (35)	Metastasis (70)
Radiomics (ZP)	0.713 (0.685-0.741)	0.762 (0.680-0.834)	0.658 (0.623-0.693)
Absorbed Dose	0.713(0.678-0.746)	0.717 (0.642-0.786)	0.696 (0.654-0.737)
Combined (Dose + ZP)	0.729 (0.702-0.758)	0.734 (0.660-0.802)	0.692 (0.653-0.723)
	Average c-index (95 % confidence intervals)		
Progression Model	All (103)	Primary HCC (35)	Metastasis (68)

Radiomics (ZSN)	0.694 (0.676-0.710)	0.565 (0.528-0.598)	0.656 (0.629-0.680)
Absorbed Dose	0.754 (0.742-0.766)	0.613 (0.585-0.635)	0.719 (0.700-0.737)
Combined (Dose+ZSN)	0.803 (0.790-0.815)	0.638 (0.610-0.661)	0.762 (0.740-0.780)

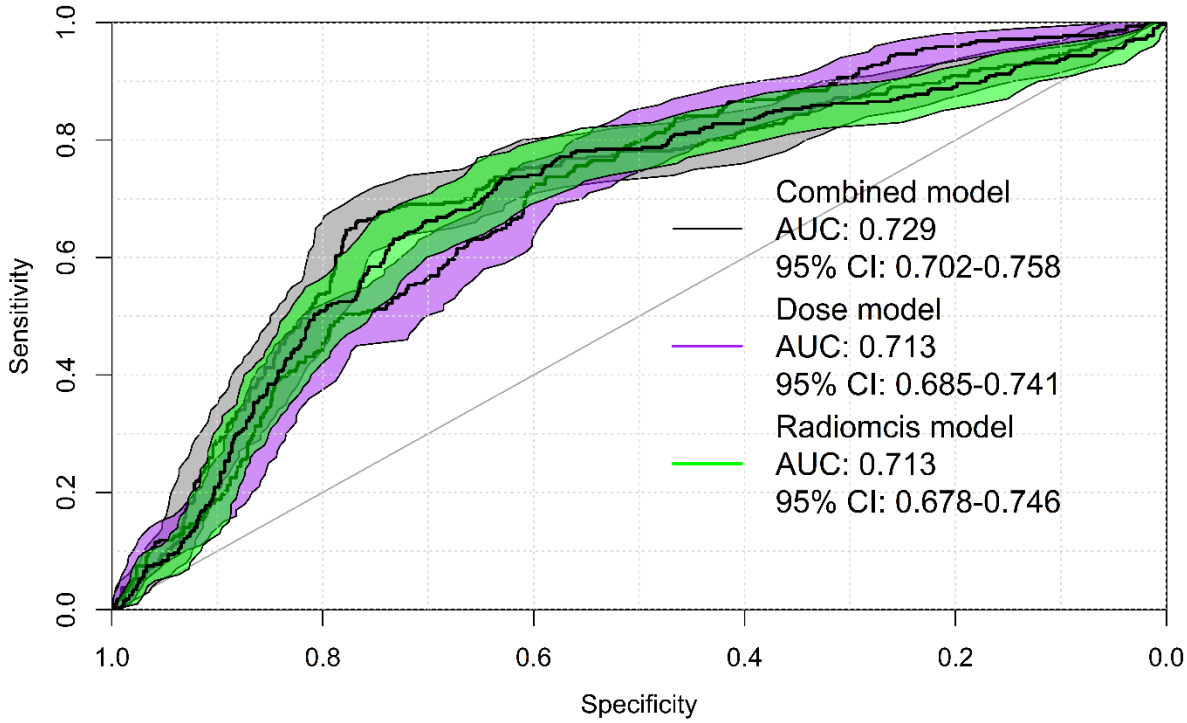


Fig. 4.5 ROC curves for overall response (OR) at first follow-up with the combined model, radiomics alone model, and dose alone model.

4.4 Discussion

Uncovering robust radiomics features is an important task for building robust models for identifying responders and non-responders and prediction of cancer progression. Thus, radiomics features extracted from repeated PET scans, different number of OS-EM PET iterations,

with/without Gaussian post-filtering were evaluated for robustness using CCC. Despite the higher noise associated with ^{90}Y PET compared with FDG PET, 15 radiomics features were identified as robust with CCC > 0.85. In general, the robust features for different scans (repeatability), OS-EM iterations 1/2 and with/without filtering largely overlap, which indicates that robust features tend to be consistent for different imaging settings. The results also showed that more features are robust to different iteration setting and less features are robust to application of Gaussian filtering. In a study of intratumor FDG PET uptake heterogeneity quantification by Hatt *et al.*, zone percentage (ZP) was found to be robust with respect to the delineation method used and the partial volume effects. This feature also demonstrated high differentiation power for prediction of response in esophageal carcinoma [42]. In a study by Doumou *et al.*, ZP presented substantial agreement across different segmentation and different levels of smoothing [43]. A study by Ashrafinia *et al.* showed that ZSN extracted from ^{99m}Tc-Sestamibi Myocardial-Perfusion SPECT (MPS) images showed high reproducibility [44]. Another recent study by Li *et al.* on FDG PET radiomics analysis, showed that ZSN is a stable feature [45]. The phantom repeatability and reproducibility study provides robust features for further radiomics modeling that has the potential to generalize to PET images reconstructed at other institutions where different reconstruction settings might have been applied. While this phantom study focused on reconstruction settings, there are other sources of variability as mentioned that we did not evaluate here, such as segmentation, interpolation, preprocessing, which are investigated in other literatures [22, 23, 33, 46] and reviewed in [47, 48].

The aim of this work is to find radiomics signature that can facilitate dose metrics in the prediction of tumor response. The final model order is small being 2 (dose+ZP and dose+ZSN), which is reasonable considering the high correlation between most radiomics features (Figure 4.2). The correlation between ZP and absorbed dose is 0.483 (p-value = 1.828e-7) and ZSN and absorbed dose is -0.057 (p-value= 0.565) (Table 4.4), which indicates that ZSN could provide more complementary information to the combined model than ZP. This is consistent with the substantial higher c-index for the combined absorbed dose and ZSN model (0.803) compared with ZSN only (0.694) and absorbed dose only (0.754) models for progression, but only slightly higher AUC for the combined absorbed dose and ZP model (0.729) compared with the ZP only (0.713) and absorbed dose only (0.713) models for OR (Table 4.5). In Fig. 4.5, the ROC curves for radiomics alone, dose alone and combined models did present some overlap. However, it still showed consistent trends in the data, that the combined model performs better than individual models. Access to larger Y-90 PET imaging data sets is required to independently validate these findings and to reach statistical significance for the improvement of the performance of the combined model over the individual models. Further studies, such as obtaining radiomics features from FDG-PET, CT, or MRI, could potentially add more complementary information and further improve the performance [49].

ZP is a feature from GLSZM matrix, quantifying the coarseness of the texture by the ratio of number of zones and number of voxels. The higher the value is, the finer the texture is, and according to our results the higher the probability the tumor will respond. Fig. 4.7 (a), (b) show example lesions with large/small ZP, that were classified as responder/non-responder; (c), (d)

show lesions with large/small ZSN, that did not progress for a long follow-up time (1174 days) and progressed in a short time (44 days). Smaller ZP values correspond to coarser appearance and worse response. In another study by Ha *et al.*, ZP was one of the features used to characterize locally advanced breast cancer [50]. The trend is consistent with what we found in our study, that larger ZP is associated with better response. ZSN measures the variability of size zone volumes in the ROIs, higher the value, larger the variance of the size zone volumes. The hazard ratio for ZSN is smaller than 1, which means the higher the ZSN, the better the lesion prognosis.

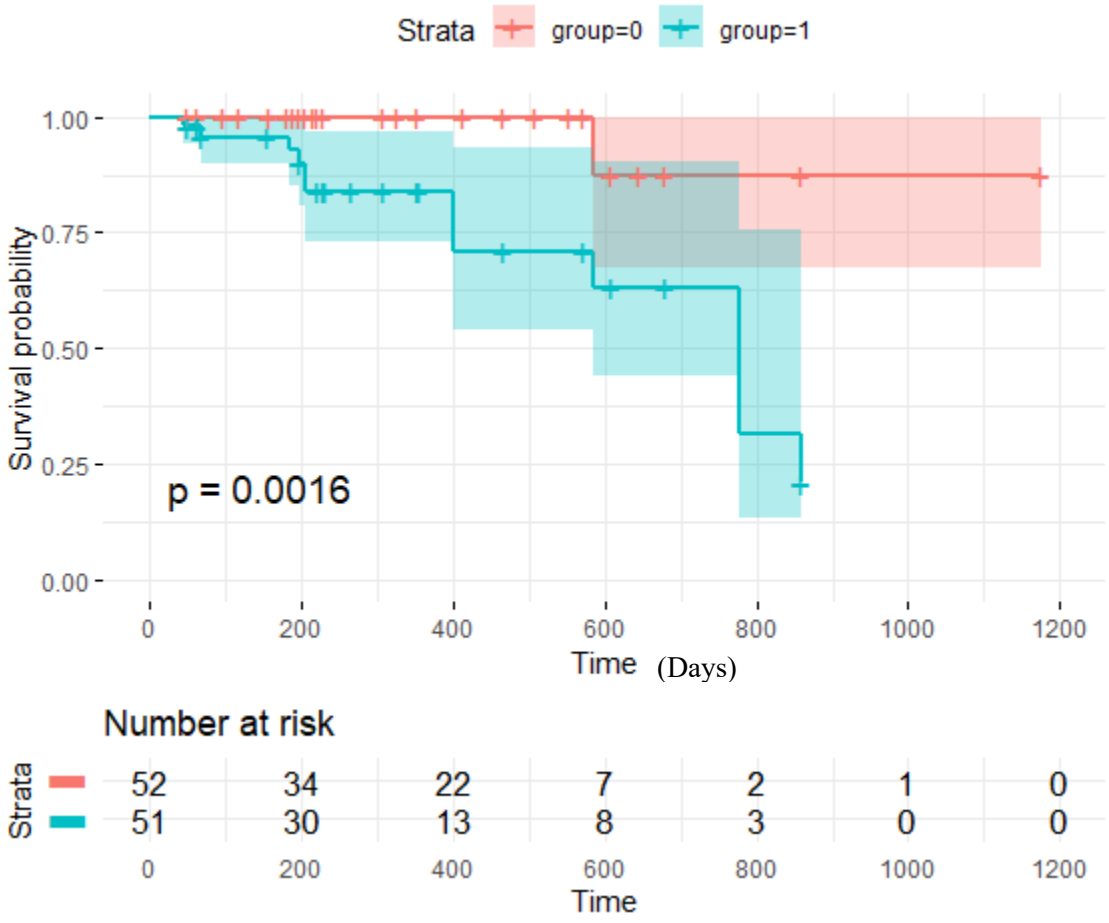
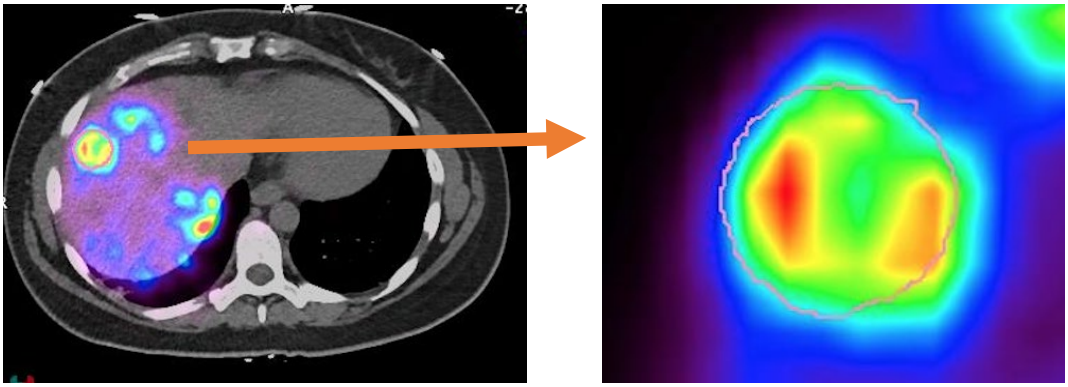
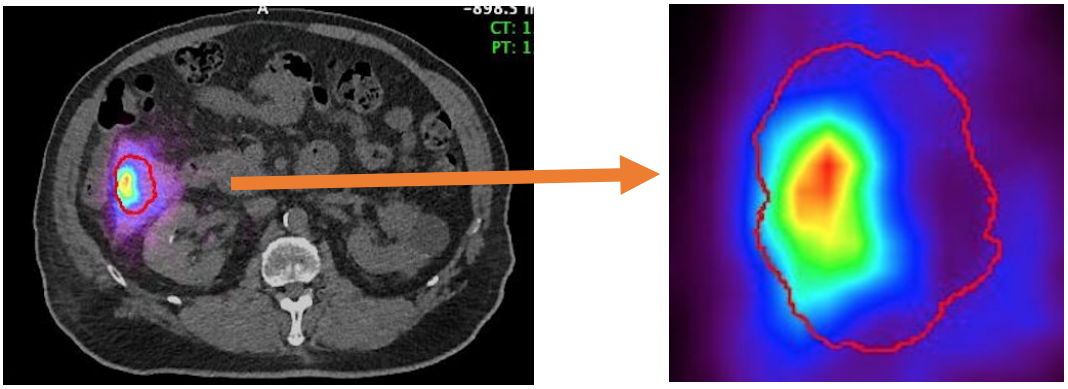


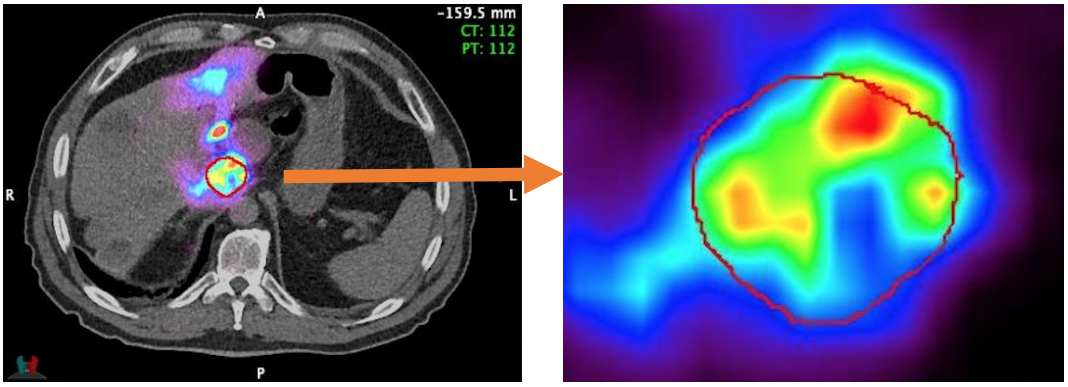
Fig. 4.6 Kaplan-Meier plot of the combined model for progression (absorbed dose + ZSN). This is the result of 10 times 5-fold cross validation, the test set for each fold were combined to evaluate the overall performance. High and low risk lesions for progression were stratified by median value of the Cox model output, with high risk group lesions having shorter time to progression, vice versa.



