UNIVERSITY OF CALIFORNIA,
IRVINE

Machine Learning in Clinical Application of Medical Imaging for Lesion Detection, Segmentation, Diagnosis, Therapy, and Prognosis Prediction

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Biomedical Engineering

by

Yang Zhang

Dissertation Committee:
Professor Min-Ying Lydia Su, co-Chair
Associate Professor Gultekin Gulsen, co-Chair
Assistant Professor Daniel Chow
Associate Professor Michelle Digman
Professor Frithjof Kruggel

2020

DEDICATION

To

my mentors

in recognition of their worth

for their dedication, support, caring, and encouragement, for starting me on my path

and

believing in my future.

"A mentor empowers a person to see a possible future and believe it can be obtained."
—Shawn Hitchcock

.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I am very thankful for all of the years I have spent working at the Center for Functional Onco-Imaging (CFOI). I feel extremely grateful to have met and worked with so many incredible people and would like to thank all the lab members for helping, training, teaching, and guiding me all these years. I would not have been able to go this far without help from so many wonderful, caring, loving people.

I owe many thanks to my amazing advisor and mentor Dr. Lydia Su. I am so grateful to have met her and consider myself lucky to be her student. I don't know many other professors who care about their students' learning, daily life, and future career as much as their research. I am grateful for her help, mentorship, and leadership in creating such a supportive, collaborative, and fun research environment. I would like to thank her for being so open and communicative, always making time to discuss problems with us no matter how busy she is. I like to thank her for being so patient, optimistic, and encouraging when I encountered a problem, especially for teaching me how to be resourceful and creative when solving problems. When I was struggling with deadlines for abstracts and papers, it was great to have her support by being with me until the last moments before submission. I truly appreciate all her help and the time that she spent in training me and her insightful advice not only on my career but also my health and personal life. She is not just a good teacher, a good advisor but also a great friend to me and all people around me. It has been such a great pleasure to work at CFOI.

I would also like to thank my mentor Dr. Joen-Hor Chen for all the guidance and support he has provided me. I like to thank him for providing valuable insight from a radiologist's perspective, which greatly changed my work from a scientist's perspective to have a real clinical impact to improve patient care. Most importantly, I benefit a lot from the great clinical resources that he provided to me, as well as his encouraging and thought-provoking comments that inspired me and taught me how to conduct research by developing and implementing new ideas. He provided me with a unique opportunity to gain a wide breadth of clinical experience while being a graduate student.

I would also like to thank Dr. Gultekin Gulsen. I like to thank him as my co-advisor at BME, as well as all help, advice, training, and support. I like to thank him for being such a kind and generous mentor and always trying his best to offer me all available resources. His great personality and very warm and caring attitude towards everyone at the center always encourage and inspire me to become a better person.

I want to give my most sincere gratitude to Dr. Frithjof Kruggel for his great help in my darkest time when I first came to US to start my graduate study. His compassionate help raised me up from my lowest point and the most depressing moment of my life. I also want to express my gratitude to Dr. Daniel Chow, Dr. Michelle Digman, and Dr. Beth Lopour for kindly serving on my committee and their valuable suggestions and great discussions.

I would also like to thank my officemates. These years staying at UCI would not have been fun and memorable without them. The support I have received from my colleagues from the CFOI has also been endearing and I am in awe of their generosity. I want to say thanks to Dr. Tiffany Kwong, Dr. Farouk Nouizi, Dr. Alex Luk, Dr. Hakan Erkol, Dr. Hon Yu, Dr. Jaedu Cho, Dr. Jie Zheng, Dr. Jessica Kwong, Yan-lin Liu, Maha Algarawi, and Michael Marks. I greatly value their friendship and deeply appreciate their belief in me.

I would like to show my appreciation to my collaborators, Dr. Ke Nie, Dr. Peter Chang, Dr. Ching-Chung Ko, Dr. Lee-Ren Yeh, Dr. Te-Chang Wu, Dr. Ning Lang, Dr. Vivian Park, Dr. Meihao Wang, Dr. Yezhi Lin, Dr. Jiejie Zhou, Dr. Qiong Ye, and Xiao Chen. Without their great help, support, and clinical resources, it would not be possible for me to have access to many great datasets related to different clinical problems. These were invaluable for my training and research.

Lastly, I would like to give a big thanks to my dad, my mom, my grandpa, and my late grandma in heaven for all their endless support and love.

CURRICULUM VITAE

Yang Zhang

EDUCATION

University of California, Irvine Irvine, CA
Ph.D. in Department of Biomedical Engineering

University of California, Irvine Irvine, CA
Master in Department of Electrical Engineering and Computer Science

Zhengzhou University Zhengzhou, China
B.S. in communication engineering

RESEARCH

Center for Functional Onco-Imaging, University of California, Irvine, CA
Graduate Student Researcher 2015-present

- Automatic Breast and Fibroglandular Tissue Segmentation in Breast MRI Using Deep Learning
- Prediction of Chemoradiation Therapy Response in Rectal Cancer Using Pre-treatment and Mid-radiation Multi-parametric MRI
- Evaluation of Variations in Radiotherapy Planning Target Volume Delineation Due to Image Contrast and Patient Motion
- 3D MRI for Quantitative Analysis of Quadrant Percent Breast Density: Correlation with Quadrant Location of Breast Cancer
- Development of Robust Texture Parameters for Characterizing Normal Breast Parenchymal Patterns
- Development of Automatic Breast and Fibroglandular Tissue Segmentation
- Prediction of Breast Cancer Malignancy and Molecular Subtypes
- Automated Localization and Segmentation of Locally-Advanced Rectal Cancer
- Differentiation of Metastatic Cancer in the Spine Originated from Lung Cancer and Other Tumors

Medical Signal and Image Processing Laboratory, University of California, Irvine, CA
Graduate Student Researcher 2013-2015

- Finite Element Method Modeling for Human Head Impact Simulation

AWARDS & HONORS

- Trainee Research Prize, Radiology Society of North America 2019

Awarded as the author of the best papers at Radiological Society of North America annual meeting

- Presidential Award for Academic Excellence, Zhengzhou University 2010
- Outstanding Student Leader, Zhengzhou University 2008

ORAL PRESENTSTIONS

1. "Improving CBCT Quality to CT Level for Adaptive Radiation Therapy using Deep-Learning with Generative Adversarial Network" Oral Presentation at Radiological Society of North America 105th Scientific Assembly and Annual Meeting, Chicago, IL, December 2019
2. "Automatic Spine Segmentation for Detection of Abnormal Vertebra and Differentiation of Benign and Malignant Fracture on CT Using Deep Learning" Oral Presentation at Radiological Society of North America 105th Scientific Assembly and Annual Meeting, Chicago, IL, December 2019
3. "Automatic Search in Breast MRI Dataset for Detection of Suspicious Lesions Using Mask RCNN", Power Pitch at the 27th International Society for Magnetic Resonance in Medicine Annual Meeting, Montreal, Canada, May 2019.
4. "Comparison of Radiomics and Deep Learning for Differentiation of Spinal Metastases Coming from Lung Cancer and Other Primary Cancers". Oral Presentation at the 27th International Society for Magnetic Resonance in Medicine Annual Meeting, Montreal, Canada, May 2019
5. "Independent Validation of U-Net Based Breast and Fibroglandular Tissue Segmentation Method on MRI Datasets Acquired Using Different Scanners". Oral Presentation at the 27th International Society for Magnetic Resonance in Medicine Annual Meeting, Montreal, Canada, May 2019.
6. "Prediction of Neoadjuvant Chemoradiation Therapy Response in Rectal Cancer Using Radiomics Compared to Deep Learning Based on Pre-Treatment and mid-RT MRI." Oral Presentation at the 27th International Society for Magnetic Resonance in Medicine Annual Meeting, Montreal, Canada, May 2019.
7. "Machine Learning for Tumor Subtype Differentiation and Neoadjuvant Chemoradiation Response Prediction." Invited Talk at MR of Cancer study group of International Society for Magnetic Resonance in Medicine, Paris, France, June 2018.
8. **"**Machine Learning for Prediction of Chemoradiation Therapy Response in Patients with Locally-Advanced Rectal Cancer (LARC) Using Pre- and Early-Treatment Follow-up Multiparametric MRI." Oral Presentation at the 26th International Society for Magnetic Resonance in Medicine Annual Meeting, Paris, France, June 2018.
9. **"**Prediction of Breast Cancer Molecular Subtypes Using Conventional Feature Extraction and Two Machine Learning Architectures Based on DCE-MRI." Power Pitch at the 26th International Society for Magnetic Resonance in Medicine Annual Meeting, Paris, France, June 2018.

PROFESSIONAL ASSOCIATION

- Radiological Society of North America
- International Society for Magnetic Resonance in Medicine
- American Association of Physicists in Medicine
- Society for Imaging Informatics in Medicine

ARTICLES IN PEER-REVIEWED JOURNAL PAPERS

1. Chen JH, Liao F, **Zhang Y**, Li Y, Chang CJ, Chou CP, Yang TL, Su MY. 3D MRI for Quantitative Analysis of Quadrant Percent Breast Density: Correlation with Quadrant Location of Breast Cancer. Acad Radiol. 2017 Jul;24(7):811-817. doi: 10.1016/j.acra.2016.12.016

2. Shin GW, **Zhang Y**, Kim MJ, Su MY, Kim EK, Moon HJ, Yoon JH, Park VY. Role of dynamic contrast-enhanced MRI in evaluating the association between contralateral parenchymal enhancement and survival outcome in ER-positive, HER2-negative, node-negative invasive breast cancer. J Magn Reson Imaging. 2018 Dec;48(6):1678-1689. doi: 10.1002/jmri.26176
3. Chen JH, **Zhang Y**, Chan S, Chang RF, Su MY. Quantitative analysis of peri-tumor fat in different molecular subtypes of breast cancer. Magn Reson Imaging. 2018 Nov;53:34-39. doi: 10.1016/j.mri.2018.06.019
4. **Zhang Y**, Chen JH, Chang KT, Park VY, Kim MJ, Chan S, Chang P, Chow D, Luk A, Kwong T, Su MY. Automatic Breast and Fibroglandular Tissue Segmentation in Breast MRI Using Deep Learning by a Fully-Convolutional Residual Neural Network U-Net. Acad Radiol. 2019 Nov;26(11):1526-1535. doi: 10.1016/j.acra.2019.01.012
5. Lang N, **Zhang Y**, Zhang E, Zhang J, Chow D, Chang P, Yu HJ, Yuan H, Su MY. Differentiation of spinal metastases originated from lung and other cancers using radiomics and deep learning based on DCE-MRI. Magn Reson Imaging. 2019 Dec;64:4-12. doi: 10.1016/j.mri.2019.02.013
6. Shi L, **Zhang Y**, Nie K, Sun X, Niu T, Yue N, Kwong T, Chang P, Chow D, Chen JH, Su MY. Machine learning for prediction of chemoradiation therapy response in rectal cancer using pre-treatment and mid-radiation multi-parametric MRI. Magn Reson Imaging. 2019 Sep;61:33-40. doi: 10.1016/j.mri.2019.05.003
7. **Zhang Y**, Chen JH, Chen TY, Lim SW, Wu TC, Kuo YT, Ko CC, Su MY. Radiomics Approach for Prediction of Recurrence in Skull Base Meningiomas. Neuroradiology 2019 Dec;61(12):1355-1364. doi: 10.1007/s00234-019-02259-0
8. Chen JH, Chan S, **Zhang Y**, Li S, Chang RF, Su MY. Evaluation of breast stiffness measured by ultrasound and breast density measured by MRI using a prone-supine deformation model. Biomark Res. 2019;7:20. doi: 10.1186/s40364-019-0171-1
9. Zhou J, **Zhang Y**, Chang KT, Lee KE, Wang O, Li J, Lin Y, Pan Z, Chang P, Chow D, Wang M, Su MY. Diagnosis of Benign and Malignant Breast Lesions on DCE-MRI by Using Radiomics and Deep Learning with Consideration of Peri-Tumor Tissue. J Magn Reson Imaging. 2020 Mar;51(3):798-809. doi: 10.1002/jmri.26981
10. Zhang J, Chen Y, Zhang Y, Zhang E, Hu HJ, Yuan H, **Zhang Y**, Su MY, Lang N. Diagnosis of Spinal Lesions Using Perfusion Parameters Measured by DCE-MRI and Metabolism Parameters Measured by PET/CT. Eur Spine J. 2020 May;29(5):1061-1070. doi: 10.1007/s00586-019-06213-9
11. **Zhang Y**, Chen JH, Lin Y, Chan S, Zhou J, Chow D, Chang P, Kwong T, Yeh DC, Wang X, Parajuli R, Mehta RS, Wang M, Su MY. Prediction of Breast Cancer Molecular Subtypes on DCE-MRI Using Convolutional Neural Network with Transfer Learning between Two Centers. In Press, European Radiology, 2020.

CONFERENCE PROCEEDINGS

1. Wang X, **Zhang Y**, Chen JH, Chan S, Su MY. Automatic Breast Tumor Segmentation Methods for Mass and Non-Mass Lesions for Quantitative Morphology and Texture Analysis. Presented at the 25th International Society for Magnetic Resonance in Medicine Annual Meeting, Honolulu, Hawaii, April 22 –27, 2017. Program Number: 4935.
2. **Zhang Y**, Chen JH, Chan S, Su MY. Co-registration of Breast MRI and CT Using Gravity Unloading. Presented at the 25th International Society for Magnetic Resonance in Medicine Annual Meeting, Honolulu, Hawaii, April 22 – 27, 2017. Program Number: 4940.
3. **Zhang Y**, Chen JH, Chan S, Yeh DC, Su MY. Development of Robust Texture Parameters for Characterizing Normal Breast Parenchymal Patterns. Presented at the 25th International Society for Magnetic Resonance in Medicine Annual Meeting, Honolulu, Hawaii, April 22– 27, 2017. Program Number: 4941.
4. **Zhang Y**, Shi L, Sun X, Niu T, Yue N, Kwong T, Chang P, Khy M, Chow D, Su MY, Nie K. Machine Learning for Prediction of Chemoradiation Therapy Response in Patients with Locally-

Advanced Rectal Cancer (LARC) Using Pre- and Early-Treatment Follow-up Multiparametric MRI. Present at the 26th International Society for Magnetic Resonance in Medicine Annual Meeting, Paris, France, June 16-21, 2018. Program Number: 0829

5. **Zhang Y**, Chan S, Chen JH, Chow D, Chang P, Khy M, Yeh DC, Wang X, Su MY. Prediction of Breast Cancer Molecular Subtypes Using Conventional Feature Extraction and Two Machine Learning Architectures Based on DCE-MRI. Present at the 26th International Society for Magnetic Resonance in Medicine Annual Meeting, Paris, France, June 16-21, 2018. Program Number: 0993.
6. **Zhang Y**, Park VY, Kim MJ, Chang P, Khy M, Chow D, Chen JH, Luk A, Su MY. Automatic Breast and Fibroglandular Tissue Segmentation Using Deep Learning by A Fully-Convolutional Residual Neural Network. Present at the 26th International Society for Magnetic Resonance in Medicine Annual Meeting, Paris, France, June 16-21, 2018. Program Number: 2420.
7. Chang P, Khy M, **Zhang Y**, Su MY, Chow D. Deep learning convolutional neural networks accurately classify genetic mutations and survival in gliomas. Present at the 26th International Society for Magnetic Resonance in Medicine Annual Meeting, Paris, France, June 16-21, 2018. Program Number: 3655.
8. **Zhang Y**, Shi L, Sun X, Niu T, Yue N, Chang P, Chow D, Khy M, Kwong T, Chen JH, Su MY, Nie K. Evaluation of MRI-Guided Radiotherapy Planning Target Volume Delineation for Colorectal Cancer Due to Variations in Image Contrast and Patient Motion. Present at the 26th International Society for Magnetic Resonance in Medicine Annual Meeting, Paris, France, June 16-21, 2018. Program Number: 3797
9. Chen JH, Chan S, **Zhang Y**, Yeh DC, Su MY, 3D MRI for Quantitative Analysis of Quadrant Percent Density (QPD): Correlation with Location of Breast Cancer Growing in Different Quadrants. Present at the 26th International Society for Magnetic Resonance in Medicine Annual Meeting, Paris, France, June 16-21, 2018. Program Number: 4324.
10. Chen JH, Chan S, **Zhang Y**, Yeh DC, Su MY. Correlation of Breast Stiffness Measured by Ultrasound with Breast Density Measured on MRI Matched by Using a Prone-Supine Deformation Model. Present at the 26th International Society for Magnetic Resonance in Medicine Annual Meeting, Paris, France, June 16-21, 2018. Program Number: 4325.
11. Chen JH, **Zhang Y**, Chan S, Yeh DC, Su MY, Quantitative Analysis of Peri-Tumor Interface Fat and the Volumetric Fat Percentage and Contrast Enhancement in Three Peri-Tumoral Shells to Differentiate Molecular Subtypes of Breast Cancer. Present at the 26th International Society for Magnetic Resonance in Medicine Annual Meeting, Paris, France, June 16-21, 2018. Program Number: 4460
12. **Zhang Y**, Shi L, Sun X, Niu T, Yue N, Chang P, Chow D, Khy M, Kwong T, Chen JH, Su MY, Nie K. Automated Localization and Segmentation of Locally-Advanced Rectal Cancer Based on T2, DWI and DCE Multi-Parametric MRI Using Deep Learning. Present at the 26th International Society for Magnetic Resonance in Medicine Annual Meeting, Paris, France, June 16-21, 2018. Program Number: 4712.
13. Lang N, **Zhang Y**, Chow D, Yu HJ, Yuan H, Su MY. Differentiation of Pulmonary From Non-Pulmonary Spine Metastases Using Conventional DCE Kinetic Analysis and Machine Learning. Present at the 26th International Society for Magnetic Resonance in Medicine Annual Meeting, Paris, France, June 16-21, 2018. Program Number: 5168
14. **Zhang Y**, Shi L, Nie K, Sun X, Niu T, Yue N, Kwong T, Chang P, Chow D, Chen JH, Su MY. Prediction of Neoadjuvant Chemoradiation Therapy Response in Rectal Cancer Using Radiomics Compared to Deep Learning Based on Pre-Treatment and mid-RT MRI. Presented at the 27th International Society for Magnetic Resonance in Medicine Annual Meeting, Montreal, Canada, May 11–16, 2019. Program Number: 0101.
15. **Zhang Y**, Chen JH, Chang KT, Park VY, Kim MJ, Chan S, Chang P, Chow D, Luk A, Kwong T, Su MY. Independent Validation of U-Net Based Breast and Fibroglandular Tissue Segmentation Method on MRI Datasets Acquired Using Different Scanners. Presented at the 27th International

Society for Magnetic Resonance in Medicine Annual Meeting, Montreal, Canada, May 11–16, 2019. Program Number: 0284.

16. **Zhang Y**, Chang KT, Chan S, Chang P, Chow D, Chen JH, Su MY, Automatic Search in Breast MRI Dataset for Detection of Suspicious Lesions Using Mask RCNN, Presented at the 27th International Society for Magnetic Resonance in Medicine Annual Meeting, Montreal, Canada, May 11–16, 2019. Program Number: 0594.
17. Zhou J, **Zhang Y**, Chang KT, Chang P, Chow D, Wang O, Wang M, Su MY. Differential Diagnosis of Benign and Malignant Breast Lesions Based on DCE-MRI by Using Radiomics and Deep Learning with Five Different Networks. Presented at the 27th International Society for Magnetic Resonance in Medicine Annual Meeting, Montreal, Canada, May 11–16, 2019. Program Number: 0596.
18. **Zhang Y**, Lang N, Zhang E, Zhang J, Chow D, Chang P, Yu HJ, Yuan H, Su MY. Comparison of Radiomics and Deep Learning for Differentiation of Spinal Metastases Coming from Lung Cancer and Other Primary Cancers. Presented at the 27th International Society for Magnetic Resonance in Medicine Annual Meeting, Montreal, Canada, May 11–16, 2019. Program Number: 1144.
19. Yeh LR, **Zhang Y**, Chen JH, Chang P, Chow D, Su MY. Differentiation of Vertebral Fracture Types using Five Different Convolutional Neural Network Approaches. Presented at the 27th International Society for Magnetic Resonance in Medicine Annual Meeting, Montreal, Canada, May 11–16, 2019. Program Number: 1372.
20. **Zhang Y**, Shi L, Sun X, Niu T, Yue N, Chen JH, Kwong T, Su MY, Nie K. Change of Radiotherapy Planning Target Volume Delineated on Pre-Treatment and mid-RT Follow-up MRI After 3-4 Weeks of Treatment. Presented at the 27th International Society for Magnetic Resonance in Medicine Annual Meeting, Montreal, Canada, May 11–16, 2019. Program Number: 2338.
21. Ko CC, **Zhang Y**, Chen JH, Chang P, Chow D, Kwong T, Su MY. Radiomics Approach for Prediction of Tumor Recurrence and Progression of Skull Base Meningioma. Presented at the 27th International Society for Magnetic Resonance in Medicine Annual Meeting, Montreal, Canada, May 11–16, 2019. Program Number: 3080.
22. **Zhang Y**, Chen JH, Chang KT, Chan S, Pan HB, Zhou JJ, Wang O, Wang M, Lydia Su MY. Development of U-Net Breast Density Segmentation Method for Fat-Sat T1-Weighted Images Using Transfer Learning from Model for Non-Fat-Sat Images. Presented at the 27th International Society for Magnetic Resonance in Medicine Annual Meeting, Montreal, Canada, May 11–16, 2019. Program Number: 4740.
23. Negrete NT, Takhtawala R, Shaver M, Kart T, **Zhang Y**, Kim MJ, Park VY, Su MY, Chow D, Chang P. Automated breast cancer lesion detection on breast MRI using artificial intelligence. Presented at America Society of Clinical Oncology Annual Meeting, May 31 - June 4, 2019, Chicago, IL. Program Number: 265341.
24. **Zhang Y**, Yue N, Su MY, Ding Y, Liu B, Zhou Y, Nie K. Improving CBCT Quality to CT Level using Deep-Learning Method for Adaptive Radiation Therapy. Presented at the 61st American Association of Physicists in Medicine Annual Meeting and Exhibition, San Antonio, Texas, July 14–18, 2019. Program Number: 45661
25. **Zhang Y**, Yeh LR, Chen JH, Lang N, Xing X, Chen Y, Wang Q, Chang P, Chow D, Yuan H, Su MY. Differentiation of Benign and Malignant Vertebral Fracture on MR Using ResNet Deep Learning Compared to Radiologist's Reading. Presented at Radiological Society of North America 105th Scientific Assembly and Annual Meeting, Chicago, IL, December 1–6, 2019. Program Number: MK399-SD-TUA7
26. Wang M, **Zhang Y**, Zhou J, Lee K, Chang KT, Chang P, MD, Wang O, Su MY. Diagnosis of Benign and Malignant Breast Lesions on DCE-MRI Using Radiomics and Deep Learning with Peri-Tumor Tissue. Presented at Radiological Society of North America 105th Scientific Assembly and Annual Meeting, Chicago, IL, December 1–6, 2019. Program Number: SSM02-05

27. **Zhang Y**, Yue N, Su MY, Ding Y, Liu B, Zhang Y, Zhou Y, Nie K. Improving CBCT Quality to CT Level for Adaptive Radiation Therapy using Deep-Learning with Generative Adversarial Network. Presented at Radiological Society of North America 105th Scientific Assembly and Annual Meeting, Chicago, IL, December 1–6, 2019. Program Number: SSM23-03
28. Lang N, **Zhang Y**, Xing X, Chen Y, Wang Q, Chang P, Chow D, Yuan H, Su MY. Automatic Spine Segmentation for Detection of Abnormal Vertebra and Differentiation of Benign and Malignant Fracture on CT Using Deep Learning. Presented at Radiological Society of North America 105th Scientific Assembly and Annual Meeting, Chicago, IL, December 1–6, 2019. Program Number: SST05-08
29. **Zhang Y**, Yeh LR, Chen JH, Lang N, Xing X, Chen Y, Wang Q, Chang P, Chow D, Yuan H, Su MY. Deep learning for the detection and differentiation of vertebral fracture. Presented at ISMRM Virtual Conference, Aug 8-14, 2020. Program Number: 0247
30. Zhou J, **Zhang Y**, Chang KT, Lee K, Wang O, Li J, Lin Y, Pan Z, Chang P, Chow D, Wang M, Su MY. Diagnosis of Benign and Malignant Breast Lesions on DCE-MRI by Using Radiomics and Deep Learning with Consideration of Peri-Tumor Tissue. Presented at ISMRM Virtual Conference, Aug 8-14, 2020. Program Number: 0566
31. Chen Y, **Zhang Y**, Zhang E, Xing X, Wang Q, Yuan H, Su, MY, Lang N. Classification of Spinal Metastases Coming from Different Primary Cancer Origin by Using Quantitative Radiomics Analysis with Multi-Class SVM. Presented at ISMRM Virtual Conference, Aug 8-14, 2020. Program Number: 1149
32. Ko CC, **Zhang Y**, Chang KT, Chen JH, Su MY. Radiomics and Machine Learning for Prediction of Recurrence in Meningiomas. Presented at ISMRM Virtual Conference, Aug 8-14, 2020. Program Number: 1907
33. Ko CC, Chang KT, **Zhang Y**, Chen JH, Su MY. Radiomics Approach for Prediction of Recurrence in Pituitary Macroadenomas. Presented at ISMRM Virtual Conference, Aug 8-14, 2020. Program Number: 1908
34. **Zhang Y**, Zhou J, Park YV, Chan S, Wang M, Kim MJ, Chang KT, Chang P, Chow D, Chen JH, Su MY. Automatic Detection and Segmentation of Breast Cancer on MRI Using Mask R-CNN Trained on Non-Fat-Sat Images and Tested on Fat-Sat Images. Presented at ISMRM Virtual Conference, Aug 8-14, 2020. Program Number: 2318
35. Zhou J, **Zhang Y**, Lee K, Chen JH, He X, Xu N, Ye S, Wang O, Li J, Lin Y, Wang M, Su MY. Comparison of Breast Cancer Diagnostic Performance Using Radiomics Models Built Based on Features Extracted from DCE-MRI and Mammography. Presented at ISMRM Virtual Conference, Aug 8-14, 2020. Program Number: 2321
36. Wang M, **Zhang Y**, Zhou J, Miu H, Xu N, He X, Ye S, Liu H, Wang O, Li J, Lin Y, Su, MY. Diagnosis of Non-Mass-Like Enhancement Lesions on DCE-MRI by Using Quantitative Radiomics and Radiologists' BI-RADS Reading. Presented at ISMRM Virtual Conference, Aug 8-14, 2020. Program Number: 2323
37. **Zhang Y**, Li W, Zhang Z, Xue Y, Chang P, Chow D, Su MY, Ye Q. Artificial Intelligence Analysis on Prostate DCE-MRI to Distinguish Prostate Cancer and Benign Prostatic Hyperplasia. Presented at ISMRM Virtual Conference, Aug 8-14, 2020. Program Number: 2394
38. **Zhang Y**, Lin Y, Chan S, Chen JH, Zhou J, Chow D, Chang P, Wang M, Su MY. Differentiation of Breast Cancer Molecular Subtypes on DCE-MRI by Using Convolutional Neural Network with Transfer Learning. Presented at ISMRM Virtual Conference, Aug 8-14, 2020. Program Number: 3526
39. **Zhang Y**, Chen JH, Lin Y, Chan S, Zhou J, Chow D, Chang P, Kwong T, Wang X, Parajuli R, Mehta RS, Wang M, Su MY. Prediction of Breast Cancer Molecular Subtypes on DCE-MRI Using Convolutional Neural Network with Transfer Learning between Two Centers. Accepted for presentation at RSNA annual meeting, 2020.

ABSTRACT OF THE DISSERTATION

Machine Learning in Clinical Application of Medical Imaging for Lesion Detection, Segmentation, Diagnosis, Therapy, and Prognosis Prediction
by

Yang Zhang

Doctor of Philosophy in Biomedical Engineering

University of California, Irvine, 2020

Professor Lydia Min-Ying Su and Professor Gultekin Gulsen, co-Chairs

Medical imaging, including computed tomography (CT), magnetic resonance imaging (MRI), mammography, ultrasound, X-ray, and nuclear medicine, is the non-invasive process utilized to create visual representations of interior organs and tissues. Medical imaging's clinical purpose is to observe health, aid in diagnosis, monitor treatment response, and perform follow-up for disease surveillance. Clinically, interpreting medical images has mostly been performed by human experts such as radiologists or physicians. However, given the wide variety in pathological conditions and the potential fatigue that can result from visual assessment of numerous images, computer-aided diagnosis or detection (CAD) algorithms have been developed and proven to be very helpful. These CAD systems can also provide various functions, such as giving quantitative measurements, extracting radiomics features, and displaying the most important information to assist radiologists' interpretation. Furthermore, it can even detect suspicious findings and present the malignancy probability using various methods, such as different markers and colors.

The maturity of radiomics analysis with machine learning has provided a very efficient method to build classification models for clinical tasks, including diagnosis, staging, and prognosis prediction. In recent years, neural network methods, a machine learning technique inspired by the human neuronal synapse system, have been widely applied in medical imaging for disease management. The increased volume and quality of digital imaging datasets has created the potential for more accurate and efficient image evaluation using fully automated computer algorithms. However, compared with other machine learning methods such as radiomics, neural networks suffer from several major limitations, including the need for a large dataset to train the deep architecture, the high demand for computing power, and the poor generalization to other datasets not considered in training.

However, during the last 5 years, neural networks have become increasingly popular and have even proven feasible for implementation in clinical practice with the growing availability of big data, enhanced computing power, and novel algorithms. There are many Artificial Intelligence (AI) companies working in this field, and new software being rapidly approved by FDA for clinical use. Deep Learning (DL) algorithms, particularly the convolutional neural networks (CNN), have become the methodology of choice for analyzing medical images. Unlike conventional CAD algorithms, such as radiomics analysis in which task-related features are designed mostly by human experts based on their knowledge about the target domains, deep learning incorporates the feature engineering steps into its learning process. That is, instead of extracting pre-defined features, deep learning only requires pre-processed input data and outcome, discovering its own characteristic information in a self-taught manner. Therefore, the burden of feature

engineering has shifted from humans to computers to generate more consistent and reliable outputs.

This thesis will feature radiomics and deep learning-based techniques developed and implemented to extract information from medical images for performing commonly needed clinical tasks, including: lesion detection, organ/tissue segmentation, tumor classification, therapy planning, therapy response prediction, and prognosis prediction.

Chapter 1. Introduction

1.1 Motivation

Radiologic imaging is critically needed in modern patient care. When a patient is presenting with symptoms and needs clinical care, very often the first task that a physician will do is to determine which imaging is required to reveal more information about the disease. Both diagnostic and therapeutic indications for radiologic imaging are expanding rapidly [1]. This rapid expansion is a consequence of the demand for more efficient, accurate, cost-effective, and less invasive treatment. Technologic advancements in radiologic imaging equipment have also fueled the utilization of imaging. Such technologic advancements include the capability to acquire higher resolution images, enabling the visualization of smaller anatomic structures and abnormalities. Based on different tissue contrast mechanisms, various modalities have been developed in the past decades in the field of diagnostic radiology [2]. Today, the mainstream modalities include radiography, fluoroscopy, mammography, digital breast tomosynthesis (DBT), ultrasound, computed tomography (CT), magnetic resonance imaging (MRI), and positron emission tomography (PET).

Let's use the diagnosis of breast cancer in MRI as an example to elaborate how machine learning is being implemented. Compared to conventional examinations, including clinical, mammography, and ultrasound, MRI has been proven to be the most sensitive (94%-100%) imaging modality in delineating tumor extent and detecting multifocal or multi-centric diseases [3]. However, despite its high sensitivity, MRI has a high false-positive rate and detects benign lesions with specificity ranging from 37%-97%, which may lead to patient anxiety, unnecessary biopsies, or over-treatment [4]. As the use of MRI increases, reading many images in a reasonable time becomes a concern. Furthermore, as an increasing

number of MRI studies are being performed in small community hospitals, the experience, and training of radiologists in interpreting MRI raises a critical problem as well [5].

Since 1980s, numerous machine learning (ML) algorithms with different mathematical bases and logical theories have been applied to perform classification tasks. For example, several computer-aided detection (CAD) systems were developed and introduced in the clinical workflow in the early 2000s. However, adverse impacts of these systems have been reported in multiple clinical studies, the most noticeably, the increase of recall rates [6, 7]. The CAD systems were also found to generate more false positives than human readers, which led to longer assessment times and additional biopsies [6]. Thus, the net benefit gained by using CAD was unclear [6]. It is expected that modern machine learning technology may help overcome the limitations of previous CAD systems, achieve higher detection accuracy, and help human readers become more productive by allowing them to shift tedious, repetitive radiology tasks to artificial intelligence (AI).

Deep learning is a new and exciting field of ML that has revolutionized many technological fields in a wide spectrum, from autonomous vehicles, discovery of new stars, DNA sequencing, to stock price prediction [8]. Indeed, the rapid rise in AI technology demonstrates enormous potential to change and influence radiological practice. Deep learning is well suited to medical "big data," and can be used to extract useful knowledge from enormous quantities of images. This new AI technology has the potential to greatly impact the radiology field from performing automatic lesion detection to suggesting differential diagnoses and composing preliminary radiology reports [9].

In my PhD research, I focus on six fields of clinical application: lesion detection, organ/tissue segmentation, differential diagnosis, treatment planning, neoadjuvant therapy

response, and prognosis prediction, as shown in **Figure 1-1**. I have performed these tasks for 6 different diseases: breast cancer, rectal cancer, prostate cancer, brain tumors, spine lesions, and COVID-19 lung lesions. For breast cancer, I developed methods for breast and fibroglandular tissue segmentation, benign and malignant diagnosis, and molecular subtype differentiation. For rectal cancer, I developed methods to predict neoadjuvant chemoradiation therapy response. For spinal diseases, different methods were developed for the differentiation of metastatic cancers coming from different primary origins, as well as diagnosis of benign vs. malignant fractures. For prostate cancer, I performed differential diagnosis to distinguish prostate cancer from benign prostatic hyperplasia (BPH), and also improved image quality for radiation treatment planning. For brain tumor, I predicted the prognosis of several different brain cancers, including meningiomas and nonfunctioning pituitary macroadenomas. For COVID-19 lung lesions, I co-registered the serial CT images to evaluate the progression of the COVID-19 infected areas inside lungs in follow-up scans compared to baseline scans.

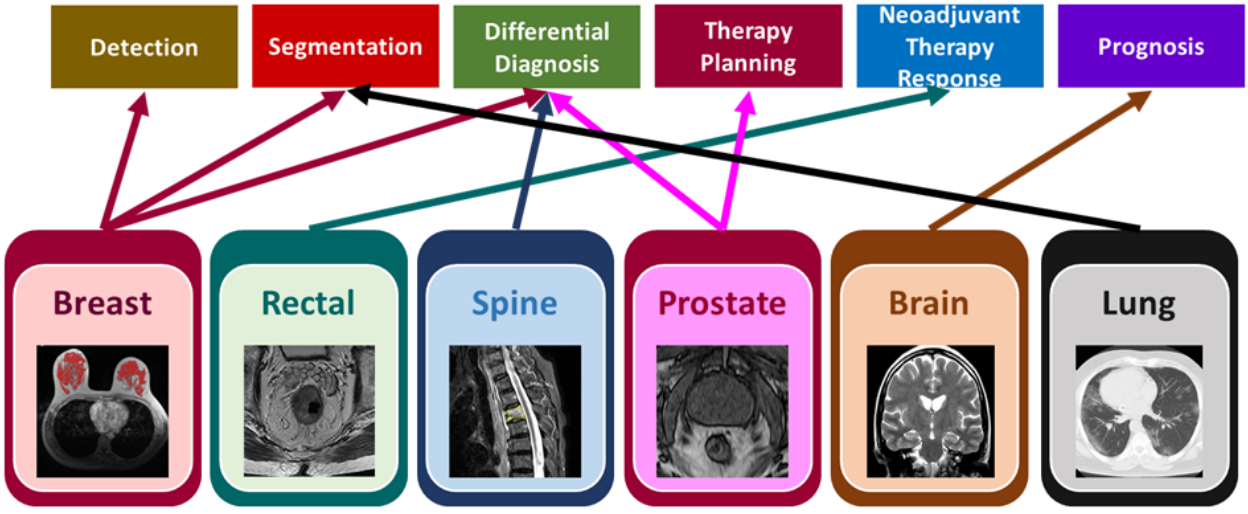

Figure 1-1: A diagram to show 6 diseases analyzed in this dissertation and corresponding application fields.

1.2 Dissertation Structure

The organization of this dissertation is listed as follows:

Chapter 2 provides an introduction to machine learning and deep learning. This chapter gives a brief background on the important concepts and current algorithms, as well as their clinical applications.

Chapter 3 describes the implementation of Mask R-CNN to search and detect suspicious breast cancers in the entire image dataset of breast MRI examinations. After the location of the suspicious lesion was detected, the malignancy probability was given to evaluate the accuracy of the detected lesion as a true positive cancer, not a false positive benign lesion. Furthermore, the tumor was segmented, and the result was compared to the ground truth to evaluate the accuracy of the detected lesion size.

Chapter 4 shows two image segmentation projects. The first project performed fully-automatic segmentation of the breast and fibroglandular tissues from the non-fat-sat and fat-sat breast MRI using U-net. The second project analyzed lung CT of patients with confirmed COVID-19 infection. Every patient had several longitudinal CT's acquired during hospitalization. The baseline images and the follow-up images were co-registered, first using the Affine registration based on the body areas, followed by the non-rigid registration based on the segmented lung areas. Through the registration, the lesions or infected areas at different locations at different follow-up times could be objectively evaluated, and further segmented for volumetric comparisons.