(a) ZP = 0.07, responder at first follow-up



(b) ZP = 0.02, non-responder at first follow-up



(c) ZSN = 340, status = no progression, time = 1174 days

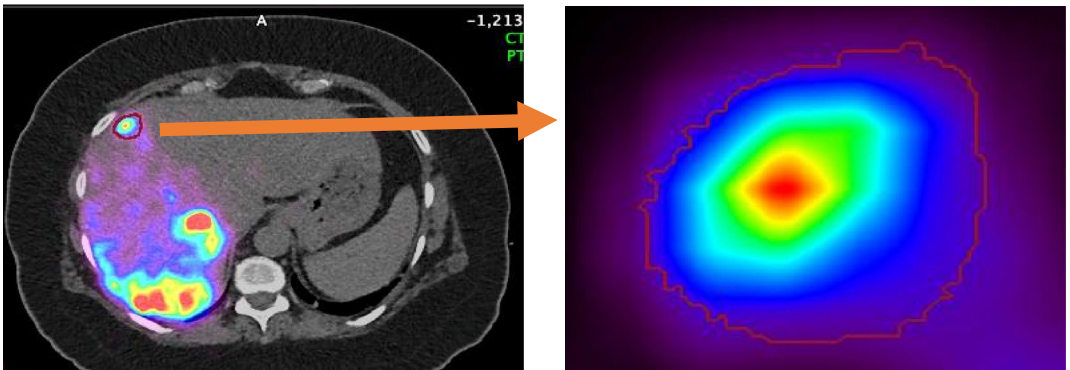


Fig. 4.7 Example 90Y PET/CT images with CT-defined lesion contours (left: PET/CT axial slice showing the anatomical position within liver, right: magnified lesion on PET). (a) Lesion with large ZP value corresponding to responder; (b) Lesion with small ZP value corresponding to non-responder; (c) lesion with large ZSN value corresponding to no progression at 1174 day; (d) Lesion with small ZSN value corresponding to progression in 44 days.

The modified LASSO method we developed was inspired by R. Bach's work on Bolasso, which showed that the Lasso selects all the variables that should enter the model with probability tending to one exponentially fast [51]. So, if we run the Lasso for multiple bootstrapped replications of a given sample, then intersecting the supports of the Lasso (i.e., non-zero coefficients) leads to consistent model selection. However, the direct application failed since the intersection of the supports lead to null for some datasets. Bunea *et al.* came up with similar variants of bootstrap enhanced LASSO (Be-LASSO) [52]. The percentage of times each predictor was selected (variable inclusion probability) was recorded and user-defined threshold (50%) was used to determine the variables. V. Abram *et al.* built upon Bunea's method of Be-LASSO [53]. Instead of user-defined probability for feature selection, they used the quantiles of the bootstrap distribution of the coefficients of variables to determine the significance of that variable. In our study, we developed a new way to select features, still based on the bootstrap LASSO. Instead of using predefined probability or the distribution quantile, we obtained a ranking of the features based on the frequency of being selected in the bootstrap, then, we performed cross validation to calculate the AUC/c-index vs. number of top features included in the model. In this way, we obtained the most parsimonious model, which is desired when small sample size is unavoidable.

In summary, absorbed dose is a strong predictor for tumor control, both in terms of OR at first follow-up and time to progression, which is consistent with recent reports [14-16]. The radiomics

feature signals the complimentary value of texture to improve the absorbed dose only model prediction. It is interesting to explore the underlying biological mechanism of the reason for higher ZP and ZSN leading to better prognosis, which should be investigated on larger dataset in the future. The two features model can be interpreted as: given the dose being fixed, the change in ZP/ZSN will help to predict tumor control (OR/progression). Using this information, additional attentions would be given to the lesions that possess lower ZP/ZSN value, which have a higher risk of failure (in terms of OR/progression), which is potentially informative for clinical decisions. Immediate prediction of response, based on radiomics features and dose metrics both of which can be derived from ^{90}Y PET/CT performed immediately after RE, has clinical utility. Instead of waiting for the first follow up morphologic imaging that typically occurs at > 2 months, the potential to predict non-responding lesions immediately after therapy would facilitate adaptive therapy to selected lesions where ^{90}Y RE is followed by further treatment such as stereotactic body radiation therapy or microwave ablation. Limitation of our study include the heterogeneous patient cohort and the small sample size. Patient ^{90}Y imaging data is scarce because post-therapy imaging is not routinely performed after RE, but studies reporting ^{90}Y SPECT/CT and PET/CT imaging is rising and is expected to become more readily available, enabling studies with larger cohorts.

4.5 Conclusion

In this study, radiomics only, absorbed dose only and combined models showed predictive ability for tumor OR and progression in ^{90}Y radioembolization patients. The final tumor OR model

consisting of the robust radiomics feature ZP and mean absorbed dose achieved a nested CV AUC 0.729 while the final progression model consisting of the robust radiomics feature ZSN and mean absorbed dose achieved a c-index of 0.803. Further validation on large external cohorts will be necessary for clinically applicable models. Nonetheless, this study showed the potential of combining ^{90}Y PET derived radiomics and absorbed dose for improved model building to predict tumor OR and progression in ^{90}Y radioembolization treatment.

4.6 References

1. Kennedy A. Radioembolization of hepatic tumors. J Gastrointest Oncol. 2014;5:178.
2. Gans JH, Lipman J, Golowa Y, Kinkhabwala M, Kaubisch A. Hepatic Cancers Overview: Surgical and Chemotherapeutic Options, How Do Y-90 Microspheres Fit in? Semin Nucl Med: Elsevier; 2019.
3. Gillies RJ, Kinahan PE, Hricak H. Radiomics: images are more than pictures, they are data. Radiology. 2015;278:563-77.
4. Aerts HJ, Velazquez ER, Leijenaar RT, Parmar C, Grossmann P, Carvalho S, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nature communications. 2014;5:4006.
5. Lambin P, Leijenaar RT, Deist TM, Peerlings J, De Jong EE, Van Timmeren J, et al. Radiomics: the bridge between medical imaging and personalized medicine. Nature reviews Clinical oncology. 2017;14:749.
6. Vallières M, Freeman CR, Skamene SR, El Naqa I. A radiomics model from joint FDG-PET and MRI texture features for the prediction of lung metastases in soft-tissue sarcomas of the extremities. Phys Med Biol. 2015;60:5471.
7. Zhang B, Tian J, Dong D, Gu D, Dong Y, Zhang L, et al. Radiomics features of multiparametric MRI as novel prognostic factors in advanced nasopharyngeal carcinoma. Clin Cancer Res. 2017;23:4259-69.
8. Li H, Zhu Y, Burnside ES, Huang E, Drukker K, Hoadley KA, et al. Quantitative MRI radiomics in the prediction of molecular classifications of breast cancer subtypes in the TCGA/TCIA data set. NPJ breast cancer. 2016;2:16012.
9. Veit-Haibach P, Buvat I, Herrmann K. EJNMMI supplement: bringing AI and radiomics to nuclear medicine. Springer; 2019.

10. Morin O, Vallières M, Jochems A, Woodruff HC, Valdes G, Braunstein SE, et al. A deep look into the future of quantitative imaging in oncology: a statement of working principles and proposal for change. International Journal of Radiation Oncology* Biology* Physics. 2018;102:1074-82.
11. Visvikis D, Le Rest CC, Jaouen V, Hatt M. Artificial intelligence, machine (deep) learning and radio (geno) mics: definitions and nuclear medicine imaging applications. Eur J Nucl Med Mol Imaging. 2019:1-8.
12. Blanc-Durand P, Van Der Gucht A, Jreige M, Nicod-Lalonde M, Silva-Monteiro M, Prior JO, et al. Signature of survival: a 18F-FDG PET based whole-liver radiomic analysis predicts survival after 90Y-TARE for hepatocellular carcinoma. Oncotarget. 2018;9:4549.
13. Gensure RH, Foran DJ, Lee VM, Gendel VM, Jabbour SK, Carpizo DR, et al. Evaluation of hepatic tumor response to yttrium-90 radioembolization therapy using texture signatures generated from contrast-enhanced CT images. Acad Radiol. 2012;19:1201-7.
14. Fowler KJ, Maughan NM, Laforest R, Saad NE, Sharma A, Olsen J, et al. PET/MRI of hepatic 90Y microsphere deposition determines individual tumor response. Cardiovasc Intervent Radiol. 2016;39:855-64.
15. Srinivas SM, Natarajan N, Kuroiwa J, Gallagher S, Nasr E, Shah SN, et al. Determination of radiation absorbed dose to primary liver tumors and normal liver tissue using post-radioembolization 90Y PET. Front Oncol. 2014;4:255.
16. Chan KT, Alessio AM, Johnson GE, Vaidya S, Kwan SW, Monsky W, et al. Prospective Trial using internal pair-production positron emission tomography to establish the yttrium-90 radioembolization dose required for response of hepatocellular carcinoma. International Journal of Radiation Oncology* Biology* Physics. 2018;101:358-65.
17. Kappadath SC, Mikell J, Balagopal A, Baladandayuthapani V, Kaseb A, Mahvash A. Hepatocellular Carcinoma Tumor Dose Response After 90Y-radioembolization With Glass Microspheres Using 90Y-SPECT/CT-Based Voxel Dosimetry. International Journal of Radiation Oncology* Biology* Physics. 2018;102:451-61.
18. Dewaraja YK, Devasia T, Kaza RK, Mikell JK, Owen D, Roberson PL, et al. Prediction of tumor control in 90Y radioembolization by logit models with PET/CT based dose metrics. J Nucl Med. 2019:jnumed. 119.226472.
19. Pasciak AS, Bourgeois AC, McKinney JM, Chang TT, Osborne DR, Acuff SN, et al. Radioembolization and the dynamic role of 90Y PET/CT. Front Oncol. 2014;4:38.
20. Willowson KP, Tapner M, Bailey DL. A multicentre comparison of quantitative 90 Y PET/CT for dosimetric purposes after radioembolization with resin microspheres. Eur J Nucl Med Mol Imaging. 2015;42:1202-22.
21. Robinson K, Li H, Lan L, Schacht D, Giger M. Radiomics robustness assessment and classification evaluation: A two‐stage method demonstrated on multivendor FFDM. Med Phys. 2019;46:2145-56.
22. Traverso A, Wee L, Dekker A, Gillies R. Repeatability and reproducibility of radiomic features: a systematic review. International Journal of Radiation Oncology* Biology* Physics. 2018;102:1143-58.

23. Zwanenburg A. Radiomics in nuclear medicine: robustness, reproducibility, standardization, and how to avoid data analysis traps and replication crisis. Eur J Nucl Med Mol Imaging. 2019;46:2638-55.
24. Waninger J, Green M, Cheze CLR, Rosen B, El IN. Integrating radiomics into clinical trial design. The quarterly journal of nuclear medicine and molecular imaging: official publication of the Italian Association of Nuclear Medicine (AIMN)[and] the International Association of Radiopharmacology (IAR),[and] Section of the Society of. 2019.
25. Zwanenburg A, Leger S, Agolli L, Pilz K, Troost EG, Richter C, et al. Assessing robustness of radiomic features by image perturbation. Sci Rep. 2019;9:1-10.
26. El Naqa I, Grigsby PW, Apte A, Kidd E, Donnelly E, Khullar D, et al. Exploring feature-based approaches in PET images for predicting cancer treatment outcomes. Pattern recognition. 2009;42:1162-71.
27. Ohri N, Duan F, Snyder BS, Wei B, Machtay M, Alavi A, et al. Pretreatment 18F-FDG PET textural features in locally advanced non–small cell lung cancer: Secondary analysis of ACRIN 6668/RTOG 0235. J Nucl Med. 2016;57:842-8.
28. Zwanenburg A, Leger S, Vallières M, Löck S. Image biomarker standardisation initiative. arXiv preprint arXiv:161207003. 2016.
29. Hatt M, Majdoub M, Vallières M, Tixier F, Le Rest CC, Groheux D, et al. 18F-FDG PET uptake characterization through texture analysis: investigating the complementary nature of heterogeneity and functional tumor volume in a multi–cancer site patient cohort. J Nucl Med. 2015;56:38-44.
30. El Naqa I. The role of quantitative PET in predicting cancer treatment outcomes. Clinical and translational imaging. 2014;2:305-20.
31. Therasse P, Arbuck SG, Eisenhauer EA, Wanders J, Kaplan RS, Rubinstein L, et al. New guidelines to evaluate the response to treatment in solid tumors. J Natl Cancer Inst. 2000;92:205-16.
32. Lin L. A Concordance Correlation Coefficient to Evaluate Reproducibility. Biometric, 45, 255-268. 1989.
33. Zhao B, Tan Y, Tsai W-Y, Qi J, Xie C, Lu L, et al. Reproducibility of radiomics for deciphering tumor phenotype with imaging. Sci Rep. 2016;6:23428.
34. Fave X, Mackin D, Yang J, Zhang J, Fried D, Balter P, et al. Can radiomics features be reproducibly measured from CBCT images for patients with non‐small cell lung cancer? Med Phys. 2015;42:6784-97.
35. Hu P, Wang J, Zhong H, Zhou Z, Shen L, Hu W, et al. Reproducibility with repeat CT in radiomics study for rectal cancer. Oncotarget. 2016;7:71440.
36. van Timmeren JE, Leijenaar RT, van Elmpt W, Wang J, Zhang Z, Dekker A, et al. Test–retest data for radiomics feature stability analysis: generalizable or study-specific? Tomography. 2016;2:361.
37. Friedman J, Hastie T, Tibshirani R. The elements of statistical learning: Springer series in statistics New York; 2001.
38. DiCiccio TJ, Efron B. Bootstrap confidence intervals. Statistical science. 1996:189-212.

39. Hochberg Y, Benjamini Y. More powerful procedures for multiple significance testing. Stat Med. 1990;9:811-8.
40. Haibo H, Yang B, Edwardo GA, Shutao L. Adaptive Synthetic Sampling Approach for Imbalanced Learning. IEEE International Joint Conference on Neural Networks, IJCNN; 2016. p. 1322-8.
41. Brooks FJ, Grigsby PW. The effect of small tumor volumes on studies of intratumoral heterogeneity of tracer uptake. J Nucl Med. 2014;55:37-42.
42. Hatt M, Tixier F, Le Rest CC, Pradier O, Visvikis D. Robustness of intratumour 18 F-FDG PET uptake heterogeneity quantification for therapy response prediction in oesophageal carcinoma. Eur J Nucl Med Mol Imaging. 2013;40:1662-71.
43. Doumou G, Siddique M, Tsoumpas C, Goh V, Cook GJ. The precision of textural analysis in 18 F-FDG-PET scans of oesophageal cancer. Eur Radiol. 2015;25:2805-12.
44. Ashrafinia S, Ghazi P, Marcus CV, Taghipour M, Yan R, Valenta I, et al. Robustness and Reproducibility of Radiomic Features in 99mTc-Sestamibi SPECT imaging of Myocardial Perfusion. Med Phys: WILEY 111 RIVER ST, HOBOKEN 07030-5774, NJ USA; 2017.
45. Li Y, Jiang J, Lu J, Jiang J, Zhang H, Zuo C. Radiomics: a novel feature extraction method for brain neuron degeneration disease using 18F-FDG PET imaging and its implementation for Alzheimer's disease and mild cognitive impairment. Ther Adv Neurol Disord. 2019;12:1756286419838682.
46. Parmar C, Velazquez ER, Leijenaar R, Jermoumi M, Carvalho S, Mak RH, et al. Robust radiomics feature quantification using semiautomatic volumetric segmentation. PLoS One. 2014;9.
47. Zwanenburg A, Vallières M, Abdalah MA, Aerts HJ, Andrearczyk V, Apte A, et al. The Image Biomarker Standardization Initiative: standardized quantitative radiomics for high-throughput image-based phenotyping. Radiology. 2020;295:328-38.
48. Avanzo M, Stancanello J, El Naqa I. Beyond imaging: the promise of radiomics. Phys Med. 2017;38:122-39.
49. Wei L, Osman S, Hatt M, El Naqa I. Machine learning for radiomics-based multi-modality and multi-parametric modeling. The quarterly journal of nuclear medicine and molecular imaging: official publication of the Italian Association of Nuclear Medicine (AIMN)[and] the International Association of Radiopharmacology (IAR),[and] Section of the Society of. 2019;63:323.
50. Ha S, Park S, Bang J-I, Kim E-K, Lee H-Y. Metabolic radiomics for pretreatment 18 F-FDG PET/CT to characterize locally advanced breast cancer: histopathologic characteristics, response to neoadjuvant chemotherapy, and prognosis. Sci Rep. 2017;7:1556.
51. Bach FR. Bolasso: model consistent lasso estimation through the bootstrap. Proceedings of the 25th international conference on Machine learning: ACM; 2008. p. 33-40.
52. Bunea F, She Y, Ombao H, Gongvatana A, Devlin K, Cohen R. Penalized least squares regression methods and applications to neuroimaging. Neuroimage. 2011;55:1519-27.
53. Abram SV, Helwig NE, Moodie CA, DeYoung CG, MacDonald III AW, Waller NG. Bootstrap Enhanced Penalized Regression for Variable Selection with Neuroimaging Data. Front Neurosci. 2016;10:344.

CHAPTER 5

Variational Autoencoder SurvivalNet Radiomics Modeling of Overall Survival for Hepatocellular Carcinoma Patients

This chapter developed a new radiomics actuarial model for liver cancer patients post radiotherapy and is based on the paper: **Wei, L.,** Owen, D., Mendiratta-Lala, M., Rosen, B., Cuneo, K., Lawrence, T. S., Ten Haken, R. K., El Naqa, I., "Variational Autoencoder SurvivalNet Radiomics Modeling of Overall Survival for Hepatocellular Carcinoma Patients." Physica Medica (2020), *submitted.*

5.1 Introduction

Liver cancer is a leading cause of cancer related deaths worldwide, with increasing incidences [1]. Hepatocellular carcinoma (HCC) is the most common type of primary liver cancer. Surgical resection and liver transplantation are used with curative intent for selected patients [2]. However, the majority of HCC patients are ineligible for surgery due to the location of the tumor or poor liver function [3]. There are several non-surgical liver-directed treatments including radiofrequency ablation (RFA), microwave ablation (MWA), trans-arterial chemoembolization

(TACE) and, more recently, stereotactic body radiotherapy (SBRT). RFA/MWA can be limited by lesion size and proximity to critical organs; TACE is non-curative, with limitations such as poor neovascularization of some tumors or portal vein involvement [4]. Recently, with the development of advanced radiotherapy delivery technologies, more precise partial liver irradiation using SBRT has become an important option for HCCs that are not suitable for resection [5, 6].

Although over 90% of tumors will be controlled by SBRT [5, 7], intrahepatic progression within the liver, remote from the treatment zone is common, with failure rates reported at 50% [3, 8]. SBRT treatment improved the overall survival rate as well, but the current survival rate is still not satisfying [9]. Radiomics may aid in developing models to predict the risk of intrahepatic progression and overall survival. Radiomics is a field of medical image analytics by which images are converted into a large number of quantitative features with subsequent datamining that relates these features to biological and clinical endpoints. Radiomics has been widely applied in cancer research and has shown to be able to capture distinct phenotypic differences and be associated with clinical prognosis in many cancer types [10-13]. Although various studies have been conducted to identify key radiomic features which predict overall survival and local control for HCC, little has been done to stratify the major intrahepatic risk after SBRT treatment. These current studies in overall survival use mostly Cox models or random survival forests.

This study focuses on pre-SBRT arterial phase contrast-enhanced computed tomography (CECT) images, involving a relatively large data set of 167 patients. The novelty of our work can be summarized as: (1) development of a comprehensive model based on radiomics (features from

both gross tumor volume (GTV) regions and liver exclusive of the GTVs (liver-GTV)), clinical features and raw CT images; (2) novel VAE based survival model combining different sources of information; (3) investigations of correlation and contribution of clinical, radiomics, image features and miRNA data, providing possible interpretation of underlying mechanism; (4) patch-based training that augmented data and improved the performance. This manuscript contributes to providing a better understanding of the HCC heterogeneity across patients, guidance for personalized HCC treatment planning in clinical practice and development of new methods that fuse conventional and deep learning based radiomic analyses.

5.2 Methods and Materials

A brief description of the workflow presented in this study is shown in Fig. 5.1. Details are provided in the following.

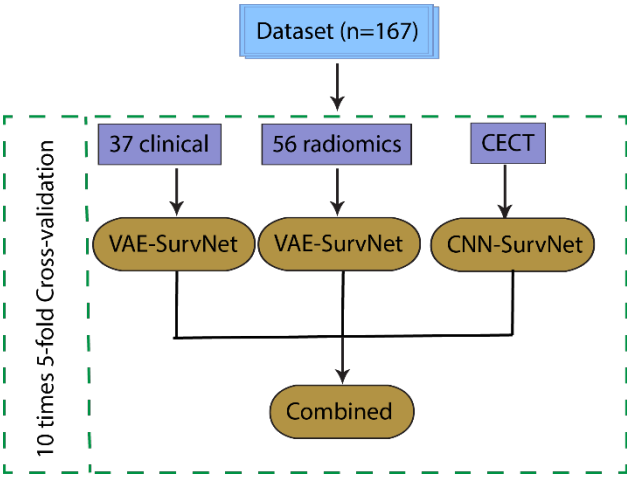


Fig. 5.1 Workflow for the modeling.

5.2.1 Patient cohort

After IRB approval, a HIPAA compliant retrospective analysis of HCC patients treated with SBRT was performed. A total of 303 HCC patients treated with SBRT were reviewed. Patients without: (1) contrast-enhanced CT (CECT) images; (2) gross tumor or liver contour in the database were excluded from analysis. A total of 167 HCC patients met the inclusion criteria. The endpoint of the intrahepatic recurrence-free survival was whether or not patients with HCC developed intrahepatic tumors after SBRT. This was defined by the presence of new tumor outside the planning target volume (PTV) of the previously treated tumor(s). The ground truths were determined by experienced clinicians based on CECTs, MRI, and relevant clinical records for each patient. The IRS and OS events were right censored if no recurrence or death till the last follow-up date.

Univariate Cox model c-index for clinical, radiomics and miRNA were investigated. Overall, 37 clinical features, 56 radiomics and 84 miRNA features were tested. P-values of 0.05 was considered significant. The radiomics features extracted in this study is the most commonly used and robust signatures based on a thorough literature review [14-18].

5.2.2 CT Images Acquisition and Processing

Arterial phase images and structure sets including liver, GTV, and liver-GTV from CECT were exported from an Eclipse treatment planning system (Varian Medical Systems Inc, Palo alto, CA). Contouring had been performed by experienced clinicians. The resolution of raw images ranged

from 0.80 to 1.37 mm in-plane with 3 mm slice thickness. In order to extract texture features from the 3D volumes, the images were resampled to isotropic voxel sizes of 1*1*1 mm to obtain rotationally invariance and also the consistency across different patients. Trilinear interpolation algorithm was used for the resampling. Gray level quantization is required for the calculation of texture features (tractability). We applied Lloyd-Max quantization. Lloyd-Max quantization is a method that tries to find a quantizer that minimizes the mean squared error (MSE) of original and new images, which can conserve most information in the images while discretizing.

5.2.3 Neural Network based Survival Model Construction

Models were trained separately using clinical, radiomics and imaging data. Then, the individual models were fused and evaluated. For comparison, Cox regression [19] was examined for survival analyses. These models were trained and evaluated by strictly splitting data into train, validation and test sets with a 10 times 5-fold scheme. The metric used was the Harrel's c-index [20], Kaplan-Meier plot for high and low risk groups for survival. The risk groups were determined by a criterion using the median value of the outputs from the survival model. Confidence intervals were calculated by bias-corrected and accelerated (BCa) bootstrap interval algorithm [21].

5.2.4 Patch-based Variational Autoencoder Survival Joint Model for Radiomics and Clinical Features

For feature selection, algorithms such as Relief-F [22], SVM-RFE [23], mRMR [24], etc., are available. However, these methods might cause selection bias and lead to over-optimistic results if the data is not split correctly. In addition, it is more tedious with two steps in the analysis - feature selection and subsequent model building. In comparison, the VAE-SurvNet methods automatically learns a latent space to represent the important signals and train the survival model in one step efficiently.

Kingma *et al.* [25] introduced the Variational Autoencoders (VAEs) that resemble the naive AEs and variational Bayesian methods. Instead of learning a function that represents the data, variational autoencoders are able to learn a probability distribution from the data. The short coming of a pure VAE for a classification problem is that it is unsupervised and the features obtained from the latent space might be irrelevant to the endpoint of interest. Thus, a supervised joint training network was designed that contains a classification part, which takes the latent space features as an input, goes through a fully-connected layer and outputs the risk probability. By this technique, the latent features learned by VAE is more specific to the desired task.

Specifically, the VAE consists of an encoder which takes the input and converts it into two latent vectors (a vector of means, μ, and a vector of standard deviations, σ) that parameterize a Gaussian distribution and a decoder that reconstructs a latent space sample z back to the original space. The loss function of the VAE model is defined by two parts: a reconstruction loss that measures how

similar is the output comparing with the input and a regularization loss determined by Kullback-Liebler divergence (KL divergence), that measures how closely the latent variables match Gaussian distributions.

Considering the i^{th} input sample x_i, the output from the encoder is a hidden representation z, which has weights and biases θ . The encoder can be denoted as $q_\theta(z|x)$. For the decoder network, a value z is denoted as input, and a reconstructed output $x^{(*)}$ is generated from some conditional distribution $p_\varphi(x|z)$, which represents the decoder network. Thus, the loss function can be expressed as follows:

$$l_i(\theta, \varphi) = -E_{z\, q_\theta(z|x_i)}[log p_\varphi(x_i|z)] + KL(q_\theta(z|x_i)||p(z)), \tag{5.1}$$

where l_i is the loss for a single data point. The first term is the reconstruction loss, which encourages the network to reconstruct the input data; the second term is the KL divergence between the encoder's distribution and the prior distribution $p(z)$, which measures how much information is lost during the compression. This term also serves as a regularizer that prevents the network from simply copying the input and leading to overfitting.

The VAE architecture was determined by minimizing the validation loss with respect to different networks (latent space dimension and layer number). To determine the network structure, first, the classification part was ignored and only the VAE part was tuned. The number of layers and nodes in the layers were grid-searched based on the loss function. Once the VAE part was fixed, the whole network (including the survival part) was jointly-trained by optimizing the total loss function that consists of VAE and survival loss. A key point here is the ratio between VAE and

survival losses, which regulates the supervised and unsupervised portions. This hyper-parameter was tuned on the training set.