Chapter 5 shows five projects involving differential diagnosis where we applied radiomics and deep learning methods to cancer imaging. First, CNN was established to differentiate the benign and malignant breast tumors, then to identify molecular subtypes

of breast cancer on MR images. Second, CNN was utilized to classify the benign and malignant vertebral fractures on MR and CT images. Third, we used traditional tumor ROI-based analysis, radiomics, and deep learning to differentiate metastatic lesions in the spine that originated from primary lung cancer and other cancers. Lastly, we established a bi-directional recurrent CNN using Convolutional Long Short Term Memory Network (CLSTM) to diagnose prostate cancer and benign prostatic hyperplasia.

Chapter 6 describes the application of deep learning for radiotherapy planning. Several unsupervised deep-learning models using different strategies were developed to improve cone-beam CT (CBCT) image quality for adaptive radiotherapy and to further validate the model on different anatomical sites.

Chapter 7 demonstrates the application of radiomics method and deep learning methods for the prediction of chemoradiation therapy response in patients with locally-advanced rectal cancer (LARC), using pre- and early-treatment follow-up multiparametric MRI. For the first time, the deep learning method, CNN, is applied to differentiate patients showing different treatment responses, including pathological complete response (pCR) vs. non-pCR, and good response vs. poor response.

Chapter 8 presents two projects involving brain tumors using brain MRI. The first project predicted the progression and recurrence of skull-based meningioma. The second project applied the same methods to predict the recurrence of nonfunctioning pituitary macroadenomas. For these two projects, we used radiomics to establish the prediction model of the progression/recurrence.

Chapter 9 concludes the research included in this thesis, and gives outlooks for further improvement in the future.

Chapter 2. Machine Learning and Deep Learning Algorithms in Image Processing

2.1 Introduction to Machine Learning

Machine learning (ML) is defined as a set of methods that automatically detect patterns in data to predict future data or enable decision making under uncertain conditions [10]. The most prominent characteristic of ML is that it is driven by data, and the decision process is accomplished with minimal human intervention. Machine learning algorithms enable computers to learn from data and improve themselves without being programmed explicitly. Computers are presented with many examples relevant to a task, and they find statistical patterns in these examples that eventually allow the system to automate the task [11]. In classical modeling problems, data is inputted into the model and results are obtained based on pre-defined rules. This is the typical formation of a forward problem, whereas machine learning is designed for inverse problems. That is, the data and results are known and the computer works to establish the model. The generated model can then be applied to new data to produce the original results [10]. Based on the formation of the algorithms, machine learning can be classified into 3 types of algorithms [10].

1) Supervised learning: This is the most widely used type of machine learning algorithm in which a system is presented with labeled data as the ground truth for training. Supervised learning generates a function that reproduces output by inferring it from training data. For this method, the training data is prepared with numerical or nominal vectors that represent the characteristics of the input data and the corresponding output

data. The goal is to approximate the mapping function extremely well that when there is a new input dataset the model can predict the output variables as accurate as possible.

2) Unsupervised learning: In unsupervised learning, a system is presented with unlabeled, uncategorized data and the system's algorithms act on the data without prior training. Since the examples are unlabeled, there is no objective evaluation of the accuracy. Though unsupervised learning encompasses many other solutions involving summarizing and explaining key features of the data, unsupervised learning is similar to a cluster analysis in statistics and focuses on the manner which composes the vector space representing the hidden structure, including dimensionality reduction and clustering [12]. The output is dependent upon the coded algorithms to undiscover the hidden information.

3) Reinforcement Learning: A reinforcement learning algorithm learns by interacting with its environment. The algorithm is rewarded when it performs correctly and penalized when it performs incorrectly. Reinforcement learning is very popular in game development.

In this dissertation, most of the machine learning algorithms are supervised learning models. Based on the function, the algorithms can be divided into three categories: classification, regression, and localization. Classification is when the output variable is a category or discrete number, such as 'red' and 'blue', or 'spam' and 'non-spam'. Regression is when the output is a continuous value. For instance, the output of the stock price prediction is continuous. Localization is designed for image processing. The algorithm will locate the objects of interest on the images. Based on the mathematics of the modeling, the algorithms can be divided into four groups. The first group is linear modeling which is the most basic method. This method was proposed long time ago and widely used for several hundreds of years. The assumption is errors of this model are subject to Gaussian

distribution. By applying the generalization or extension, we can get a generalized linear model, such as logistic regression [10], and support vector machine (SVM) [13-16]. The second group is naïve Bayes method [17]. The assumption is that all of the data are conditionally independent. The third group is K-Nearest Neighbor (KNN) [10]. This method is based on the metrics in the Euclidean space. The last group is Neural Network. Currently neural network is the most popular method in AI field.

A linear model followed by a non-linear activation function will form a perceptron [8, 11, 18]. A neural network is made up of a lot of perceptrons, and it resembles the multilayered human cognition system. The more perceptrons the model has, the greater the power of the network. When the number of network layers becomes very deep, the algorithm becomes very powerful, which is widely known as 'Deep Learning'. Due to its accuracy and efficiency, deep learning has attracted a lot of attention and increased popularity for its use in big healthcare data, where it has exhibited impressive performances in mimicking humans in various fields, including medical imaging.

A typical task in radiology is to detect structural abnormalities and classify them into disease categories. While computer-aided detection (CAD) systems have been developed and introduced in clinical workflow to aid clinicians, deep learning technology has the potential to overcome the limitations of previous CAD systems, achieving greater detection accuracy and increasing productivity by allowing human readers to shift repetitive radiology tasks to AI [2, 12, 19]. Currently all advanced CAD system has utilized deep learning algorithms [12, 19]. However, deep learning has not been widely used in medical applications yet, mainly due to the need of a large dataset to train the model, as well as the need of transfer learning to re-tune the developed model for different settings. When the

dataset is not suitable for deep learning analysis, radiomics provides an alternative strategy, by extracting many features and using machine learning statistical methods to build predictive models, as in the next section [20, 21].

2.2 Radiomics

"Radiomics" involves the high-throughput extraction of quantitative imaging features with the intent of creating mineable databases from radiological images [20-22]. It is based on that such profound analyses and mining of image feature data will reveal quantitative predictive or prognostic associations between images and medical outcomes.

The goal of radiomics is to convert images into mineable data with high fidelity and high throughput. However, with the current state of radiomics, image features have to be extracted automatically and with high throughput, putting a high premium on novel machine learning algorithm development. The radiomics enterprise can be divided into five processes with definable inputs and outputs, each with its own challenges that needs to be overcome: (a) image acquisition and reconstruction, (b) image segmentation and rendering, (c) feature extraction and feature qualification, (d) data selection, (e) processing and linking to outcomes.

Feature extraction and feature processing play key roles in this process [23]. Many studies have focused on imaging feature engineering in the hopes to find features that reflect the patients' pathological information. However, due to the origins and noisy level of the medical images, finding the best features is not a straightforward task.

2.2.1 Feature Extraction

Imaging features are evaluated from the Regions of Interest (ROI) on images. Currently, the popular features analyzed from different imaging modalities in clinical applications can be grouped into Kinetic Features, Pharmacokinetic Features, Morphological Features, and Texture Features [24-26].

(a) Kinetic Features

Kinetic features describe the temporal change of the signal intensities through parameters obtained directly from the time-intensity curve (TIC). They are model-free, heuristic, parameters directly calculated from the entire time course, or from different phases, such as wash-in, wash-out, etc. The commonly analyzed kinetic features include initial area under the curve (IAUC), relative enhancement ratio, enhancement slope, time to peak, basal signal, perfusion index, sum of intensities difference (SOD), etc. [24, 27, 28].

(b) Pharmacokinetic Features

Pharmacokinetic features reflect physiological parameters of tissues and are calculated based on mathematical models according to a model-based strategy [28-31]. They include extracellular extravascular space (EES), plasma space, and transfer constants between the plasma space and the EES. For more complex kinetic models, pharmacokinetic features also include permeability flux, extraction fraction, and capillary transit time [28-31].

(c) Morphological Features

Morphological features describe the shape and topology of the region of interest, e.g. manually drawn or computer segmented lesions. The commonly analyzed morphological

features include area, circularity, compactness, complexity, perimeter, radial length, smoothness, roughness, sphericity, eccentricity, volume, rectangularity, solidity, speculation, convexity, curvature, edge, etc. [24, 32-34]. Many other can also be extracted.

(d) Textural Features

Textural features are commonly used to extract the information related to the intensity distribution patterns or geometric structures on medical images. There are many definitions of texture features as a function of local spatial variation in the intensity of voxels. Texture analysis matrices (GLCM, GLRLM, GLSZM, NGTDM) can be applied to calculate features representing a wide range of patterns associated with heterogeneity. Due to the large number of possible features, texture features play a significant role in radiomics [25, 35-38].

(e) Clinical Features

Clinical features can provide crucial information needed for making important clinical classifications, e.g. for diagnosis, therapy selection, or prognosis prediction, which can be extracted from patient's medical records [39]. The imaging parameters extracted from different MR sequences or maps are often combined with clinical features to form a multi-dimensional model for a single subject. Machine learning (ML) algorithms can be utilized for the combination of clinical and imaging features.

2.2.2 Feature Classification

Several algorithms were commonly used, e.g. support vector machine (SVM), logistic regression (LR), random forest (RF), artificial neural network (ANN), fuzzy C means [2, 10].

(a) Support vector machine

The support vector machine (SVM) is the most popular classification algorithm, and typically exhibits the highest performance ranks for most classification problems, given its advantages of regularization and convex optimization [13-16]. The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space (N is the number of features) that distinctly classifies the data points. In SVM, different kernel functions are applied to transform the original data into specific feature space to select support vectors. Hyperplanes are decision boundaries that help classify the data points. Data points falling on either side of the hyperplane can be attributed to different classes. Also, the dimension of the hyperplane depends upon the number of features. Support vectors are data points that are closer to the hyperplane and influence the position and orientation of the hyperplane. Using these support vectors, we maximize the margin of the classifier. Due to the utilization of the hyperplane, the classification performance is relatively better than other methods [15]. Also, this strategy can overcome the overfitting issue during training. But due to the complicated settings, the required training dataset needs to be larger compared to using other methods.

(b) Logistic Regression

Linear and logistic regression systems are widely used due to their simple architecture [10]. The parameters of linear regression are estimated to ensure the best fit of the straight line in the data space. Logistic regression employs the logistic function to differentiate binomial distributions and is usually used as a classifier. The strategy of logistic function is very simple. The output of the linear model is applied to sigmoid function. All values are

nonlinear rescaled to the range between 0 and 1. Logistic regression is one of the simplest methods in ML. With very few inputs, a relatively general model can be established.

(c) Radom Forest

Random forest consists of a large number of individual decision trees that operate as an ensemble [40]. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model's prediction. A lot of random subsets of features and inputs are utilized to build a large number of relatively uncorrelated trees, which operate as a committee and will outperform any of the individual constituent models [40, 41]. With the uncorrelated trees, the importance of the features can be estimated. Also, this ensemble strategy will never encounter the overfitting problem.

Random forest algorithm runs efficiently on large data set and can handle thousands of inputs without inputs deletions. Also, this method can automatically balance the errors when facing unbalanced data[41].

In every decision tree grown in the forest, the out-of-bag cases are selected, and the number of votes cast for the correct class is counted. Next we randomly permute the values of a specific feature in the out-of-bag cases and put these cases down the tree. Subtracting the number of votes for the correct class in the feature-permuted out-of-bag data from the number of votes for the correct class in the untouched out-of-bag data, the average of this number over all trees in the forest is the raw importance score for this feature. Then features with the highest importance scores can be selected, and features not contributing in providing useful information are eliminated. In this method, the number of examples can be much higher than the number of dimensionalities in the discriminative hyperplane, to eliminate overfitting.

(d) Fuzzy C-Means

Unlike those supervised algorithms mentioned above, fuzzy C-means (FCM) is an unsupervised learning algorithm which does not need corresponding labels. FCM is a data clustering technique where a dataset is grouped into a specific number of clusters with each data point in the dataset belonging to every cluster with a certain degree [42, 43]. FCM algorithm is very similar to K-Means clustering algorithms, and also called as Soft K-Means. K-mean algorithm cluster each data point to a centroid based on the minimized Euclidean distance. But Fuzzy C-means algorithm assigns each data point a weighting associated with a particular cluster [10]. For the results, each data point will get a list of probabilities which can be summed up to 1. Fuzzy-C means will tend to run slower than K means, since it's actually doing more work. Each point is evaluated with each cluster, and more operations are involved in each evaluation. K-Means just needs to do a distance calculation, whereas fuzzy c means needs to do a full inverse-distance weighting.

2.3 Deep Learning

One specific type of the machine learning algorithms is Neural Network. A linear model followed by a non-linear activation function will form a perceptron. Neural Network is a network which is made up of a lot of perceptrons, which resembles the multilayered human cognition system [8, 10, 18, 44, 45].

Artificial intelligence (AI) aims to mimic cognitive, intensive tasks via complex computational models trained on top of existing datasets. A computational model trained using the input from expert readers (radiologists) can automatically perform many clinical

tasks currently done by radiologists based on visual reading, e.g. localize and segment lesions for diagnosis, staging and therapy response evaluation, and provide a potential solution to many clinical problems. Novel AI technologies, such as deep learning models, have been exploited in recent years with impressive results [12, 24, 44, 46-51]. In this section, several popular deep learning architectures will be described, including Artificial Neural Network, Convolution Neural Network, Recurrent Neural Network, U-Net, Residual Neural Network, Long Short term Memory, and Generative Adversarial Network.

2.3.1 Artificial Neural Network

The basic element of neural network is perceptron which is a simple generalized linear model. It takes an input, aggregates it (weighted sum) and returns 1 only if the aggregated sum is more than some threshold else returns 0. The first part is a linear combination of the input vectors, and the second part is a non-linear function, as shown in Error! Reference s ource not found.**1.**

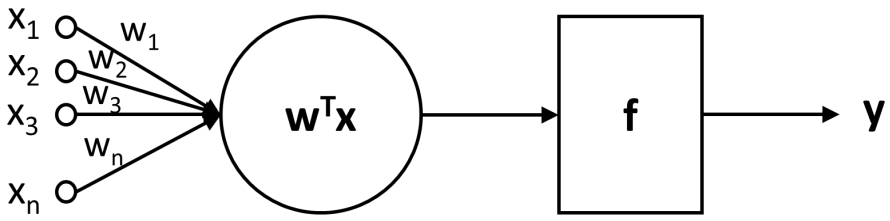

Figure 2-1: Diagram of perceptron. x is the input vector, and w is the weights to be fitted. After obtaining the weighted summation, the non-linear function f is applied to get the output y.

A three-layer back-propagation neural network, known as multi-layer perceptron (MLP) artificial neural network (ANN) was utilized to obtain optimal classifiers. The three-layer topology has an input layer, one hidden layer, and an output layer. From Universal

approximation theorem [52], a feed-forward network with a single hidden layer containing a finite number of neurons can approximate continuous functions. The number of nodes in the input corresponds to the number of input variables. The number of hidden nodes is usually determined by a number of trial-and-error runs. Different neural network architectures with hidden nodes were tested. A stochastic gradient decent with the mean squared error function was used as the learning algorithm. The optimal architecture was chosen as the one for which the validation error was the lowest. With the determined number of hidden nodes, both the learning rate and the momentum coefficient were varied during network training to ensure a high probability of global network convergence. For training, criteria for convergence was met with a root mean squared error less than or equal to a small value or a large number of iterations. With the determined number of hidden nodes, both the learning rate, and the momentum coefficient, were varied during network training to ensure a high probability of global network convergence.

2.3.2 Convolution Neural Network

Convolutional neural networks (CNNs) based deep learning approaches can learn feature representations automatically from training data. The multiple layers of the CNNs aims to process the imaging data with different levels of abstractions, enabling the machine to navigate and explore large datasets and discover complex structures and patterns that can be used for prediction [53].

A CNN is a special kind of neural networks that has been widely applied to a variety of pattern recognition problems, such as computer vision, speech recognition, etc. The CNN was first inspired by Hubel et al. [54] and continually implemented by many researchers.

Some successful implementations of CNN are NeoCognitron [55], LeNet5 [56], HMAX [57], AlexNet [48], GoogLeNet [58], ResNet [59], etc. Different from the conventional machine learning methods, feature extraction, selection, classification procedures are combined into one structure and can be completed by various convolution operations in different layers.

This work focuses on two dimensional convolutional neural networks in particular. The basic idea of CNN is to build invariance properties into neural networks by creating models that are invariant to certain inputs transformation. This idea originates from a problem that often occurs in the feed forward neural networks, especially multilayer feed forward neural network (MLP). The problem is that all MLP layers are fully connected to each other. This removes the spatial information of the inputs which are needed for the computational efficiency.

Unlike ordinary neural networks, CNN has a special architecture. The architecture of CNN usually is composed of a convolutional layer and a sub-sampling layer as presented in **Figure 2-2.** The convolutional layer implements a convolution operation, and the sub-sampling layer implements a sub-sampling operation, known as a pooling. CNN is built based on three basic ideas, i.e., local receptive fields, weight sharing, and pooling.

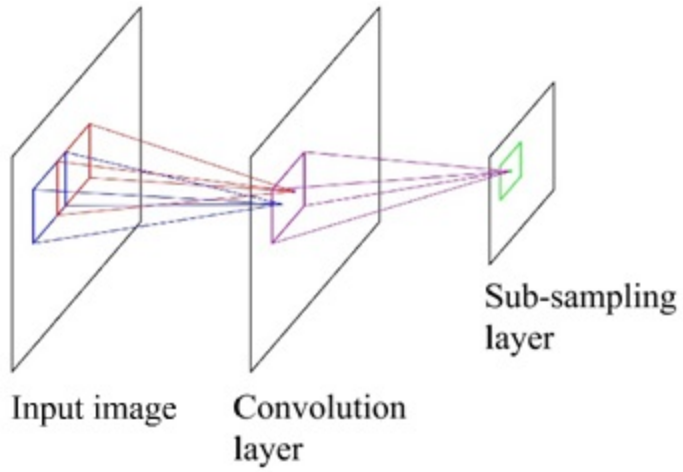

Figure 2-2: Structure of convolution operations.

(a) Local receptive fields

For the feed forward neural network, the input of every neuron is fully connected to all the hidden nodes in the next layer. In contrast, for CNN, each neuron in a hidden layer is only connected to a small field of the previous layer, which is called a local receptive field. For example, if the field has a 3 × 3 area, a neuron of the first convolutional layer corresponds to 9 pixels of the input layer. **Figure 2-3** illustrates the local receptive field by the blue box, representing the neuron in the input layer mapped to the single highlighted pixel in the convolutional layer.

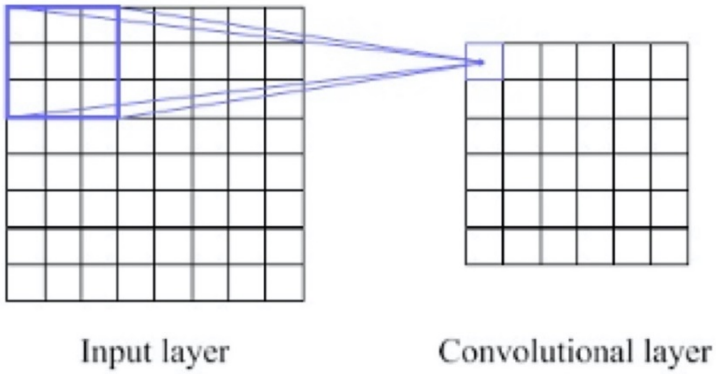

Figure 2-3: The local receptive fields mapping to convolutional layer

(b) Weight Sharing

In the convolutional layer, the neurons are organized into multiple parallel hidden layers, also known as feature maps. Each pixel in a feature map is connected to a local receptive field. For every feature map, all neurons share the same weight parameter that is known as a filter or kernel. This is known as weight sharing. For instance, **Figure 2-4** shows the 5 resultant feature maps obtained when an input of 32 × 32 pixels is trained by a convolutional layer with a 3 × 3 filter and 5 feature maps.

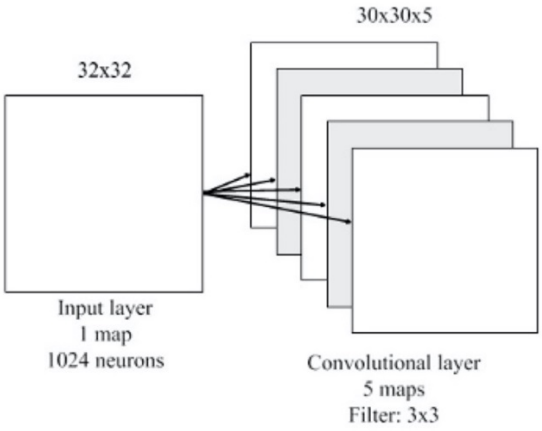

Figure 2-4: An example of a convolutional layer. The filter size is 3 × 3 and 5 feature maps.

(c) Pooling

As described earlier, a CNN contains not only convolutional layers, but sometimes also pooling layers. When there is a pooling layer, it is usually used immediately after a convolutional layer. It means the outputs of the convolutional layer are the inputs to the pooling layer of the network, as shown in **Figure 2-5**. The idea of a pooling layer is to generate translation invariant features by computing statistics of the convolution activations from a small receptive field that corresponds to the feature map. The size of a small receptive field in the pooling layer depends on the pooling size or kernel pooling. **Figure 2-6** illustrates the origin of the pooling layer if there are more than one feature map.

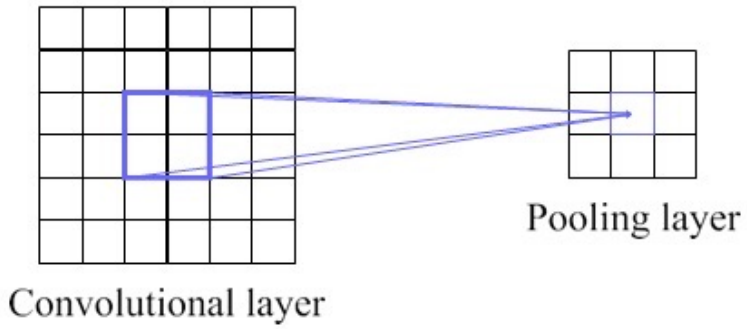

Figure 2-5: Pooling operation between layers

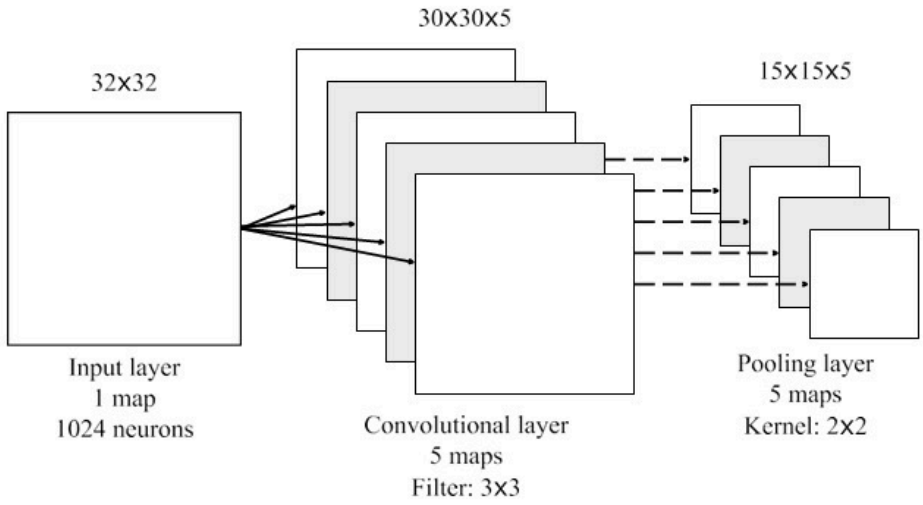

Figure 2-6: An example to show the pooling layer after convolutional layer

(d) Activation Function

Each convolutional operation and pooling are then followed by a nonlinear activation function, σ. Many activation functions have been proposed over the years, however recent studies have shown that the rectified linear (ReLU) activation function has many advantages, including stable gradients at the extreme values of optimization [60, 61]. The ReLU operation is defined simply by:

$$x_l = max(C_l, 0)$$

where the l-th activation map x_l represents the convolutional output, C_l, described above with threshold at zero. Stacking serial convolutional and nonlinear activation functions allows a CNN to model high-order complex feature representations in a mathematically efficient form.

Final classification error was determined using a softmax log-loss function, defined by:

$$y = -\sum_{l}\left(x_{lc} - log\sum_{d=1}^{D} e^{X_{ld}}\right)$$

where the loss, y, is calculated by subtracting the l-th activation map of the ground-truth class, c, with the sum of the softmax normalized (exponential function) values of the remaining class dimensions, D.

2.3.3 Recurrent Neural Network

CNN is designed for image processing. Recurrent Neural network (RNN) are popular models that have shown great promise in many sequential input applications [8, 12, 18, 44, 45]. The key idea of RNN is to make use of sequential or temporal information. In traditional neural network, all inputs are independent of each other and enter the system simultaneously. RNNs perform the same task for every element of a sequence, with the output being depended on the previous computations [62]. RNN is capable of memorizing the information from information about what has been calculated so far. **Figure 2-7** shows the diagram of RNN. In the figure xx, the hidden state S works as the memory of the network which captures the previous information. The output O is calculated based on the memory at time t. The weight sharing scheme in RNN is U, V, and W. During the running period, RNN shares the same parameters (U, V, W above) across all steps. This reflects the

fact that we are performing the same task at each step, just with different inputs. This greatly reduces the total number of parameters we need to learn.

Currently, RNN is a very promising architecture in a lot of fields and have shown great success in many Natural Language Processing (NLP) tasks [8], such as machine translation, text generating, speech recognition, and so on.

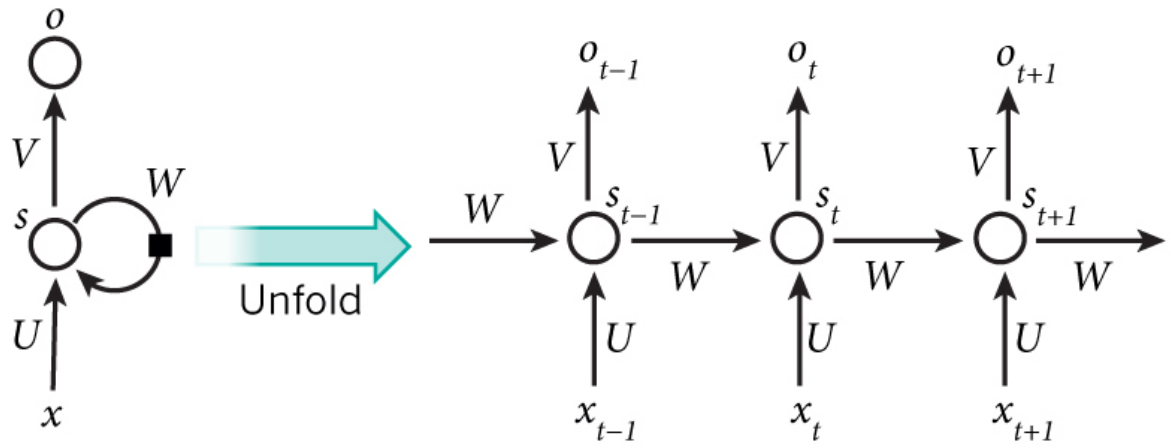

Figure 2-7: Diagram of Recurrent Neural Network [8]

The above diagram shows an RNN being unrolled (or unfolded) into a full network. By unrolling we simply mean that we write out the network for the complete sequence. For example, if the sequence we care about is a sentence of 5 words, the network would be unrolled into a 5-layer artificial neural network, one layer for each word. The formulas that govern the computation happening in an RNN are as follows:

x_t is the input at time step t. For example, x_1 could be a one-hot vector corresponding to the second word of a sentence.

s_t is the hidden state at time step t. It's the "memory" of the network. s_t is calculated based on the previous hidden state and the input at the current step: $s_t=f(Ux_t + Ws_{t-1})$. The

function *f* usually is a nonlinearity such as tanh or ReLU. s_{-1}, which is required to calculate the first hidden state, is typically initialized to all zeroes.

o_t is the output at step *t*. For example, if we wanted to predict the next word in a sentence it would be a vector of probabilities $o_t = softmax(Vs_t)$.

2.3.4 Long Short term Memory

One issue of the general RNN architecture shown in **Figure 2-8** is long-term dependencies [62]. This means the output at any time points can reply on the information or inputs from all old inputs. For the long sequential inputs, the later outputs can gather all of the information from the very beginning inputs which works as a crucial feature of RNN [8, 18, 62]. However, during the backpropagation of the training process, if the new input is added into the system, fewer information can be processed due to the gradients are difficult to be modified. This is called as Gradient Vanishing. As the gradient is back-propagated to earlier layers, repeated multiplication may make the gradient infinitively small. However, in practice, the present output usually depends on the close context, which means closer inputs and states should be weighted more than other. LSTM can be modeled as

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci} \circ c_{t-1} + b_i)$$

$$f_t = \sigma\left(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf} \circ c_{t-1} + b_f\right)$$

$$c_t = f_t \circ c_{t-1} + i_t \circ \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c)$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co} \circ c_{t-1} + b_o)$$

$$h_t = o_t \circ \tanh(c_t)$$

where ° denotes the Hadamard product.

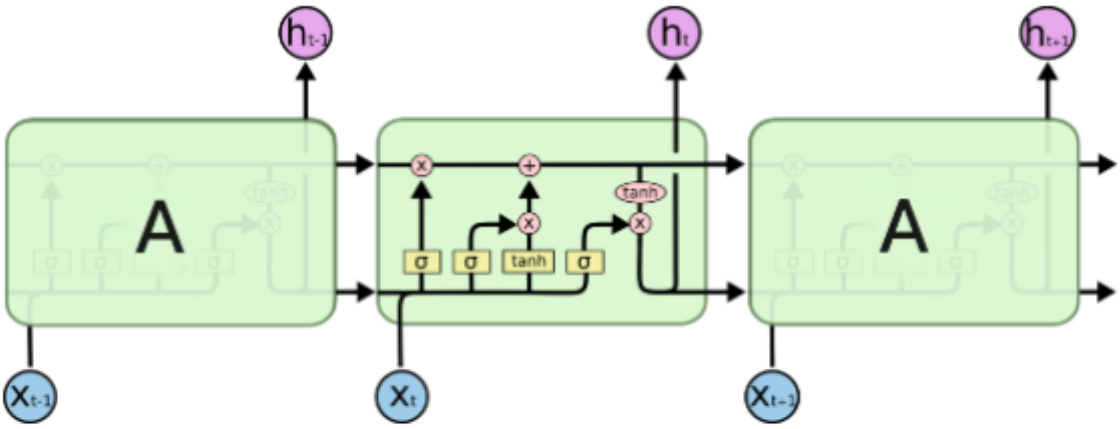

Figure 2-8: Diagram of long short term memory network from [62]

For general purpose of sequence modeling, LSTM as a special RNN structure which has proven to be capable of a lot of sequential models in many applications [62]. LSTM used a critical scheme is a memory cell which serves as a conveyor belt connecting time series and acts as an accumulator. In LSTM, three gates work to control and weight different inputs and hidden states. The first gate is called input gate which controls the extent that current input and past hidden states have on affecting the current cell state. The second gate is called forget gate that is a sigmoid function. The forget gate decides what information is going to be thrown away or dampened from the current cell state. The last gate is the output gate which gather all states and inputs to get the current outputs. Thus, the forget gate makes LSTM focus more on the recent memory with forgetting old information. The output is a filtered version of the current cell state controlled by the output gat and pushed through a tanh function to scale the output value between the range between -1 and 1. With a relatively complicated architecture, the performance of LSTM has been proven to be

improved in many applications. In this dissertation, LSTM has been utilized in several projects to process the image data in a time series, e.g. the pre- and post-contrast images acquired using dynamic-contrast-enhanced MRI (DCE-MRI).

2.3.5 Convolutional LSTM

LSTM is designed for value-based inputs. All of parameters or weights are vectors and the input-to-state and state-to-state transitions are all linear combinations. This is accomplished by the fully-connected layers. In our research, most of the inputs are images [63]. To adjust this, the fully-connected layers are replaced by convolutional layers. Also, the weights become convolutional kernel which reduce the number of parameters.

$$i_t = \sigma(W_{xi} * X_t + W_{hi} * H_{t-1} + W_{ci} \circ C_{t-1} + b_i)$$

$$f_t = \sigma\big(W_{xf} * X_t + W_{hf} * H_{t-1} + W_{cf} \circ C_{t-1} + b_f\big)$$

$$C_t = f_t \circ C_{t-1} + i_t \circ \tanh(W_{xc} * X_t + W_{hc} * H_{t-1} + b_c)$$

$$o_t = \sigma(W_{xo} * X_t + W_{ho} * H_{t-1} + W_{co} \circ C_{t-1} + b_o)$$

$$H_t = o_t \circ \tanh(C_t)$$

2.3.6 Residual Network

According to the universal approximation theorem, with enough capacity, 3 layers fully connected layers can represent any functions. For complicated applications, the number of the layers would be increased to a large number and the architecture can be very massive. Thus, the network is prone to overfitting the input data. Increasing network depth does not

work by simply stack layers together. A large-scale network is very difficult to train because of the overfitting as well as vanishing gradient problem. As the network goes deeper, the corresponding performance get saturated, or even starts degrading.

The core idea of ResNet is 'Residual connection' which skips several convolution layers which can be considered as an identity mapping [59, 64]. By this way, some information from previous layers can be kept so the gradient vanishing problem can be avoided during the training process. Also, wither few layers, the performance can reach a satisfactory level. This means the architecture requires fewer inputs to avoid overfitting. The basic block of ResNet is shown in **Figure 2-9.**

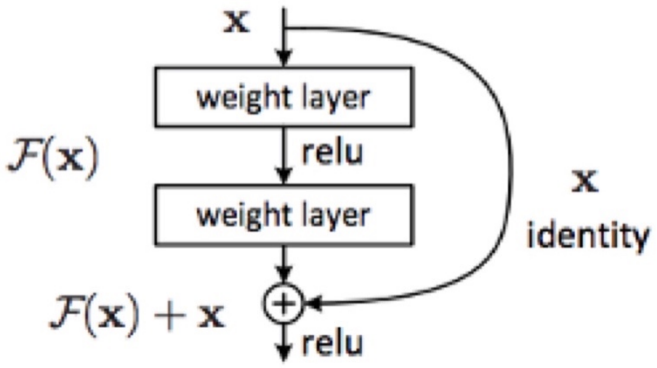

Figure 2-9: Residual blocks from [59]

The ResNet consists of several residual blocks. There are a lot of different predefined ResNet architectures, as shown in

Table 2.1. Each ResNet block is either two layers deep (used in small networks like ResNet 18, 34) or 3 layers deep (ResNet 50, 101, 152) [59]. In this dissertation, ResNet50 is utilized. In ResNet50, each 2-layer block is replaced in the 34-layer net with this 3-layer bottleneck block. They use option 2 for increasing dimensions. This model has 3.8 billion trainable parameters.

Table 2.1: Different Resnet Architectures from [59]

layer name	output size	18-layer	34-layer	50-layer	101-layer	152-layer
conv1	112×112	7×7, 64, stride 2				
conv2_x	56×56	3×3 max pool, stride 2				
		$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$
conv3_x	28×28	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times8$
conv4_x	14×14	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix}\times6$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times6$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times23$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times36$
conv5_x	7×7	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$
	1×1	average pool, 1000-d fc, softmax				
FLOPs		1.8×10^9	3.6×10^9	3.8×10^9	7.6×10^9	11.3×10^9

2.3.7 U-Net

As we mentioned before, segmentation is a main application of machine learning. For the previous architecture, the output is a single scalar so the previous architectures cannot be utilized to segment the images. In Image Segmentation, the machine has to partition the image into different segments, each of them representing a different entity. The output should be able to label all pixels on the images. Thus, the segmentation can be viewed as the pixel-wised classification.

One advantage of CNN is that the imaging features extracted from the images are flexible and can be adaptive to the applications. To obtain the proper features, the features in different spectrum should be considered. The architecture of U-net consists of three sections: The contraction, The bottleneck, and the expansion section. The contraction section is made of many contraction blocks. Each block takes an input applies two 3×3 convolution layers followed by a 2×2 max pooling. The number of kernels or feature maps

after each block doubles so that architecture can learn the complex structures effectively. The bottommost layer mediates between the contraction layer and the expansion layer. It uses two 3×3 CNN layers followed by 2×2 up convolution layer. From the combination of the contraction and expansion sections, the whole architecture looks like a 'U', as shown in **Figure 2-10**.

The key of U-net is the contraction section, one image can be sub-sampled to a small feature map. Meanwhile, different imaging features can be calculated from feature maps with different sizes. These features contain high-frequency components as well as low-frequency components, which means the details of the images and the outline of the images can be processed at the same time. If the segmented objects are rough, the low-frequency components will be weighted more. If the segmented objects contain some detailed information, the high-frequency should be focused more.

Similar to contraction layer, it also consists of several expansion blocks. Each block passes the input to two 3×3 CNN layers followed by a 2×2 upsampling layer. Also after each block number of feature maps used by convolutional layer get half to maintain symmetry. However, every time the input is also get appended by feature maps of the corresponding contraction layer. This action would ensure features that are learned while contracting the image will be used to reconstruct it. The number of expansion blocks is as same as the number of contraction block. After that, the resultant mapping passes through another 3×3 CNN layer with the number of feature maps equal to the number of segments desired.