Radiomics and clinical features are 1D vectors with 56 and 37 variables for each sample, while the CT image input is 3D matrix, which was resized to (224,224,48) to be fed into the CNN network. Another important technique we used is the patch-based training, which can augment the data and improve the performance. The random crop dimension we used is (80,80,40), which was determined by experiments.

5.2.5 Neural Network based Survival Analysis

Cox proportional hazard model (CPH) is the most commonly used survival analysis method to explore the relationships between patients' covariates and the survival time. It assumes that the log-risk of failure is a linear combination of the covariates. The hazard function is represented as the formula below:

$$\lambda(t|x) = \lambda_0(t)\exp\big(h(x)\big), h(x) = \beta^T x, \tag{5.2}$$

$h(x)$ is a linear function of variables x. The weights β are tuned by optimizing the Cox partial likelihood, which is the product of the probability at each event time T_i that the event has occurred to subject i, given the subject is still at risk at time T_i, as shown below:

$$L_c(\beta) = \prod_{i:E_i=1} \frac{\hat{r}_\beta(x_i)}{\sum_{j\in R(T_i)} \hat{r}_\beta(x_j)} = \prod_{i:E_i=1} \frac{\exp(\hat{h}_\beta(x_i))}{\sum_{j\in R(T_i)} \exp(\hat{h}_\beta(x_j))}, \tag{5.3}$$

T_i is the duration, E_i is the event indicator, and x_i is the input feature for subject i. The risk set $R(T_i) = (i: T_i \geq t)$ is the set of patients that are still at risk at time t.

However, this assumption might be too simplistic for complex relationships. To model nonlinearity of the features and the risk of failure, NNs are used in Katzman *et al.*'s work, called the DeepSurv [26]. Instead of using $h(x)$ as shown in Eqn. (5.1), NN was used to estimate the log-risk function, with the output giving $\hat{h}_\beta(x_i)$, where β are the NN parameters. Similarly, the objective function of the NN is still the Cox partial likelihood:

$$l(\beta) = -\frac{1}{N_{E=1}}\left\{\log\left(\prod_{i:E_i=1}\frac{\exp(\hat{h}_\beta(x_i))}{\sum_{j\in R(T_i)}\exp(\hat{h}_\beta(x_j))}\right) + \lambda\cdot\|\beta\|_2^2 + \gamma\cdot\sum_{:E_i=1}\left|\hat{h}_\beta(x_i)\right|\right\}, \tag{5.4}$$

$N_{E=1}$ is the number of patients that are not censored and contribute to the log-likelihood loss calculation. Last two terms are penalty that aim to regularize the loss function, with the first term being L2 penalty and third term being a penalty for the prediction to restrain its value not to deviate too much and cause overflow during training.

For the modeling of $\hat{h}_\beta(x_i)$, the original work used pure MLP, while in our study, VAE architecture was applied with two advantages, (1) the latent features could be obtained; (2) the partial likelihood in Eqn.(5.3) for survival and the VAE loss function (KL divergence and reconstruction binary cross entropy loss) jointly training makes the generated model more robust. The total loss function in the NN is thus:

$$l_{\text{total}} = l_{\text{vae}} + \tau\cdot l_{\text{Cox}}, \tag{5.5}$$

τ is a weight that balances the two parts of losses, which is tuned on the training set.

5.3 Results

The univariate analysis of clinical and radiomics for the IRS and OS endpoints are shown in table 5.1 and 5.2 (miRNA were not included in the table for simplicity). There are 13 significant clinical variables: Number of Active Liver Lesions at Time of Treatment (c-index 0.557, p-value 0.009), Total Number Fractions (c-index 0.558, p-value 0.009), Pre-RT ICGR15 (c-index 0.568, p-value 0.003), Na (pre-treatment) (c-index 0.620, p-value 0.005), Albumin (g/DL) (c-index 0.636, p-value 0.000), Total bilirubin (mg/dL) (c-index 0.601, p-value 0.000), MELD (c-index 0.574, p-value 0.031), MELD-Na (c-index 0.596, p-value 0.009), Child-Pugh (c-index 0.626, p-value 0.000), ALBI Raw Score (c-index 0.624, p-value 0.000), Alkphos CTCAE Liver Toxicity Grade (c-index 0.549, p-value 0.004), treated (c-index 0.556, p-value 0.016), platelet (c-index 0.535, p-value 0.028). There are 6 significant radiomics features: correlation_gtv_32 (c-index 0.547, p-value 0.033), ZSN_liver_gtv_8 (c-index 0.586, p-value 0.008), GLN_liver_gtv_16 (c-index 0.585, p-value 0.014), ZSN_liver_gtv_16 (c-index 0.596, p-value 0.015), ZSN_liver_gtv_32 (c-index 0.578, p-value 0.047) and SZHGE_liver_gtv_64 (c-index 0.578, p-value 0.047). There are 3 significant miRNA features: hsa-let-7i-5p (c-index 0.689, p-value 0.021), hsa-miR-10b-5p (c-index 0.621, p-value 0.013) and hsa-miR-660-5p (c-index 0.652, p-value 0.047.

Due to the small size of patients that having miRNA data, we didn't train models for genetic information. Instead, the correlation of miRNA and other significant features were investigated. There are 3 miRNA + 13 clinical + 6 radiomics features, in total 22 features. hsa-let-7i-5p is significantly correlated with GLN_liver_gtv_16 (-0.5, p-value:0.023), hsa-miR-10b-5p is

significantly correlated with total bilirubin (-0.4, p-value:0.035). In general, it was found that the three miRNAs are significantly correlated with GLN, SZHGE, ZSN with different gray levels.

For multivariate analysis, the radiomics, clinical and CT raw image individual models results are summarized in table 5.3. The average c-indexes for test sets are 0.554 (0.531-0.577), 0.599 (0.581-0.617) and 0.546 (0.519-0.573) for radiomics, clinical and combined models, respectively using Cox models. The average c-indexes for test sets are 0.579 (0.544-0.621), 0.629 (0.601-0.643), 0.581 (0.553-0.613) and 0.650 (0.635-0.683) for radiomics, clinical, CT image input and combined models, respectively using neural networks. The Cox model cannot handle image input. The combined models for NN outperformed the clinical models alone, which indicates the value of complementary information that imaging can provide. The VAE and CNN network architectures are shown in Fig. 5.2. We used random crop to augment the CT image input network. Different strategies were applied, such as transfer learning, it turns out the performance were all pretty similar. Thus, we used the basic CNN structure for the CT image data. In order to show the effectiveness of the CT image NN model, random image data were fed into the same CNN-SurvNet. The c-index was random results. Example patient input and random noise data are shown in Fig. 5.3. Notice that the architecture presented here might not be the optimal structure, however, based on our experiments, the performance is not sensitive to the structure, and the goal of this work is to show the concept that VAE-SurvNet model possess predictive power, and not to find the optimal solution.

Table 5.1 Univariate Analysis for clinical variables

Clinical Variables	c-index for OS	p-value for OS
Sex	0.484	0.811
Age	0.526	0.589
Pre-Tx Cirrhosis (0=no, 1=yes)	0.530	0.059
Portal Vein Thrombosis (0=no, 1=yes)	0.536	0.105
Number of Active Liver Lesions at Time of Treatment	**0.557**	**0.009**
Total Number Fractions	**0.558**	**0.009**
Total EQD2	0.584	0.054
Tx Break? 0=no, 1=yes	0.523	0.772
Time Btwn First and Final Fractions (Days)	0.533	0.180
Pre-RT ICGR15	**0.568**	**0.003**
Tumor_Volume	0.498	0.707
LIVER-GTV Volume (cc)	0.502	0.928
LIVER-GTV Mean Dose (Gy) LQ: α/β=2.5	0.481	0.800
Treatment-related complication (0=no, 1=yes)	0.485	0.646
ECOG PS (pre-treatment)	0.506	0.854
Na (pre-treatment)	**0.620**	**0.005**
Creatinine (pre-treatment; mg/dL)	0.530	0.621
Albumin (pre-treatment; g/DL)	**0.636**	**0.000**

ALT (pre-treatment; IU/L)	0.500	0.656
Alkphos (pre-treatment; IU/L)	0.547	0.269
Total bilirubin (pre-treatment; mg/dL)	**0.601**	**0.000**
Protime with INR (pre-treatment; s)	0.605	0.131
AFP (pre-treatment, 0=<2.0)	0.574	0.989
MELD (baseline)	**0.574**	**0.031**
MELD-Na (baseline)	**0.596**	**0.009**
Child-Pugh (baseline)	**0.626**	**0.000**
Barcelona score (HCC ONLY) 0=0, A=1, B=2, C=3, D=4	0.534	0.473
ALBI Raw Score (Baseline)	**0.624**	**0.000**
ALBI Raw Score (Baseline)	0.521	0.291
AST CTCAE Liver Toxicity Grade (pre-tx)	0.515	0.459
ALT CTCAE Liver Toxicity Grade (pre-tx)	0.504	0.729
Alkphos CTCAE Liver Toxicity Grade (pre-tx)	**0.549**	**0.004**
Total bilirubin CTCAE Liver Toxicity Grade (pre-tx)	0.504	0.794
Treated	**0.556**	**0.016**
PLATELET_pre	**0.535**	**0.028**
HEMATOCRIT_pre	0.574	0.074
ABSLYMPH_pre	0.489	0.379

ABSNEUT_pre	0.484	0.811

Table 5.2 Univariate Analysis for radiomics variables

Regions	Gray levels	Radiomics Features	c-index for OS	p-value for OS
GTV	8	Correlation	0.545	0.056
		GLN	0.539	0.969
		HGRE	0.525	0.275
		SZE	0.540	0.840
		GLN	0.566	0.660
		ZSN	0.581	0.911
		SZHGE	0.504	0.239
	16	Correlation	0.542	0.056
		GLN	0.538	0.983
		HGRE	0.517	0.606
		SZE	0.506	0.222
		GLN	0.560	0.686
		ZSN	0.571	0.925
		SZHGE	0.520	0.813
	32	Correlation	**0.547**	**0.033**
		GLN	0.534	0.987

		HGRE	0.509	1.000
		SZE	0.500	0.310
		GLN	0.566	0.559
		ZSN	0.564	0.913
		SZHGE	0.499	0.877
	64	Correlation	0.542	0.055
		GLN	0.468	0.998
		HGRE	0.504	0.766
		SZE	0.493	0.429
		GLN	0.560	0.595
		ZSN	0.559	0.706
		SZHGE	0.517	0.499
Liver-GTV	8	Correlation	0.542	0.078
		GLN	0.544	0.106
		HGRE	0.521	0.663
		SZE	0.530	0.911
		GLN	0.524	0.680
		ZSN	**0.586**	**0.008**
		SZHGE	0.511	0.325
	16	Correlation	0.523	0.148

	GLN	0.533	0.221
	HGRE	0.508	0.896
	SZE	0.527	0.246
	GLN	**0.585**	**0.014**
	ZSN	**0.596**	**0.015**
	SZHGE	0.489	0.431
32	Correlation	0.537	0.153
	GLN	0.539	0.252
	HGRE	0.512	0.990
	SZE	0.544	0.149
	GLN	0.578	0.196
	ZSN	**0.578**	**0.047**
	SZHGE	0.511	0.314
64	Correlation	0.537	0.096
	GLN	0.536	0.116
	HGRE	0.511	0.917
	SZE	0.540	0.180
	GLN	0.567	0.319
	ZSN	0.551	0.178
	SZHGE	**0.532**	**0.031**

Table 5.3 C-indexes for radiomics, clinical, raw image CNN and combined models.

	Radiomics	Clinical	Image	Combined
Cox	0.554 (0.531-0.577)	0.599 (0.581-0.617)	NA	0.546 (0.519-0.573)
NN	0.579 (0.544-0.621)	0.629 (0.601-0.643)	0.581 (0.553-0.613)	0.650 (0.635-0.683)

Fig. 5.2 VAE-SurvNet and CNN-SurvNet structure.

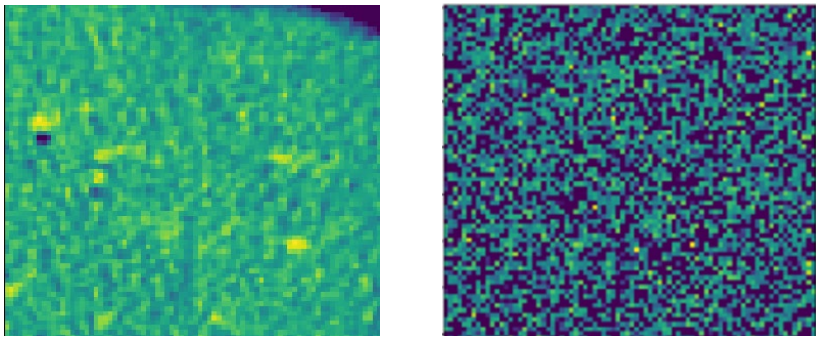


Fig. 5.3 Example slice of cropped patient CT image (left) and random input (right).

5.4 Discussion

Due to the challenges associated with the heterogeneity of livers among different patients and the complicated etiologic factors associated with HCC, limited work has been done for the HCC prognosis analysis. Cozzi *et al.* [27] conducted a retrospective study of 138 HCC patients treated with VMAT for the prediction of overall survival and local control. They applied univariate and logistic regression for clinical response and Cox regression hazards model for survival analysis on clinical and radiomic features, which showed significant prediction performance. However, the features were extracted from non-contrast-enhanced images, which usually suffer from poor image quality, especially for liver with disease. Zhou *et al.* [14] developed a CT-based radiomics signature for preoperatively predicting the early recurrence of HCC using the LASSO algorithm. They built a radiomic and clinical combined model with AUC of 0.836 for the prediction of early recurrence. It focused on patients who underwent hepatectomy and didn't consider survival analysis. Kiryu *et al.* [28] investigated the relationship of texture features with filtration at different filter levels and the prognosis of HCC 5-year overall survival and disease-free survival using

preoperative non-contrast enhanced CT images. They showed the KM curves for OS and DFS were significantly different between patient groups dichotomized by cut-off values for all CT texture features. Bakr *et al.* [29] explored noninvasive biomarkers of microvascular invasion in patients with HCC (28 patients) using quantitative image features extracted from contrast-enhanced CT. Chaudhary *et al.* [30] conducted a deep learning study using multi-omics features to identify survival subgroups of HCC. The model provides two subgroups with significant survival differences and model fit of c-index 0.68.

Compared to the studies above, this study assessed the prediction potential of radiomic features extracted from contrast-enhanced CT pre-treatment images, the original images, and pre-treatment clinical factors for risk assessment of intrahepatic progression of HCC in the liver elsewhere from the primary tumor site(s) and overall survival using neural networks. The radiomics prediction models showed modest performance in our experiments. The possible reasons are: (1) we used a relatively strict validation framework that adopted repeated nested cross-validation; (2) survival analysis is not uncommon to have lower results than classification; (3) Overall survival is a complex target to predict. CT images might not have sufficient predictive power; (4) the data size is too small to learn the underlying mechanism; (5) the absolute value might fluctuate for different datasets. Nonetheless, the contribution of this work comes in three ways: (1) It showed the complementary information from images that could help the clinical factors; (2) We proposed a novel VAE-Survnet that could combine multi-omics features including raw images, which outperformed the traditional Cox modeling. Another interesting result is the significant correlation between miRNA and radiomics features. Various studies have suggested that heterogeneity of

tumors is associated with genomic heterogeneity and tumoral microenvironment, thus plays an important role in the cancer prognosis [31-33], which is found in our work as well.

Although this work is able to provide preliminary guidance for the treatment planning based on the pre-treatment data, future work on adaptation of treatment plans (e.g., dose distribution) that customize better to the patient need to be investigated. Though we have conducted strict cross-validation to evaluate the performance, these identified biomarkers and clinical factors warrant further validation in large external and multi-institutional prospective studies to be applied to personalized treatment planning for HCC patients.

5.5 Conclusion

A new graph-based feature selection method was developed that enables efficient data reduction of large-scale radiomics analysis of liver cancer imaging data. Robust survival models were built based on supervised learning of imaging and clinical features for risk assessment of HCC progression elsewhere in the liver and overall prognosis. Texture features, VAE features in combination with clinical factors showed promise for HCC recurrence-free and overall survival predictions, which can be used to personalize liver cancer treatment.

5.6 References

1. Jemal A, Siegel R, Ward E, Hao Y, Xu J, Murray T, et al. Cancer statistics, 2008. CA Cancer J Clin. 2008;58:71-96.
2. Llovet JM, Burroughs A, Bruix J. Hepatocellular carcinoma. Lancet. 2003;362:1907-17.
3. Kwon JH, Bae SH, Kim JY, Choi BO, Jang HS, Jang JW, et al. Long-term effect of stereotactic body radiation therapy for primary hepatocellular carcinoma ineligible for local ablation therapy or surgical resection. Stereotactic radiotherapy for liver cancer. BMC Cancer. 2010;10:475.
4. Raza A, Sood GK. Hepatocellular carcinoma review: current treatment, and evidence-based medicine. World journal of gastroenterology: WJG. 2014;20:4115.
5. Wahl DR, Stenmark MH, Tao Y, Pollom EL, Caoili EM, Lawrence TS, et al. Outcomes after stereotactic body radiotherapy or radiofrequency ablation for hepatocellular carcinoma. J Clin Oncol. 2016;34:452.
6. Schaub SK, Hartvigson PE, Lock MI, Høyer M, Brunner TB, Cardenes HR, et al. Stereotactic body radiation therapy for hepatocellular carcinoma: current trends and controversies. Technol Cancer Res Treat. 2018;17:1533033818790217.
7. Feng M, Suresh K, Schipper MJ, Bazzi L, Ben-Josef E, Matuszak MM, et al. Individualized adaptive stereotactic body radiotherapy for liver tumors in patients at high risk for liver damage: a phase 2 clinical trial. JAMA oncology. 2018;4:40-7.
8. Ohri N, Tomé WA, Romero AM, Miften M, Ten Haken RK, Dawson LA, et al. Local control after stereotactic body radiation therapy for liver tumors. International Journal of Radiation Oncology* Biology* Physics. 2018.
9. Wang C-y, Li S. Clinical characteristics and prognosis of 2887 patients with hepatocellular carcinoma: A single center 14 years experience from China. Medicine. 2019;98.
10. Gillies RJ, Kinahan PE, Hricak H. Radiomics: images are more than pictures, they are data. Radiology. 2015;278:563-77.
11. Vallières M, Freeman CR, Skamene SR, El Naqa I. A radiomics model from joint FDG-PET and MRI texture features for the prediction of lung metastases in soft-tissue sarcomas of the extremities. Phys Med Biol. 2015;60:5471.
12. Tseng H-H, Wei L, Cui S, Luo Y, Ten Haken RK, El Naqa I. Machine learning and imaging informatics in oncology. Oncology. 2018:1-19.
13. Antropova N, Huynh BQ, Giger ML. A deep feature fusion methodology for breast cancer diagnosis demonstrated on three imaging modality datasets. Medical physics. 2017;44:5162-71.
14. Zhou Y, He L, Huang Y, Chen S, Wu P, Ye W, et al. CT-based radiomics signature: a potential biomarker for preoperative prediction of early recurrence in hepatocellular carcinoma. Abdominal Radiology. 2017;42:1695-704.
15. Shan Q-y, Hu H-t, Feng S-t, Peng Z-p, Chen S-l, Zhou Q, et al. CT-based peritumoral radiomics signatures to predict early recurrence in hepatocellular carcinoma after curative tumor resection or ablation. Cancer Imaging. 2019;19:11.

16. Ji G-W, Zhu F-P, Xu Q, Wang K, Wu M-Y, Tang W-W, et al. Radiomic Features at Contrast-enhanced CT Predict Recurrence in Early Stage Hepatocellular Carcinoma: A Multi-Institutional Study. Radiology. 2020:191470.
17. Peng J, Qi X, Zhang Q, Duan Z, Xu Y, Zhang J, et al. A radiomics nomogram for preoperatively predicting prognosis of patients in hepatocellular carcinoma. Translational Cancer Research. 2018;7:936-46.
18. Guo D, Gu D, Wang H, Wei J, Wang Z, Hao X, et al. Radiomics analysis enables recurrence prediction for hepatocellular carcinoma after liver transplantation. Eur J Radiol. 2019;117:33-40.
19. Cox DR. Regression models and life‐tables. Journal of the Royal Statistical Society: Series B (Methodological). 1972;34:187-202.
20. Harrell FE, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. Jama. 1982;247:2543-6.
21. Efron B. Better bootstrap confidence intervals. Journal of the American statistical Association. 1987;82:171-85.
22. Kira K, Rendell LA. A practical approach to feature selection. Machine Learning Proceedings 1992: Elsevier; 1992. p. 249-56.
23. Guyon I, Weston J, Barnhill S, Vapnik V. Gene Selection for Cancer Classification using Support Vector Machines. Machine Learning. 2002;46:389-422. doi:10.1023/a:1012487302797.
24. Ding C, Peng H. Minimum redundancy feature selection from microarray gene expression data. Journal of bioinformatics and computational biology. 2005;3:185-205.
25. Kingma DP, Welling M. Auto-encoding variational bayes. arXiv preprint arXiv:13126114. 2013.
26. Katzman JL, Shaham U, Cloninger A, Bates J, Jiang T, Kluger Y. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. BMC Med Res Methodol. 2018;18:24.
27. Cozzi L, Dinapoli N, Fogliata A, Hsu W-C, Reggiori G, Lobefalo F, et al. Radiomics based analysis to predict local control and survival in hepatocellular carcinoma patients treated with volumetric modulated arc therapy. BMC Cancer. 2017;17:829. doi:10.1186/s12885-017-3847-7.
28. Kiryu S, Akai H, Nojima M, Hasegawa K, Shinkawa H, Kokudo N, et al. Impact of hepatocellular carcinoma heterogeneity on computed tomography as a prognostic indicator. Sci Rep. 2017;7:12689. doi:10.1038/s41598-017-12688-7.
29. Bakr SH, Echegaray S, Shah RP, Kamaya A, Louie J, Napel S, et al. Noninvasive radiomics signature based on quantitative analysis of computed tomography images as a surrogate for microvascular invasion in hepatocellular carcinoma: a pilot study. Journal of Medical Imaging. 2017;4:041303.
30. Chaudhary K, Poirion OB, Lu L, Garmire LX. Deep learning–based multi-omics integration robustly predicts survival in liver cancer. Clinical Cancer Research. 2018;24:1248-59.
31. Aerts HJ, Velazquez ER, Leijenaar RT, Parmar C, Grossmann P, Carvalho S, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nature communications. 2014;5:4006.

32. Segal E, Sirlin CB, Ooi C, Adler AS, Gollub J, Chen X, et al. Decoding global gene expression programs in liver cancer by noninvasive imaging. Nature biotechnology. 2007;25:675.
33. Vallières M, Kay-Rivest E, Perrin LJ, Liem X, Furstoss C, Aerts HJ, et al. Radiomics strategies for risk assessment of tumour failure in head-and-neck cancer. Sci Rep. 2017;7:10117.

CHAPTER 6

Multimodality Approach using Deep Attention Convolutional Neural Networks for Intrahepatic Recurrence Localization of Liver Cancer Post-SBRT

This chapter developed a new radiomics algorithm for localization of intrahepatic failure and is based on the paper: **Wei, L.,** Owen, D., Mendiratta-Lala, M., Rosen, B., Cuneo, K., Lawrence, T. S., Ten Haken, R. K., El Naqa, I., " Multimodality Approach using Deep Attention Convolutional Neural Networks for Intrahepatic Recurrence Localization of Liver Cancer Post-SBRT." (2020), *under processing.*

6.1 Introduction

Hepatocellular carcinoma (HCC) is the most common type of primary liver cancer and the fourth leading cause of cancer-related death worldwide [1]. Surgery either a partial resection or a liver transplantation remains the standard treatment for curative purpose [2]. However, only 15% to 30% are candidates due to the location of the tumor or underlying liver dysfunction [3, 4]. There are several non-surgical liver-directed therapies including radiofrequency ablation (RFA), microwave ablation (MWA), trans-arterial chemoembolization (TACE) and, more recently,

stereotactic body radiotherapy (SBRT). Previously, RT was used cautiously due to the narrow window when trading off between tumor control and radiation-induced liver disease (RILD) [4]. Recently, with the development of advanced radiotherapy delivery technologies, more precise partial liver irradiation using SBRT enables highly conformable dose distributions with a rapid dose drop off for HCCs that are not suitable for resection [4, 5].

Although over 90% of tumors will be controlled by SBRT [5, 6], intrahepatic progression within the liver, remote from the treatment zone is common, with failure rates reported at 50% [3, 7]. Deep learning has seen dramatic development recently, which enables computers to automatically capture complicated patterns in the datasets. In particular, convolutional neural networks (CNNs), a type of neural network, have surpassed human performance in computer vision on the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) [8]. Enormous amounts of multimodality imaging data containing valuable signals and information has been generated during cancer patient heath care. Deep learning has been widely applied in medical image analysis, ranging from classification [9, 10], image registration/reconstruction/synthesis/segmentation [11-14], survival analysis [15]. In this study, we take advantage of the advanced deep learning algorithms to predict the recurrent HCC tumor Couinaud segment using multimodality pre-SBRT contrast enhanced CT and T1 weighted MR images as well as the 3D dose distribution. The novelty of our work can be summarized as: (1) using neural networks (VoxelMorph) to automatically obtain liver Couinaud segment masks; (2) development of novel Attention Gating U-Net model to predict intrahepatic recurrence location of tumors; and (3) investigations of contributions for multimodality images and the correlation to treatment dose.

6.2 Methods and Materials

6.2.1 Patient cohort

After IRB approval, a HIPAA compliant retrospective analysis of HCC patients treated with SBRT was performed. A total of 303 HCC patients treated with SBRT were reviewed. Patients without: (1) T1-weighted MR images; (2) contrast-enhanced CT (CECT) images; (3) gross tumor or liver contour in the database were excluded from the analysis. A total of 102 HCC patients met the inclusion criteria. The endpoint of intrahepatic recurrence-free survival was selected whether or not patients with HCC developed intrahepatic tumors after SBRT. This was defined by the presence of new tumor outside the planning target volume (PTV) of the previously treated tumor(s). The ground truths were determined by experienced clinicians based on CECTs, MRI, and relevant clinical records for each patient, indicating the recurrence(s) is (are) located in which segment(s). Liver segment is based on Couinaud classification, which divides the liver into eight functionally independent segments by the vascular branches, as shown in Fig. 6.1.

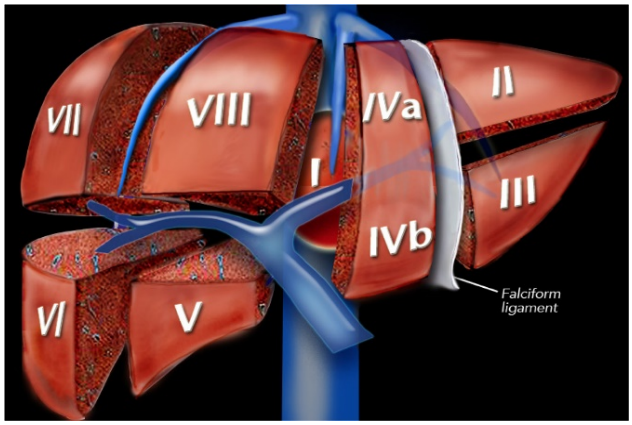


Fig. 6.1 Couinaud segment for liver.