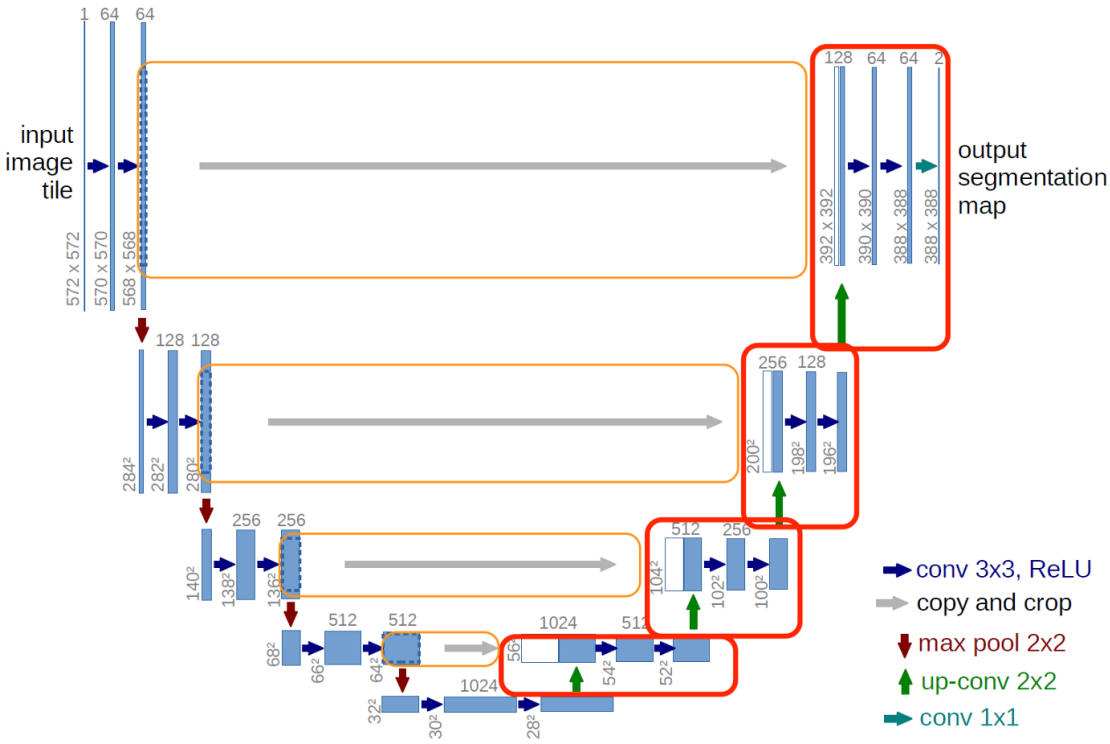

Figure 2-10: An example U-net Diagram

2.3.8 Generative Adversarial Network

Unlike all of the architectures mentioned above, Generative Adversarial Network (GAN) is an unsupervised learning method that involves automatically discovering and learning the regularities or patterns in input data in such a way that the model can be used to generate or output new examples that plausibly could have been drawn from the original dataset. The goal of GAN is to automatically generate data from the noise to get the objective results. GAN has two parts: one is generator, and another is discriminator. The generator model that we train to generate new examples, and the discriminator model that tries to classify examples as either real (from the domain) or fake (generated). The training process of the two parts looks like a zero-sum game. So we call it as 'adversarial'. The

discriminator tries to discriminate the real and fake and the generator tries to produce fake results which are much more similar to real results.

In this dissertation, the GAN model we used is pixel-to-pixel GAN which is closer to supervised learning and designed for image processing. The input images have their corresponding outputs. During the training process, the generator tries to produce the fake images based on the input. Meanwhile, the fake outputs and the real outputs are put into the discriminator. The discriminator works as a classifier to distinguish fake results.

The most important part of the training process is to balance the generator and the discriminator. Due to the origins of the architectures or inputs, the prediction abilities of generator and discriminator are different. If the learning rate of the generator is faster than discriminator, the discriminator will be fooled easily and give up being smart. Then the generator will randomly produce many results without any guidance. If the learning rate of the generator is slower than discriminator, the discriminator cannot be fooled, and the generator will lose the training direction then always produces random regardless of the results of discriminator.

In this dissertation, we used U-net as discriminator to improve the image qualities. However, the traditional metrics, such as mean square errors and cross entropy, cannot correctly reflect the similarity between the outputs and the labels. Thus, the best way to measure the U-net output is the discriminator which works as an adaptive loss function. Besides pix2pix GAN, there are some other GAN architecture's, such as cycle GAN and stack GAN, which have been proven to be successful in some fields.

2.4 Algorithm Implementation

The implementation of radiomics and deep learning can be summarized as image preprocessing, algorithm selection, deep learning configuration and overfitting regularization.

Image Preprocessing

The prediction performance might be various depending on the image quality, especially for CT images. The first step to process CT images is de-noising using Gaussian filters. The filter parameter should be determined by the noise level. For high quality images and filtered CT images, we should do normalization for each slice. One normalization method is to scale all pixel levels from 0 to 1. But this method is heavily affected by outliers. Another method which we used in this dissertation is to set mean as zero and standard deviation as 1. The normalized images can be easily processed in radiomics and deep learning.

Another preprocessing is data augmentation which is also used in overfitting regularization [65]. Unlike de-noising, one easy way to do data augmentation is adding noise. This is to differentiate the augmented images with each other. But the higher noise level can break the tissue textures which increases the difficulty when extracting imaging features. Another augmentation method is affine transformation. Affine transformation including 3 procedures: translation, rotation and scaling. The affine transformation can keep image textures. For segmentation project, affine transformation should be limited to protect the spatial information.

Machine Learning Algorithms Selection

There are numerous algorithms can be used to establish feature selection and classification models. Depends on different applications, different algorithms can lead to different results to meet the requirements of the projects setting. Thus, it's very important to select the appropriate algorithm for specific research. Besides CNN, the most popular algorithms which can always give best performance are SVM and random Forest based on my experience in this dissertation.

Random forest algorithm is fast and scalable. It works better for multi-class tasks and has a very good tolerance for the outliers. Also, the random forest is a resemble method of decision trees. When random forest is applied to select features, the predictors are out-of-bag permuted. This strategy can perfectly solve the issue about unbalanced data. From the general dataset, random forest shows slightly better performance [23]. Their main disadvantage is that they easily overfit, but that's common for ensemble methods.

For imaging data, SVM always give high accuracy [10]. Support vector machine is a special kind of linear model with specific kernel [15, 16]. The kernel in SVM works as a transform which maps input parameters into a different feature space where the transformed data can be divided more obviously. The kernel used in this paper is Gaussian kernel. Other classification models, such as logistic regression and decision tree, work in the original features space. Thus SVM is capable to reach higher accuracy. Meanwhile, the cost function of SVM allows margins between different groups. This can improve the robustness of the model and avoid overfitting during the training process. For this study,

with limited case number, SVM is a best option to balance the variance and bias of the input data.

In the feature selection process, when a few features show dominant capabilities in distinguishing different classes, random forest is a suitable method because of its interpretable and explainable nature [40]. In contrast, if more uncorrelated features should be combined together and some hidden information should be explored for the final classification tasks, SVM will be a better option. Usually, SVM can be utilized to select parameters from a large number of features, and to combine them to reach a high accuracy. Each of the selected features, by itself, cannot contribute much to the final results (i.e. not a dominating feature), but combining all of them is powerful [16]. However, SVM is memory-intensive in our sequential selection scheme. Usually, this process will take several days on a single CPU.

Deep Learning Configuration

Currently, deep learning algorithm can always obtain satisfactory performance for the medical imaging applications in many fields. However, the training of algorithms is very challenging when trying different architectures to explore the potential results. The training process is achieved by updating the weights of the network in response to the errors the model makes on the training dataset. Updates are made to continually reduce this error until either a good enough model is found, or the learning process gets stuck and stops [8, 18]. Theoretically, deep learning models can be thought to learn by navigating a non-convex error surface [44]. As we all known, there exists very mature strategies for convex optimization but challenge to obtain the proper results for non-convex problems.

The optimization algorithm we used in all mentioned projects is backpropagation. Backpropagation refers to a technique from calculus to calculate the derivative (e.g. the slope or the gradient) of the model error for specific model parameters, allowing model weights to be updated to move down the gradient. However, the final performance can be influenced by a lot of factors, such as initialization, learning rate, cost function designing and optimizer. When we start to design the deep learning algorithm, first and most elementary factor should be considered is cost function which must can appropriately reflect the measurement of the learning errors. This step defines the searching direction of the learning process. Some popular cost functions include cross entropy, mean square error, hinge loss, etc. which should be selected based on the requirements of the applications. After determined the cost function, the architecture can be chosen. Meanwhile, the number of the trainable parameters must be considered to control overfitting. Usually, the training cases should 3 or 5 time more than the number of the trainable parameters [10, 18]. From my experience, this parameter scale cannot really avoid overfitting during training. So, some useful methods have been applied to avoid the overfitting. With training and validation datasets, the other hyper-parameters can be adjusted, including initialization, learning rate, optimizer and some hyper-parameters in the overfitting-avoiding methods. The aim of this procedure is to get convergence and cannot obtain successful results under some circumstances. With a better optimized system, the convolutional layer can perfectly extract imaging information to represent the proper characteristics of the aim of the applications.

Overfitting

Overfitting is always the most difficult problem when training a deep learning algorithm, especially for medical images application due to the much smaller case number compared to natural images and cannot reach a satisfactory level. For example, ImageNet used more than 14 million natural images and could achieve the surprising performance [66]. For medical images, the number of inputs is limited by the patient number, equipment qualities, and even some legal or ethics issues. To overcome this disadvantage, there are many methods to deal with the overfitting during the training process.

A commonly used method is data augmentation, in which the volume of the input dataset can be increased manually [8]. Two popular methods to augment the data is random affine transformation and/or adding background noise.

Another technique to avoid overfitting is to add regularization terms to the cost functions [10]. The original cost function is derived from the maximum likelihood (ML) estimation. If some priors of the parameters are determined beforehand, the cost function is derived from the maximum a posterior (MAP) estimation. If the prior belongs to Gaussian distribution, the normalization terms to be added is L2 norm, named as Bridge. If the prior belongs to Laplacian distribution, the normalization terms to be added is L1 norm, named as Lasso. Usually Lasso can lead to sparser parameters during training processing. These normalization terms can regularize the training process based on the regularization coefficient which is pre-defined based on the inputs.

Recently, a method, called Dropout [61], has been widely applied to the majority of deep learning applications, including this proposal. Dropout is a technique where randomly selected neurons are ignored or “dropped-out” during training. This process is random.

Consequently, this means that their contribution to the activation of downstream neurons is temporally removed on the forward pass and any weight updates are not applied to the neuron on the backward pass. The effect is that the network becomes less sensitive to the specific weights of neurons. This in turn results in a network that is capable of better generalization and is less likely to overfit the training data.

Transfer learning become popular to improve the deep learning performance. Some research indicated that transfer learning can be utilized to reduce the input case number [19, 67-71]. With a pre-trained model, the smaller inputs are used to fine-tune the network weights. To obtain the best performance, all of these methods should be searched and tried concurrently.

Chapter 3. Automatic Detection of Breast Cancer on MRI

Detection is an emerging technique in machine learning. In medical images, a detection technique is performed to identify the areas where the patients' lesions are located as box coordinate. In this chapter, the detection method is implemented to localize the breast cancers on the MR images. Detection and treatment of breast cancer in its early stage is very important to increase patient survival [72].

3.1 Motivation and Clinical Applications

Breast MRI is a well-established clinical imaging modality for diagnosis of breast cancer. Compared to mammography and ultrasound, dynamic contrast-enhanced (DCE) MRI is the most sensitive modality for lesion detection and diagnosis, and also for screening of high-risk women and evaluating response to neoadjuvant chemotherapy [73-76]. In the clinical setting, the evaluation is done by radiologists' visual interpretation. The suspicious abnormality should be identified first, and then further characterized. Since many sequences with thin slices were acquired to cover the entire breast with hundreds of images, it would take some time and effort for a radiologist to carefully evaluate the entire dataset. Therefore, the reading was usually done with the assistance of computer-aided diagnosis (CAD) software, that generates subtraction images, maximum intensity projection (MIP), color-coded DCE wash-out maps and DCE time course, etc., and displays them together on the workstation for evaluation. The morphological and the temporal information were interpreted by a radiologist and combined to determine the level of malignancy based on BI-RADS categories [77].

The diagnostic sensitivity and specificity of breast MRI can be affected by several factors, e.g. radiologists' experience [78, 79], magnetic field strength [80] and DCE-MRI protocol [81-83]. The current CAD system displayed essential information to improve workflow efficiency and diagnostic accuracy, especially for patients with multiple lesions or satellite lesions [84]. Many studies have further attempted to characterize the abnormal lesions and give a final diagnosis [85-88]. Most of them applied computer algorithms to extract features and build a diagnostic model, but not very successful due to the limited information provided by pre-defined features [89]. For developing a fully-automatic CAD system, the first required task is to detect abnormal lesions, which is rarely reported.

In recent years, artificial intelligence (AI) algorithms, particularly deep learning, have demonstrated remarkable progress in medical image analysis for performing many clinical tasks. Convolutional Neural Network (CNN) is a common deep learning method that can be applied to give probability of malignancy for identified lesions [12] [90]. It can be further applied to perform search in the entire MRI dataset to detect abnormal lesions [89, 91, 92]. Patch-based CNN is used to discriminate whether each patch (small portion of images) belongs to lesion or not [91, 93-95]. Another approach uses CNN [96, 97] or Mask Regional-Convolutional Neural Network (R-CNN) [98] to search the whole image or feature map to detect and localize the lesion.

The purpose of this study was to implement Mask R-CNN to search and detect suspicious lesions in breast MR images [99]. The architecture provides a flexible and efficient framework for parallel evaluation of region proposal (attention), object detection and segmentation [100-102]. After the location of the lesion is detected, the tumor is further segmented, and the result is compared to the ground truth.

3.2 Subjects and Image Dataset

Patients and Datasets

The Institutional Review Board approved this retrospective study and requirement for informed consent was waived. Only patients with confirmed breast cancer that presented as mass lesions were studied. A dataset obtained from one hospital with 241 patients (mean age 49 y/o, range 30–80 y/o) was used for training. Another dataset from a different hospital with 98 patients (mean age 49 y/o, range 22-67 y/o) was used for testing.

MR Imaging Protocols

For the training dataset, breast MRI was performed on a 3T scanner (Trio-Tim, Siemens, Erlangen, Germany). The DCE-MRI consisted of 7 frames, including one pre-contrast and six post-contrast acquisitions using non-fat-sat sequence, with TR/TE= =280/2.6 msec, flip angle=65^{o}, matrix= =512×343, field of view=34cm, and slice thickness=3mm. Since the lesion laterality was known in each patient, only the breast with cancer was analyzed. The testing dataset was acquired using a 1.5 Tesla scanner (Magneton Skyra, Siemens Medical Solutions, Erlangen, Germany). DCE-MRI was acquired using a fat-suppressed three-dimensional fast low angle shot (3D-FLASH) sequence with one pre-contrast and four post-contrast frames, with TR/TE=4.50/1.82 msec, flip angle=12°, matrix size=512x512, field of view=32cm, and slice thickness=1.5 mm.

Ground Truth Segmentation

The tumor was segmented on the subtraction images (post-contrast Frame-3 subtracting pre-contrast Frame-1) using the fuzzy c-means (FCM) clustering algorithm

[103]. A square ROI was placed on maximum intensity projection to indicate the location. The tumor within the selected ROI was enhanced using an un-sharp filter with a 5x5 kernel constructed using the inverse of the two-dimensional Laplacian filter. FCM algorithm was applied to obtain the membership map of all voxels indicating the likelihood of each voxel belonging to the tumor or the non-tumor cluster. The ground truth was verified by an experienced radiologist and corrected if necessary. Based on the segmented tumor mask, the smallest bounding boxes covering the lesions were computed to evaluate the deep learning detection algorithm.

3.3 Mask R-CNN Architecture

The deep learning detection algorithm was implemented using a custom architecture derived from the Mask R-CNN (22), shown in **Figure 3-1.** Firstly, various pre-defined shape and distribution of bounding boxes were placed to identify potential abnormality on the entire image. Then the bounding boxes were ranked based on the likelihoods. Several bounding boxes on each slice with highest probabilities were extracted to generate region proposals to locate specific regions. These composite region proposals were pruned using non-maximum suppression and used as input into a classifier to determine whether it belonged to breast lesions. For the positive tumor detection, a segmentation network was added to determine the tumor boundary with binary masks. The image features for various parallel detection, classification and instance segmentation were extracted from the backbone network. In this study, we used ResNet101 as the feature pyramid network (FPN) to work as the backbone [59]. In residual network, the learning was implemented by the bottle-neck block which started with one 1x1 convolutional layer to extract a specific

number of feature maps, then connected with a 3x3 convolutional layer, and lastly connected with one 1x1 convolutional layer. In ResNet101 network, there were 33 residual blocks, and the parameters were initialized using ImageNet. The number of input channel was 3, including the contrast-enhanced and pre-contrast image of the lesion side, and the contrast-enhanced image of the contralateral normal side. The inputs from the FPN bottom-up pathway were added to the feature maps of the top-down pathway using a projection operation to match matrix dimensions as shown in **Figure 3-1.**

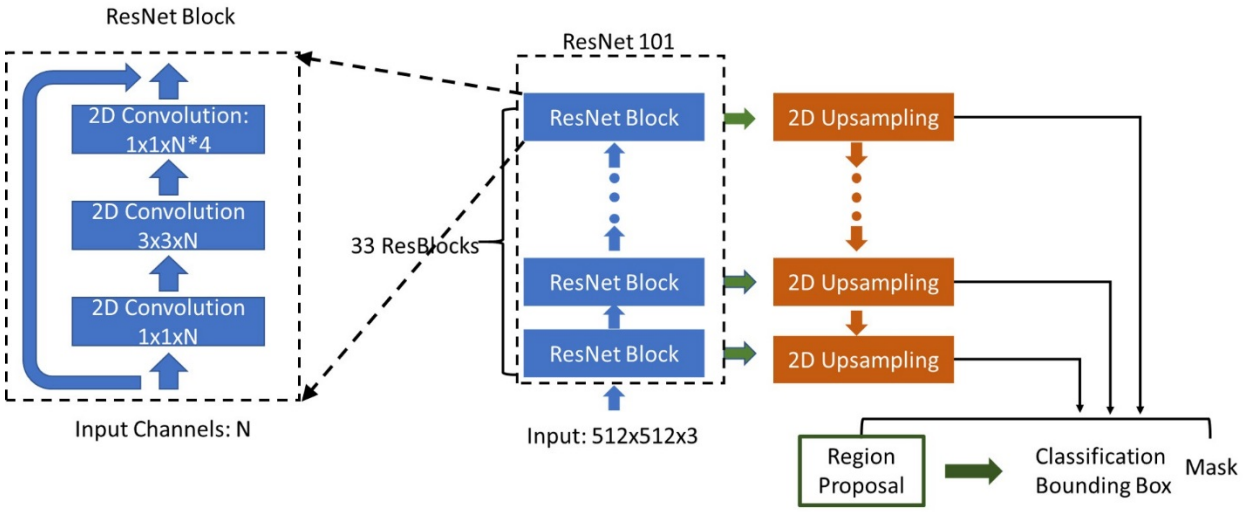

Figure 3-1: Mask R-CNN architecture. Hybrid 3D-contracting (middle block) and 2D-expanding (right block) fully convolutional feature-pyramid network architecture used for the mask R-CNN backbone. The architecture incorporates both traditional 3 x3 filters as well as bottleneck 1x1–3x3–1x1 modules (left block). The contracting arm is composed of 3D operations and convolutional kernels. The number of input channel is 3.

All images were resized to in-plane resolution matrix of 512x512. The pixel intensities on each slice were normalized to mean=0 and standard deviation=1. The mask R-CNN architecture was trained using 128 sampled ROI on one image. The ratio of positive samples and negative samples was fixed at 1:3. The top 256 proposals from FPN were pruned using non-maximum suppression, which could be used for the bounding box

regression. The anchors spanned 4 scales (128×128, 64×64, 32×32, 16×16) and 3 aspect ratios (1:1, 1:2, 2:1) [99]. The final loss function was focal loss including a term for L2 regularization of the network parameters [104]. All models were trained with Adam optimization. The learning rate was set to 0.0001 with momentum term 0.5 to stabilize training [105]. In the training dataset, 10-fold cross-validation was used to evaluate the performance. Ninety percent of the data was randomly assigned into the training cohort while the remaining 10% was used for validation. This process was repeated 10 times until each study in the entire dataset was used for validation once. After fine-tuning, the final trained network was applied to the independent dataset for testing. Since it was not reliable to detect very small lesion < 3 mm, if a lesion was detected only on a single slice without involving any of the neighboring slices, it was dismissed. This study was implemented in Python 3.6 using the open-source TensorFlow 1.4 library (Apache 2.0 license) [106]. Experiments were performed on a GPU-optimized workstation with four NVIDIA GeForce GTX Titan X cards (12GB, Maxwell architecture). Inference benchmarks for speed were determined using a single-GPU configuration.

3.4 Evaluation of Tumor Location and Segmentation

Intersection over Union (IoU), defined as the ratio between the prediction result and the ground truth, was utilized to evaluate the accuracy of the predicted tumor bounding boxes. The prediction was true positive if IoU was ≥ 0.5. The case with IoU < 0.5 was false negative. On the image slice which did not contain lesion, if no bounding box was detected the prediction was true negative; if any lesion was detected it was false positive. For each

true positive lesion, the segmented tumor was compared to ground truth using the Dice Similarity Coefficient (DSC), and the overall accuracy based on all pixels.

3.5 Detection Results

Determination of Three Inputs into Network

The architecture allows 3 input channels. Firstly, the pre-contrast image, post-contrast image and the subtraction image of the diseased breast were used as inputs. **Figure 3-2** shows an example, in which the parenchymal enhancements in bilateral breast are identified as possible lesions. When the post-contrast image was replaced by the subtraction image from the contralateral normal breast, the symmetry could be used to eliminate the false detection of bilateral parenchymal enhancements. Therefore, the three inputs were determined as pre-contrast (used to identify chest region), and the subtraction images from the diseased breast and the contralateral normal breast.

Performance in Training Dataset

For the training set, the performance was evaluated using 10-fold cross-validation. There were a total of 1,469 positive slices containing lesions, and 9,135 negative slices without lesions. Based on the IoU, there were 1,245 true positive, 7,834 true negative, 1,301 false positive, and 224 false negative cases. The sensitivity of tumor detection was 0.85, the specificity was 0.86, and the overall accuracy was 0.86. In the 1,245 true positives, the tumor was segmented and compared to the ground truth to calculate the DSC. The mean value was 0.82, ranging from 0.64 to 0.97 in 10-fold cross-validation. **Figure 3-3**, **Figure 3-4** and **Figure 3-5** show three examples illustrating different detection results.

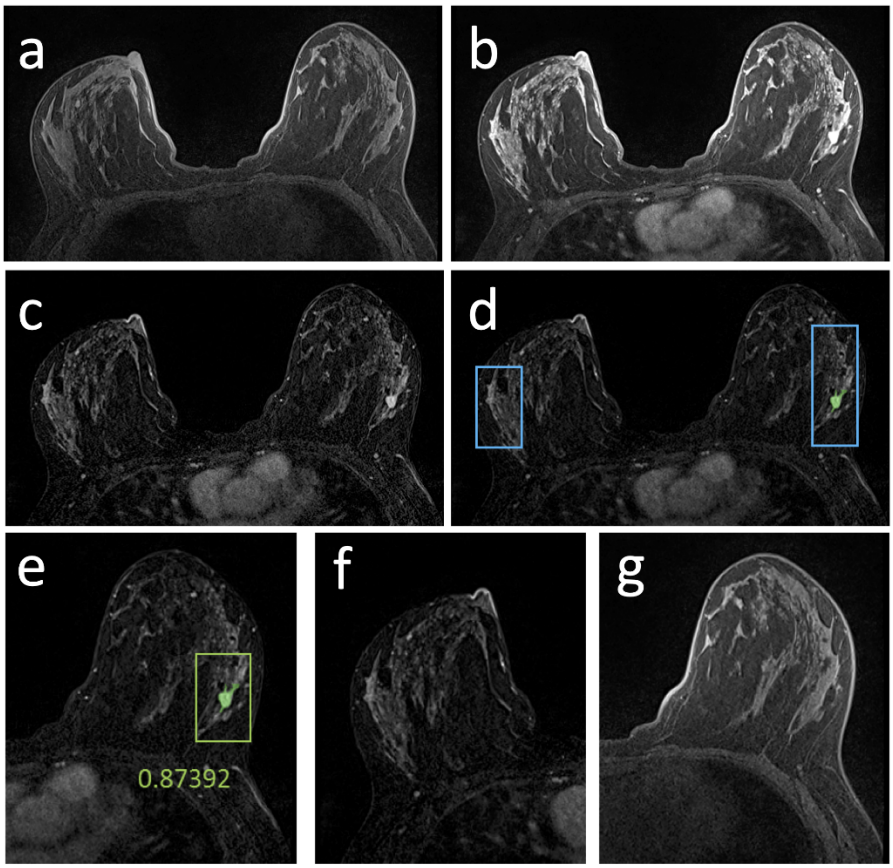

Figure 3-2: One case example from a 62-year-old patient with a small mass lesion, who also shows strong parenchymal enhancement in both breasts. (a) Pre-contrast image; (b) The 2nd post-contrast image; (c) The subtraction image; (d) Tumor detection result searched by the algorithm when using pre-contrast, post-contrast and subtraction images as inputs. Two large blue boxes are the detection output from Mask R-CNN. The box in the left breast (right side of image) correctly encloses the cancer, but it also contains the surrounding parenchymal enhancement and much larger than the size of the cancer. Another blue box in the right breast (left side of image) wrongly detects the parenchymal enhancement, thus a false positive result. (e-g) When the subtraction image, contralateral subtraction image, and pre-contrast images are used as inputs, the small cancer is correctly diagnosed with probability=0.87.

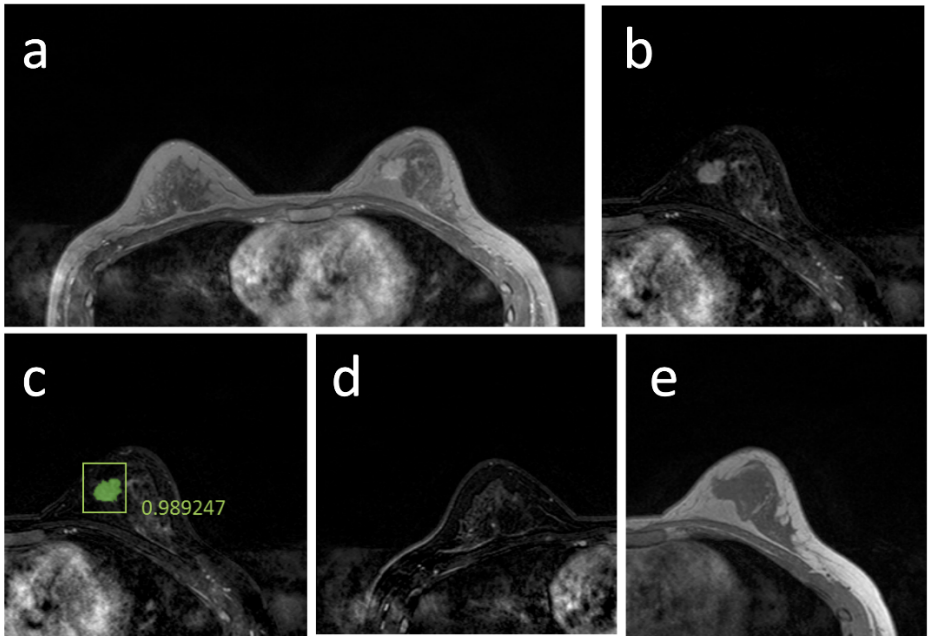

Figure 3-3: True positive case example from a 41-year-old patient with a strongly enhanced mass lesion. (a) Post-contrast image; (b) The subtraction image; (c) Tumor detection result searched by the algorithm. The segmented tumor is highlighted by green color, and used as the ground truth. The red box is the output from Mask R-CNN, which correctly detects the location of the cancer with probability=0.99, a true positive result. (d) The subtraction image of the contralateral normal breast. (e) The pre-contrast image, used as one input to identify the breast region, so the enhancements from the heart can be excluded. (c-e) are used as the 3 inputs into the Mask R-CNN.

Performance in Testing Dataset

The testing dataset had 8,832 slices, 1,568 positive and 7,264 negative slices. The model developed from the training set was directly applied to evaluate the performance. There were 1,254 true positive, 5,396 true negative, 1,895 false positive, and 314 false negative slices. The sensitivity was 0.80, the specificity was 0.74, and the overall accuracy was 0.75. In true positive cases, the mean DSC of the segmented tumor was 0.79.

Factors Associated with False Detection

To understand the possible factors leading to the false predictions, we further analyzed tumor size, tumor enhancement, parenchymal enhancement and tumor locations in the

different diagnostic group. The small tumor was difficult to be detected, which was the main reason for false negative prediction. The mean tumor area calculated from all slices was significantly larger in true positive compared to false negative groups (area 355 mm^2 vs. 42 mm^2, $p<0.01$). Although the entire image, including the chest region with contrast enhancement from the heart, was used in the search, there were only a few false positives inside the chest region. For the tumor segmentation, the difference between the predicted tumor and the ground truth was mainly coming from the parenchymal enhancement, especially for cases with severe field inhomogeneity from strong bias field.

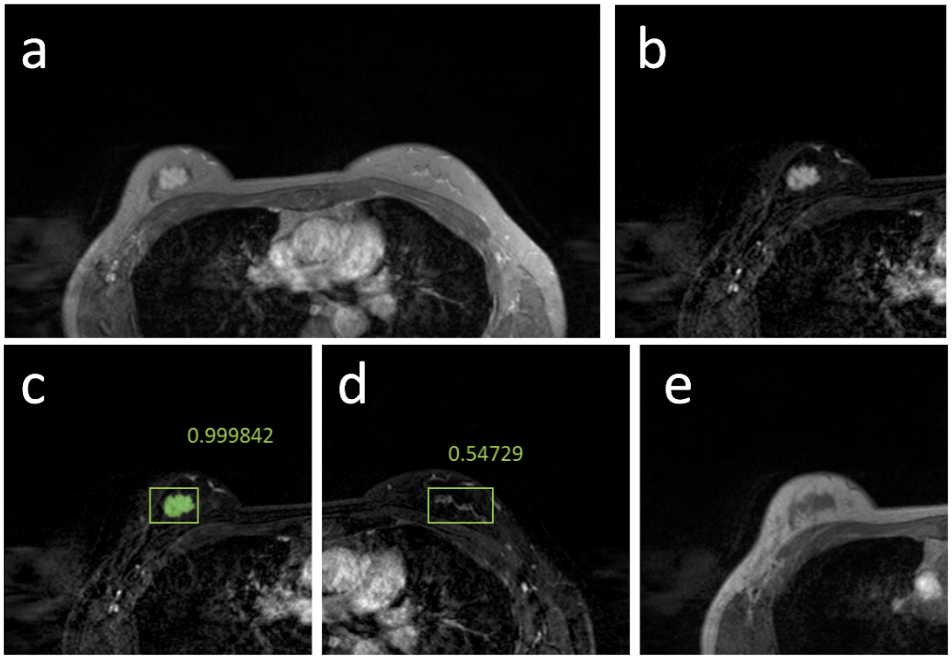

Figure 3-4: True positive and false positive case example from a 39-year-old patient with a strongly enhanced mass lesion. (a) Post-contrast image; (b) The subtraction image; (c) Tumor detection result searched by the algorithm. The green box is the output from Mask R-CNN, which correctly detects the location of the cancer with probability=0.99. (d) The subtraction image of the contralateral normal breast. In this breast, an area with probability=0.54 is detected, a false positive result. (e) The pre-contrast image, used as one input to identify the breast region.

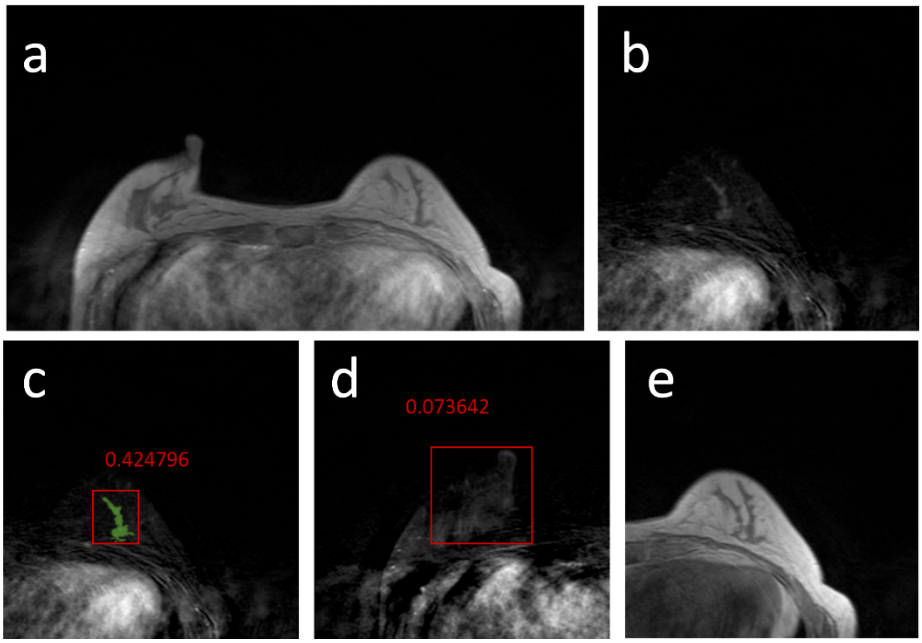

Figure 3-5: False Negative case example from a 57-year-old patient with a mildly enhanced, but pathologically confirmed cancer. (a) Post-contrast image; (b) The subtraction image of the breast which contains the tumor; (c) Tumor detection result searched by the algorithm. The segmented tumor is highlighted by green color, and used as the ground truth. The red box is the result of Mask R-CNN, with the malignant probability=0.42, a false negative result. (d) The subtraction image of the contralateral normal breast. A true negative result, with probability=0.08, is marked as an example. (e) The pre-contrast image, used as one input to identify the breast region.

3.6 Summary and Discussion

In this chapter, we implemented a fully automatic deep learning method using Mask R-CNN for detection of breast cancer by searching the entire set of MR images. Many studies have investigated the value of machine learning, including radiomics and deep learning, for differentiation of benign and malignant lesions, but they all focused on characterization of already identified abnormal lesions. The detection was a much more challenging task, especially in MRI where many images were acquired to cover the entire breast. The results showed that Mask R-CNN was a feasible method that achieved the mean accuracy of 0.86 in the training dataset, and 0.75 in the independent testing dataset. In

detected lesions, the segmented tumor was also in good agreement with the ground truth, with DSC of 0.82 in training dataset and 0.79 in testing dataset.

The chest region includes the enhancements from the heart. While it is very easy for a human reader to dismiss them, the task is difficult for the computer. One commonly used strategy is to segment the breast first and only perform the search within the breast [97], but this requires one more pre-processing step and not easy to achieve a clean breast segmentation. Deep learning offers a fully-automatic strategy. We demonstrated that by including the pre-contrast image as one input, which clearly demarcated the background and chest region, it provided anatomic information and helped to dismiss enhancements from the heart in the chest region. The results also demonstrated that by including the contralateral subtraction image as one input, it helped to eliminate false positives coming from parenchymal enhancements, as shown in Figure 2. Using bilateral breast symmetry as reference is very important in radiologists' visual interpretation, and it can be implemented in deep learning as well, by including the contralateral breast as one input.

Deep learning is an emerging method which has been shown capable of searching and detecting abnormalities in pathology images. For example, Bejnordi et al. [107] used multiple AI methods to detect breast cancer lymph node metastases on pathology whole-slide images. For radiology, the earliest application is for detecting pneumonia on chest X-Ray [108]. For breast lesions, most studies were for 2D mammography, and then extended to DBT that consisted more images in one dataset. Kooi et al. applied patch-based method to mammography [94], which divided the whole image into many small portions for local recognition. Samala et al. applied the method to DBT by using a pre-trained model from mammography [95]. Besides patch-based method, another feasible method is weakly

supervised learning. Kim et al. utilized this method to detect and localize lesion from the 4-view digital mammograms [96], similar to the reading of radiologists in clinics. A residual neural network was implemented with input of 4 views. The feature maps before the global pooling layers were extracted to give the probability maps, indicating the detected lesion location and the level of suspicion. After the lesion was identified, it could be further segmented and characterized to make a diagnosis as benign or malignant. This streamlined procedure has been implemented as a commercial product. Ribli et al. implemented the faster R-CNN algorithm using VGG16 as backbone network to detect lesions on digital mammograms [98], and reached sensitivity of 0.9. The strategy was similar to our Mask R-CNN.

For lesion detection on MRI, because many images were acquired with different pulse sequences, it was much more challenging compared to mammography and DBT. Wang et al. [109] designed a Siamese Network to detect metastatic lesions in the spine using the patch-based method. Dalmış et al. also applied the patch-based method to localize breast cancers on DCE-MRI [91]. The candidate areas for patch extraction were first identified by U-net, and then a Siamese neural network was applied for detection, using the 3D patch and the symmetrical patch from the contralateral breast as inputs. This method obtained a sensitivity of 0.83 for mass tumors. The weakly supervised learning has also been implemented to predict the presence of cancer in DCE-MRI by Zhou et al. [97]. A dense net was applied within the segmented breast areas, and the suspicious lesion locations were calculated from the feature maps. Based on the detection results, conditional random field was employed to estimate the tumor boundary, but the DSC was only 0.51. Our presented method using Mask R-CNN applied the region proposal network

to search suspicious regions within the entire image, which has been shown as a sensitive approach, as in [98]; and also, including the contralateral normal breast as one input could utilize the symmetry to improve specificity, as in [91]. This Mask R-CNN has been applied to search, detect, and diagnose brain hemorrhage on head CT, and achieved very high accuracy of 0.97 [100].