6.2.2 Images Acquisition and Processing

Arterial phase images and structure sets including liver, GTV, and liver-GTV from CECT were exported from an Eclipse treatment planning system (Varian Medical Systems Inc, Palo alto, CA). Contouring of the CT images had been performed by experienced clinicians. The resolution of raw images ranged from 0.80 to 1.37 mm in-plane with 3 mm slice thickness. 3D dose distributions for SBRT treatment were exported from Eclipse, which were co-registered to the CT images. For patients who didn't have MR images in Eclipse, MR images were exported from McKesson and imported into Eclipse and registered to the CT images using the built-in function of the software.

6.2.3 Obtaining Couinaud Segments by Unsupervised Deformable Image Registration

The recurrence ground truth is segment-wise binary vector of length 9 (the eight Couinaud segment), with 0 being no recurrence for the segment and 1 being recurrence. In order to predict the recurrence of HCC tumors for each segment, it is necessary to obtain the Couinaud segments for each liver. However, since Couinaud segments are based on vascular branches, the variability across different patients is large. It is very time-consuming and subjective to delineate across different people to manually obtain these segments. Thus, we proposed a neural network based unsupervised deformable registration method to obtain these segments.

First, a liver atlas with manually obtained Couniaud segments was acquired from 3D slicers (http://www.slicer.org) [16]. The deformable registration was then utilized to register the CT altas to each patient liver. The Couinaud segments were then transferred to patient liver CT images. We

applied VoxelMorph framework for the deformable, pairwise image registration. The parameterized registration function was learnt using a convolutional neural network (CNN), with input being the atlas CT image and the patient liver CT and output being atlas CT registered to patient CT. The objective function is similar to traditional registration algorithms, which tries to minimize the dissimilarity between the intensities of the two input images and penalize the spatial variation of the deformation as well. Let m, f denote the two input images (moving and fixed). First, all the images were resampled to 224*224*48 before being fed to the network. The registration function can be denoted as $g_\theta(f, m) = u$, where θ is the CNN parameters. Thus, the mapping was formed by $\varphi = Id + u$, I is the identity transform. The loss function is shown as below:

$$Loss_{\text{total}} = Loss_{\text{similarity}}(f, m^{\circ}\varphi) + \lambda L_{\text{smooth}}(\varphi), \tag{6.1}$$

The cross-correlation (CC) is used for similarity loss between $f, m^{\circ}\varphi$, since it is more robust to intensity variations [17]. A higher CC value means a better alignment - $Loss_{\text{similarity}}(f, m^{\circ}\varphi) = -CC(f, m^{\circ}\varphi)$. The similarity loss alone might lead to non-smooth φ, so a smoothing term shown below was added:

$$L_{\text{smooth}}(\varphi) = \sum_{p} \|\nabla u(p)\|^2, \tag{6.2}$$

where p is the voxel of the input volumes.

Fig. 6.2 shows the overview of this method. Stochastic gradient descent (SGD) was used to find optimal parameters. Fig. 6.3 shows the network structure, which is built upon a Unet backbone

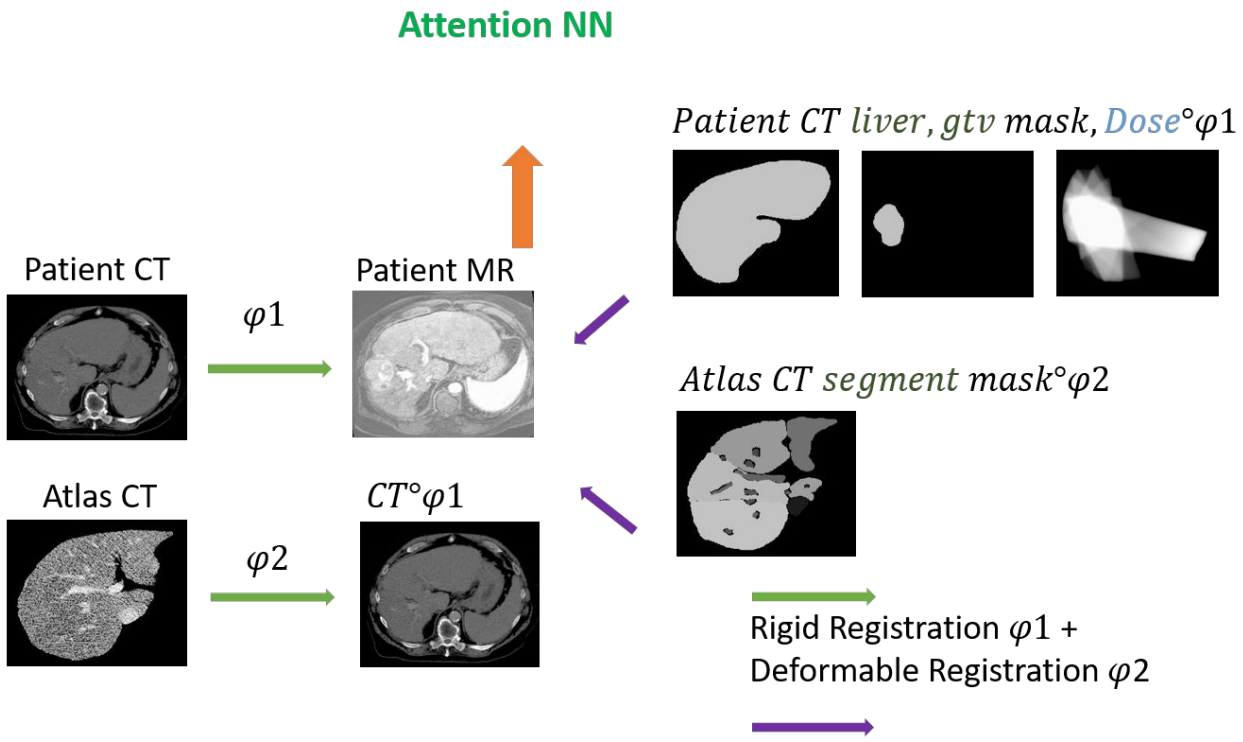


Fig. 6.2 Overview of the registration process.

with skip connections. Atlas and patient CT images were concatenated with the dimension of 224*224*48*2. The convolution kernel size is 3, and stride is 2 to reduce the spatial dimensions.

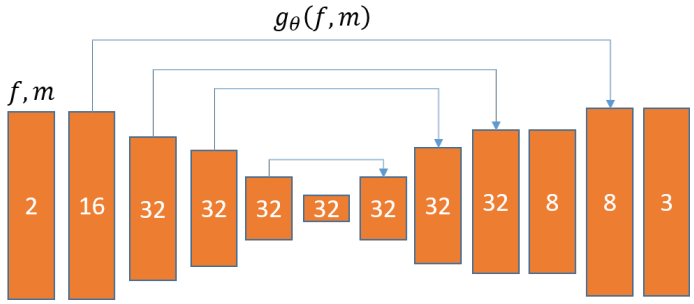


Fig. 6.3 VoxelMorph architecture.

6.2.4 Attention Neural Network for Recurrence Segment Prediction

An attention neural network was applied to predict the recurrence location of the HCC segment-wise for tumor information localization. Attention gates (AGs) are widely used in natural language processing (NLP) [18], image captioning [19], etc. In Oktay's work, a novel self-attention gating module was added to the U-Net framework to reduce the false-positive predictions for small objects that shows large shape variations [20]. Attention coefficients, $\alpha_i \in [0,1]$, can identify salient image regions and prune feature responses to preserve only the activations relevant to the specific task. AGs give the element-wise multiplication of input feature-maps and attention coefficients: $\hat{x}_{i,c}^l = x_{i,c}^l \cdot \alpha_i^l$, where *i, c* and *l* are the spatial, channel and layer index. A single scalar

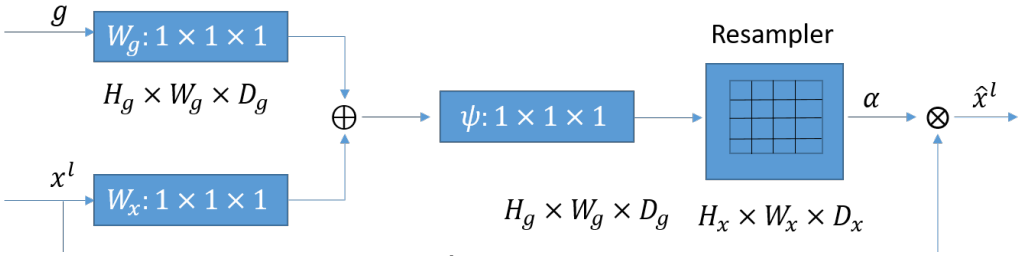


Fig. 6.4 Structure of the AGs with input x^l being scaled by α, which is learnt by both the coarser signal from g and the activations from x.

attention value is computed for each pixel vector with a length of c. The AG is formulated as below:

$$\hat{x}_i^l = \sigma_2(\psi^T \left(\sigma_1\left(W_x^T x_i^l + W_g^T g_i + b_g\right)\right) + b_\psi), \tag{6.3}$$

where σ_2 is the sigmoid function, $\sigma_1$is ReLU activation. AGs are parameterized by W_x^T, W_g^T, b_g, and b_ψ. The structure of the AG is shown in Fig. 6.5. The AGs were then incorporated into the conventional U-Net structure to highlight the salient regions, which is shown in Fig. 6.4. Deep-supervision was also used in the network to help the hidden feature-maps to be discriminative at each image scale [21]. The basic network architecture will be similar to the standard one used in the U-Net framework. The additional deep feedback is brought in by associating a companion local output with each hidden layer, which acts as a kind of feature regularization and results in faster convergence in practice. The U-Net architecture is shown in Fig. 6.5. The output is a 2 channel 3D probability map with the original size as input. One of output channel was multiplied with the Couinaud segment masks to obtain scores for each segment to indicate the risk of recurrence. Binary cross-entropy with logits was applied to the segment scores to calculate the loss and back propagate. The other channel was used to segment the original tumor by calculating the mean squared error loss. This is an auxiliary task that takes advantage of the prior knowledge (where the primary tumor(s) was(were)) and let the network (AGs) learn the critical regions and help with our main task of recurrence prediction.

The exponential growth in the use of more than one imaging modality for diagnostic, therapeutic and prognostic purposes in noninvasive and quantitative cancer studies has facilitated the

development of multimodality techniques. The intuition of multimodality imaging is that different types of images can provide complementary information and combining them may result in more complete characterization of the disease (e.g., a tumor). In this study, CECT, MR and 3D dose distributions were available and co-registered. Individual models for each modality were developed using the described AGs network. Concatenated 3-channel input model was also trained to see if these images could provide complimentary information to the recurrence prediction.

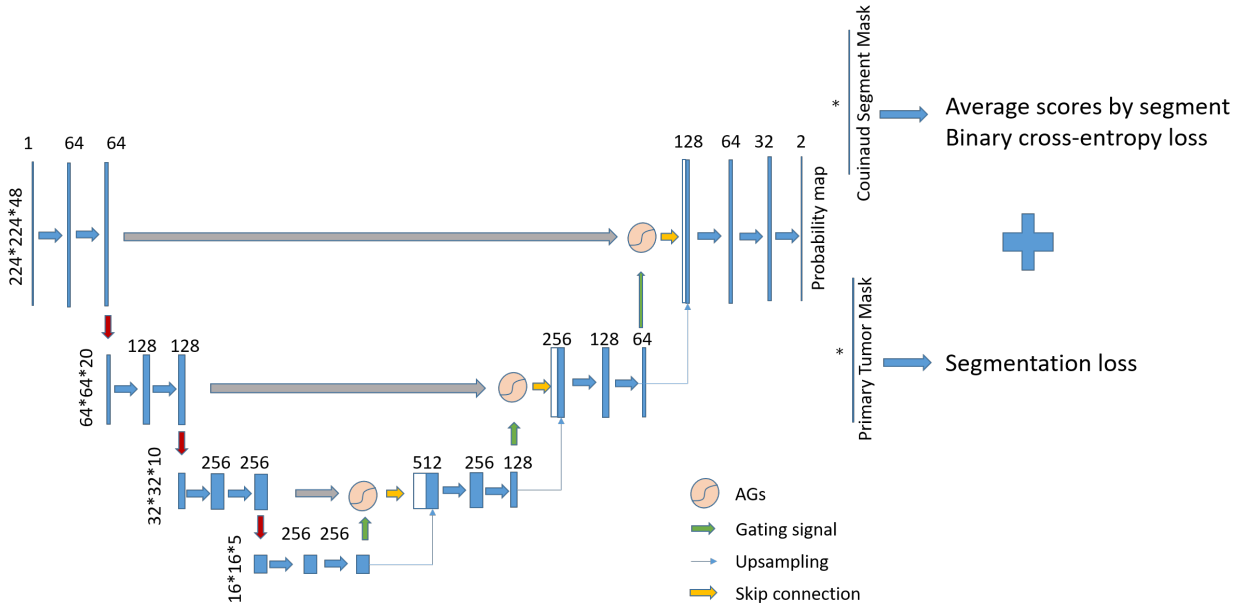


Fig. 6.5 Attention U-Net segmentation model. Attention gates (AGs) filter the features propagated through the skip connections.

Lastly, for outcome prediction, which requires much more data than end-to-end tasks, we used several image transformations to augment the data, including shifting with range [0.1,0.1], rotation within 15 degrees, scaling with range [0.7,1.3] and random flipping probability of 0.5. The learning rate is 1e-5, with l2 weight decay of 1e-6, and batch size 16.

6.3 Results

59 of the 102 patients developed intrahepatic recurrence elsewhere. There are 85 Couinaud segments that had recurred tumors out of the total 9*102=918 segments. The plot in Fig. 6.6 showed the distribution of the recurred tumors for each of the 9 segments.