For most object detection algorithms, the high sensitivity is usually associated with high false positive. The Mask R-CNN is not a single shot algorithm, and can increase specificity [99]. In all of the selected regions, they were ranked to extract those with high probabilities. Then the bounding boxes were regressed to generate lesion masks. Furthermore, if a lesion was only detected on a single slice without involving neighboring slices, it was smaller than 3 mm and unlikely to be a true lesion. These additional processing steps could improve specificity while maintaining a reasonable sensitivity. The drawback was that, compared to other architectures, the training became much more complicated, and might take longer and need more training cases.

The major limitation was the small case number and the unbalanced data. For each patient, the number of positive imaging slices containing the lesion was much smaller than the negative slices, which was an inherent problem for lesion detection on MRI. The unbalanced input might lead to unstable training. Second, the training dataset was non-fat-sat images, and the testing dataset was fat-sat images; although this represented a realistic clinical scenario, not an optimal setting for evaluating the performance of the developed model. For different datasets, such as images acquired using different protocols or different MR systems, transfer learning could be applied, i.e. to use part of the testing dataset to re-tune the developed model, which can be implemented when more cases are available.

Lastly, only cases with well-defined mass lesions were analyzed. For non-mass-like enhancements, it was very difficult to be detected by any computer algorithms.

In conclusion, we developed a Mask R-CNN method for detection and segmentation of breast cancer, by searching the entire set of breast MRI. The algorithm allowed the search on the whole image without prior breast segmentation, and reached an accuracy of 0.86 in the training dataset. The inclusion of the pre-contrast image and the contralateral subtraction image as inputs could help to eliminate false positives coming from the heart and the normal parenchymal enhancements, and achieved a high specificity of 0.86. In the detected lesions, the DSC of the segmented tumor compared to ground truth was 0.82. The model could be applied to fat-sat images acquired using a different MR system, and achieved accuracy of 0.75 and DSC of 0.79. The results suggest that Mask R-CNN has a potential to be further optimized for detection of breast lesions in MRI, which can then be integrated with other algorithms to develop a fully-automatic, deep learning-based, breast MRI diagnosis system.

Chapter 4. Segmentation of Breast and Fibroglandular Tissue and COVID-19 Lung Infection Lesions

4.1 Automatic Breast and Fibroglandular Tissue Segmentation on MRI

4.1.1 Motivation and Clinical Applications

Breast density is an established risk factor for the development of breast cancer. Measurement of breast density is mostly performed on two-dimensional (2D) mammography. While two quantitative volumetric analysis tools (Volpara and Quantra) are commercially available to measure dense tissue volume, studies have found that they tend to underestimate the percent breast density in women with dense breast [110, 111]. Furthermore, differences between Volpara and Quantra alone have been found to be as high as 14% [112]. A fundamental limiting factor of all mammography-based density quantification methods is the characteristic 2D overlapping tissues on mammography.

Breast MRI is an established clinical imaging modality for high-risk screening, diagnosis, pre-operative staging and neoadjuvant therapy response evaluation. The most common clinical indication was diagnostic evaluation (40.3%), followed by screening (31.7%) [111]. Passage of the breast density notification law has had a major impact on MRI utilization. Basically the law required that women with dense breast need to be informed of: 1) they have dense breast; 2) the breast density may limit the efficacy of mammography screening; 3) a high breast density is associated with increased breast cancer risk; and 4) other imaging methods can be used for supplemental screening. After the law in California went into effect on April 1, 2013, the use of MRI screening increased from 8.5% to 21.1% in non-high-risk women [113]. Furthermore, as early results of the

abbreviated MRI protocols are promising, this may reduce the cost of MRI for patients allowing for wider use in women with dense breasts and women with mild to moderate cancer risk for screening [114].

The increasing popularity of breast MRIs have led to the fast accumulation of large breast MRI database. This offers a great opportunity to address some clinical questions regarding the use of breast density, e.g. whether the volumetric density can be incorporated into risk models to improve the prediction accuracy [115], or be used as a surrogate biomarker to predict hormonal treatment efficacy [116, 117]. Since MRI is a three-dimensional (3D) imaging modality with distinctive tissue contrast, it can be used to measure the fibroglandular tissue (FGT) volume. However, because many imaging slices are acquired in one MRI, an efficient, objective, and reliable segmentation method is needed. Various semi-automatic [103] and automatic [118-120] breast MRI segmentation methods have been developed in T1 weighted [43] or Dixon-based images [43, 121]. Some operator interventions and post-processing manual corrections may be needed, which are subjective and time-consuming. Therefore, despite of the great progress so far, the efficiency and accuracy need to be further improved for clinical use of MR-measured density. A fully-automatic method that can achieve a high accuracy will be very helpful for exploring and implementing the application of quantitative breast density in clinical settings.

In recent years, deep learning algorithms have been widely applied for classification applications, and they also provided an efficient method for organ and tissue segmentation, including the brain [122, 123], head and neck [124], chest and heart [125, 126], abdomen and pelvis [127-129], breast [130-132], and bone and joint [133]. Since most medical images have high resolutions, patch-based approach is commonly employed for

segmentation, where images are divided into small patches with a specified size as the input of the neural network [109, 132, 134]. This method can fully utilize the local information of the focused area. However, for large structures like the entire organ, a large receptive field for pixel classification is required [135]. The Fully-Convolutional Residual Neural Network (FC-RNN), commonly noted as U-net, is another algorithm that can search a large area [124, 130, 131, 133]. and has been shown suitable for segmenting the whole breast and FGT [130, 131]. Dalmış et al. first applied deep learning for breast MRI segmentation, and demonstrated improved efficiency over an atlas-based method [131].

The purpose of this study was to develop and validate a deep-learning segmentation method based on the U-net architecture, first for breast segmentation within whole image, and then for FGT segmentation within the breast on non-fat-sat T1-weighted MRI. The developed model using a training dataset was tested in independent validation datasets acquired using four different MR systems. Then we applied FC-RNN, or U-net, for segmentation of breast and FGT on fat-sat images. Two datasets from different hospitals were used, one for training the other for independent testing. In addition, the benefit of transfer learning was investigated. The previous model developed for segmentation of non-fat-sat images was used as the basis, and re-trained for fat-sat images. The results obtained without and with transfer learning were compared.

4.1.2 Subjects and Image Dataset

Non-fat-sat Training Dataset

The initial dataset used for training included 286 patients with unilateral estrogen receptor positive, HER2-negative, lymph node-negative invasive breast cancer (median

age, 49 years; range, 30–80 years), as reported in a recent publication [136]. In this study only the contralateral normal breast was analyzed. MRI was performed on a 3T Siemens Trio-Tim scanner (Erlangen, Germany), and the pre-contrast T1-weighted images without fat suppression were used for segmentation. The Institutional Review Board approved this retrospective study and requirement for informed consent was waived.

Non-fat-sat Independent Validation Datasets

The validation dataset included 28 healthy volunteers (age 20–64, mean 35 years old), as described in a previous paper [137]. These women were recruited to participate in a non-contrast breast density study. Each subject was scanned using four different MR scanners in two institutions, including GE Signa-HDx 1.5T, GE Signa-HDx 3T (GE Healthcare, Milwaukee, WI), Philips Achieva 3.0T TX (Philips Medical Systems, Eindhoven, Netherlands) and Siemens Symphony 1.5T TIM (Siemens, Erlangen, Germany). Non-contrast T1-weighted images without fat suppression were used for segmentation. Since both left and right breasts were normal, they were analyzed separately, so there was a total of 56 breasts. The validation was done using the 56 breasts acquired by each scanner first, and then using all 224 breasts acquired by all 4 scanners together. With a cases number of more than 200, it should be sufficient to do independent validation.

Fat-sat Training Dataset

The fat-sat training dataset had 126 women (mean age 48.5 y/o, range 22-67 y/o) with unilateral cancer. MRI was performed using a 1.5T scanner (Magneton Skyra, Siemens Medical Solutions, Erlangen, Germany) with a 16-channel Sentinelle breast coil. Dynamic

contrast-enhanced (DCE)-MRI was acquired using a fat-suppressed three-dimensional fast low angle shot (3D-FLASH) sequence with one pre-contrast and four post-contrast frames, with TR/TE=4.50/1.82 msec, flip angle=12°, matrix size=512x512, field of view=32 cm, and slice thickness=1.5 mm. The spatial resolution was 0.6x0.6x1.5mm. The pre-contrast, fat-suppressed T1W imaging sequence was used for analysis. In this study, only the contralateral normal breast was used for segmentation.

Fat-sat Independent Validation Dataset

The fat-sat testing dataset had 40 women (mean age 44 y/o, range 33–70 y/o) from another medical institution, also with unilateral cancer. The MRI was performed for diagnosis or pre-operative staging. For the fat-sat testing set, MRI was done using a 3T scanner (Magnetom Skyra, Siemens Medical Solutions, Erlangen, Germany) with a 16-channel Sentinelle breast coil. The pre-contrast, fat-suppressed T1W imaging sequence used for density analysis was also acquired using the 3D-FLASH sequence, with TR/TE =4.36/1.58 msec, flip angle =10°, matrix size =384×288, field of view =30 cm, and slice thickness=1.0 mm.

Ground Truth Segmentation

The ground truth was generated using a template-based automatic breast segmentation method [118]. In most breast MR scans, while breasts presented very different shapes and sizes, the chest region including the lung and the heart could be detected at similar locations with similar shape and intensity. These features were used to locate and segment out the chest region to isolate the breast. After the breast was

segmented, the next step was to differentiate FGT from fat. A correction method combined Nonparametric Nonuniformity Normalization (N3) and Fuzzy C-means (FCM) algorithms was used to correct the field inhomogeneity (bias-field) within the imaging region [138]. After the bias-field correction, K-means clustering was used to separate FGT from fatty tissues on pixel levels, with the number of clusters determined by the operator (KTC) who was a research physician and had one year of experience in performing breast segmentation. Since our group has been devoting to the development of breast MRI segmentation methods since 2008 [139] and many papers have been published, the operator knew the most likely clusters number to be used to accurately segment the fibroglandular tissue. In some cases, due to issues of tissue contrast, the mostly applied clusters number might need to be modified to produce the most accurate segmentation results. The segmentation results were then inspected by a radiologist, who had 12 years of experience in interpreting breast MR images, and if necessary, manually corrected. The manual correction, if needed, usually happened in the upper and lower margin of the breast tissue and in the breast areas showing inhomogeneous signal intensity. Not all subjects needed the correction. For those studies which needed correction, the number of slices ranged from 1 to 5 in each subject. This template-based segmentation has a very good reproducibility. The average inter-reader variability of breast and FGT were 3.7% and 3.9%, respectively [139]. The results were used as the ground truth for neural network training and independent validation.

4.1.3 Deep Learning Using U-net Architecture

The goal was to use U-net to separate three-class labels on each MR image, including (1) fat tissue and (2) FGT inside the breast, and (3) all non-breast tissues outside the breast. The first U-net was used to segment the breast from the entire image. Then, within the obtained breast mask, the second U-net was used to differentiate fat and FGT. The left and right breasts were separated using the centerline, and a square matrix containing one breast was cropped and used as the input. The pixel intensity on the cropped image was normalized to z-score maps (mean=0, and standard deviation = 1).

The U-net is a fully connected convolutional residual network (**Figure 4-1**) [135], which consists of convolution and max-pooling layers at the descending part (the left component of U), and convolution and up-sampling layers at ascending part (the right component of U). In the down-sampling stage, the input image size is divided by the size of the max-pooling kernel size at each max-pooling layer. In the up-sampling stage, the input image size is increased by the operations, which are performed and implemented by convolutions, where kernel weights are learned during training. The arrows between the two components of the U show the incorporation of the information available at the down-sampling stage into the up-sampling stage, by copying the outputs of convolution layers from descending components to the corresponding ascending components. In this way, fine-detailed information captured in descending part of the network is used at the ascending part. The output images share the same size of the input images.

In this study, there were four down-sampling and four up-sampling blocks. In each down-sampling block, two convolutional layers with a kernel size of 3 × 3 were each followed by a rectified-linear unit (ReLU) for nonlinearity [60], and then followed by a

max-pooling layer with 2 × 2 kernel size. In the up-sampling blocks, the image was up-convolved by a factor of two using nearest neighbor interpolation, followed by a convolution layer with a kernel size of 2 × 2. The output of the corresponding down-sampling layer was concatenated. Then, two convolutional layers, each followed by a ReLU, was applied to this concatenated image.

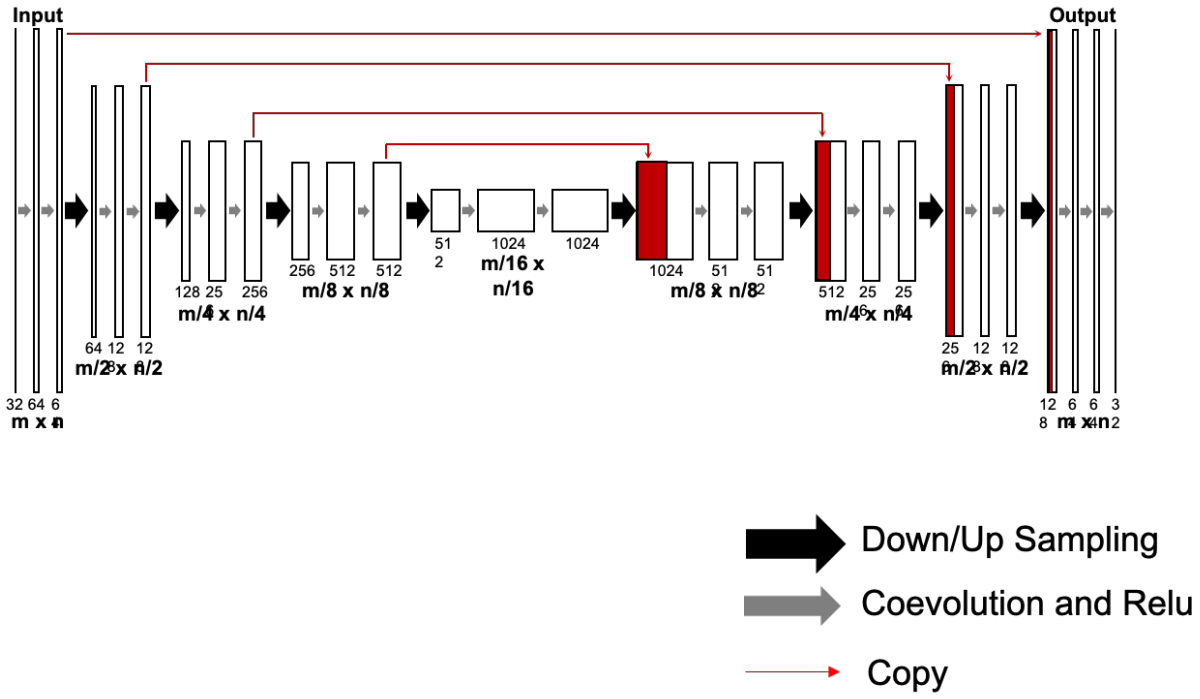

Figure 4-1: Architecture of the Fully-Convolutional Residual Neural Network (FC-RNN), or U-net. The U-net consists of convolution and max-pooling layers at the descending phase (the initial part of the U), the down-sampling stage. At the ascending part of the network, up-sampling operations are performed, which are also implemented by convolutions, where kernel weights are learned during training. The arrows between the two parts show the incorporation of the information available at the down-sampling steps into the up-sampling operations. The input of the network is the normalized image and the output is the probability map of the segmentation result.

4.1.4 Training Configuration and Transfer Learning

Firstly we used 286 non-fat-sat images to pre-train the established U-net. During the training process, the He initialization method was used to initialize the weights of the

network and the optimizer was Adam with a learning rate = 0.001 [105]. Finally, a convolutional and a sigmoid unit layer was added to produce probability maps for each class which correspond to the input image size. A threshold of 0.5 was utilized to determine the final segmented mask. The training processes included a total of 60,000 iterations before convergence. L2 regularization was used to prevent overfitting. Also, some background noise was added into the original images to do the image augmentation. Software code for this study was written in Python 3.5 using the open- source TensorFlow 1.0 library (Apache 2.0 license). Experiments were performed on a GPU-optimized workstation with a single NVIDIA GeForce GTX Titan X (12GB, Maxwell architecture).

Then the weights of the trained model using the 286 non-fat-sat images were saved, as the initial model to re-tune parameters for training fat-sat images using transfer learning [68]. For comparison, another model was trained directly using the He initialization method, which was a popular method commonly used for CNN training [140]. The initial weights differ in range depending on the size of the layers, and the He method provides a controlled initialization for faster and more efficient gradient descent.

4.1.5 Evaluation

In the initial training set of 286 patients, a 10-fold cross-validation was used to evaluate the performance of the U-net model. The final model was developed by training the 286 patient dataset with the hyperparameters which were optimized from the 10-fold cross-validation. The obtained model was then applied to segment the MRI of 28 healthy volunteers in the independent validation datasets. The ground truth for each case was available for comparison, and the segmentation performance was evaluated using the Dice

Similarity Coefficient (DSC) and the overall accuracy based on all pixels[141]. For example, the pixel accuracy of FGT segmentation was the correct classified pixel number over all pixel number of FGT. The algorithm was tested using 10-fold cross validation, so 10 accuracies could be calculated. The mean accuracy was the mean value of these 10 values. In addition, the Pearson's correlation was applied to evaluate the correlation between the U-net prediction output and the ground truth volume.

In the training set of 126 patients, the segmentation performance was evaluated using 10-fold cross-validation. The ground truth of each case was used for evaluation of the segmentation performance, by calculating DSC and the overall accuracy based on all pixels. Then a final model was developed using the hyperparameters optimized from these 10-fold cross-validation runs in the training dataset, and applied to the independent testing dataset of 40 patients. To evaluate the training efficiency of the transfer learning, models were developed using different number of training cases, 10, 20 … 110, to 126, and the obtained results were compared. Each developed model were applied to the testing dataset and to obtain corresponding DSCs.

4.1.6 Results

Non-fat-sat Segmentation

In the 10-fold cross-validation performed in the training dataset, the DSC range for breast segmentation was 0.83-0.98 (mean 0.95±0.02) and accuracy range was 0.92-0.99 (mean 0.98±0.01). For the FGT segmentation, the DSC range was 0.73-0.97 (mean 0.91±0.03) and accuracy range was 0.87-0.99 (mean 0.97±0.01). **Figure 4-2** and **Figure 4-3** show the segmentation results from two women with different breast morphology and

density. The correlation between the U-net prediction output and ground truth for breast volume and FGT volume are shown in **Figure 4-4.**

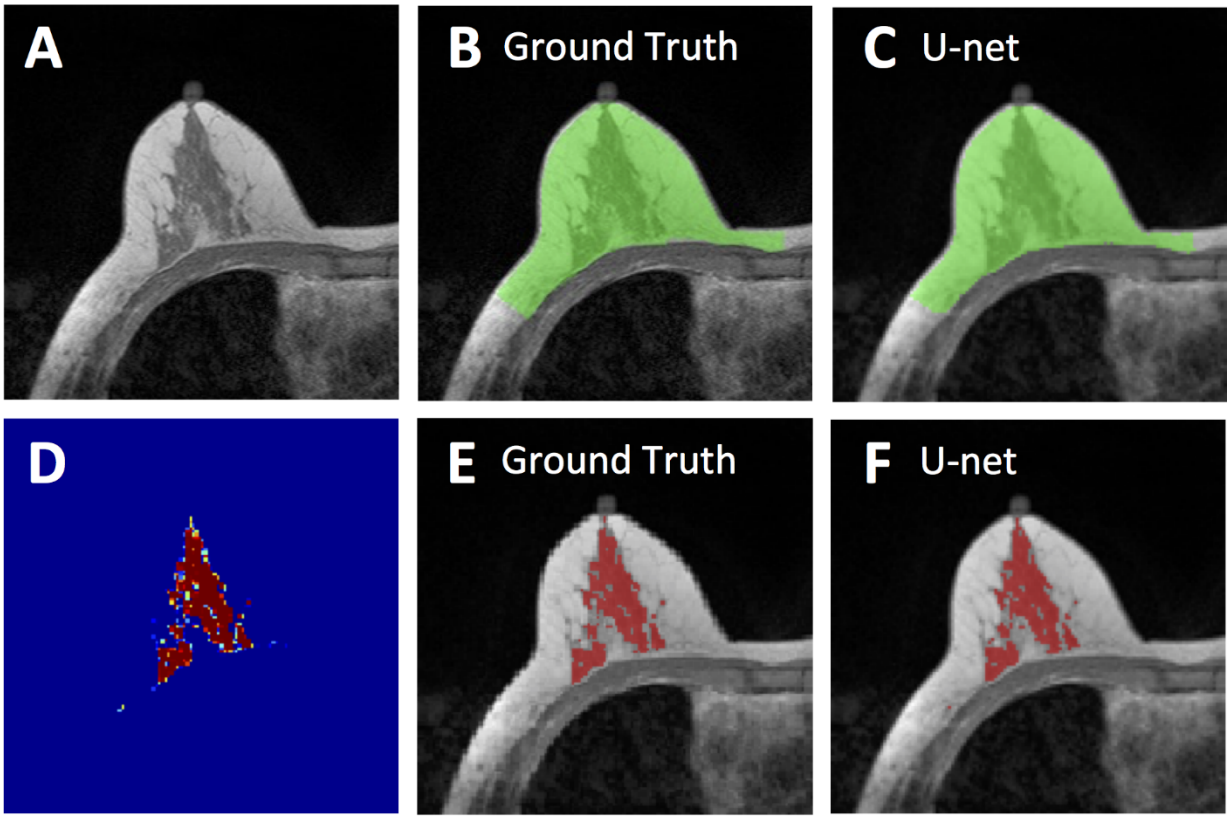

Figure 4-2: Segmentation results from a 62-year-old woman with moderate breast density**.** A: The original non-fat-suppressed T1-weighted image. B: The ground truth breast segmentation result obtained by using template-based method, shown in green. C: The breast segmentation result generated by U-net (green). D: The generated FGT probability map by the U-net. E: The ground truth FGT segmentation result within the breast obtained by using K-means clustering after bias-field correction (shown in red). F: The FGT segmentation result generated by U-net (red). For breast segmentation, DSC is 0.99 and accuracy is 0.99. For FGT segmentation, DSC is 0.97 and accuracy is 0.99.

The final model obtained from the training set was applied to the independent datasets acquired from the 28 healthy women using 4 different scanners. The processing time for one case was within 10s. The DSC and accuracy for each scanner was calculated separately, and then combined for all 4 scanners together. The results are shown in **Table 4.1**.

Table 4.1: The dice similarity coefficient (DSC) and the accuracy for the segmentation of breast and FGT in different MR scanners.

		GE 1.5T	GE 3T	Philips 3T	Siemens 1.5T	All MRI
Dice Similarity Coefficient						
Breast	Mean ± stdev	0.86 ± 0.06	0.87 ± 0.04	0.86 ± 0.05	0.87 ± 0.06	0.86 ± 0.05
	Range	0.56 – 0.95	0.54 – 0.95	0.50 – 0.95	0.58 – 0.97	0.50 – 0.97
FGT	Mean ± stdev	0.84 ± 0.05	0.81 ± 0.07	0.86 ± 0.05	0.84 ± 0.07	0.83 ± 0.06
	Range	0.61 – 0.96	0.53 – 0.94	0.64 – 0.94	0.61 – 0.94	0.53 – 0.96
Accuracy						
Breast	Mean ± stdev	0.95 ± 0.03	0.92 ± 0.03	0.92 ± 0.03	0.96 ± 0.04	0.94 ± 0.03
	Range	0.73 – 0.98	0.72 – 0.98	0.69 – 0.98	0.73 – 0.99	0.69 – 0.90
FGT	Mean ± stdev	0.92 ± 0.03	0.93 ± 0.03	0.93 ± 0.04	0.93 ± 0.04	0.93 ± 0.04
	Range	0.74 – 0.98	0.71 – 0.97	0.75 – 0.97	0.74 – 0.97	0.71 – 0.98

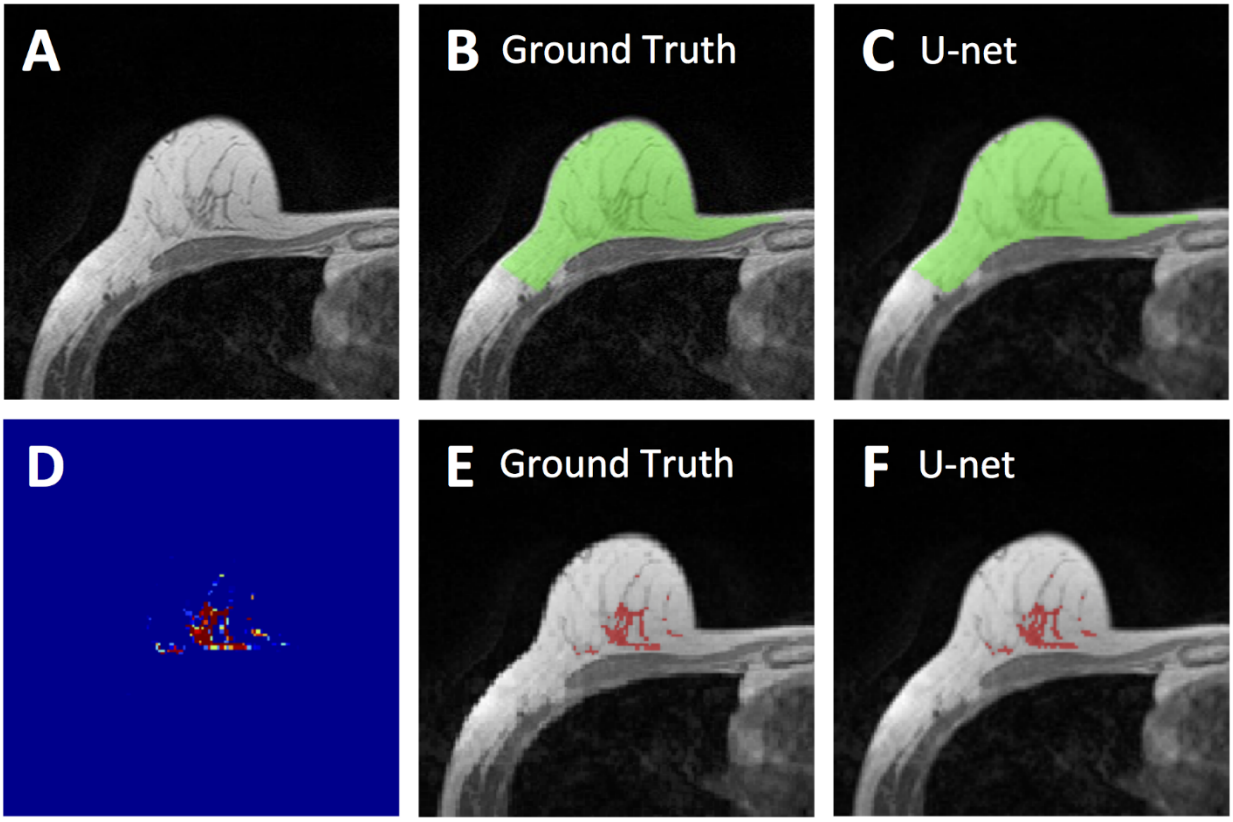

Figure 4-3: Segmentation results from a 55-year-old woman with fatty breast. A: The original non-fat-suppressed T1-weighted image. B: The ground truth breast segmentation result obtained by using template-based method, shown in green. C: The breast segmentation result generated by U-net (green). D: The generated FGT probability map by the U-net. E: The ground truth FGT segmentation result within the breast obtained by using K-means clustering after bias-field correction (shown in red). F: The FGT segmentation result generated by U-net (red). For breast segmentation, DSC is 0.99 and accuracy is 0.99. For FGT segmentation, DSC is 0.94 and accuracy is 0.98.

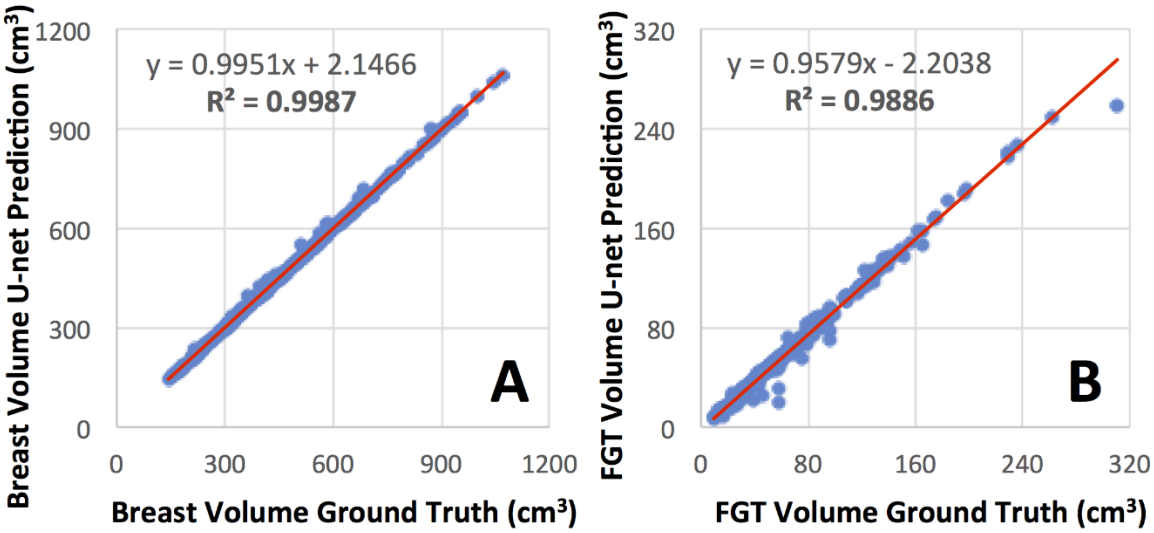

Figure 4-4: Correlation of breast volume (A) and FGT volume (B) between the ground truth obtained by using the template-based segmentation and the U-net prediction.

Figure 4-5 and **Figure 4-6** illustrate the segmentation results of two women with different breast morphology. The correlation between the U-net prediction output and ground truth for breast volume is shown in **Figure 4-7.**The obtained results for four different scanners were similar. The correlation coefficient r was high, in the range of 0.96-0.98. In each figure, the fitted line was very close to the unity line, and the slope was close to 1. The segmentation result for FGT volume is shown in **Figure 4-8.** The FGT segmentation results for MRI acquired using 4 different scanners were similar. The correlation coefficient r was very high, in the range of 0.97-0.99. However, using the unity line as reference, the U-net segmented FGT volume was lower compared to the ground truth, as in the two case examples demonstrated in **Figure 4-5** and **Figure 4-6.**

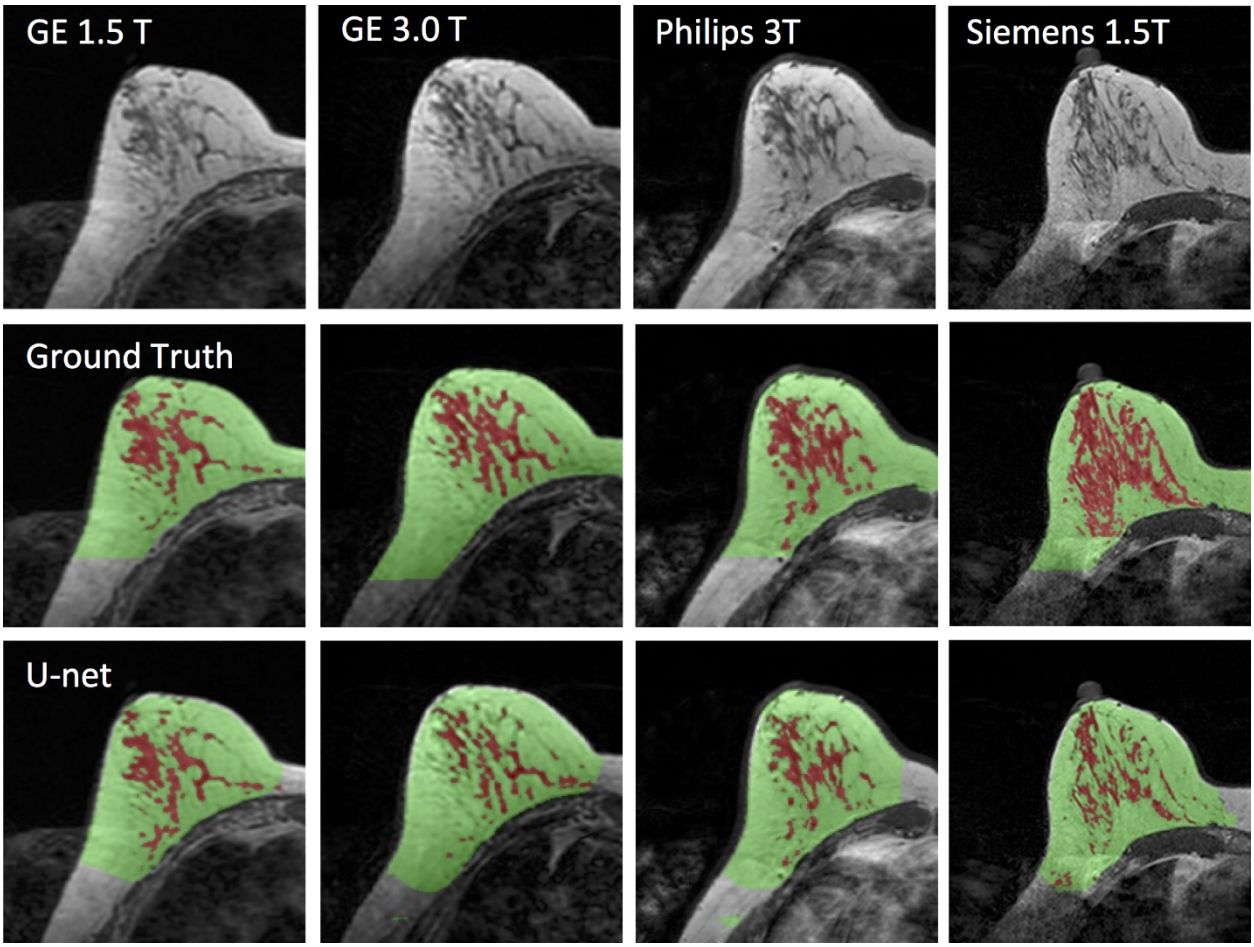

Figure 4-5: Images of a 43-year-old woman with heterogeneous breast morphology acquired using the GE 1.5T, GE 3.0T, Philips 3.0T, and Siemens 1.5T systems. The top row shows the original images. The center row shows the ground truth obtained by using the template-based segmentation method. The bottom row shows the U-net prediction results. The FGT volume segmented by U-net is smaller compared to the ground truth.

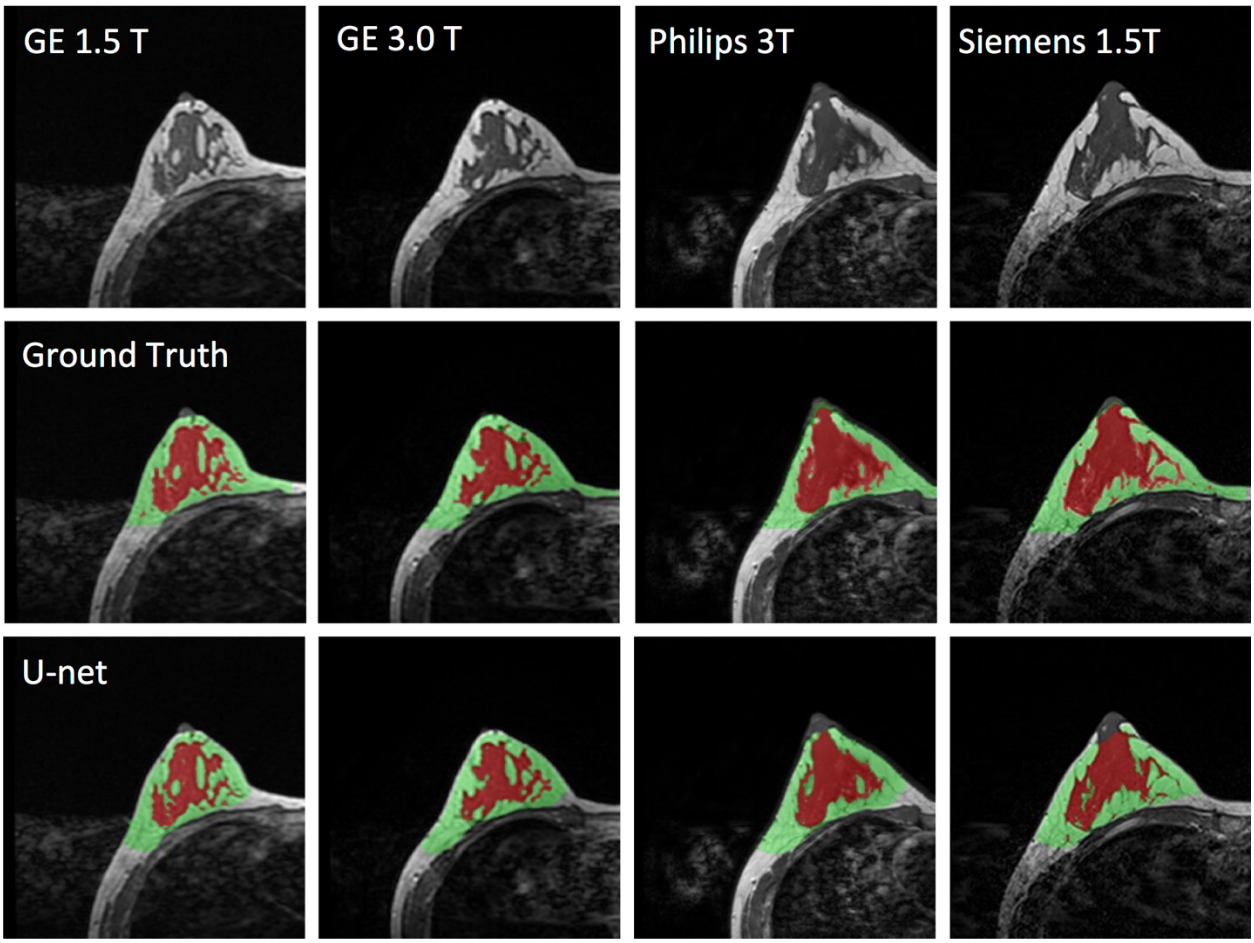

Figure 4-6: Images of a 29-year-old woman with dense breast acquired using the GE 1.5T, GE 3.0T, Philips 3.0T, and Siemens 1.5T systems. The top row shows the original images. The center row shows the ground truth obtained by using the template-based segmentation method. The bottom row shows the U-net prediction results.

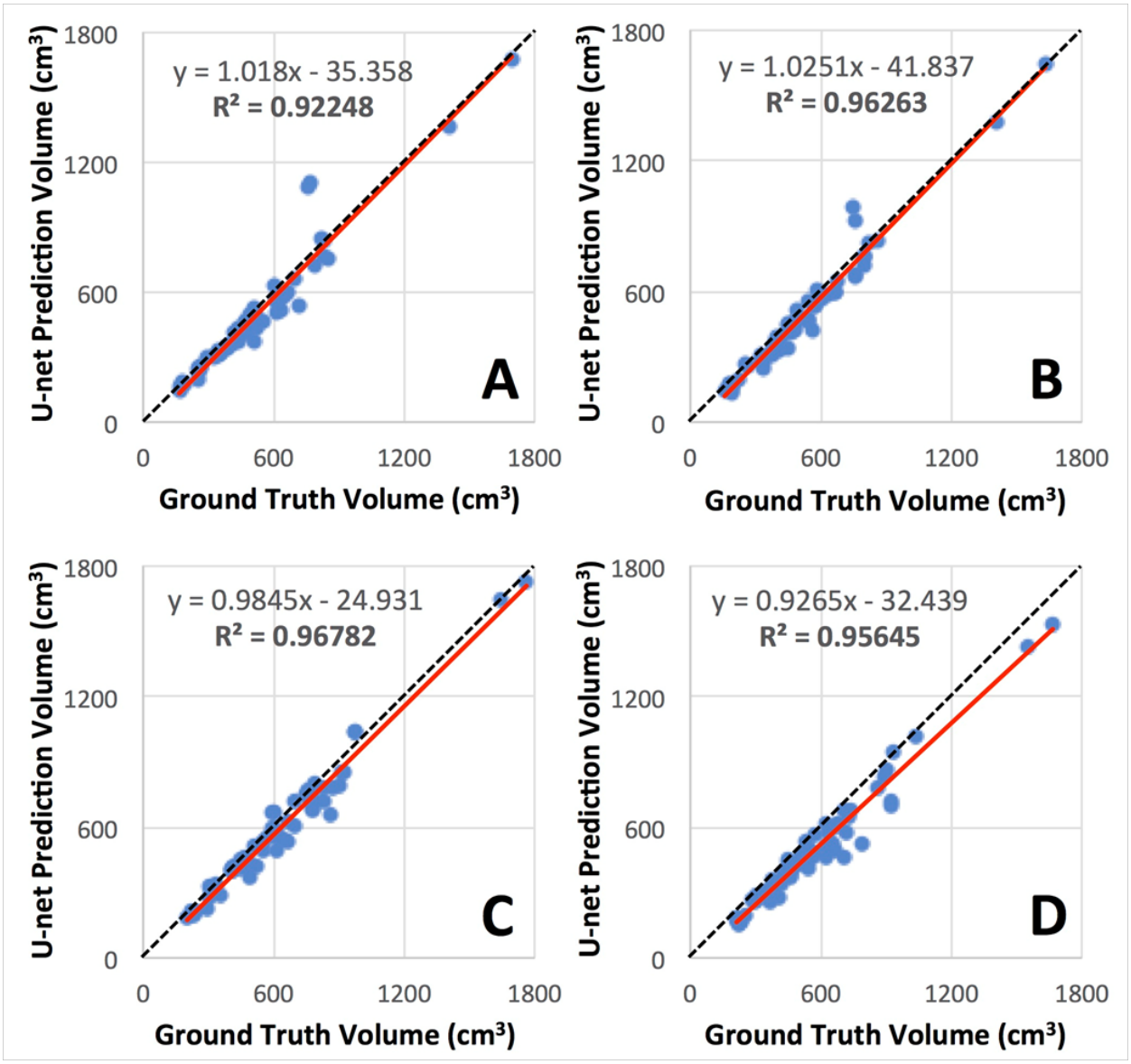

Figure 4-7: Correlation of breast volume between the ground truth obtained from the template-based segmentation method and the U-net prediction**.** (A) GE 1.5 T, (B) GE 3T, (C) Philips 3T, (D) Siemens 1.5T. The red line is the trend line, and the dashed black line is the unity line as reference.

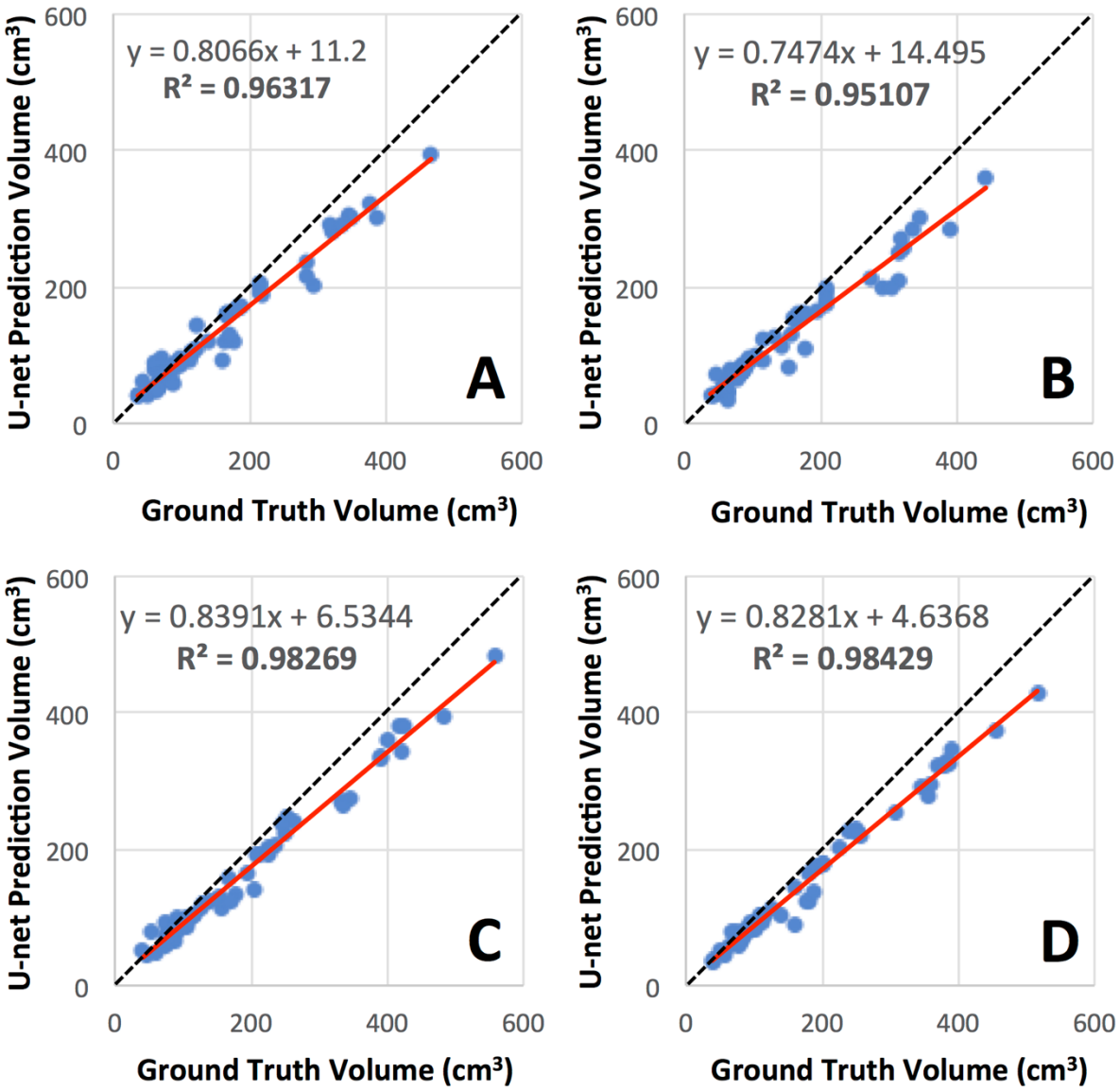

Figure 4-8: Correlation of FGT volume between the ground truth obtained from the template-based segmentation method and the U-net prediction. (A) GE 1.5 T, (B) GE 3T, (C) Philips 3T, (D) Siemens 1.5T. The red line is the trend line, and the dashed black line is the unity line as reference. Volume segmented by U-net is smaller compared to ground truth.

Effect of Transfer Learning from Non-fat-sat to Fat-sat training images

Figure 4-9 illustrate the segmentation results from four women with different breast morphology and density. It was obvious that the FGT segmentation results were very similar between the two approaches. By direct training using the He initialization without

TL, the mean DSC in the 10-fold cross-validation for breast segmentation was 0.95±0.03. The range in the 10-fold runs was 0.94-0.97, suggesting that the model was robust and could achieve a high accuracy in all runs. For pixel-based analysis, the mean accuracy was 0.97±0.04 (10-fold run range 0.95-0.98). For FGT segmentation, the mean DSC was 0.80±0.11 (range 0.75-0.89) with mean accuracy of 0.86±0.03 (range 0.81-0.90). When the model from non-fat-sat was used for initialization, the performance was better. For breast segmentation, mean DSC was 0.97±0.02 (range 0.96-0.98) with mean accuracy of 0.97±0.01 (range 0.96-0.97). For the FGT segmentation, the mean DSC was range 0.86±0.08 (range 0.74-0.90) with mean accuracy of 0.90±0.05 (range 0.87-0.96). All segmentation results are summarized in **Table 4.2** for comparison. The correlation between the U-net prediction output and ground truth for breast and FGT volume are shown in **Figure 4-11**. As noted, there was a high correlation (R2>0.90) for both the training and testing datasets. However, when carefully comparing the segmentation results case by case, we did see mild degree of inconsistency between U-net and ground truth in some cases. **Figure 4-10** shows four women with inconsistent segmentation results of FGT between U-net and ground truth.

Table 4.2: The dice similarity coefficient (DSC) and accuracy in the Training Set and Testing Set by using the U-net model developed with and without transfer learning

Dataset		**Dice Coefficient**		**Accuracy**	
		Range	**Mean**	**Range**	**Mean**
Training Set	Breast	0.96-0.99	0.97	0.95-0.99	0.97
	Fibroglandular	0.33-0.96	0.86	0.53-0.98	0.90
Testing Set (Transfer Learning)	Breast	0.72-0.98	0.89	0.82-0.98	0.91
	Fibroglandular	0.38-0.97	0.81	0.48-0.98	0.86
Testing Set (No Transfer Learning)	Breast	0.69-0.98	0.83	0.79-0.98	0.89
	Fibroglandular	0.34-0.95	0.81	0.52-0.98	0.87

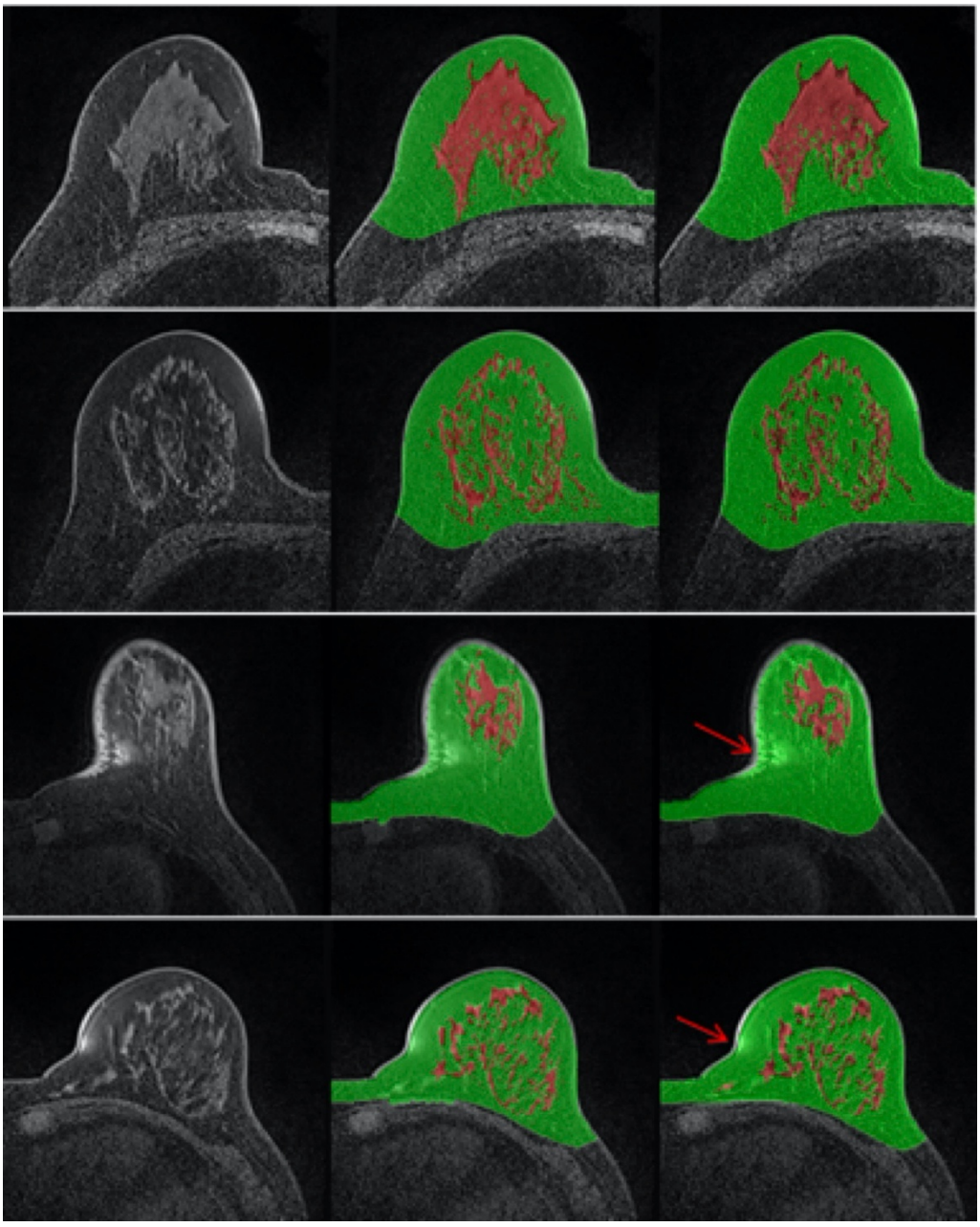

Figure 4-9: Four representative cases of different breast size and parenchymal patterns showing accurate FGT segmentation using AI compared to the ground truth. Left column: original image; central column: ground truth of breast and FGT segmentation; right column: segmentation results using AI. Lower two panels show two cases with susceptibility artifact. Despite of the artifact of bright signal intensity (arrows) similar to FGT, AI can still recognize and exclude it.

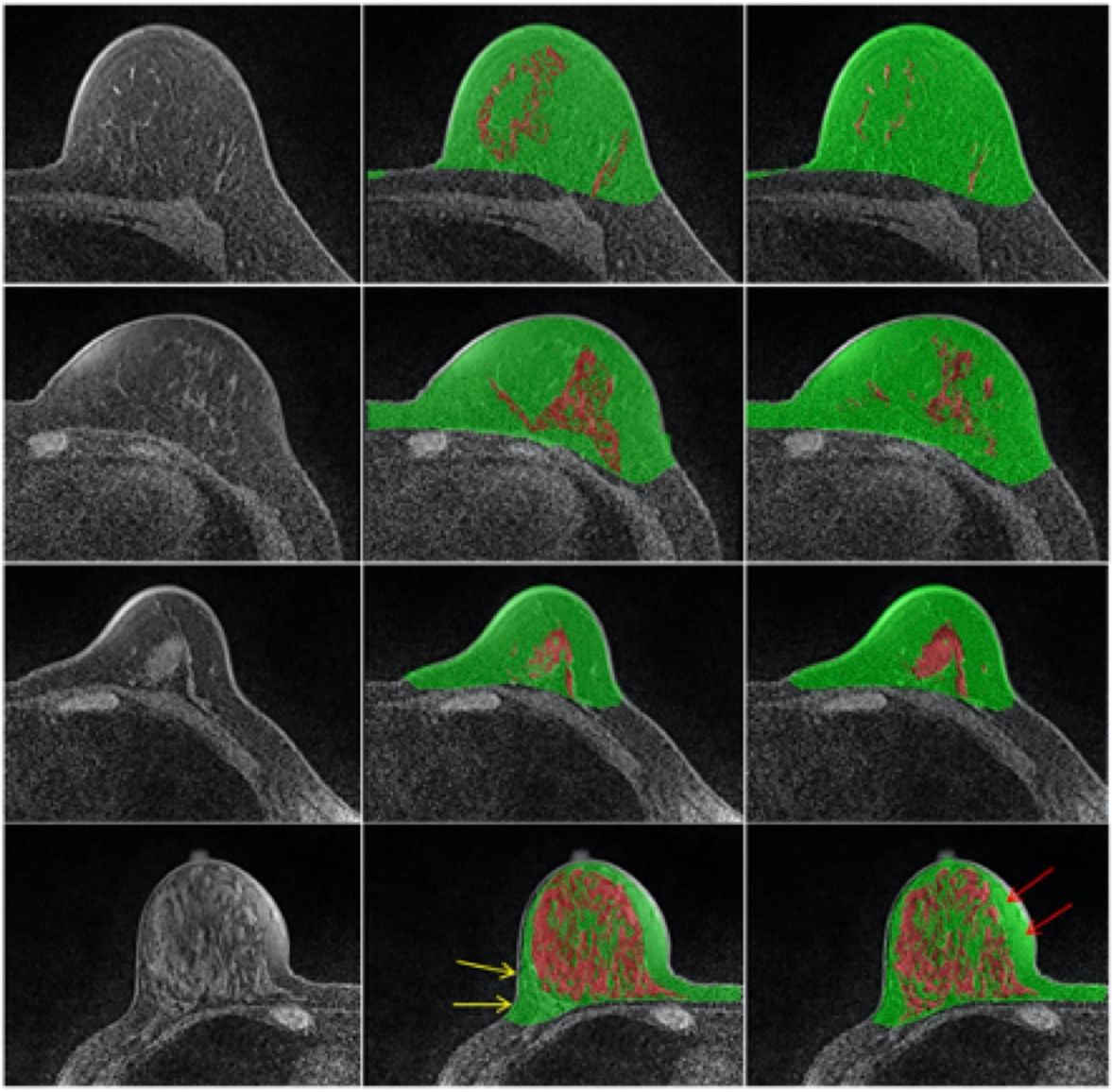

Figure 4-10: Four cases of inconsistent FGT segmentation between AI and the ground truth**.** Left column: original image; central column: ground truth of breast and FGT segmentation; right column: segmentation results using AI. Upper and middle upper (second) panels show that the FGT results from ground truth are over-segmented compared to the original image. Obviously, the results from AI are more accurate. Middle lower (third) and lower panels show that the FGT results of the ground truth are under-segmented compared to the original image. Note the under-segmented FGT in the lower margin (yellow arrows) of the lower panel. Note also the incomplete suppression of the fat signals (red arrows) which are recognized and excluded by AI.

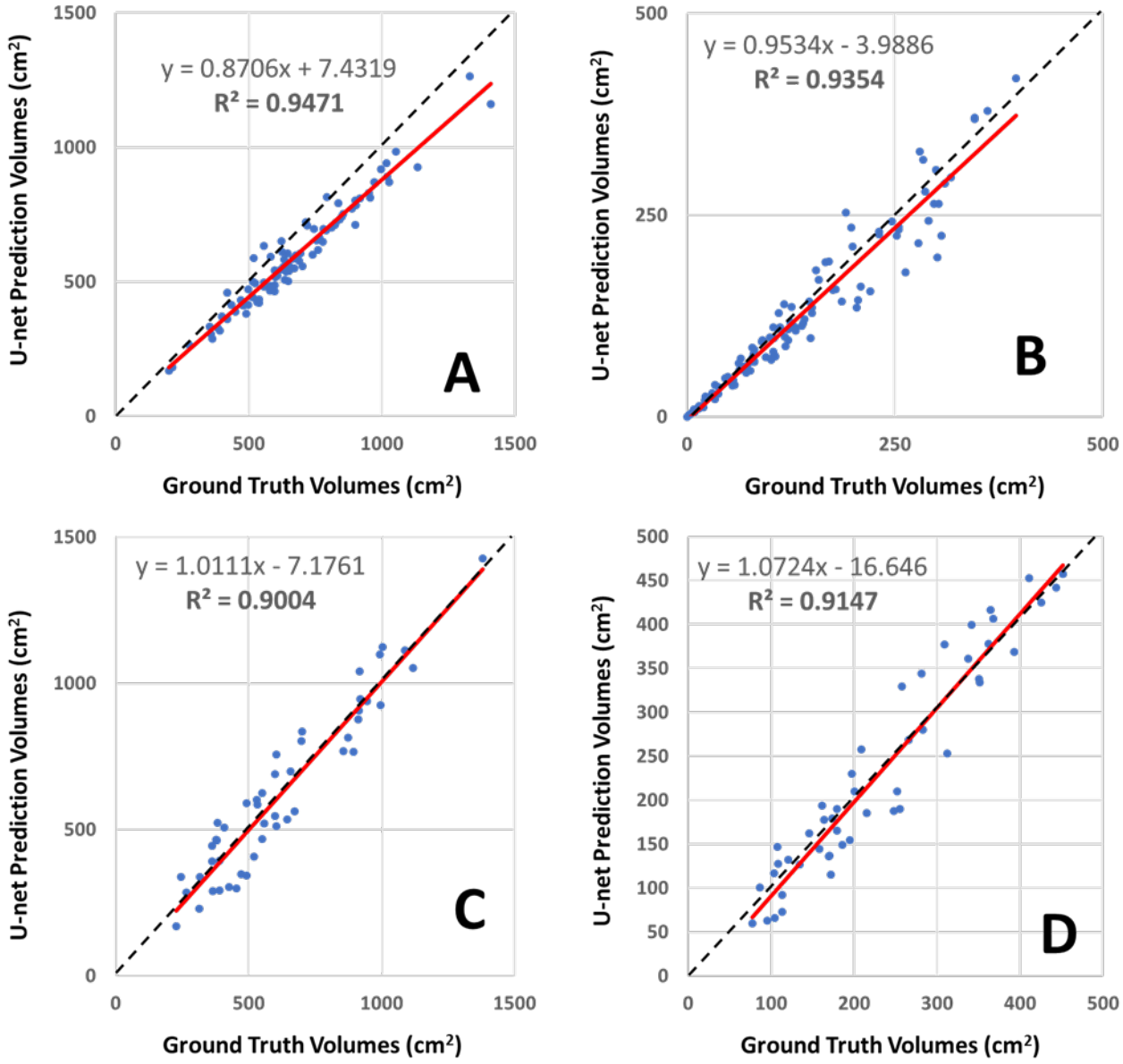

Figure 4-11: Correlation of breast volume between the ground truth obtained from the template-based segmentation method and the U-net pre-diction**.** (A) Training Data Breast Volumes, (B) Training Data FGT Volumes, (C) Testing Data Breast Volumes, (D) Testing Data FGT Volumes. The red line is the trend line, and the dashed black line is the unity line.

Segmentation Performance in Fat-sat Testing dataset

When the developed model from fat-sat training dataset without transfer learning was applied to the testing dataset, the mean DSC for breast segmentation was 0.83±0.06, with

mean accuracy of 0.89±0.03. For the FGT segmentation, the mean DSC was 0.81±0.1 with mean accuracy of 0.87±0.07. When using the model developed with transfer learning was applied, the performance in the testing dataset was slightly improved for breast segmentation, showing mean DSC of 0.89±0.06 and mean accuracy of 0.91±0.03. For the FGT segmentation, the mean DSC was 0.81±0.08 with mean accuracy of 0.86±0.05.

Efficiency of Transfer Learning

To evaluate the efficiency of training without and with TL, the performances of models developed using different number of training cases, 10, 20 … to 126, were compared. The results are shown in **Figure 4-12**. Without TL, DSC was low when the training case number was small. When sufficient number of cases was used for training (>30 or breast segmentation, and >80 for FGT segmentation), the achieved DSC could reach those trained with TL, only slightly lower for breast segmentation and the same for FGT segmentation.

4.1.7 Summary and Discussion

In this chapter, a deep-learning method based on the U-net architecture [135], for breast and FGT segmentation on non-fat-sat and fat-sat MRI was implemented. For non-fat-sat segmentation, to objectively test the performance of the developed method, we used independent validation datasets from MRI acquired using four scanners at two different institutions. The results showed that for both the breast and the fibroglandular tissue segmentation, high accuracy was achieved (0.98±0.01 and 0.97±0.01, respectively). When the model was applied to independent datasets for validation, the performance was also very good (accuracy >0.92). For fat-sat segmentation, a dataset from one hospital was used

for training and another dataset from a different hospital was used for independent testing. A model developed previously for a non-fat-sat image dataset was used as the basis for re-training, or transfer learning, to investigate its benefit [142].

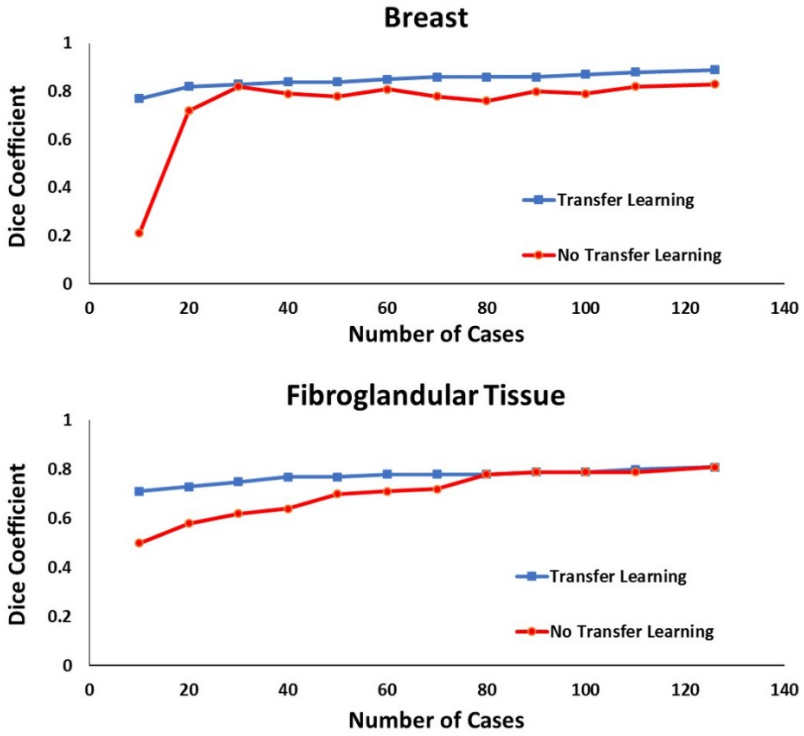

Figure 4-12: The plot of DSC in the testing dataset by using the model developed with different number of training cases from 10, 20, … to 126, with and without transfer learning. When the training case number is small, DSC is low. When sufficient number of cases is used for training (>30 or breast segmentation, and >80 for FGT segmentation), the achieved DSC with and without transfer learning is comparable, only slightly better with transfer learning for breast segmentation.

Transfer learning (TL) is a popular approach in deep learning where pre-trained models are used as the starting point on computer vision tasks [68]. The results showed that the DSC for breast segmentation was very high in the training dataset, with the mean of 0.95 without TL and 0.97 with TL. In testing dataset, the DSC was also satisfactory, with the mean of 0.83 without TL and 0.89 with TL. For FGT segmentation, it was more difficult

compared to the breast segmentation, and the DSC was in general lower. In the training dataset, the mean DSC was 0.80 without TL, and 0.86 with TL. In the testing dataset, the mean DSC was 0.81. The results suggest that deep learning segmentation using U-net is feasible to perform fully automatic segmentation for the breast and FGT and yield reasonable accuracy compared to the ground truth. Meanwhile, TL could be applied to improve the segmentation accuracy compared to the direct training using the He initialization method [140]. Particularly the training efficiency could be improved with TL, thus not requiring a large number of input data to get satisfactory performance. The results suggest that when the number of training cases is limited, applying TL can help to develop a good model and achieve a higher accuracy.

Unlike 2D mammography, 3D MRI provides genuine volumetric assessment of the FGT for quantification of breast density, thus it may be used to assess small changes in density over time following hormonal or chemotherapy [117, 143]. Three-dimensional MR density can also be used to study breast symmetry [144], and peritumoral environment [145]. Additionally, 3D MR breast and FGT segmentation method is necessary for the quantitative measurement of background parenchymal enhancement (BPE) [146, 147], which has shown to be related to the aggressiveness of the tumor, treatment response, prognosis, and breast cancer risk [148]. Quantitative measurement of dense tissue volume may also be incorporated into risk prediction models to improve the accuracy of breast cancer risk predicted for each individual woman. Currently, some models have already included mammographic density as a risk factor. The value of MR density has also been proven in two large scale studies [149, 150]. King et al [149] specifically states an association of increased FGT on MRI and breast cancer risk. Because of accurate fully

automatic FGT segmentation in T1-weighted imaging, the quantified assessment of BPE is possible, and both King and Dontchos' work [149, 150] shows that increased quantified BPE is associated with increased breast cancer risk. With more and more large MRI datasets gradually becoming available, this will allow studies to investigate whether the inclusion of MR volumetric density into the risk models outperforms other models. Since a very large dataset needs to be analyzed, an efficient segmentation tool that can provide precise information about breast density is required.

In recent years, machine learning has been widely applied for organ/tissue segmentation on MRI, including breast and FGT segmentation. Wang et al. applied support vector machine (SVM) algorithm to T1W, T2W, proton density (PD), and Dixon sequences, and obtained the overlap ratios around 93%-94% for FGT segmentation [151]. Although the result was very good, the requirement of 4 different MR sequences was not practical in the clinical breast MRI protocol. Convolutional neural network (CNN) has become an important tool in the image processing and computer vision research. Among the different approaches, U-net is a powerful algorithm which can extract different classes of information related to different tissues in a large field, thus suitable for the breast segmentation [135, 152]. It has been applied for breast and FGT segmentation on non-fat-sat images [130, 131, 142, 152, 153]. Dalmış et al. segmented breast and FGT using a dataset of 66 pre-contrast T1weighted MR [131]. The U-net was trained for two 2-class classification to sequentially separate breast first, followed by fat and FGT; as well as one 3-class classification to segment breast, fat and FGT simultaneously. The average DSC values for FGT segmentation obtained from the 3-class classification, two 2-class classification, and atlas-based methods were 0.850, 0.811, and 0.671, respectively, demonstrating the

superior performance of U-net over atlas-based method. This study did not have independent testing datasets.

All these studies showed consistent results, demonstrating the good performance of U-net for segmentation on non-sat-fat MR images, which had higher signal-to-noise ratio (SNR), higher tissue contrast, and fewer image artifacts compared to fat-sat images, thus easier for segmentation. There were few studies reporting the application of CNN for FGT segmentation on fat-sat MR images. In [152], Fashandi et al. used 70 patients with fat-suppressed MR and non-fat-suppressed MR to train various U-net models to segment the breast, but not going further to segment FGT within the breast. Similarly, very high DSC's were obtained for breast segmentation, with the highest of 0.96 when multi-channel inputs combing all images were used in 3D convolutions in U-net. Ha et al. applied 3D U-net to segment sagittal view fat-suppressed T1W images of 137 patients, and achieved DSC of 0.95 and 0.81 for breast and FGT segmentation, respectively [130]. The reported DSCs of breast and FDT segmentation were similar to our results. The U-net developed by Ha et al. utilized 3D convolutions and the evaluation was done by cross-validation of training dataset, without testing using an independent dataset.

Form the results of the non-fat-sat MR segmentation, 286 cases were used as the initial training dataset, and ten-fold cross-validation was used to adjust the hyperparameters of the neural networks. One noticeable problem, generally seen in this study, showed that the FGT was under-segmented by the U-net (**Figure 4-8**). From other literature, the issue of the underestimation in FGT segmentation has not been addressed. Hence, this should not be the flaw of U-net. As this trend was consistent for all 4 scanners, this appeared to be a systematic bias problem, and not due to sporadic variations. The

ground truth in FGT segmentation was performed by the operator, who had to select the cluster numbers to differentiate FGT from fat based on their image intensities. For example, when using a total cluster number of 6 with 3 for FGT, the segmentation appeared reasonable. However, when using a total cluster number of 5 with 2 for FGT, the segmentation quality was very likely to be reasonable as well, but this would result in a lower FGT volume. As shown in the two case examples illustrated in **Figure 4-5** and **Figure 4-6**, the U-net segmented FGT volume was lower; however, when visually inspecting the segmentation results separately, both appeared reasonable. Therefore, although U-net FGT volume was lower than manually segmented volume, this did not mean that there was an error. In fact, we believe that the fully automatic method using deep learning can provide an objective method not affected by the operator's judgment, and it has a potential to replace the semi-automatic method and eliminate the operator's input.

In a study by Chang et al. [154], FGT segmentation was performed using a computer-assisted clustering method on 38 patients with both fat-sat and non-fat-sat images, and showed 5% difference in the segmented FGT volume on average. The result was not surprising, due to their different image quality and tissue contrast. The quality of fat-sat images might be affected by many factors, including MR systems (such as magnetic field strength, transmitting RF field inhomogeneity or inaccuracy, B_1 shimming, receiver breast coil, fat-sat pulse sequence) and the variation in different patients (body shape, breast size, tissue composition, etc.). In general, any factor leading to signal variability can result in tissue misclassification, thus inaccurate FGT segmentation [155]. In our FGT segmentation results, although the mean DSC was greater than 0.8, the range was pretty wide, with the lowest in 0.3-0.4. These extreme cases had poor image quality and low SNR, which often led

to low tissue contrast between fat and FGT and difficult to be differentiated. In these cases, clustering algorithm also had difficulty to differentiate and segment tissues, thus might not provide a ground truth, and the low DSC could not be interpreted as failure of U-net. For breast segmentation, some cases also had a low DSC in the range of 0.7. For extremely fatty breast with a good fat suppression, the SNR of breast tissue can be very low and indifferentiable from the background, as also demonstrated in [154]. Despite of these problems, for diagnostic purposes, the enhanced tumors can be easily identified on fat-sat images without the need of generating subtraction images, thus it is more popular than non-fat-sat images [156]. The capability of an efficient and accuracy method for segmentation of breast and FGT on fat-sat images will provide helpful information to explore its clinical application in improving the accuracy of risk prediction models [157, 158] and evaluating therapy response [117, 159].

Another strength of deep learning was the ability to handle field inhomogeneity, or bias-field. Intensity inhomogeneity often presented as a smooth intensity variation across the image is mainly due to poor radio frequency (RF) coil uniformity, gradient-driven eddy currents, and patient's anatomy inside and outside the field of view [160]. For conventional segmentation algorithms, retrospective correction methods including filtering [161], or bias field estimation [138], were commonly used. However, for medical images with high noise level or severe intensity inhomogeneity, this problem could not be completely eliminated. In our experience, the images at caudal and cranial ends of an MRI volume often had a low signal intensity, and the bias-field correction was very important for segmentation on these slices. Our results showed that U-net methods were minimally affected by the bias field, although no specific bias-field correction was applied as a prior

step. This indicated that U-net was able to learn the bias field and make corrections. However, other studies [130] found that bias field correction would improve the segmentation results and had shown specific examples. Ha et al. [130] used smaller dataset and different modality. Thus the importance of the bias field correction cannot be evaluated based on different datasets. We believe if the inhomogeneity is very high or higher than the mean intensity, bias field correction will definitely improve the results.

There were some limitations in this study. First, only two datasets, each acquired using a consistent breast MRI DCE sequence, were analyzed. The trained model may not be applicable to images acquired using a different MRI system or with a different imaging protocol. However, as demonstrated here, for future application in other datasets, the model developed in this study can be used as the basis for transfer learning to develop a specific model for each dataset. Another limitation was the implementation of U-net based on 2D slices. To fully utilize the morphological information, 3D convolution should be employed. However, the 3D analysis will need many more trainable parameters which require more training cases.