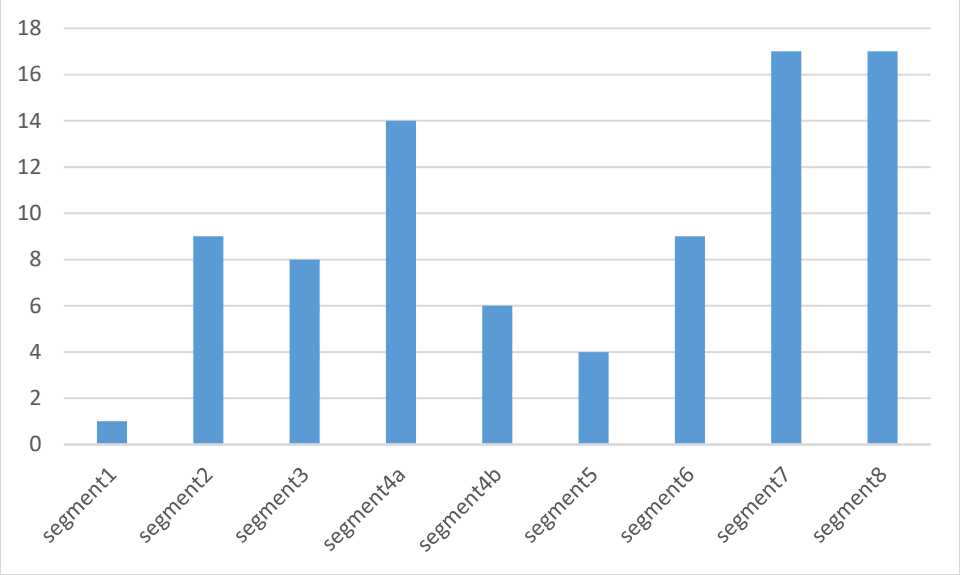


Fig. 6.6 Number of recurred cases in each segment.

Since there is no ground truth for the Couinaud segments, the Dice coefficients for deformable registration between atlas and patient CT images were calculated. The mean Dice similarity coefficient is 0.80 (std: 0.09), which is satisfying considering the large variation in liver anatomy for different patients. Example registration/segmentation result is shown in Fig. 6.7, from left to right are patient CT, atlas CT, moved atlas CT, segments, and transformation field. The total time used for registration and segmentation for 102 patients was less than 2 minutes, which is

significantly shorter than conventional deformable image registration methods. The loss function for VoxelMorph contains two parts: reconstruction and smoothing. The change of total, reconstruction and smoothing losses versus epoch numbers was shown in Fig. 6.8.

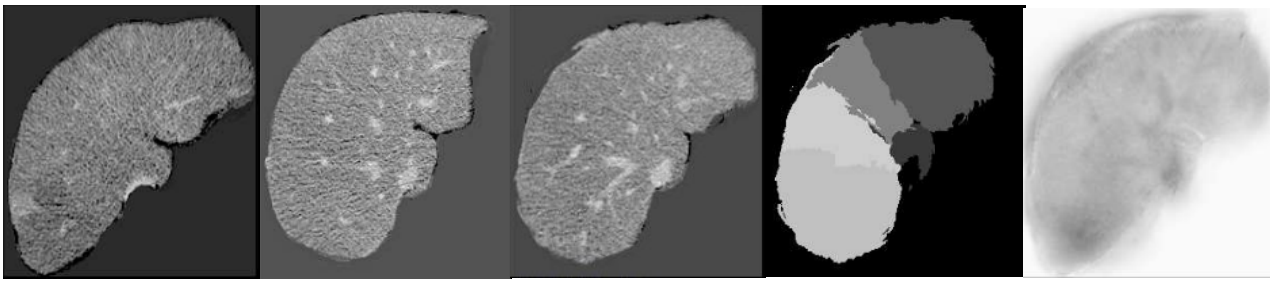


Fig. 6.7 Example registration results: from left to right – patient CT, atlas CT, moved atlas CT, segments, and transformation field.

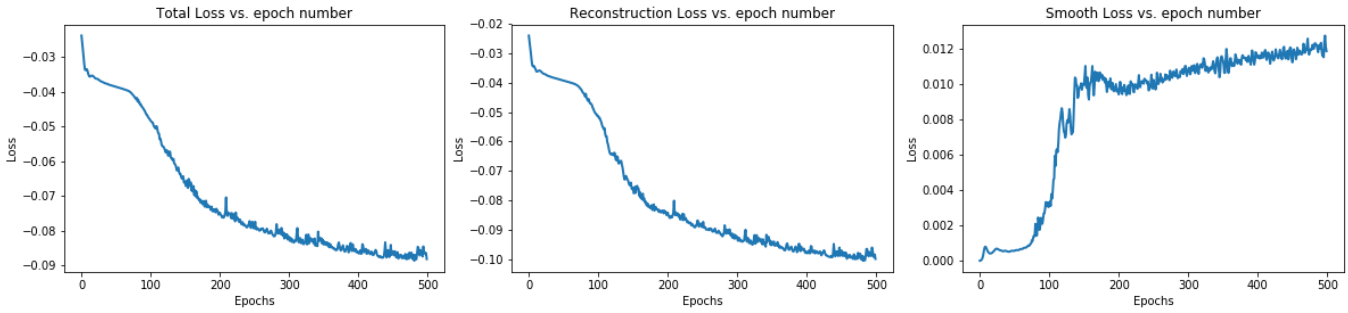


Fig. 6.8 Loss vs. epochs for VoxelMorph NN. From left to right: total, reconstruction, and smooth losses.

The individual models using CT, MR and dose were trained and the ROC curves are shown in shown in Fig. 6.9, which showed the predictive power for the ACNN model with AUC of 0.676 (95% CI: 0.538-0.814), 0.608 (95% CI: 0.500-0.740), 0.670 (95% CI: 0.541-0.799) and 0.686 (95% CI: 0.574-0.797) as computed by the Delong test for CT, MR, dose and combined models, respectively.

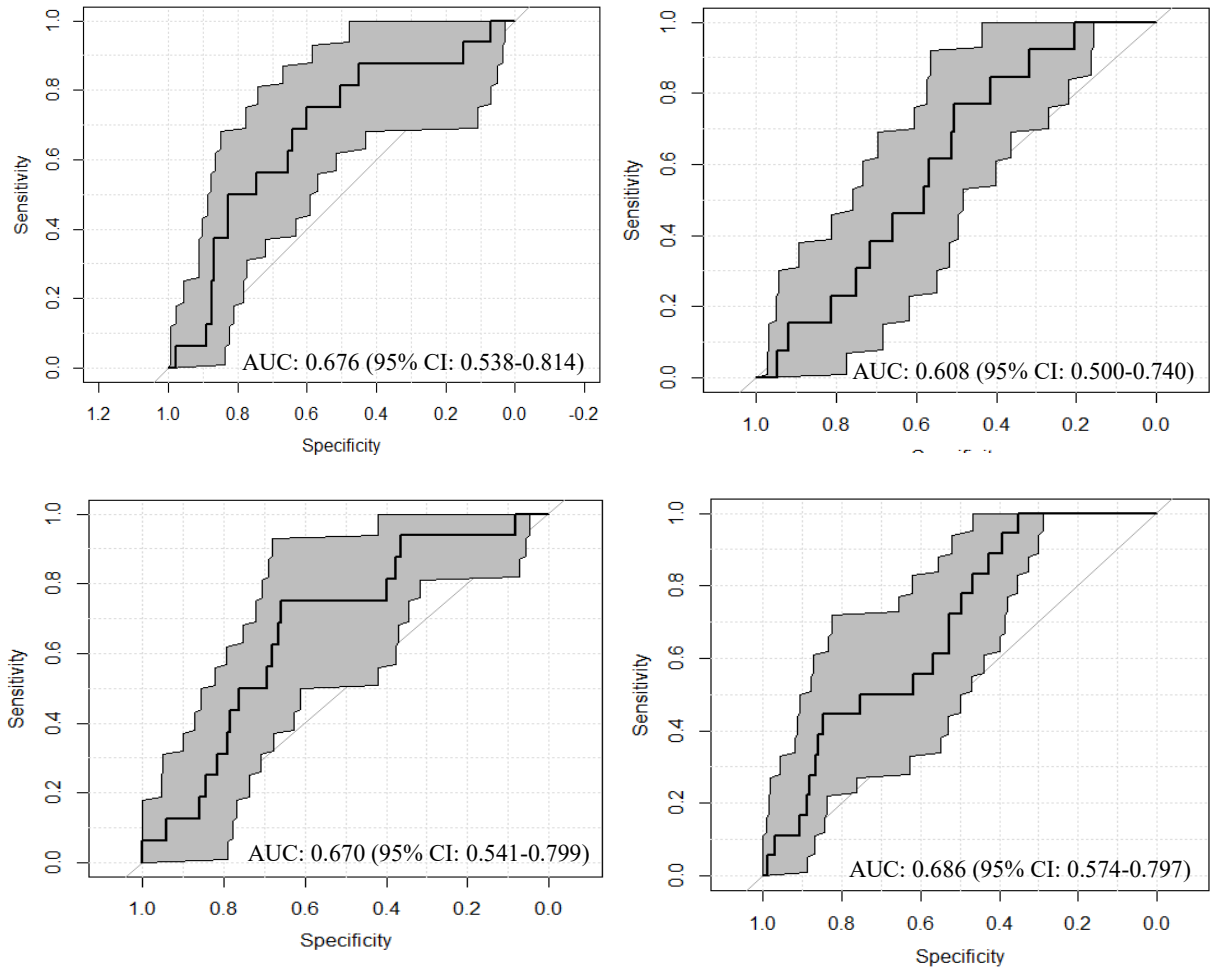


Fig. 6.9 ROC curves for CT, MR, dose and combined models.

6.4 Discussion

59 out of the 102 enrolled patients have developed intrahepatic recurrence, which is consistent with the high progression rates for post-treated HCC patients. Interestingly, the recurred location

based on this dataset showed high rates in segments 7, 8, and 4a, while relatively low rates for segments 1, 5 and 4b. Segment 1 is anatomically small in volume, which might contribute to the low rate. These results need to be further investigated in larger dataset to be confirmed later.

In terms of the Couinaud segment results, since there is no ground truth for this endpoint. We examined the Dice coefficients for fixed and transformed atlas image to help evaluate the segmentation. For the ACNN network, the CT-based model gave the largest AUC value for individual models, while the MR model gave the smallest AUC, which is contradict to our intuition since MR provide better soft tissue contrast. The possible reason might be that MR images need to be standardized with more care compared with CT images, which we will continue working on in the near future. Though the combined model performed better than the individual models, it is not significantly better. It is another work that we will continue, since now the three images were only concatenated and fed into the network. Better representation of these multimodality images should be investigated.

Though the prediction is significantly better than random guess, in order to apply this work to clinic practice, the interpretation is needed as well, which is next step in this work. Specifically, we are investigating gradient-based [22] and deconvolution-based [23] methods to obtain a salient map to help understand what critical regions in the images contribute to the endpoint. Other future work is to extend this work to the treatment planning system to avoid functional liver and boost dose to the high risk regions. In addition, based on visual checking of these images, we found there are some cases that the recurred tumor had appeared as low Lirad score lesions even before the

treatment. The model might capture these suspicious signals and make the prediction. This assumption needs to be investigated later.

6.5 Conclusion

Using the VoxelMorph network, we are able to register atlas images to patient CT images efficiently and accurately. Based on this work, obtaining Couinaud liver segment automatically becomes possible. Then, ACNN was trained to predict the recurrence location of the post SBRT HCC patients, which showed significant results which has the potential to be used in clinic to help realization of precision radiation therapy.

6.6 References

1. Akinyemiju T, Abera S, Ahmed M, Alam N, Alemayohu MA, Allen C, et al. The burden of primary liver cancer and underlying etiologies from 1990 to 2015 at the global, regional, and national level: results from the global burden of disease study 2015. JAMA oncology. 2017;3:1683-91.
2. Llovet JM, Burroughs A, Bruix J. Hepatocellular carcinoma. Lancet. 2003;362:1907-17.
3. Kwon JH, Bae SH, Kim JY, Choi BO, Jang HS, Jang JW, et al. Long-term effect of stereotactic body radiation therapy for primary hepatocellular carcinoma ineligible for local ablation therapy or surgical resection. Stereotactic radiotherapy for liver cancer. BMC Cancer. 2010;10:475.
4. Schaub SK, Hartvigson PE, Lock MI, Høyer M, Brunner TB, Cardenes HR, et al. Stereotactic body radiation therapy for hepatocellular carcinoma: current trends and controversies. Technol Cancer Res Treat. 2018;17:1533033818790217.
5. Wahl DR, Stenmark MH, Tao Y, Pollom EL, Caoili EM, Lawrence TS, et al. Outcomes after stereotactic body radiotherapy or radiofrequency ablation for hepatocellular carcinoma. J Clin Oncol. 2016;34:452.

6. Feng M, Suresh K, Schipper MJ, Bazzi L, Ben-Josef E, Matuszak MM, et al. Individualized adaptive stereotactic body radiotherapy for liver tumors in patients at high risk for liver damage: a phase 2 clinical trial. JAMA oncology. 2018;4:40-7.
7. Ohri N, Tomé WA, Romero AM, Miften M, Ten Haken RK, Dawson LA, et al. Local control after stereotactic body radiation therapy for liver tumors. International Journal of Radiation Oncology* Biology* Physics. 2018.
8. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Adv Neural Inf Process Syst; 2012. p. 1097-105.
9. Nie D, Lu J, Zhang H, Adeli E, Wang J, Yu Z, et al. Multi-channel 3D deep feature learning for survival time prediction of brain tumor patients using multi-modal neuroimages. Sci Rep. 2019;9:1-14.
10. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017;542:115-8.
11. Qin C, Schlemper J, Caballero J, Price AN, Hajnal JV, Rueckert D. Convolutional recurrent neural networks for dynamic MR image reconstruction. IEEE Trans Med Imaging. 2018;38:280-90.
12. Osokin A, Chessel A, Carazo Salas RE, Vaggi F. GANs for biological image synthesis. Proceedings of the IEEE International Conference on Computer Vision; 2017. p. 2233-42.
13. de Vos BD, Berendsen FF, Viergever MA, Sokooti H, Staring M, Išgum I. A deep learning framework for unsupervised affine and deformable image registration. Med Image Anal. 2019;52:128-43.
14. Li X, Dou Q, Chen H, Fu C-W, Qi X, Belavý DL, et al. 3D multi-scale FCN with random modality voxel dropout learning for intervertebral disc localization and segmentation from multi-modality MR images. Med Image Anal. 2018;45:41-54.
15. Zhang Y, Lobo-Mueller EM, Karanicolas P, Gallinger S, Haider MA, Khalvati F. CNN-based survival model for pancreatic ductal adenocarcinoma in medical imaging. BMC Med Imaging. 2020;20:1-8.
16. Fedorov A, Beichel R, Kalpathy-Cramer J, Finet J, Fillion-Robin J-C, Pujol S, et al. 3D Slicer as an image computing platform for the Quantitative Imaging Network. Magn Reson Imaging. 2012;30:1323-41.
17. Avants BB, Epstein CL, Grossman M, Gee JC. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. Med Image Anal. 2008;12:26-41.
18. Lin Z, Feng M, Santos CNd, Yu M, Xiang B, Zhou B, et al. A structured self-attentive sentence embedding. arXiv preprint arXiv:170303130. 2017.
19. You Q, Jin H, Wang Z, Fang C, Luo J. Image captioning with semantic attention. Proceedings of the IEEE conference on computer vision and pattern recognition; 2016. p. 4651-9.
20. Oktay O, Schlemper J, Folgoc LL, Lee M, Heinrich M, Misawa K, et al. Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:180403999. 2018.
21. Lee C-Y, Xie S, Gallagher P, Zhang Z, Tu Z. Deeply-supervised nets. Artificial intelligence and statistics; 2015. p. 562-70.

22. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: Visual explanations from deep networks via gradient-based localization. Proceedings of the IEEE international conference on computer vision; 2017. p. 618-26.
23. Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. European conference on computer vision: Springer; 2014. p. 818-33.

CHAPTER 7

Discussion and Future Perspective

7.1 Current challenges and recommendations

7.1.1 Radiomics and model fitting issues

Outcome modeling for medical images suffers from the insufficient data obstacle, with limited samples, noisiness in the labeling, unknown complicated and underlying mechanisms, etc. Though there is some low hanging fruit, end-to-end image analysis, such as image registration/segmentation, which is less prone to overfitting than outcome modeling that has been embedded in commercial softwares already. The reality is that image-based outcome modeling is still an essential part for personalized health care and more needs to be done to achieve its promise. Among these is better data management (sharing, storing, labeling, etc.), which is necessary but is beyond our scope here. Hence, the focus here has been on modifying and developing new algorithms that can capture the most relevant and useful signals out of such small and noisy data. It is unclear how to decide which kind of methods is better when we deal with a specific task. Although quite time-consuming to obtain, hand-crafted features take advantage of domain knowledge and tend to outperform CNN methods when the available dataset is too small and noisy. In addition, we can choose feature selection and classification models that contain less fitting

parameters to avoid overfitting, rather than the CNN methods that usually have thousands or even more parameters to fit. Another advantage is that engineered features are easier to interpret, since we know what features we extract and feed into the model. CNN methods are attracting more and more attention, with the advantage of avoiding tedious feature engineering and higher prediction accuracy. For small dataset, transfer learning, data augmentation and GAN have been implemented to amplify the dataset size and have shown promising results in image segmentation, disease detection and endpoint prediction tasks. Transfer learning is using (part of) pretrained models from other datasets to initialize the current neural network on the target dataset. However, this is not always helpful, especially in the case of outcome modeling, where the cross-task relationships are ambiguous. For the project in Chapter 5, we have experimented with various pretrained models, none of them showed better results than training from scratch. The reason might be that the source dataset (e.g., ImageNet) might be too different from our medical images for the task at hand. The features learnt thus may not transferrable. Another direction to think about is the combination of traditional features and CNNs directly applied to images. Features carefully designed can provide experts experience that might be hard to learn by CNN based on limited samples, while CNN can extract other important information missed by hand-crafted features till larger datasets become available.

7.1.2 Repeatability and Reproducibility issues

For CT, inter-scanner variability of image features produces differences in extracted features that are comparable to the variability in patient images acquired by the same scanner [1]. The choice of methods of reconstruction, such as filtered back projection or iterative algorithm, also affect radiomic features [2]. Smoothing of the image and reducing the slice thicknesses can improve reproducibility of CT-extracted features [3, 4].

In PET imaging, textural features are sensitive to different acquisition modes [5, 6], reconstruction algorithms, and their user-defined parameters such as the number of iterations, the post-filtering level, input data noise, matrix size, and discretization bin size [7, 8].

Radiomic features extracted from MRI scans depend on the field of view, field strength, reconstruction algorithm and slice thickness. Results of the DCE MRI depend on the contrast agent dose, method of administration, and the pulse sequence used. The radiomic features extracted from DW-MRI depend on acquisition parameters and conditions as k-space trajectory, gradient strengths and b-values. The repeatability of MR-based radiomic features has been investigated [9] using a ground truth digital phantom of brain glioma patients and an MRI simulator capable of generating images according to different acquisition (field strength, pulse sequence, arrangement of field coils) and reconstruction methods. It was found that some features are subject to small changes, compared with clinical effect size.

7.1.3 Standardization and harmonization

Although research in the field of radiomics has drastically increased over the past several years, there still remains a lack of reproducibility and validation of current radiomic models. There are currently no guidelines and standard definitions for radiomic features and for constructing these features into clinical models. Current initiatives are underway to improve standardization and harmonization in radiomic studies.

As a part of radiomic signature validation, there are efforts to explore distributed feature sharing and model development across contributing institutions [10]. A key component in this exercise is the assessment and redressal of batch effects [11] and confounding variables across contributing sites, so as to ameliorate systematic yet unmeasured sources of variation. Another key component is the use of methods to harmonize data as well as model parameters across study sites, with the intent of meaningful comparisons across clinical population [12]. Such efforts are necessary to enable the widespread and generalizable development of models that are transportable across institutions. In addition to the careful calibration and stability analysis of radiomic features within predictive models, there is also a need for ensuring model robustness through approaches like noise injection [13]. Adversarial training approaches from neural networks can have value in the modern deep learning modeling area by incorporating not only positive examples but negative ones too [14]. The workflow for computing features is complex and involves many steps, often leading to incomplete reporting of methodological information (e.g., texture matrix design choices and gray-

level discretization methods). As a consequence, few radiomics studies in the current literature can be reproduced from start to end.

To accelerate the translation of radiomics methods to the clinical environment, the Image Biomarker Standardization Initiative (IBSI) [15] has the goal to provide standard definitions and nomenclature for radiomic features, reporting guidelines, and to provide benchmark datasets and values to verify image processing and radiomic feature calculations.

7.1.4 Interpretability issues

It is recognized that machine learning algorithms tend to generally trade interpretability for better prediction. Hence, clinicians are still reluctant to embrace these methods as part of their clinical practice, because they have long been perceived them as “black boxes”, meaning that it is difficult to determine how they arrive at their predictions. For example, it is difficult to understand deep neural networks due to the large number of interacting, non-linear parts [16, 17]. To improve interpretability of radiomics for the clinician, methods based on graph approaches can be utilized [18], and in the context of deep learning better visualization tools are being developed such as feature maps highlighting regions of the tumor that impact the prediction of the deep learning classifier are also being proposed [17].

7.2 Future perspectives

7.2.1 Interpretable radiomics

Models giving good prediction and good representation though necessary are not sufficient for medical practice, especially for those using deep learning models or other complex methods. For the current thesis work, we mainly focused on outcome modeling that gives answers to what, when and where questions for tumors. The next step should emphasize better understanding of the model decisions and gain more insights into how these models operate. Various techniques have been proposed for this purpose, such as DeConvNet, which is composed of deconvolution and unpooling layers identifying pixel-wise class labels and predict segmentation masks [19, 20], Gradient-based, e.g., Grad-CAM, which uses the gradients of any target concept flowing into the final convolutional layer to produce a coarse localization map highlighting the important regions in the image for predicting the particular concept [21]. These models are post-hoc ones that achieve interpretability by sampling instances and labeling them with the trained neural networks [22]. Since we build imaging-based neural network models to predict what will happen to the target and healthy tissues, it will be beneficial to investigate these post-hoc interpretability models in the future.

7.2.2 Advanced modeling: Graph Neural Networks

However, there are also some criticism towards the post-hoc explanation models. Black box models with explanations can lead to an overly complicated decision pathway that is ripe for human error [23]. Marrying interpretable graph models with neural networks (graph neural

networks) have recently emerged in the machine learning and other related areas, and demonstrated superior performance in various problems. Graph models encode the structural information to model the relations among entities, and furnish more promising insights underlying the data. Despite some successes of these embedding methods, many of them suffer from the limitations of the shallow learning mechanisms and might fail to discover the more complex patterns behind the graphs. Combining GNN with deep learning has the advantage of not only encoding structural prior knowledge to the model, but also the large capacity of deep learning frameworks. In addition, graphical model is better in interpretation. GNN is also superior with heterogeneous input of patient data – the collection of imaging, genetic, clinical features for integrating such data.

7.2.3 Clinical translation: Functional liver avoidance treatment planning

This topic is an extension of the localization prediction project (Chapter 6). The rule when making treatment plans for patients is to give the target (tumor) the desired amount of dose, while sparing the normal tissue as much as possible. Considering the high recurrence rate of HCC, if we can predict the location of recurrence, we can further modify the treatment regimens to account for suspicious/risky locations and improve the patient prognosis post treatment.

7.3 References

1. Mackin D. Measuring Computed Tomography Scanner Variability of Radiomics Features. Investig. radiology 50, 757–65. 2015.

2. Kim H, Park CM, Lee M, Park SJ, Song YS, Lee JH, et al. Impact of reconstruction algorithms on CT radiomic features of pulmonary tumors: analysis of intra-and inter-reader variability and inter-reconstruction algorithm variability. PLoS One. 2016;11.
3. Leijenaar RT, Carvalho S, Velazquez ER, Van Elmpt WJ, Parmar C, Hoekstra OS, et al. Stability of FDG-PET Radiomics features: an integrated analysis of test-retest and inter-observer variability. Acta Oncol. 2013;52:1391-7.
4. Zhao B, James LP, Moskowitz CS, Guo P, Ginsberg MS, Lefkowitz RA, et al. Evaluating variability in tumor measurements from same-day repeat CT scans of patients with non–small cell lung cancer. Radiology. 2009;252:263-72.
5. Desseroit M-C, Tixier F, Weber WA, Siegel BA, Le Rest CC, Visvikis D, et al. Reliability of PET/CT shape and heterogeneity features in functional and morphologic components of non–small cell lung cancer tumors: a repeatability analysis in a prospective multicenter cohort. J Nucl Med. 2017;58:406-11.
6. Galavis PE, Hollensen C, Jallow N, Paliwal B, Jeraj R. Variability of textural features in FDG PET images due to different acquisition modes and reconstruction parameters. Acta Oncol. 2010;49:1012-6.
7. Lu L, Lv W, Jiang J, Ma J, Feng Q, Rahmim A, et al. Robustness of radiomic features in [11 c] choline and [18 f] fdg pet/ct imaging of nasopharyngeal carcinoma: Impact of segmentation and discretization. Mol Imaging Biol. 2016;18:935-45.
8. Bailly C, Bodet-Milin C, Couespel S, Necib H, Kraeber-Bodéré F, Ansquer C, et al. Revisiting the robustness of PET-based textural features in the context of multi-centric trials. PLoS One. 2016;11.
9. Yang F, Dogan N, Stoyanova R, Ford JC. Evaluation of radiomic texture feature error due to MRI acquisition and reconstruction: a simulation study utilizing ground truth. Phys Med. 2018;50:26-36.
10. AlBadawy EA, Saha A, Mazurowski MA. Deep learning for segmentation of brain tumors: Impact of cross‐institutional training and testing. Med Phys. 2018;45:1150-8.
11. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007;8:118-27.
12. Chang K, Balachandar N, Lam C, Yi D, Brown J, Beers A, et al. Distributed deep learning networks among institutions for medical imaging. J Am Med Inform Assoc. 2018;25:945-54.
13. Zur RM, Jiang Y, Pesce LL, Drukker K. Noise injection for training artificial neural networks: A comparison with weight decay and early stopping. Med Phys. 2009;36:4810-8.
14. Li S, Chen Y, Peng Y, Bai L. Learning more robust features with adversarial training. arXiv preprint arXiv:180407757. 2018.
15. Zwanenburg A, Leger S, Vallières M, Löck S. Image biomarker standardisation initiative. arXiv preprint arXiv:161207003. 2016.
16. Yosinski J, Clune J, Nguyen A, Fuchs T, Lipson H. Understanding neural networks through deep visualization. arXiv preprint arXiv:150606579. 2015.

17. Sankar V, Kumar D, Clausi DA, Taylor GW, Wong A. SISC: end-to-end interpretable discovery radiomics-driven lung cancer prediction via stacked interpretable sequencing cells. arXiv preprint arXiv:190104641. 2019.
18. Luo Y, McShan D, Ray D, Matuszak M, Jolly S, Lawrence T, et al. Development of a fully cross-validated Bayesian network approach for local control prediction in lung cancer. IEEE transactions on radiation and plasma medical sciences. 2018;3:232-41.
19. Noh H, Hong S, Han B. Learning deconvolution network for semantic segmentation. Proceedings of the IEEE international conference on computer vision; 2015. p. 1520-8.
20. Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. European conference on computer vision: Springer; 2014. p. 818-33.
21. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: Visual explanations from deep networks via gradient-based localization. Proceedings of the IEEE international conference on computer vision; 2017. p. 618-26.
22. Laugel T, Lesot M-J, Marsala C, Renard X, Detyniecki M. The dangers of post-hoc interpretability: Unjustified counterfactual explanations. arXiv preprint arXiv:190709294. 2019.
23. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence. 2019;1:206-15.