In summary, we presented deep-learning approaches based on the U-net architecture for breast and FGT segmentation on MRI. This method showed good segmentation accuracy, and there was no need to do the post-processing correction. With further refinement of the methodology and validation, this deep learning-based segmentation method may provide an accurate and efficient means to quantify FGT volume for evaluation of breast density.

4.2 COVID-19 Lung Infection Segmentation via Co-Registration of Serial Chest CT

4.2.1 Background and Motivation

The coronavirus disease 2019 (COVID-19), caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), is an ongoing pandemic. The number of people infected by the virus is increasing rapidly all over the world. This has led to great public health concern in the international community, as the World Health Organization (WHO) declared the outbreak to be a Public Health Emergency of International Concern (PHEIC) on January 30, 2020 and recognized it as a pandemic on March 11, 2020 [162, 163].

Recent studies reported that the possible pathological mechanism in COVID-19 lung infection is caused by diffuse alveolar damage and inflammatory exudation, which is similar to histologic findings seen in SARS-CoV-2 pneumonia [164, 165]. The pathological evolution during the course of infection in COVID-19 has not been clarified, and the disparity of such changes in patients with different clinical severities are largely unknown.

To evaluate the patient's response and investigate the potential problems after treatment, close follow-up using imaging can provide great information related to the progression of COVID-19 and the response to treatment. Considering the long incubation period of this disease and its popular infectivity, it is challenging to find an appropriate procedure to monitor the course of the disease. Chest CT, especially high-resolution CT (HRCT), can detect small areas of ground glass opacities (GGO) [166], and therefore, is a promising imaging tool for longitudinal monitoring the disease. Before the polymerase chain reaction (PCR) diagnostic test became widely available, chest x-ray and CT were used in China during the early break out to identify patients who might have been infected and need

to be isolated to control the transmission. Compared to plain x-ray, Chest CT has played a pivotal diagnostic role in the assessment of patients with COVID-19, which can also be used for follow-up assessment and evaluation of disease evolution [167]. Although in the US, chest CT was not used or diagnosis, and also not recommended to follow the lung infections, we are fortunate to obtain a unique dataset from our collaborators in China. Every patient had several longitudinal CT's acquired during hospitalization. Quantitative analysis of the CT scans can provide an automatic and objective estimation of the disease burden, facilitating and expediting imaging interpretation during the COVID-19 pandemic [168].

Recently, several quantitative analysis methods have been proposed to detect patients infected with COVID-19 via radiological imaging [169, 170]. Although plenty of image analysis systems have been proposed to provide assistance in diagnosing COVID-19 in clinical practice, there are limited studies related to follow-up evaluation using CT [168]. To evaluate the progression of the lung infections, the image segmentation algorithms can be utilized. However, the detection and precise segmentation of COVID-19 infection on CT is a very challenging task, due to the high variation in texture, size and position of infections on many CT slices. For example, consolidations are tiny/small, which easily results in the false-negative detection on a whole CT slice.

Although many patients have been infected by COVID-19 in the world, CT was not commonly used and it is difficult to collect sufficient labeled datasets for training machine learning models for performing quantitative analysis. Visual evaluation of changes between two CT scan is subjective, and its validity may depend on the radiologists' experience which is known to have a high variation and low ICC [171]. To solve this problem, in this project we developed a registration method between the baseline images and the follow-up

images, so the lesions can be compared in the co-registered lung areas. The method included two steps, first using the Affine registration based on the body areas, and then followed by the non-rigid registration based on the segmented lung areas. Through the registration, the lesions or infected areas at different locations at different follow-up times could be objectively evaluated, and further segmented for volumetric comparisons. The registration results were evaluated using mean square errors (MSE).

4.2.2 Subjects and CT Protocol

48 Patients, 32 male and 16 female, with COVID-19 who underwent chest CT in the radiology department of The First Affiliated Hospital of Wenzhou Medical University from January 24 to March 2, 2020, were enrolled in this retrospective study. The age range is 21-93 years old, with the mean of 53±14. Inclusion criteria were: (a) positive SARS-Cov-2 nucleic acid in double swab tests (within an interval of 2 days, real time RT- PCR) and (b) without confirmation of another viral infection. Of them, 33 patients received the first follow-up scan, 29 received the second follow-up scan, and 11 received the third follow-up scan. All patients had positive lesions on CT. The average duration between the onset of symptom and the initial CT scan is 6 days.

Non-contrast chest CT examinations were performed with two CT scanners (GE LightSpeed VCT 64-Slice, GE Healthcare, USA; Phillips Brilliance 16-Slice, Phillips, Netherland). The patients were scanned in supine position during inspiratory breathhold. The scanning range was from apex to the base of lungs. Scanning parameters were as follows: tube voltage 20 kV, tube current 50-70 mAs, pitch 1~1.5 mm, matrix 512×512, slice thickness 5 mm. Reconstruction was performed with slice thickness of 5 mm, a lung

window with a width of 1,500 HU and a level of -500 HU, and a mediastinal window with a width of 400 HU and a level of 40 HU. All images were reviewed by 2 radiologists. The region of interest (ROI) for COVID-19 lesions were manually outlined on each slice containing the infected areas inside both lungs. The total lesion volume in each patient was calculated, and the results from all patients at 4 CT scans are listed in **Table 4.3**.

Table 4.3: The distribution of COVID-19 lesion volume at 4 different CT scans

	Range (cc)	25^{th} (cc)	Median (cc)	75^{th} (cc)
Baseline (N=48)	0.1 - 1064	2.5	30.9	85.6
First F/U (N=33)	0.1 - 773	4.9	22.9	51.9
Second F/U (N=29)	2.3 - 731	8.4	15.3	47.6
Third F/U (N=11)	1.2 - 649	1.3	2.5	15.3

Co-Registration

The imaging matching was completed by two steps. The first step is to apply Affine registration using the whole body areas between the baseline images and follow-up images. The next step is to fine-tune the registration results using non-rigid Demons registration algorithm based on the segmented lung areas. Therefore, as a preprocessing step, the lung segmentation should be completed first.

Firstly, for each case, all of the CT slices were combined to obtain a 3D volume. Then the middle slice in both the axial and coronal directions were extracted. On the middle slices, a threshold of HU was utilized to identify the tissues inside the body areas but outside the lung areas. The surrounded air areas were considered as the lung areas. With identified lung areas on the middle slices in axial and coronal directions, the active contours algorithms were applied to segment the lung areas in the entire 3D volume. This technique, also called

snakes, is an iterative region-growing image segmentation algorithm [172]. Active contours can be defined as the process to obtain deformable models or structures with constraints and forces in an image for segmentation. Contour models describe the object boundaries or any other features of the image to form a parametric curve or contour. The desired contour is obtained by defining the minimum of the energy function. Deforming of the contour is described by a collection of points that finds a contour. This contour fits the required image contour defined by minimizing the energy function [172].

To complete the registration, the first step is to apply Affine registration between the baseline images and follow-up images based on the entire body areas. Then, the next step is to fine-tune the registration results using the segmented lung areas.

The lung registration was performed automatically using an intensity based non-rigid registration, Demons algorithm. Demons' method applies a diffusion process to deform the rectum mask generated from the previous slice to the current slice, based on the distribution of intensities by iteratively minimizing the energy function, *E*, as shown [173]:

$$E(u) = \|F - M \circ (T + u)\|^2 + {\sigma_n}^2 |F - M|^2 \|u\|^2$$

where *M*, the moving image, is the segmented slice with the defined rectal mask that is to be deformed to segment *F*, the adjacent fixed image slice through an image transformation represented by the symbol ◦. This symbol means "apply the transformation (T+u) to M" or "deform M by the field (T+u)". For each iteration, the deformation field, *T*, is updated such that $T = T + u$, where u is the update factor; σ_n is the image noise ratio coefficient. Thus, the lung mask for the unsegmented slices is obtained by applying the correct transformation field to the mask of the moving image, M, on adjacent slices. The transformation field was found by solving for u by minimizing the energy function and given by [174]:

$$u = \frac{(M \circ T - F)\nabla F}{[|\nabla F|^2 + (M \circ T - F)^2]}$$

this process stopped when u was sufficiently small (u<10^{-3}).

For evaluating the registration performance, the corresponding Mean Square Errors of the lung areas were calculated as:

$$\mathbf{MSE} = \frac{\mathbf{1}}{\boldsymbol{N}} \sum_{\boldsymbol{N}} \|\mathbf{Moving\ Image} - \mathbf{Referece\ Image}\|_{\mathbf{2}}^{\mathbf{2}}$$

where N is the total number of lung area voxels in one patient.

After registration, the lesion areas on the baseline images were mapped to the follow-up images using the estimated geometric transformation matrix obtained from the co-registration, so the change can be visually compared.

4.2.3 Evaluation and Results

The co-registration was applied to all patients who received follow-up scans. **Figure 4-13, Figure 4-14** and **Figure 4-15** show the MSE distributions on different follow-up scans, first F/U, second F/U, and third F/U compared to the baseline (B/L) scan, respectively. The MSE between the F/U and B/L images was first calculated after the Affine registration based on the whole body areas, and then after the non-rigid Demons registration focusing on the lung areas. Between the first F/U and B/L, the mean MSE calculated within the lung areas was 9,974 after Affine registration, which was decreased to 8,142 after completing the second step of non-rigid registration focusing on the lung area. The smaller MSE after the second step indicates that the non-rigid Demons can significantly improve the registration quality, with $p < 0.001$.

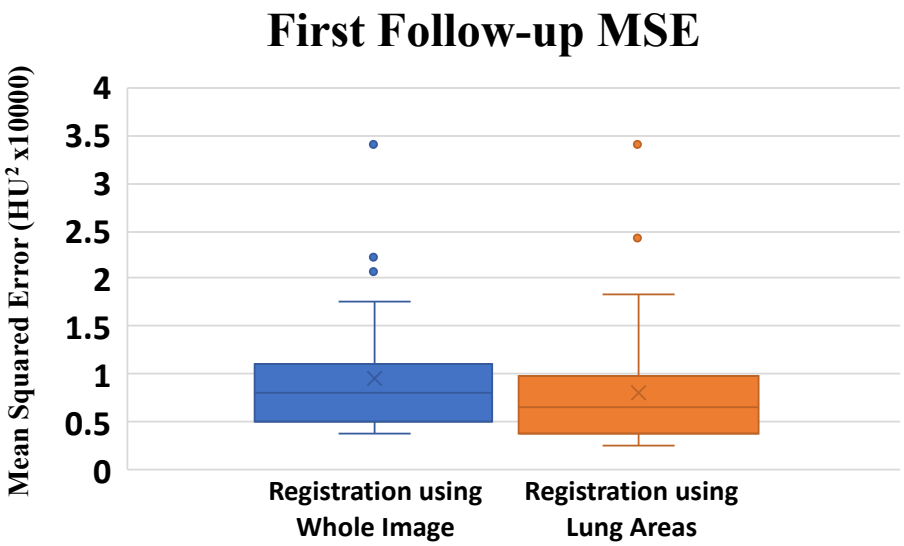

Figure 4-13: Box plot of Mean Square Errors (MSE) distributions between the baseline images and the first follow-up images, calculated using 33 patients who have the first F/U scan. The blue box shows the MSE calculated in lung after applying the first step Affine registration based on the whole body areas, with the mean of 9,974. The orange box shows the MSE after completing the second step of non-rigid registration focusing on lung areas, which is significantly decreased to the mean value of 8,142, with $p<0.001$. The results indicate that the non-rigid algorithm can further improve registration in the lung areas.

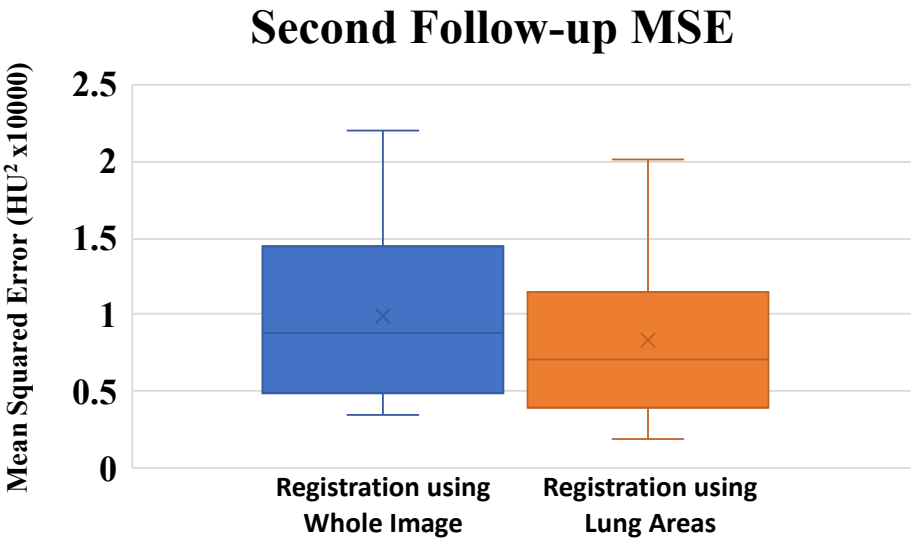

Figure 4-14: Box plot of Mean Square Errors (MSE) distributions between the baseline images and the second follow-up images, calculated using 29 patients who have the second F/U scan. The blue box shows the MSE after applying the first step Affine registration based on whole body areas, with the mean of 9,819. The orange box shows the MSE after completing the second step of non-rigid registration, which is decreased to the mean value of 8,343, indicating a significant improvement ($p<0.001$).

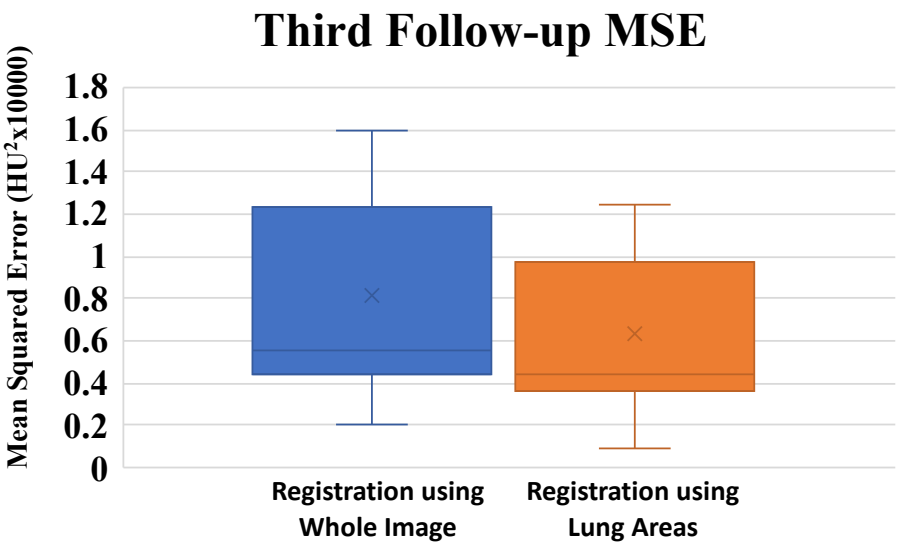

Figure 4-15: Box plot of Mean Square Errors (MSE) distributions between the B/L and the third F/U images, calculated using 11 patients who have the third F/U scan. The blue box shows the MSE after applying the first step Affine registration based on whole body areas, with the mean of 8,175. The orange box shows the MSE after completing the second step of non-rigid registration, which is decreased to the mean value of 6,358, indicating a significant improvement ($p<0.001$).

The registration of two case examples are shown in **Figure 4-16** and **Figure 4-17**. After the registration was completed and the transformation matrix was obtained, the lesion contour drawn on the B/L images were mapped to the F/U images by using the transformation matrix obtained from the registration procedure. **Figure 4-18** and **Figure 4-19** show the original lesion and the transformed lesion. Since non-rigid algorithm was applied, one concern was how much the lesion was deformed, and whether it could still be used to evaluate disease progression. In 33 patients with B/L and first F/U scans, the median change was 1%, but some cases may show > 10% change. The initial results suggest that the proposed registration method may be applied to provide visual comparison of lesions on the B/L and F/U CT, but not good for evaluating volumetric changes. More verification studies are needed. Another application of the developed registration method is to provide correspondence of lesions between B/L and F/U scans,

which may be very helpful to aid in segmentation of lesions on F/U scans based on the initial extent of disease on the B/L scan.

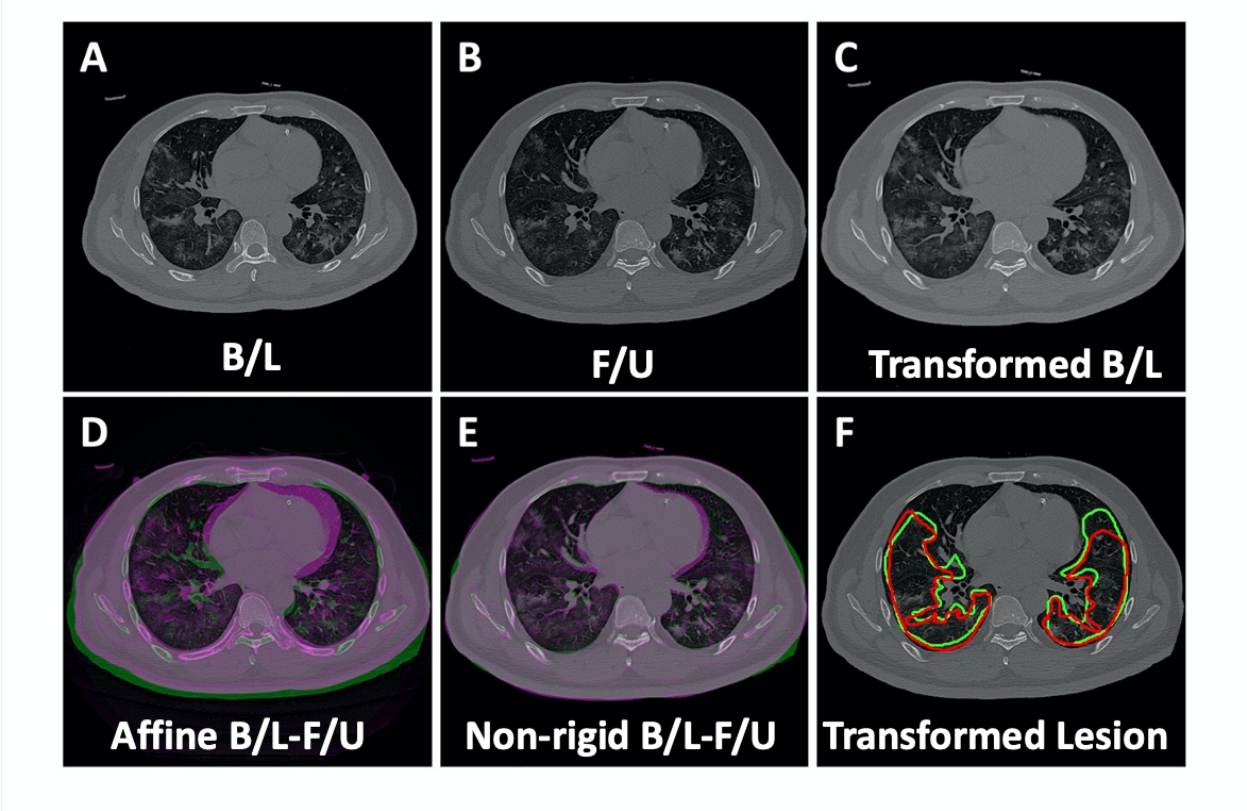

Figure 4-16: An example from a 56-year-old female patient. A: Baseline image. B: The first Follow-up image. C: Final transformed baseline image to match the F/U after completing the two-step Affine and Demons algorithms. D: Comparison between the transformed baseline image after the first-step Affine registration and the original follow-up image by overlay. When the signal intensity on F/U is higher than on B/L, the pixel is labeled using green color; when the intensity on F/U is lower than on B/L, it is labeled using purple color. E: Comparison between the final transformed baseline image and the follow-up image. It can be seen that the difference is minimum and the lung areas are well matched. F: Overlay of the transformed B/L lesion (red contour) and the labeled F/U lesion (green) on the F/U image. For this patient, the total lesion volume is 773 cc on B/L, and 785 cc on F/U, showing stable disease between the first F/U and B/L.

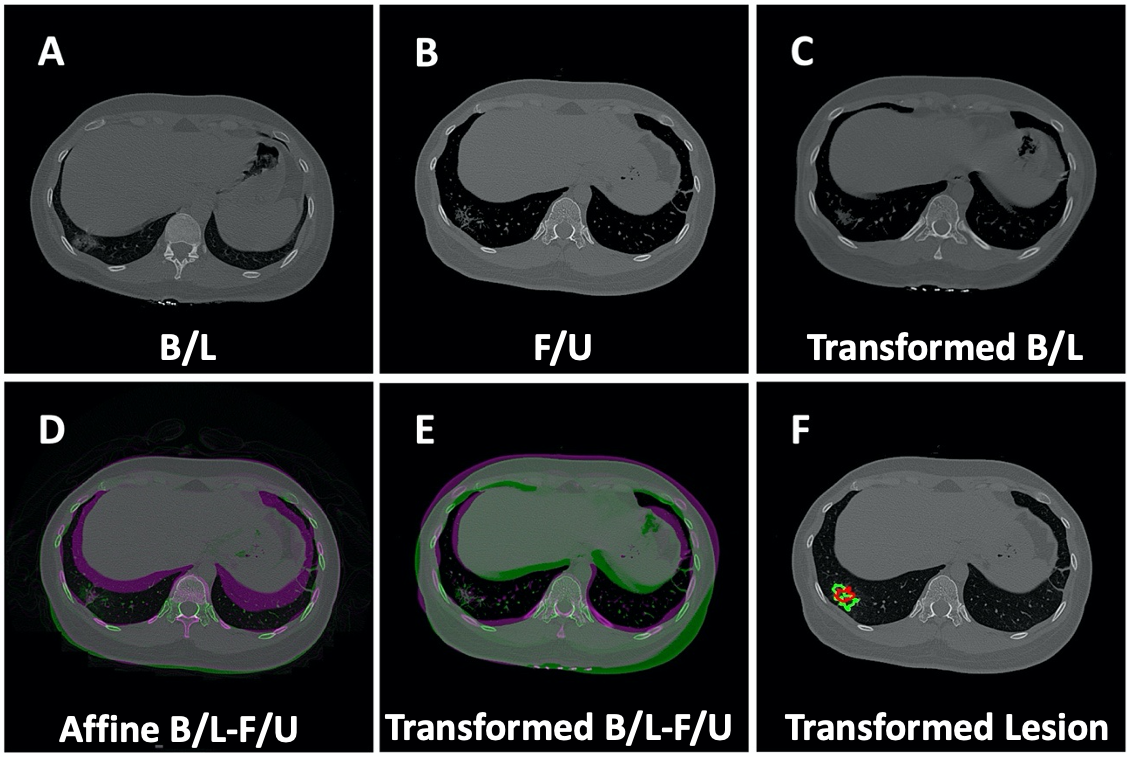

Figure 4-17: An example from a 44-year-old male patient. A: Baseline image. B: The first Follow-up image. C: Final transformed baseline image to match the F/U after completing the two-step Affine and Demons algorithms. D: Comparison between the transformed baseline image after the first-step Affine registration and the original follow-up image by overlay. When the signal intensity on F/U is higher than on B/L, the pixel is labeled using green color; when the intensity on F/U is lower than on B/L, it is labeled using purple color. E: Comparison between the final transformed baseline image and the follow-up image. It can be seen that the difference is smaller and the lung areas are better matched. F: Overlay of the transformed B/L lesion (red contour) and the labeled F/U lesion (green) on the F/U image. For this patient, the total lesion volume is 28 cc on B/L, and 23 cc on F/U, slightly lower volume at F/U but still a stable disease between the first F/U and B/L.

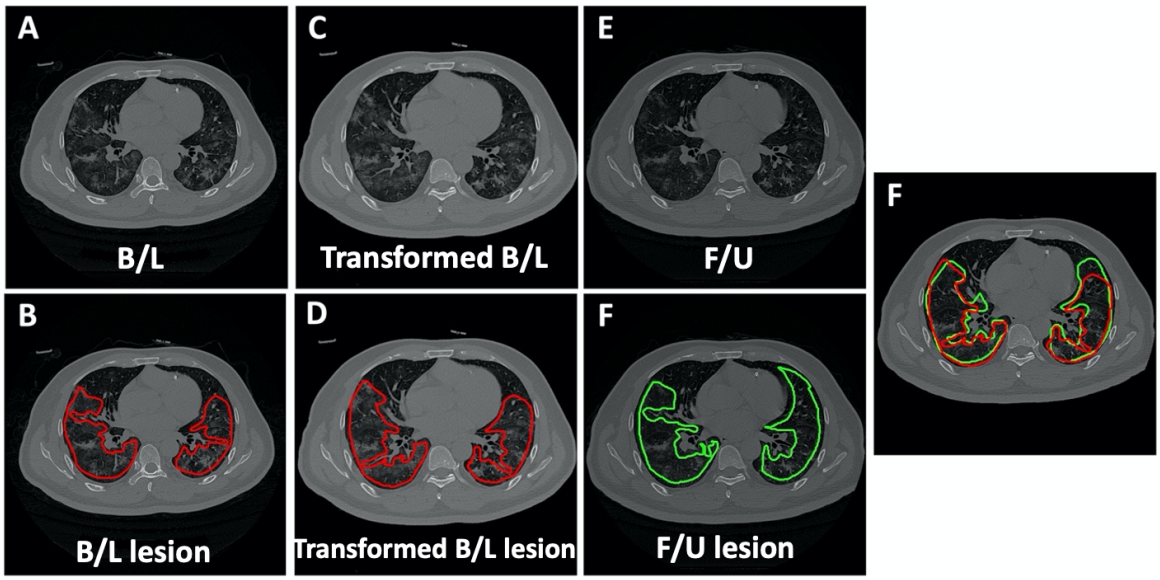

Figure 4-18: The example of 56-year-old female patient shown in **Figure 4-16**. **A:** The original baseline image. **B:** The manually labeled lesion contour (red) overlaid on the B/L image. **C:** Transformed baseline image to match F/U. **D:** The transformed lesion overlaid on the transformed B/L image. **E:** The first F/U image. **F:** The manually labeled lesion contour (green) overlaid on the F/U image. **G**: The transformed B/L lesion and F/U lesion contours are both overlaid on the F/U images, which allows direct comparison of the change. The total lesion volume is 773 cc on B/L, and 785 cc on F/U, showing a stable disease.

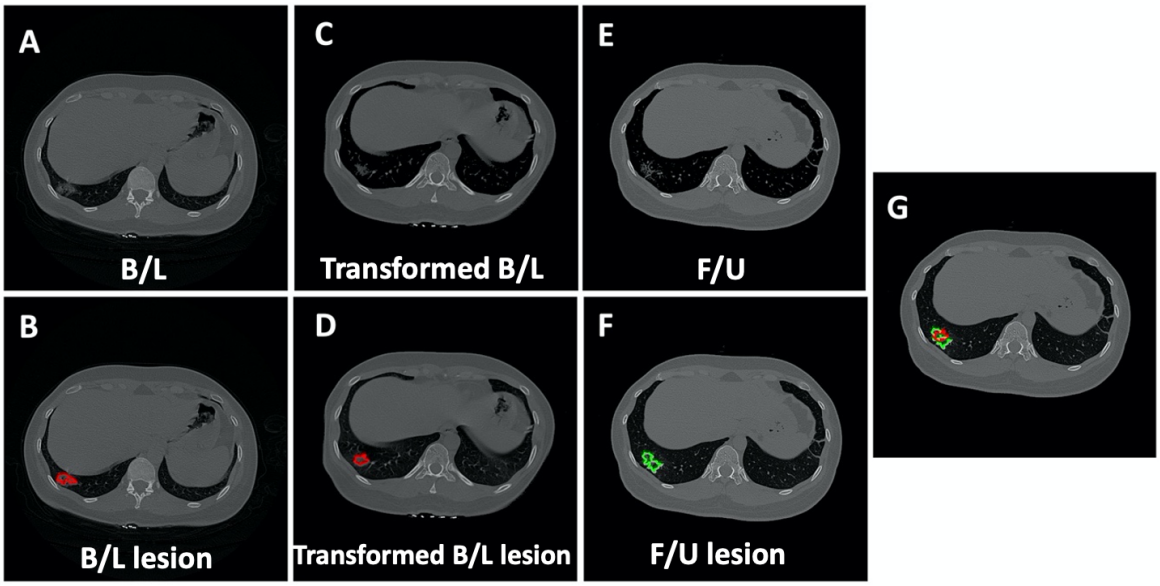

Figure 4-19: The example of 49-year-old female patient shown in **Figure 4-17**. **A, C, E**: The original B/L, transformed B/L to match F/U, and original F/U images. **B, D, F**: The lesion contour overlaid on A, C, E. **G**: The transformed B/L lesion and F/U lesion contours are both overlaid on F/U images, which allows direct comparison of the change. The total lesion volume is 28 cc on B/L, and 23 cc on F/U, slightly decreased but still a stable disease.

The patient shown in **Figure 4-18** had large, diffuse, infection areas, and the total lesion volume was 773 cc on B/L and 785 cc on F/U, showing a stable disease. The patient shown in **Figure 4-19** had the most typical covid-19 lesions presented as ground-glass opacities (GGO). The total lesion volume was 28 cc on B/L and 23 cc on F/U, slightly decreased at F/U but still considered as a stable disease. For each patient who had follow-up scans, the change of the lesion volume was calculated, and shown as the waterfall plot in **Figure 4-20** (for the first F/U compared to B/L) and **Figure 4-21** (for the second F/U compared to B/L). **Figure 4-22** compares the changes from patients who had the first and the second follow-up scans. The results demonstrate that most lesions are stabilized or show a substantial regression and only a few show progression; also that when there is a substantial decrease in lesion volume on the first F/U, the lesion will remain stable in a

regressing state on the second F/U. Several cases show different trends on the first and second F/U, and they may need to be carefully evaluated again. The manual labeling of the lesion contour is very subjective, and is expected to have a high variation.

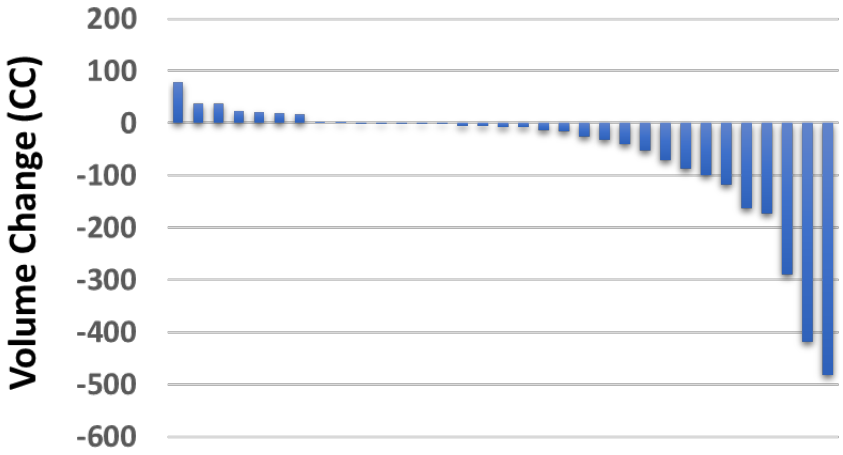

Figure 4-20: The waterfall plot to show the lesion volume change in 33 patients on their first F/U images compared to the B/L. Since many lesions are small, the absolutely change in volume is used. Most lesions are stabilized or show a substantial regression, and only 7 cases show progression.

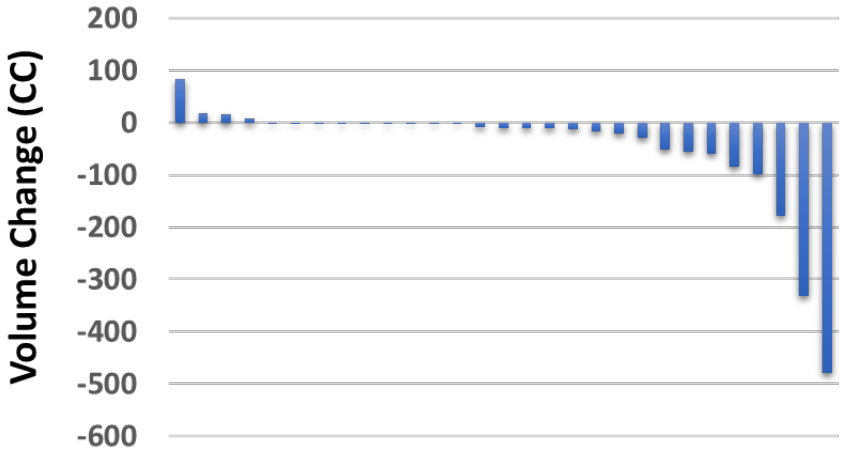

Figure 4-21: The waterfall plot to show the lesion volume change in 29 patients on their second F/U images compared to the B/L. Most lesions are stabilized or show a substantial regression, and only 4 cases show progression.

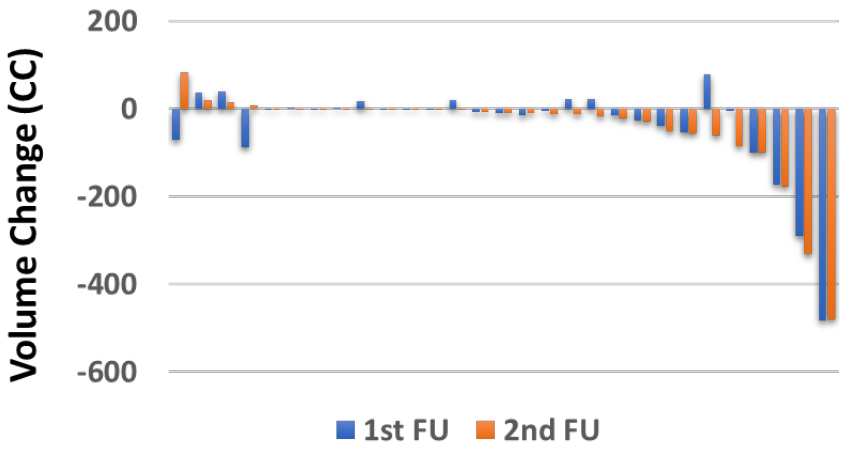

Figure 4-22: The waterfall plot to compare the volume changes from 29 patients who have the first and the second follow-up scans. The results demonstrate that most lesions are stabilized or show a substantial regression and only a few show progression; also that when there is a substantial decrease in lesion volume on the first F/U, the lesion will remain stable in a regression state on the second F/U. Several cases show different trends on the first and second F/U, and they may need to be carefully evaluated again. The manual labeling of the lesion contour is very subjective, and is expected to have a high variation.

4.2.4 Summary and Discussion

The COVID-19 is a devastating disease that has spread all over the world. CT imaging has played an important role in the initial outbreak to help fighting against COVID-19, including detecting suspicious infections and evaluating the severity of pneumonia in the lung. While many studies have reported imaging findings related to COVID-19 infection in the lung and other affected organs (e.g. heart, brain, etc.), and there was an effort to develop a similar reporting system for diagnosis named CO-RADS [175], similar to breast (BI-RADS), prostate (PI-RADS), liver (LI-RADS) cancers, imaging is still under-utilized. With more understanding about the manifestation of the disease, as well as more approved drugs for treatment, imaging can play a more important role in characterization of the disease, selection of appropriate treatment strategies, and monitoring treatment response.

In this study, we evaluated the longitudinal changes of pneumonia severity in different types of COVID-19 lesions at baseline and follow-up images. The lesion ROI was obtained from the radiologists' manual drawing, which is subjective and also heavily influenced by the image quality and the radiologist's experience. Therefore, the focus of the study is not to evaluate the clinical changes of lesions during the monitoring period, rather is to use this unique dataset to develop registration methods to allow direct comparison of lesions by overlaying the lesions on baseline and follow-up imaging studies performed at different times together. The registration was done using two steps, first by Affine registration based on the whole body areas, and then by non-rigid registration using Demons algorithm based on the segmented lung areas. After completing the registration, the transformation matrix was obtained, which was then used to map the baseline lesion onto the follow-up scans, so the changes can be visually compared. Since non-rigid algorithm was applied, one concern was how much the lesion was deformed and whether it could still be used to evaluate disease progression. In 33 patients with B/L and F/U scans, the median change was 1%, but some cases may show > 10% volumetric change, thus it is not reliable to evaluate the volumetric changes based on deformed lesions.

The initial results suggest that the proposed registration method may be applied to help visual evaluation of longitudinal changes of COVID-19 lesions in the lung. However, this was just an initial feasibility study. Many studies can be further developed, e.g. to perform automatic segmentation for different lesion types, to segment lesions on F/U scans based on the initial disease on the B/L scan, and to develop user-friendly tools for treatment response monitoring.

It is worth noting that imaging only provides partial information about patients with COVID-19. It is important to combine imaging data with clinical manifestations and laboratory examination results to help the screening, detection, diagnosis, and therapy monitoring of COVID-19 related to diseases. AI has a great capability in fusing these multi-disciplinary information for performing clinical tasks, or to improve the efficiency and precision of physicians, especially when working in physically and mentally-demanding environment with a heavy workload.

Chapter 5. Differential Diagnosis for Lesions in the Breast, Prostate and Spine

One of the most fundamental problems in computer vision and pattern recognition is classification [10, 55]. Image classification is the one of the most popular application of machine learning in which many algorithms can be utilized. In general, the process of image classification is to extract image features then classify the extracted features [20-22]. Therefore, how to extract image features and analyze image features is the key point of image classification. The traditional classification methods extract the pre-defined imaging features to represent an image. The extracted features are selected and trained to form a classifier using different machine learning classification algorithms. However, this method cannot explore sufficient and proper information from original images and heavily depends on the image quality. Meanwhile, different algorithms can lead to different performance considering the applications [20].

Different from radiomics method, the deep learning method combines the process of image feature extraction and classification on one network [46, 53]. The high-level features representation of deep learning has proven to be superior to hand-crafted low-level features and mid-level features and achieved good results in image recognition and image classification. This concept lies at the basis of the deep learning model (network), which is composed of many layers (such as convolutional layers and fully connected layers) that transforms input data (e.g. images) to outputs (e.g. classification result) while learning increasingly high-level features. The main advantage of the deep learning is that it can automatically learn data-driven (or task-specific), highly representative and hierarchical

features and performs feature extraction and classification on one network, which is trained in an end-to-end manner.

In this chapter, we applied radiomics and deep learning methods into cancer images. First, CNN was established to differentiate the benign and malignant breast tumors, then to identify molecular subtypes on MR images [176]. Second, CNN was utilized to classify the benign and malignant vertebral fractures on MR and CT images. Third, we used radiomics and deep learning to differentiate metastatic lesions in the spine originated from primary lung cancer and other cancers, to traditional hot-spot ROI analysis [177]. Last, we established a bi-directional Convolutional Long Short term Memory (CLSTM) network to diagnose the prostate cancer and benign prostatic hyperplasia.

5.1 Diagnosis of Benign and Malignant Breast Lesions on DCE-MRI by Using Radiomics and Deep Learning

5.1.1 Motivation and Clinical Application

Breast MRI is an important imaging modality for screening, diagnosis and pre-operative staging of breast cancer [178, 179]. Many benign lesions also show strong contrast enhancements, and may lead to false positive diagnosis, unnecessary biopsies or over treatment. With increasing screening and preoperative MRI performed, particularly in community settings [180], an efficient way for characterization of the enhancing lesions is important to improve diagnostic accuracy.

Conventional diagnosis made by radiologists is mainly based on evaluation of the morphological features and the DCE time course, which is subjective and varies with

radiologists' experience. This problem was well recognized, and many computer-aided-diagnosis (CAD) methods have been developed and reported in the literature in the last two decades [103, 181-184]. In addition to providing quantitative parameters related to shape, internal heterogeneity and DCE kinetics, the CAD features were further related to BI-RADS descriptors [103, 182], and used to build separate diagnostic models for mass and non-mass-like enhancements, respectively [183, 184]. With the advance in computer technology, extracting large data from medical images using automatic algorithms becomes feasible; and "radiomics", which allows high-throughput extraction of tremendous amount of quantitative information from radiographic images, emerged [21, 22]. Texture and histogram features based on MR images have potential to provide noninvasive imaging biomarkers to aid in breast cancer diagnosis, prognosis and treatment response evaluation [185, 186]. The radiomics signatures are also related to molecular biomarkers and subtypes, and can aid in patients' management using precision medicine approach [187, 188].

Convolutional Neural Network (CNN) is a common deep learning method applied to analyze photographic, pathological and radiographic images, and reported to have great potential in various clinical tasks such as segmentation, abnormality detection, disease classification and diagnosis [46]. Deep learning has been applied to detect and diagnose breast cancer on mammography, and shows promising results for mass lesions that are comparable to accuracy of radiologists [90, 94, 189-192]. Breast MRI acquires multiple sets of images with varying tissue contrast, and DCE-MRI further acquires images at different times with varying signal intensities that need to be considered, which makes implementation of deep learning algorithms more challenging, and rarely reported [67,

193, 194]. Truhn et al. investigated the diagnostic performance of benign and malignant lesions in MRI using radiomics and deep learning [194]. In their study, the input box was much greater than the size of small lesions, which contained the suspicious lesion with a large amount of peri-tumor and normal tissues, and might affect the diagnostic performance.

The tumor microenvironment is known to play a very important role in growth and invasion of tumor [195, 196], and peri-tumor tissue has been shown to provide helpful information for diagnosis and prediction of prognosis [197-201]. However, how the peri-tumor tissue should be evaluated has not been well studied [200]. The main goal of this study is to evaluate the diagnostic accuracy of breast lesions detected on DCE-MRI with CNN, by using 5 different sizes of input boxes containing tumor with different amount of peri-tumor tissues to evaluate the impact of per-tumor in diagnostic performance. For comparison with the deep learning results, the diagnosis was also done with the whole tumor ROI-based analysis and radiomics.

5.1.2 Subjects and Image Dataset

Patients

A total of 133 patients were included in this study, including 84 patients with a total of 91 malignant cancers (mean age 51±10), and 50 patients with a total of 62 benign lesions (mean age 45±11). One patient had a malignant and a benign lesion. All lesions were confirmed by histologically examination, listed in **Table 5.1**. These cases were selected from consecutive patients receiving breast MRI for diagnosis from January 2017 to May 2018, before biopsy or any treatment. Since one major purpose of this study was to

evaluate the peri-tumor tissues surrounding the lesion, a well-defined tumor boundary was needed, and thus only mass lesions that were visible on contrast-enhanced images were included. For independent testing, the newer cases performed from June to Dec 2018 were collected, by using the same criteria. This study was approved by the ethics committee of our hospital, and informed consent was waived.

Table 5.1: The pathological subtypes in malignant and benign groups in training and testing datasets

Pathology Type	Training Dataset	Testing Dataset
Malignant	N=91	N=48
Invasive Ductal Cancer	75 (82%)	34 (70%)
Ductal Carcinoma In-Situ	11 (12%)	9 (20%)
Other Invasive Cancer	5 (6%)	5 (10%)
Benign	N=62	N=26
Adenosis	31 (50%)	13 (50%)
Fibroadenoma	15 (24%)	8 (32%)
Other Benign Lesions	16 (26%)	5 (18%)

MRI Protocol and Tumor Segmentation

All patients underwent MRI on a 3T scanner (GE SIGNA HDx) using an 8-channel breast coil. The dynamic contrast-enhanced (DCE) scan was acquired using the volume imaging for breast assessment (VIBRANT) sequence, with TR=5 ms; TE=2 ms; FA=10°; slice thickness=1.2 mm; FOV=34×34cm^2; matrix size=416×416. The contrast agent, 0.1 mmol/kg gadopentetate dimeglumine (Magnevist; Bayer Schering Pharma), was intravenously

injected. The DCE series consisted of 6 frames: one pre-contrast (F1) and 5 post-contrast (F2-F6). A radiologist reviewed the images, and indicated the location and the slice range that contained the tumor, and then the tumor ROI was automatically segmented on contrast-enhanced maps by using the fuzzy-C-means (FCM) clustering algorithm with 3D connected-component labeling, as described previously [42, 103]. A second radiologist repeated the segmentation again and the obtained features were compared to test the reproducibility by using intra-class-coefficient (ICC).

5.1.3 ROI-based and Radiomics Analysis

Three heuristic DCE parametric maps were generated according to:

Wash-in Signal Enhancement (SE) Map = [(F2-F1) / F1]

Maximum Signal Enhancement (SE) Map = [(F3-F1) / F1]

Wash-out Slope Map = [(F6 – F3) / F3]

The generated DCE parametric maps were inspected to make sure no motion artifact. The examples from a benign fibroadenoma and a malignant invasive ductal cancer are shown in **Figure 5-1** and **Figure 5-2**, respectively. On each parametric map, 20 Gray Level Co-occurrence Matrix (GLCM) texture features [25], and 13 histogram-based parameters were calculated, with a total of 99 quantitative pixel-wised imaging features. The tumor segmentation was done on each 2-D slice, and they were rendered into a 3-D space with isotropic voxel resolution for extracting the 3D texture features. The intra-class-coefficient (ICC) between the two readers was 0.91±0.11, showing a high reproducibility.

After the features were extracted for all cases, they were properly normalized to mean=0 and standard deviation=1. The random forest algorithm with bootstrap-

aggregated decision trees was applied to select features to build an optimal diagnostic model [41]. The first step was to select important ones and rank the discriminating significance of all features, by using a total of 1,000 trees. During the permutation process, each feature and case could be extracted hundreds of times. The curvature test was implemented during the process of parameter tuning to select uncorrelated features. The significance of each feature was determined based on the decrease of classification accuracy when this feature was removed. The diagnostic performance was tested using 10-fold cross-validation, which could avoid over-fitting and also improve the general applicability of the developed model. The diagnostic model was built by logistic regression, first by using the top 20 features, and then by removing the lowest one, two, three … one by one. The AUC started to decrease substantially after removing 5 features; therefore, the final model was built with 15 features. The detailed radiomics analysis and model-building procedures are described in a recent publication [177].

Five whole tumor ROI-based parameters, including the 1-D tumor size, 3-D tumor volume, mean Wash-in SE ratio, mean Maximum SE ratio, and mean Wash-out slope were calculated. The mean values in the malignant and benign groups of the training and testing datasets are shown in **Table 5.2**. Three ROI-based parameters that gave the best classification performance were selected to train a logistic model for diagnosis. Then, these three ROI-based parameters and 15 radiomics features were used to build a combined diagnostic model.

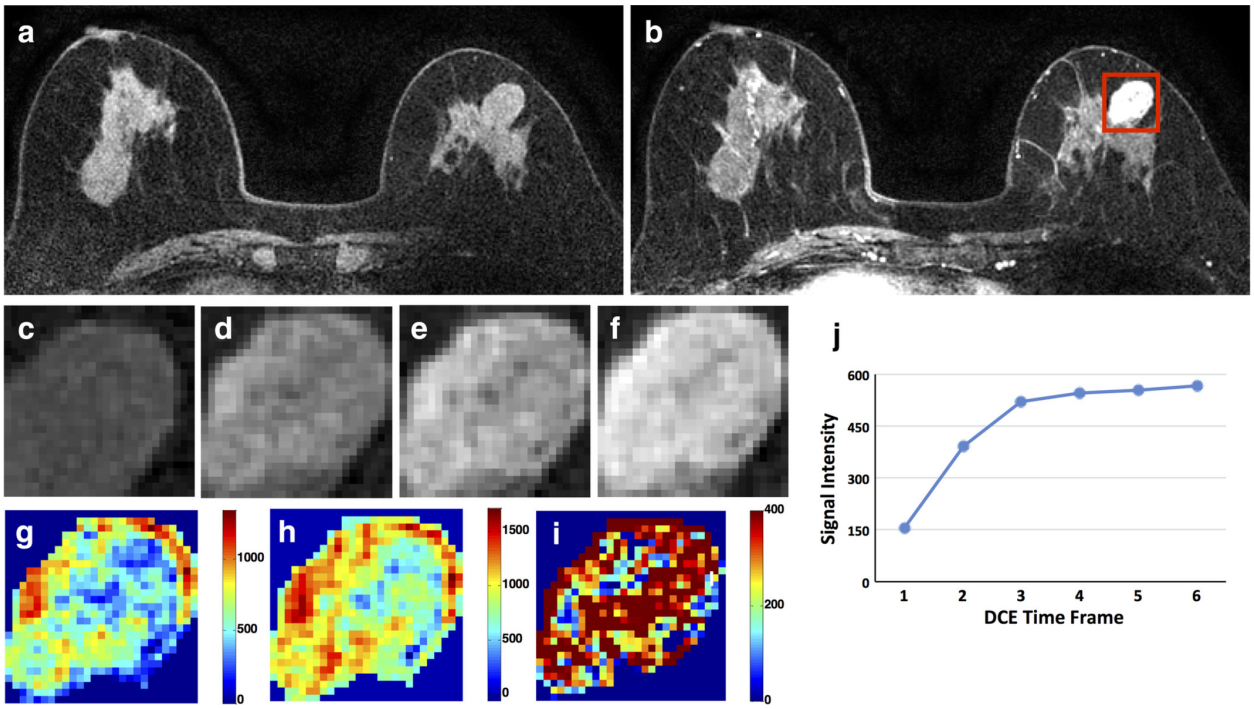

Figure 5-1: A 66-year-old patient with a benign fibroadenoma showing smooth boundary. (a) F1 Precontrast image. (b) The F2 postcontrast image. The red square box is the smallest bounding box. The zoom-in smallest bounding box containing the tumor. (c) The F1 precontrast image. (d) The F2 postcontrast image. (e) The F3 postcontrast image. (f) The last F6 postcontrast image, showing persistent enhancement with increased intensity over time. (g) The washin signal enhancement map F2-F1. (h) The F3-F1 signal enhancement map. (i) The washout F6-F3 map. (j) The DCE time course shows a persistent enhancement pattern from F1 to F6. The predicted malignancy probability is 0.69 for ROI-model (wrong), 0.20 for radiomics (correct), 0.23 for ROI + radiomics (correct), 0.36 for per-slice CNN (correct), 0.51 for per-lesion CNN (wrong based on threshold of 0.5). There are a total of 14 slices for this case, and only one slice has malignancy probability >0.5.

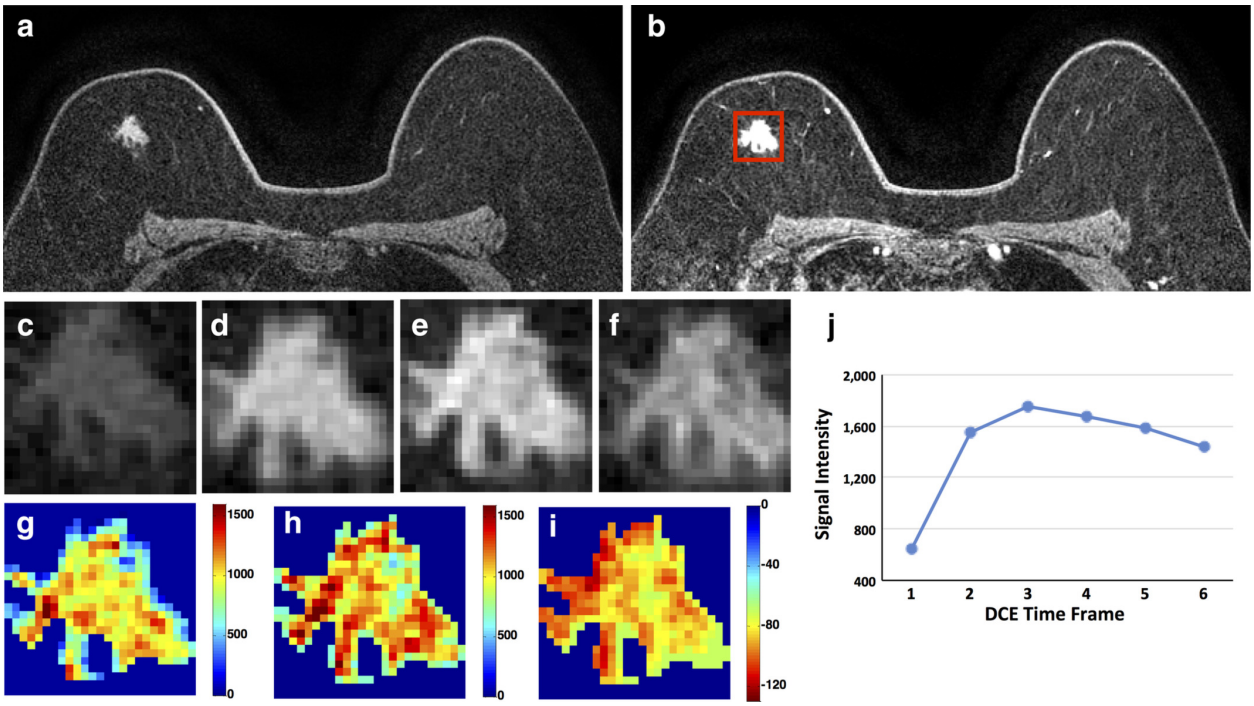

Figure 5-2: A 68-year-old patient with a malignant invasive ductal cancer showing lobulated shape and spiculated margin**.** (a) F1 Precontrast image. (b) The F2 postcontrast image. The red square box is the smallest bounding box. The zoom-in smallest bounding box containing the tumor. (c) The F1 precontrast image. (d) The F2 postcontrast image. (e) The F3 postcontrast image. (f) The last F6 postcontrast image, showing washout DCE pattern with decreased intensity after reaching maximum in F3. (g) The washin signal enhancement map F2-F1. (h) The maximum F3-F1 signal enhancement map. (i) The washout F6-F3 map. (j) The DCE time course shows a typical washout pattern, reaching maximum in F3, followed by decreased intensity from F4 to F6. The predicted malignancy probability is 0.83 for the ROI model, 0.97 for radiomics, 0.97 for ROI + radiomics, 0.97 for per-slice CNN, 0.99 for per-lesion CNN (all correct)

Table 5.2: The whole tumor ROI-based parameters in malignant and benign groups

	Training Dataset		Testing Dataset	
	Malignant (N=91)	Benign (N=62)	Malignant (N=48)	Benign (N=26)
Age	51±10	45±11	49±7	45±7
1-D size (cm)	2.01±0.70	1.44±0.62	1.94±0.86	1.19±0.78
3D Volume (cm^3)	3.74±3.09	1.09±1.46	4.16±3.25	1.13±1.60
Wash-in SE ratio	1.61±0.80	1.15±0.65	1.43±0.75	1.22±0.83
Max SE ratio	2.16±0.96	1.79±0.82	2.07±1.04	1.63±0.75
Wash-out slope	-0.03±0.14	0.09±0.16	-0.02±0.12	0.05±0.09

5.1.4 Deep learning Algorithm Implementation

Deep learning was applied to automatically differentiate the two groups, by using ResNet50 architecture. The conventional convolutional neural network (CNN) learns features using large convolutional network architectures; and in contrast, the ResNet tries to extract residual features, as subtraction of features learned from input of that layer, using "skip connections" [59]. The ResNet50 architecture contains one 3x3 convolutional layer, one max pooling layer, and 16 residual blocks. Each block contains one 1x1 convolutional layer, one 3x3 convolutional layer and one 1x1 convolutional layer. The residual connection is from the beginning of the block to the end of the block. The output of the last block was connected to a fully-connected layer with sigmoid function to give the prediction. The methods were similar to those used in Haarburger et al. [67, 194].

The analysis was done by using three DCE parametric maps as inputs. For each case, the smallest square bounding box containing the entire tumor was generated. This was done by projecting the segmented tumor ROI's from all slices together, and the smallest square box covering the projected boundary was generated. In order to evaluate the diagnostic role of peri-tumor tissues, 5 different input boxes were used, including 1) the tumor alone by setting all outside tumor pixels in the box as zero, 2) the smallest bounding box, 3) enlarged box by 1.2 times, 4) enlarged box by 1.5 times, and 5) enlarged box by 2.0 times. The same box was used for all slices in one case. The input boxes of two benign cases are illustrated in **Figure 5-3**, and those of two malignant cases are shown in **Figure 5-4**.

The bounding box was resized to 75x75 as input into the networks. All tumor slices were used as independent inputs, and the dataset was further augmented 20 times by using random affine transformations. The loss function was cross entropy [45]. The

training was implemented using the Adam optimizer fixed to 0.001 [105]. Parameters were initialized using ImageNet [66]. The L2 regularization was performed to prevent over-fitting of data by limiting the squared magnitude of the kernel weights. Additionally, an early stopping strategy was used, in which the same epoch number was applied to all folds in cross validation. The classification performance was evaluated using 10-fold cross-validation, and each case had only one chance to be included in the testing group. According to the predicted malignancy probability for each slice, the results from all cases were combined to generate the ROC curve.

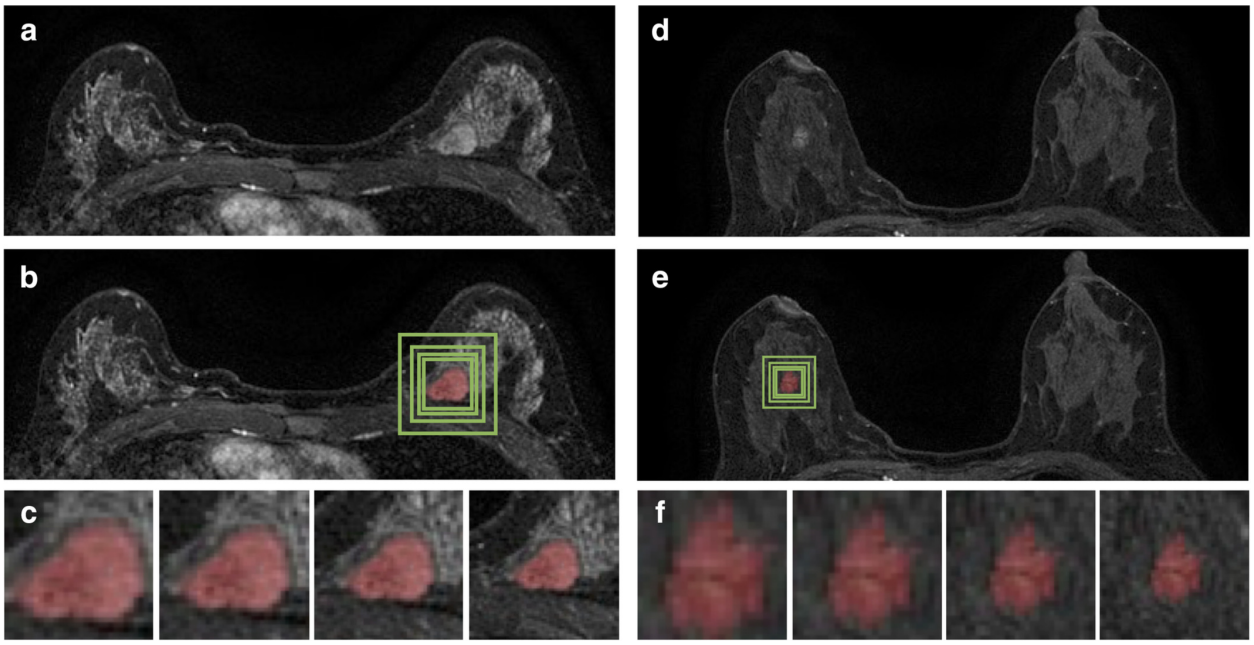

Figure 5-3: Two benign cases. A 41-year-old patient with a benign fibroadenoma showing smooth boundary**.** (a) The F3 postcontrast image. (b) The green box is the smallest square bounding box, and 1.2, 1.5, and 2 times expanded larger boxes. (c) The zoom-in image of the smallest, 1.2, 1.5, and 2 times boxes showing tumor with different amount of peritumor tissues. The predicted malignancy probability is 0.47 for the ROI model, 0.08 for radiomics, 0.10 for ROI + radiomics, 0.29 for per-slice CNN, 0.37 for per-lesion CNN (all correct). (d–f) A 54-year-old patient with a benign fibroadenoma showing low enhancement with indistinct boundary. The predicted malignancy probability is 0.28 for the ROI model, 0.02 for radiomics, 0.02 for ROI + radiomics, 0.29 for per-slice CNN, 0.29 for per-lesion CNN (all correct).

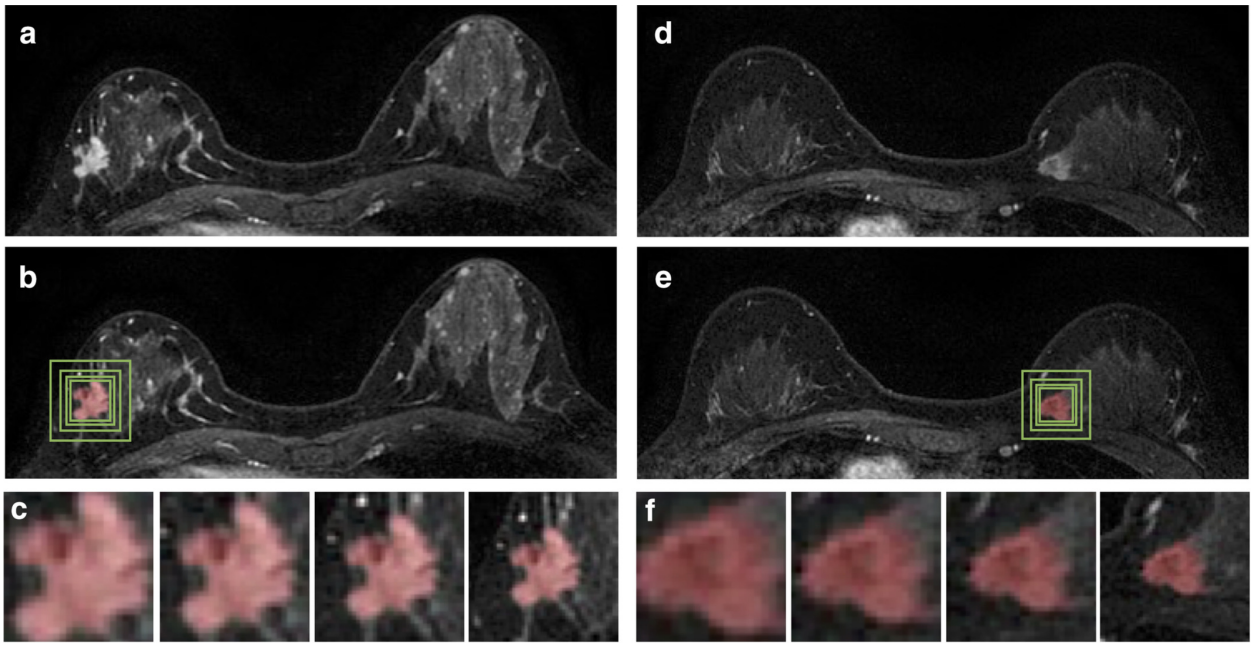

Figure 5-4: Two malignant cases. A 44-year-old patient with an invasive ductal cancer showing lobulated shape and spiculated margin**.** (a) The F3 postcontrast image. (b) The green box is the smallest square bounding box, and 1.2, 1.5, and 2 times expanded larger boxes. (c) The zoom-in image of the smallest, 1.2, 1.5, and 2 times boxes showing tumor with different amount of peritumor tissues. The predicted malignancy probability is 0.61 for the ROI model, 0.89 for radiomics, 0.90 for ROI + radiomics, 0.98 for per-slice CNN, 0.98 for per-lesion CNN (all correct). (d–f) A 41-year-old patient with an invasive ductal cancer with a clear medial boundary. The predicted malignancy probability is 0.41 for the ROI model, 0.29 for radiomics, 0.38 for ROI + radiomics (wrong prediction), 0.83 for per-slice CNN, 0.99 for per-lesion CNN (correct prediction).

The prediction results based on 2D slices meant each slice had its own diagnostic probability. For per-lesion diagnosis, the highest probability among all slices of one lesion was considered. Using this definition could increase the false positive rate, and to investigate this, the results obtained using per-slice and per-lesion basis were compared.

5.1.5 Evaluation and Results

Statistical Analysis

The statistical analysis was performed with SPSS 16.0, with $P<0.05$ was considered significant. ROC analysis was performed using the predicted malignancy probability for each slice, and the AUC obtained using 5 different input boxes was compared by using the DeLong test. For making benign vs. malignant diagnosis, the malignancy probability of 0.5 was used as the threshold to calculate the sensitivity, specificity, and overall accuracy.

Benign and Malignant Case Examples

Figure 5-1 shows the DCE images (F1, F2, F3 and F6), segmented tumor, three parametric maps, and the mean DCE time course of a benign fibroadenoma. For this slice, the malignancy probability predicted by per-slice ResNet was 0.36, correctly diagnosed as benign. However, for the whole lesion, one out of the total of 14 slices had the highest malignancy probability of 0.51 and that was assigned to this lesion, leading to a wrong per-lesion ResNet diagnosis. **Figure 5-2** shows the results of an invasive ductal carcinoma, and all models correctly diagnose this lesion as malignant. **Figure 5-3** illustrates two benign fibroadenomas, one with smooth boundary and the other showing a low enhancement. All models correctly diagnose both lesions as benign. **Figure 5-4** illustrates two invasive ductal cancer, with lobulated shape, spiculated and indistinct margin. Most models gave correct diagnosis, except the radiomics model for Figure. 4D case.

ROI-based Volume and Mean DCE Parameters

When using three parameters, including 3D tumor volume, wash-in SE ratio and wash-out slope, to build the diagnostic model, the overall diagnostic accuracy was 76%. The diagnostic sensitivity, specificity and accuracy are summarized in **Table 5.3**. The model

developed from the training dataset was applied to the testing dataset, and the accuracy was 67%.

Radiomics Analysis

The results are shown in **Table 5.3**. The plot of the malignancy probability based on the final radiomics model is shown in **Figure 5-5**. The diagnostic accuracy was 84 %. When combining the three whole tumor ROI-based parameters and the 15 selected radiomics features together, the accuracy was improved to 86%. When applying these models to the testing dataset, the accuracy was 78% for radiomics, and 77% for ROI+radiomcis.

Table 5.3: The diagnostic sensitivity, specificity and the overall accuracy using models built by ROI-based volume and DCE parameters, radiomics, and ResNet50 deep learning, with a fixed threshold of malignancy probability=0.5

	Training Dataset (10-fold Cross-Validation)				Independent Dataset		
	Sensitivity %	Specificity %	Accuracy %	AUC	Sensitivity %	Specificity %	Accuracy %
ROI Volume + DCE	77% (70/91)	74% (46/62)	76% (116/153)	0.82	71% (34/48)	62% (16/26)	67% (50/74)
Radiomics	91% (83/91)	73% (45/62)	84% (128/153)	0.91	85% (41/48)	65% (17/26)	78% (58/74)
ROI + Radiomics	91% (83/91)	77% (48/62)	86% (131/153)	0.91	83% (40/48)	65% (17/26)	77% (57/74)
CNN, Per-Slice Basis							
ResNet (Tumor Alone)	95% (1285/1358)	74% (362/488)	89% (1647/1846)	0.97	84% (848/1022)	66% (190/289)	79% (1038/1311)
ResNet (Smallest Box)	95% (1286/1358)	94% (460/488)	95% (1746/1846)	0.98	86% (879/1022)	79% (226/289)	84% (1105/1311)
ResNet (1.2 Times Box)	99% (1338/1358)	86% (419/488)	95% (1757/1846)	0.99	78% (801/1022)	70% (202/289)	77% (1003/1311)
ResNet (1.5 Times Box)	84% (1146/1358)	68% (334/488)	80% (1480/1846)	0.86	73% (741/1022)	66% (190/289)	71% (931/1311)
ResNet (2.0 Times Box)	90% (1217/1358)	67% (326/488)	84% (1543/1846)	0.71	67% (687/1022)	59% (171/289)	65% (858/1311)
CNN, Per-Lesion Basis							
ResNet (Tumor Alone)	100% (91/91)	61% (38/62)	84% (129/153)	N/A	94% (45/48)	62% (16/26)	82% (61/74)
ResNet (Smallest Box)	99% (90/91)	79% (49/62)	91% (139/153)	N/A	94% (45/48)	81% (21/26)	89% (66/74)
ResNet (1.2 Times Box)	100% (91/91)	60% (37/62)	84% (128/153)	N/A	85% (41/48)	81% (21/26)	84% (62/74)
ResNet (1.5 Times Box)	97% (88/91)	37% (23/62)	73% (112/153)	N/A	85% (41/48)	54% (14/26)	74% (55/74)
ResNet (2.0 Times Box)	99% (90/91)	24% (15/62)	69% (105/153)	N/A	79% (38/48)	46% (12/26)	54% (40/74)

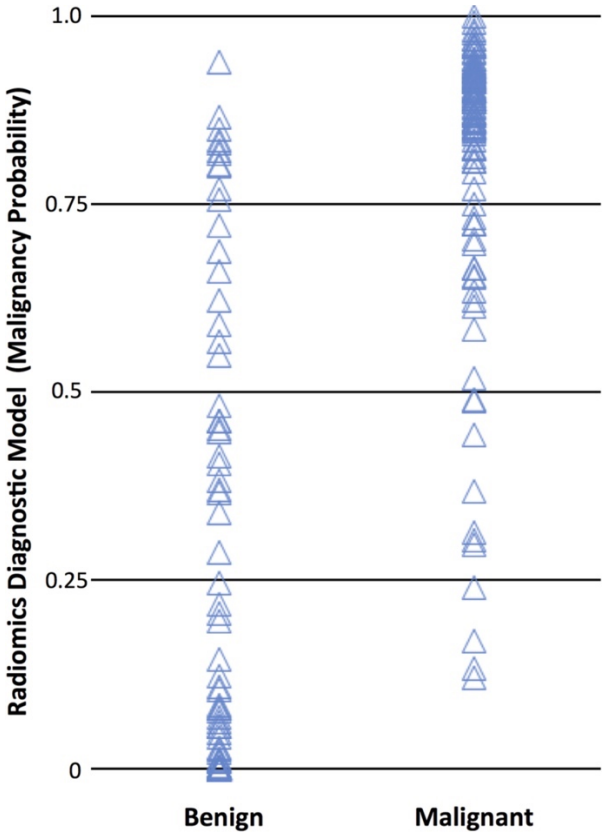

Figure 5-5: The malignancy probability calculated using the radiomics diagnostic model in the malignant and benign groups. Based on the threshold of 0.5, the overall diagnostic accuracy is 84%. Of the total of 91 malignant and 62 benign cases, True Positive = 83 cases, True Negative = 45 cases, False Negative = 8 cases, False Positive = 17 cases.

Deep Learning Analysis Using ResNet50

The results using 5 different input boxes were compared. The mean tumor volumetric percentage was 34% in the smallest bounding box, and that decreased to 28%, 23%, 17% in 1.2, 1.5, and 2.0 times boxes, respectively. In ROC analysis using the predicted per-slice malignancy probability, the AUC was 0.97±0.03 (range 0.93-0.99) for tumor alone, 0.98±0.03 (range 0.90-0.99) for smallest bounding box, 0.99±0.01 (range 0.97-0.99) for 1.2 times box, 0.86±0.07 (range 0.76-0.92) for 1.5 times box, 0.71±0.06 (range 0.63-0.81) for 2.0 times box. The AUC of tumor alone, the smallest bounding box and 1.2 times box was comparable (0.97-0.99), and when the input box was enlarged to 1.5 and 2.0 times, the AUC was significantly

decreased to 0.86 and 0.71, respectively ($p<0.01$ using DeLong test). The ROC curves obtained using these 5 different input boxes are shown in **Figure 5-6**.

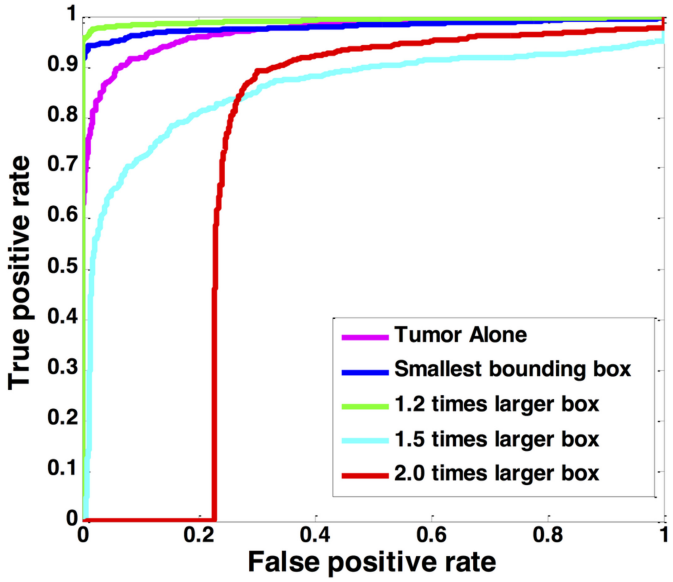

Figure 5-6: The ROC curves generated by using the predicted per-slice malignancy probability of the entire training dataset using ResNet50, with five different input methods: tumor alone, smallest bounding box, 1.2, 1.5, and 2.0 enlarged boxes.

According to the per-slice results, the highest probability was used to make per-lesion diagnosis, using the threshold of 0.5. The results are also shown in **Table 5.3**. When using the tumor alone, the sensitivity was 91/91=100%, the specificity was 38/62=61%, with the overall accuracy of 84%. When using the smallest bounding box, the sensitivity was 90/91=99%, the specificity was 49/62=79%, with the overall accuracy of 91%. The results showed that when considering adjacent peri-tumor using the smallest bounding box, the false positive case was decreased from 24/62 to 13/62, and that improved the specificity from 61% to 79% and the accuracy from 84% to 91%. When using the enlarged boxes with more peri-tumor tissue, the prediction accuracy became worse and worse as the box became bigger and bigger. The accuracy for the per-lesion diagnosis in the testing dataset was comparable, 89% when using the smallest bounding box, and worse for larger boxes.

Per-Lesion Diagnosis Based on Different Malignancy Probability Threshold

The diagnostic results of the four illustrated case examples are given in the figure legends. For the imaging slice shown in **Figure 5-1**, the predicted malignancy probability using ResNet was 0.36, correctly diagnosed as benign. However, for the whole lesion, one out of the total of 14 slices had the highest malignancy probability of 0.51, leading to a wrong malignant diagnosis according to the threshold of 0.5. If the threshold was set higher, this case could be correctly diagnosed. In order to investigate the trade-off between sensitivity and specificity, the results obtained with varying threshold from 0.5 to 0.7 were compared, listed in **Table 5.4**. As expected, increasing the threshold value could improve the specificity, with decreased sensitivity. By using the threshold of 0.5, 0.55, 0.6, 0.65, and 0.7 in the testing dataset, the specificity was 81%, 81%, 92%, 92%, and 100%, with accuracy of 89%, 89%, 81%, 78%, and 53%, respectively.

Table 5.4: The per-lesion diagnostic results obtained using the model built by ResNet50 deep learning with the smallest bounding box, based on different threshold of malignancy probability varying from 0.5 to 0.7

Malignancy Probability	Training Dataset (91 Malignant, 62 Benign)			Testing Dataset (48 Malignant, 26 Benign)		
Threshold ≥	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy
0.50	99%	79%	91%	94%	81%	89%
0.55	98%	95%	97%	94%	81%	89%
0.60	98%	97%	97%	75%	92%	81%
0.65	98%	100%	99%	71%	92%	78%
0.70	95%	100%	97%	27%	100%	53%

5.1.6 Summary and Discussion

In this section we evaluated the diagnostic performance of breast mass lesions detected on DCE-MRI using ROI-based, radiomics, and deep learning methods with different input box sizes. The accuracy was 76% using ROI-based parameters, 84% using radiomics, and 86% using combined ROI+radiomics. In deep learning using ResNet50 with the smallest bounding box, the accuracy was improved to 91%. The results obtained in the testing dataset using newer cases were comparable, showing exactly the same trend with a lower accuracy. Previous studies have shown that the peritumor environments contain important information related to the aggressiveness of the tumor, reflecting lymphovascular invasion and angiogenesis [197, 198], composition of lipid and edema [199-201], or mammary field cancerization [202, 203], and that can be used for prediction of diagnosis or prognosis. In this study we used different sizes of bounding box as inputs to evaluate their diagnostic role. In per-lesion diagnosis, the accuracy was the highest (91%) when using the smallest bounding box, and that decreased to 84% using tumor alone and 1.2 times box, and further decreased to 73% for 1.5 times box and 69% for 2.0 times box. In all 5 input methods, the sensitivity was very high (97-100%), so the accuracy was mainly driven by the specificity. Since the highest malignancy probability in a lesion was used; if any slice had probability >0.5, that lesion was considered as malignant, as shown in the **Figure 5-1** case example (a benign fibroadenoma mis-diagnosed as malignant because one out of 14 slices had the malignancy probability of 0.51). For diagnostic purposes, a high sensitivity is desired, and our results show that the specificity and overall accuracy can be optimized by carefully selecting the input box size, that is, the amount of peri-tumor tissue taken into consideration. As the size of the box increases, the performance becomes worse

and worse, which might be due to the diluted information by containing too much normal tissue, as well as the degraded input image resolution into the neural networks.

The role of peri-tumor at various distances away from the tumor in predicting tumor aggressiveness has been investigated before. Shin et al. applied a shell-based method and reported that the apparent diffusion coefficient (ADC) of proximal peritumoral stroma could differentiate between low-risk and high-risk breast cancer, but not the middle or the distal peritumoral stroma [200]. Fan et al. also applied a similar method and found proximal peritumoral stroma could differentiate between low and high Ki-67 breast cancer groups [204]. The tissues further away from the tumor boundary contained less information associated with the tumor, thus could be interpreted as "normal"; however, there was no definition of the cut-off distance that could be used to classify tissues into "peri-tumor" vs. "normal".

In addition to deep learning, we also performed diagnosis using traditional tumor ROI-based model and the more sophisticated radiomics model. Since malignant tumors are more likely to be bigger and showing the wash-out DCE pattern with stronger enhancements, using a simple ROI-based model could achieve decent prediction accuracy, 76% in our study. Radiomics could evaluate the internal heterogeneity by using texture and histogram analysis, and the accuracy was improved to 84%, with 14 of 15 selected features from texture. Our accuracy was comparable to that Truhn et al., who reported the AUC of 0.78-0.81 for radiomics [194]. In another study by Whitney et al. to differentiate between benign and Luminal A breast cancer, the AUC was 0.68 using maximum linear size, and 0.73 using radiomics features [205]. Since the radiomics features were extracted from the segmented or manually contoured tumor according to the precise boundary, the margin

might not be well evaluated. Kooi et al. [94] used an expanded area to compute the margin contrast on mammography, which may be implemented to evaluate whether it can improve the diagnostic accuracy of lesions detected on MRI.

In our deep learning, ResNet50 was used as the architecture of the convolutional neural network. Deep learning with various CNN architecture has been applied to differentiate benign and malignant mass lesions on mammography [90, 94, 189-192]. Chougrad et al. used three different CNN, and reported that ResNet50 could reach convergence during optimization process faster than VGG, and obtain a good accuracy [191]. Our ResNet50 method was similar to ResNet18 and ResNet34 used in Haarburger et al. and Truhn et al. [67, 194]. In our study, each slice was used as individual input, and L2 norm regularization, dropout and data augmentation were applied to control overfitting. In per-slice analysis using 10-fold cross-validation, the AUC's were > 0.90 in all runs, suggesting that the trained model was robust and not over-fitted. In ResNet, since it was pre-trained with photographs with RGB colors, only 3 sets of images can be used in input channel. Haarburger et al. investigated various combinations and found that the pre-contrast F1, post-contrast F3 and subtraction (F2-F1) gave the best accuracy [24]. In the present study we used three generated DCE parametric maps as inputs, (F2-F1)/F1 and (F3-F1)/F1 with (F6-F3)/F3 to take the DCE wash-out pattern into account. As T2-weighted images also provide very helpful diagnostic information, other CNN architecture that can consider more sets of images can be investigated in the future.

Two other studies also investigated the application of deep learning for cancer diagnosis on breast MRI. In an earlier study, Antropova et al. [193] used three images as input. In another study [206], they trained a long short-term memory (LSTM) network

which could consider the entire temporal sequences acquired in DCE-MRI, and achieved a significantly improved AUC to 0.88 for differentiation of benign and malignant lesions. In their study, the ROI was selected to cover the segmented lesion, similar to our smallest bounding box. Another paper by Zhou et al. [97] applied weakly supervised 3D deep learning, by using the entire segmented breast as input to predict the presence of benign vs. malignant lesions inside, and obtained AUC of 0.859. However, the main novelty in that study was to localize the lesion, not for diagnosis of detected lesions. Two review papers by Reig et al. [207] and Sheth et al. [92] gave comprehensive information and new research direction about the application of AI and machine learning for analysis of breast MRI.

This study has several limitations. First, the dataset was quite small, especially for deep learning. For medical image analysis using deep learning, it was usually done by using each slice as an independent input, and the dataset was further enhanced with augmentation, and lastly appropriate methods such as L2 norm regularization and dropout were used to avoid overfitting. Therefore, the CNN results were usually compared to ROI-based and radiomics results as a proof-of-concept, not aiming to be used directly as a diagnostic model. Second, the highest malignancy probability among all slices of one lesion was assigned to that lesion; although this could lead to a high sensitivity, it was at the expense of decreased specificity. How to incorporate the predicted per-slice probabilities from all slices with an optimal weighting to yield the per-lesion probability needs to be investigated. Third, in order to investigate the impact of peri-tumor tissue, we only included mass lesions that had a clear boundary in this study. It is known that diagnosis of mass lesions is easier and can achieve a higher accuracy compared to non-mass-like (NML) enhancements. For NML, the tumorous tissues and stroma are mixed, and it is difficult to

define the boundary for investigating the role of peri-tumor. Since a clean dataset with well-enhanced mass lesions was analyzed, the developed diagnostic models in this study may not be applicable to other datasets. Nonetheless, the models developed in deep learning may provide a basis to be applied to other datasets through proper transfer learning, which is an efficient strategy commonly used in clinical implementation of AI-based diagnostic tools.

As a summary, we applied ROI-based, radiomics, and deep learning methods to diagnose mass lesions detected on MRI. The results obtained using 5 different input boxes considering different amount of peri-tumor tissues were compared. It was shown that deep learning can achieve better diagnostic accuracy compared to ROI-based or radiomics models to differentiate benign from malignant lesions. The results also showed that using the smallest bounding box that included small amount of peri-tumor tissue adjacent to the tumor had better accuracy compared to using tumor alone or larger input boxes. Although the accuracy of AI-based methods was inferior to that of experienced radiologists [97, 194]; however, this kind of research is needed to make continuing progress, and hope it will become mature in the near future to provide fully-automatic analysis for diagnosis. As many breast MRI is performed in the community setting, the radiologists there may not be well trained to achieve a very high accuracy, and the AI-based diagnostic tools will provide a great help. Automatic, computer-aided, diagnosis using artificial intelligence is emerging, and our study may contribute in development of such diagnostic tools to be used in clinical settings in the near future.

5.2 Prediction of Breast Cancer Molecular Subtypes on DCE-MRI Using Convolutional Neural Network with Transfer Learning between Two Centers

5.2.1 Motivation and Clinical Application

Breast cancer is a heterogeneous group of disease with different phenotypes, and each subtype has different treatment strategy and prognosis. In the standard clinical practice, the status of the hormonal receptor (HR) and human epidermal growth factor receptor 2 (HER2) are evaluated to decide the appropriate treatments, including the use of hormonal therapy and HER2 targeting therapy. Microarray studies have shown that the morphological and clinical heterogeneity of breast cancer has a molecular basis [208]. Breast MRI can accurately reveal the 3-dimensional high spatial-resolution features of the disease, and is a well-established imaging modality routinely used for diagnosis, pre-operative staging and surgical planning [209]. With technological advances in imaging analysis, computer-aided diagnosis (CAD) and radiomics provide efficient methods to extract quantitative features for diagnosis, and they can also be used for molecular subtype differentiation [187, 210-213]. While most studies extract imaging features from the tumor, it has been shown that features extracted from the peri-tumoral parenchyma outside the tumor also contain useful information [213, 214].

After quantitative features were extracted, various classification methods including logistic regression [211, 213, 214], support vector machine (SVM) [212, 214], naïve Bayes model [215] and artificial neural network [216] that could deal with a large number of parameters were applied to build the classification model. While these methods have

yielded promising results, since they relied on pre-determined imaging features, the results were dependent on the choice of computer algorithms as well as the contrast variations and image quality. As such, the developed model might be specific to the analyzed dataset and not generally applicable. In the last several years, deep learning using the Convolutional Neural Network (CNN) have been applied for diagnosis and classification of breast lesions on MRI. In contrast to CAD and radiomics that extract specific features to carry out the classification task, CNN uses the raw image and performs the end-to-end learning for classification. The methods have been used for differentiation of benign and malignant lesions and achieved a high accuracy [176, 194, 206]. They have also been used for multi-class molecular subtype differentiation, which was a much more challenging task compared to diagnosis and in general had a lower accuracy [217-219]. More sophisticated deep learning networks that can fully utilize all information contained in multi-parametric MRI may help.

The purpose of this study was to apply deep learning networks to differentiate three breast cancer molecular subtypes on MRI, including HR positive and HER2 negative (HR+/HER2-), HR negative and HER2 negative (i.e. triple negative, TN) and HER2 positive (HER2+). The smallest bounding box containing the tumor and the proximal peri-tumor tissue was used as the input. A conventional CNN and a recurrent network using convolutional long short-term memory (CLSTM) that could consider the temporal information in DCE-MRI were applied, and the obtained results were compared. An independent testing dataset acquired using a different MR scanner from another hospital was used to evaluate the applicability of the model developed from the training dataset.

Then, the model was re-tuned by transfer learning to investigate its utility for general implementation in different clinical settings.

5.2.2 Subjects and Image Dataset

Patients

This was a retrospective study by retrieving breast cancer cases diagnosed by MRI from two different institutions. The inclusion criteria were consecutive patients receiving MRI for diagnosis or pre-operative staging, and who had surgery with histologically confirmed cancer and molecular subtypes. Only cases presenting as mass lesions with a clear boundary were further selected for this study, in order to minimize the uncertainty in the defined tumor area. The exclusion criteria were patients receiving neoadjuvant treatment such as chemotherapy or hormonal therapy. The molecular subtypes were obtained from the medical record, based on the examination results of immunohistochemical staining and fluorescence in situ hybridization (FISH) from the surgical specimen. The training dataset was obtained from one hospital from Aug 2013 to Dec 2014 performed on a Siemens 1.5T system, with a total of 99 patients (65 HR+/HER2-, 24 HER2+, 10 TN). The mean age was 48 years old (range 22 to 75), and the mean tumor size was 2.6 cm (range 0.4 to 5.0 cm). The independent testing cases were collected from another hospital performed on a GE 3T system. The testing dataset-1 was collected from Jan 2017 to May 2018, with a total of 83 patients (54 HR+/HER2-, 19 HER2+, 10 TN). The mean age was 51 years old (range 24 to 82), and the mean tumor size was 2.0 cm (range 0.7 to 3.5 cm). The testing dataset-2 included later cases collected from June to Dec 2018, with a total of 62 patients (37 HR+/HER2-, 15 HER2+, 10 TN). The mean age was 49 years old (range 33 to 72), and the

mean tumor size was 2.1 cm (range 0.5 to 5.3 cm). The study was approved by the Institutional Review Board and the requirement of informed consent was waived.

MR Imaging Protocol

Only the dynamic-contrast-enhanced (DCE) images were used for analysis. The training dataset was scanned on a 1.5 Tesla scanner (Siemens Magneton Skyra, Erlangen, Germany) with a 16-channel Sentinelle breast coil. DCE-MRI was acquired using a fat-suppressed three-dimensional fast low angle shot (3D-FLASH) sequence with one pre-contrast and four post-contrast frames, with TR/TE=4.50/1.82 msec, flip angle=12°, field of view=32x32 cm, matrix size=512×512 and slice thickness=1.5 mm. The spatial resolution was 0.6×0.6×1.5 mm, and the temporal resolution was 180 seconds for each DCE frame. The contrast medium 0.1 mmol/kg Omniscan® (GE Healthcare, New Jersey, USA,) was administered at the beginning of the second acquisition. The testing dataset was done on a 3T scanner (GE SIGNA HDx, Milwaukee, WI) using a dedicated 8-channel bilateral breast coil. The DCE images were acquired using the volume imaging for breast assessment (VIBRANT) sequence also with fat-suppression, with TR/TE=5/2 msec, flip angle=10°, field of view=34×34 cm, matrix size=416×416 and slice thickness=1.2 mm. The DCE series consisted of one pre-contrast and five post-contrast frames. The spatial resolution was 0.8×0.8×1.2 mm, and the temporal resolution was 130 seconds for each DCE frame. The contrast agent, 0.1 mmol/kg Magnevist® (Bayer Schering Pharma, Berlin, Germany), was injected after the pre-contrast images were acquired.

5.2.3 3D Tumor Segmentation and Preprocessing

The tumor was segmented on the contrast enhancement maps generated by subtracting pre-contrast images from post-contrast images taken at the 2nd DCE frame, using the fuzzy-C-means (FCM) clustering algorithm [42]. The segmentation was performed by two radiologists with 15 and 8 years of experience interpreting breast MRI. The range of slices containing the tumor was decided, and then a rectangle box covering the lesion shown on maximum intensity projection (MIP) was drawn. On each slice, FCM was applied to determine the tumor pixels, and then three dimensional connected-component labeling and hole filling was applied to finalize the tumor ROI. **Figure 5-7** and **Figure 5-8** show DCE images from two patients, with the segmented tumor ROI. Since only mass lesions with a clear boundary were included in this study, the segmentation could be done with computer algorithms, without the need of manual correction. After segmentation, tumor ROI's on all slices were projected together, and the smallest square bounding box covering them was determined as the input for deep learning analysis, as illustrated in [216].

5.2.4 CNN and CLSTM Architectures

For deep learning, each slice was used as an independent input. The cropped frame was resized to 32 x 32. In the training and testing dataset, the images were normalized in the same way to mean=0 and standard deviation=1, so their differences could be handled by standardization. The entire set of DCE images were normalized together so the change of signal intensity could be considered.

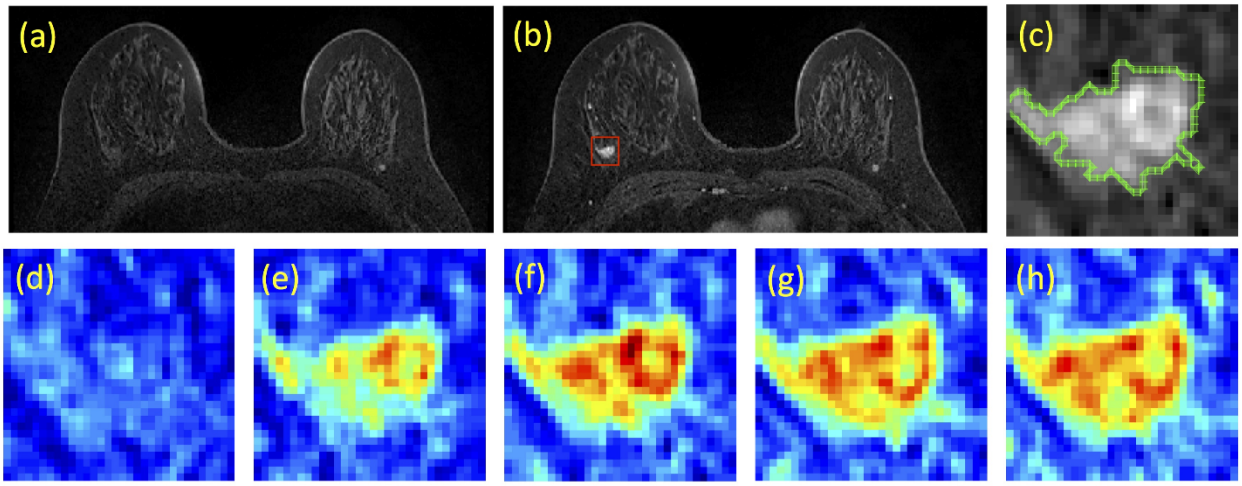

Figure 5-7: A case example from a 53-year-old woman with triple negative breast cancer in the right breast. (a) Pre-contrast image, (b) Post-contrast image, (c) The zoom-in image of the lesion with outlined tumor boundary obtained from segmentation. The square box is centered at the centroid of the tumor. (d-h) Color-coded DCE images at 5 time frames, one pre-contrast and 4 post-contrast, normalized using the same signal intensity scales.

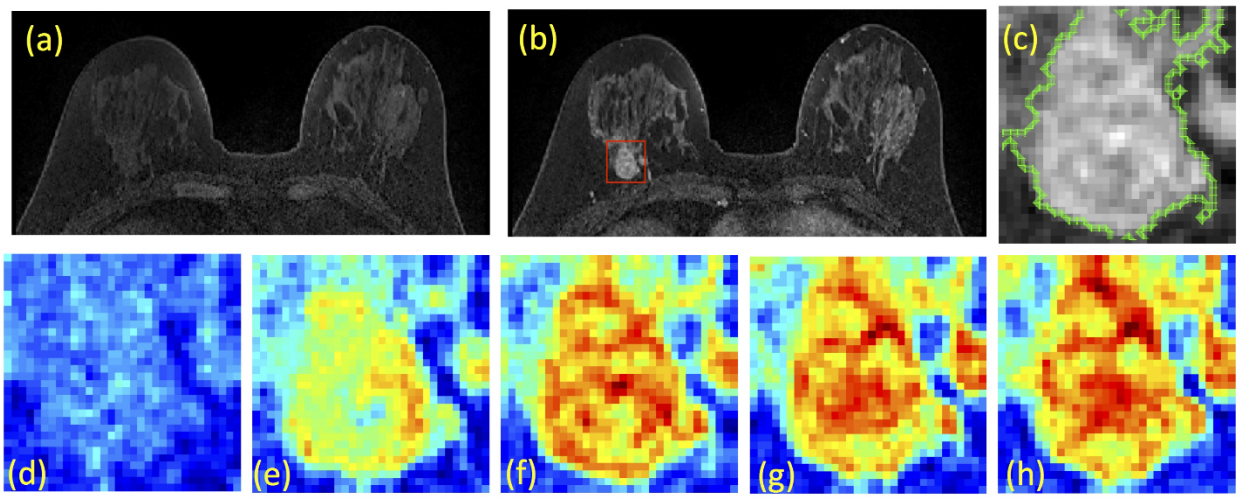

Figure 5-8: A case example from a 48-year-old woman with Hormonal-positive and HER2-negative breast cancer in the right breast. (a) Pre-contrast image, (b) Post-contrast image, (c) The zoom-in image of the lesion with outlined tumor boundary obtained from segmentation. The square box is centered at the centroid of the tumor. (d-h) Color-coded DCE images at 5 time frames, one pre-contrast and 4 post-contrast, normalized using the same signal intensity scales. Although this patient has moderate breast parenchymal enhancement (BPE), the lesion boundary is clearly visible and can be segmented with computer algorithms.

The conventional CNN architecture is shown in **Figure 5-9**. All 5 sets of pre- and post-contrast images were used together, with the input size of 32×32×5. Detailed methods using this CNN were reported in Chang et al. [220]. In brief, the architecture used 7 layers and the size of convolution kernel was 3x3. The stride number of the 2nd, 4th and 6th convolution layers in the output transformation was 2, which reduced the spatial resolution to one fourth the size of the input feature map. Instead of max-pooling, this allowed the network to learn down-sampling parameters and facilitated gradient preservation during back-propagation [220]. After each convolution layer, we used rectified linear units (ReLU), which could lead to faster training and sparse representations. The training was implemented using the Adam optimizer. In the training dataset, the parameters were initialized using the heuristic approach with the "He initialization method" [140]. L2 regularization was implemented to prevent over-fitting by limiting the squared magnitude of the kernel weights. Additionally, an early stopping strategy was used to control over-fitting, in which the same echo number was applied to all folds in cross validation. The learning rate for the Adam optimizer was fixed to 0.001 [105].

Another network, the convolutional long short term memory (CLSTM), was applied to track the temporal information of the changed signal intensity in the DCE time sequence [63], by inputting the 5 DCE datasets into the network one by one, shown in **Figure 5-10**. CLSTM is a recurrent neural network (RNN) and has convolutional layers to implement the input transformations and recurrent transformations. This architecture can extract spatial features as well as temporal features from a series of images acquired in chronologic order. The same input box used in conventional CNN was used for CLSTM, but the size became

32×32×1 instead of 32×32×5. The output was the three subtypes, and the accuracy was calculated using cases that were correctly predicted to the HR+/HER2-, HER2+, and TN groups.

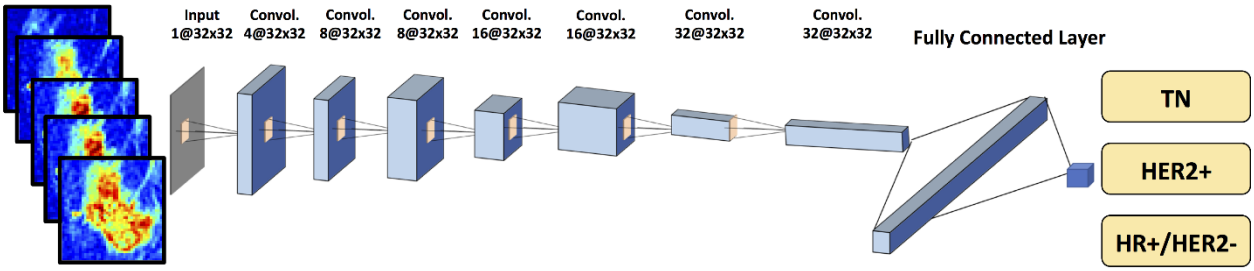

Figure 5-9: Diagram of convolutional neural network (CNN) architecture. The architecture uses 7 serial convolutional 3 x 3 filters followed by the ReLU nonlinear activation function. Dropout at 50% is applied to all convolutional and fully-connected layers after the second layer. Feature maps are down sampled to 25% of the previous layer by convolutions with a stride length of two. The number of the input channels is 5. The number of activation channels in deeper layers is progressively increased from 8 to 16 to 32 to 64. Softmax is used as the activation function of the last fully connected layer.

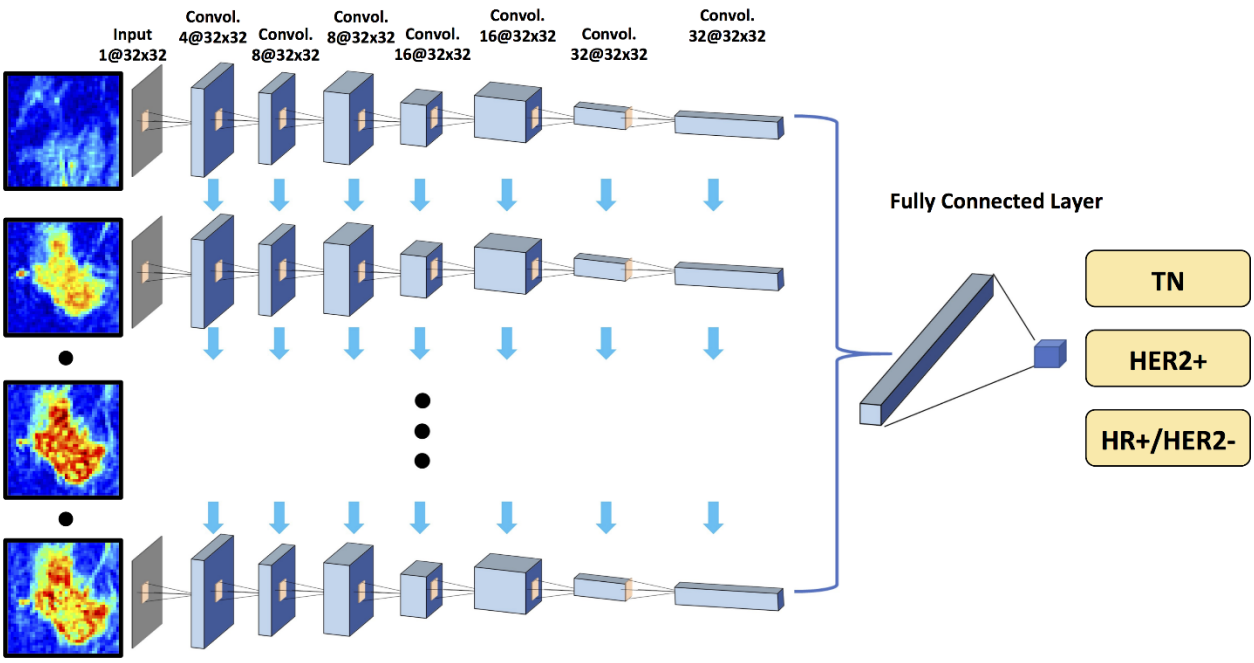

Figure 5-10: Diagram of convolutional long short term memory network (CLSTM) architecture. The architecture uses 7 serial convolutional LSTM layers via 3x3 filters followed by the ReLU nonlinear activation function. Five sets of pre-contrast and post-contrast DCE images are used as inputs. The configuration of the dropout and down sampling are the same as in Figure 3. The number of the input channels is one. Five sets of pre-contrast and post-contrast DCE images are used as inputs, by adding them one by one into the CLSTM network. The number of activation channels in deeper layers is progressively increased from 4 to 8 to 16 to 32. The last dense layer is obtained by flattening the convolutional output feature maps from all states. Softmax is used as the activation function of the last fully connected layer.

5.2.5 Model Evaluation and Transfer Learning

The first model was developed using the training dataset with 10-fold cross-validation. Each case had one chance to be included in the validation group. The results were pooled together, and the range and mean accuracy obtained using CNN and CLSTM were reported. In addition to 3-way subtype classification, the binary classification was performed to generate ROC curves.

After the model was developed, it was directly applied to the testing dataset-1 and dataset-2 for evaluation. Then, in order to consider datasets acquired using different

settings, transfer learning was applied to fine-tune the parameters and develop another model specific to the testing dataset. In the fine-tuning, the weights of the trained network from the training dataset were used as the initial values, instead of using the random He initialization method during the back-propagation process. The transfer learning was done using the testing-1 cases for training with 10-fold cross-validation, and evaluated on testing-2; and then reversely done using testing-2 for training and evaluated on testing-1. This alternative approach could be used to evaluate the robustness of the transfer learning method.

5.2.6 Results

Prediction Accuracy Using CNN and CLSTM

All results are listed in **Table 5.5**. When using the conventional CNN, the mean prediction accuracy in the training dataset obtained using 10-fold cross-validation was 0.79 (range 0.73-0.89). When using CLSTM that considered the temporal information in the DCE series, the mean prediction accuracy in the training dataset was improved to 0.91 (range 0.83-0.95). When the developed classification model was directly applied to the testing datasets, the accuracy was much lower. In Testing-1, the accuracy was 0.52 using CNN model and 0.44 using CLSTM model. In Testing-2, the accuracy was 0.47 using CNN model and 0.39 using CLSTM model. These results showed that the developed model from the training dataset acquired using a different scanner could not be applied to the testing dataset.

Table 5.5: Accuracy to classify three molecular subtypes in Training and Testing datasets using CNN and CLSTM

Dataset	Process	CNN	CLSTM
Training Dataset	Initial Training*	**0.73-0.89 (0.79)**	**0.83-0.95 (0.91)**
Testing Dataset-1	Testing Using the First Trained Model	**0.52**	**0.44**
	Second Training Using Transfer Learning*	**0.85-0.95 (0.91)**	**0.79-0.88 (0.83)**
	Testing Using the Second Model from Transfer Learning of Dataset-2	**0.82**	**0.76**
Testing Dataset-2	Testing Using the First Model	**0.47**	**0.39**
	Second Training Using Transfer Learning*	**0.82-0.89 (0.85)**	**0.74-0.87 (0.82)**
	Testing Using the Second Model from Transfer Learning of Dataset-1	**0.78**	**0.74**

***** The accuracy in the training process is evaluated using 10-fold cross-validation, and the range (mean) is shown

Binary Prediction Accuracy

In addition to 3-way classification in the training dataset, the binary prediction was performed to differentiate HR+/HER2- vs. others; TN vs. non-TN; and HER2+ vs. HER2-. The ROC curves obtained using CNN and CLSTM are shown in **Figure 5-11**.

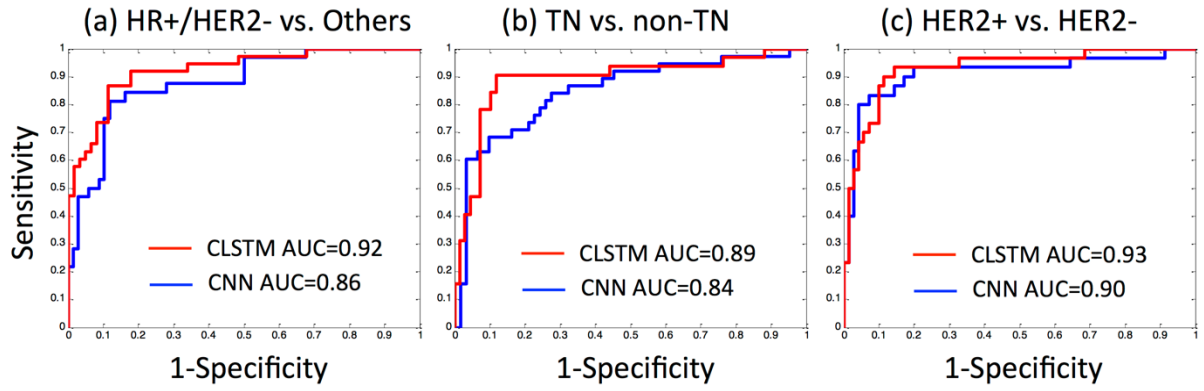

Figure 5-11: The ROC curves for binary molecular subtype classification in the Training dataset obtained using CNN and CLSTM. (a) HR+/HER2- vs. others, (b) TN vs. non-TN, (c) HER2+ vs. HER2-.

The accuracy, sensitivity, specificity and AUC are summarized in **Table 5.6.** The results were in general consistent with the 3-way classification performance, showing a higher accuracy when using CLSTM than CNN.

Table 5.6: Binary molecular subtype classification performance in the Training dataset using CNN and CLSTM

CNN	Accuracy	Sensitivity	Specificity	AUC
HR+/HER2- (N=65) vs. others (N=34)	0.81	0.79	0.82	0.86
TN (N=10) vs. non-TN (N=89)	0.76	0.71	0.79	0.84
HER2+ (N=24) vs. HER2- (N=75)	0.80	0.73	0.83	0.90
CLSTM				
HR+/HER2- (N=65) vs. others (N=34)	0.90	0.89	0.91	0.92
TN (N=10) vs. non-TN (N=89)	0.89	0.82	0.92	0.89
HER2+ (N=24) vs. HER2- (N=75)	0.92	0.90	0.93	0.93

Prediction Accuracy with Transfer Learning

By using the initial trained model as the basis, the parameters were re-tuned in the testing datasets using transfer learning, also evaluated using 10-fold cross-validation. When using CNN, the mean accuracy in re-training of Testing-1 was 0.91 (range 0.85-0.95), and that could be applied to Testing-2 to improve accuracy from 0.47 to 0.78. When using CLSTM, the re-training mean accuracy in Testing-1 was 0.83 (range 0.79-0.88), and that also greatly improved accuracy in Testing-2 from 0.39 to 0.74. Similarly, when using the Testing-2 for re-training, the developed model could be applied to Testing-1 and improved the accuracy from 0.52 to 0.82 using CNN, and from 0.44 to 0.76 using CLSTM. The

improvement is summarized in **Table 5.5**. The second model developed using transfer learning could improve accuracy by 0.31 and 0.30 using CNN, and 0.35 and 0.32 using CLSTM, overall greater than 30%.

5.2.7 Summary and Discussion

Machine learning methods, including radiomics and deep learning, have potential to provide a comprehensive evaluation of the heterogeneous tumor known to be associated with underlying tumor biology [221]. In this study, we applied deep learning to predict three breast cancer molecular subtypes: HR+/HER2-, HER2+ and TN breast cancers that have different treatment strategies. A conventional CNN and a recurrent CLSTM network were used. In the training dataset, the CLSTM that could consider the changing signal intensity in the DCE series achieved a higher mean accuracy of 0.91 compared to the mean of 0.79 by using the conventional CNN. In the independent testing, it was clear that the developed models could not be directly applied. The achieved accuracy was low, only in the range of 0.39-0.52. When transfer learning was applied, the re-tuned model using a subset of testing cases could increase the accuracy in the remaining cases to the range of 0.74-0.82, showing greater than 30% in improved accuracy. This study elaborates how the AI methods developed using one training dataset can be implemented in a different clinical setting, e.g. images acquired using different protocols, different scanners or in different hospitals. Although the approach using transfer learning was trivial, yet few studies have actually implemented the transfer learning and demonstrated how it worked using well-characterized datasets.

Breast cancer molecular subtypes are very important for choosing the optimal treatments. In the present study we used 3 subtypes based on ER, PR and HER2. Other more sophisticated, genomics-based, methods such as PAM50 could be used to provide more detailed genetic make-up, but the classification using Luminal-A, Luminal-B, HER2-enriched and basal-like are known to be closely related to these molecular biomarkers that have direct therapeutic implications. HER-2 targeting agents, Trastuzumab and Pertuzumab, are included in the treatment for HER-2 positive cancer. Long-term (5-10 years) hormonal therapy such as tamoxifen and aromatase inhibitors are used for HR positive cancer to prevent recurrence. For the TN cancers, they are more aggressive and no targeted therapy, and thus, more aggressive chemotherapy is usually given to achieve a better outcome. While these molecular markers can be evaluated from tissues obtained in biopsy or surgery, it is subject to the tissue sampling bias problem. Breast MRI contains rich information, which may be used for differentiation of molecular subtypes, by using images acquired at the time of diagnosis for a thorough assessment of the entire tumor.

For breast DCE-MRI, the pattern of the DCE kinetics (or, signal intensity time curve) is known to provide important information for lesion diagnosis, which can be taken into consideration in deep learning architecture using various strategies [67, 176, 193, 194, 206]. To consider the full spectrum of this time-dependent intensity information, CLSTM was developed to process the DCE images set by set, as in a previous study [177]. The CLSTM is similar to Long Short-Term Memory (LSTM) network reported by Hochreiter et al. [62], which is a Recurrent Neural Network (RNN) used for processing time series and text. In the CLSTM used here, the input transformations and recurrent transformations are both convolutional. This modification makes the recurrent strategy more suitable for

applications in image analysis. By using the recurrent strategy, the temporal features contained in the time order of the 5 DCE pre- and post-contrast MRI sets can be fully explored, and achieved a higher accuracy compared to conventional CNN (0.91 vs. 0.79).

Several studies have applied machine learning, including radiomics and deep learning, to differentiate breast cancer subtypes. Xie et al. applied machine learning methods based on radiomics features extracted from the DCE and DWI images, and showed the best accuracy of 72.4% to classify 4-IHC subtypes, and much higher at 91% when only considering the binary differentiation between TN vs. non-TN [218]. Ha et al. applied a deep learning method using residual neural network for subtype differentiation, and reached 70% accuracy and AUC of 0.85 [219]. Zhu et al. applied several different CNN architectures, including GoogleNet, VGG and CIFAR, to analyze DCE-MRI and achieved the best accuracy of 0.65 [217]. All these studies only analyzed a single-institutional dataset, and the reported accuracy was comparable to our result obtained with convention CNN in the training dataset. In an extensive literature search, there has not been any study that included a second independent dataset for testing, as done in our study. In addition to MRI or other breast images, H&E stained histologic images also contain rich information, and present a great opportunity for deep learning-based analysis for various breast cancer subtype classification, as demonstrated in [222, 223].

The term “transfer learning” is used broadly, which is often referring to pre-training. Usually, the pipeline of CNN classification contains 2 stages. First, a network is pre-trained by a natural image dataset to obtain the weights of the trainable parameters, e.g. ImageNet, which is a set of network weights pre-trained by a large public natural image dataset. Next, the training dataset in the intended application is used to fine-tune the pre-trained network

to achieve the best performance. For example, Nishio et al. [224], applied VGG16 to differentiate benign nodule, primary lung cancer and metastatic lung cancer on lung CT. The network was initialized using ImageNet. The accuracy was 62.3% and increased to 68% with transfer learning. Two other studies by Yuan et al. [71] and Byra et al. [225], also applied a similar strategy using fine-tuned CNN with pre-trained ImageNet, and achieved higher accuracy for prostate and breast lesion classification compared to using other methods without transfer learning. Another strategy, as demonstrated in Samala et al. [69], designed a CNN pre-trained by mammography dataset to classify breast lesions on digital breast tomosynthesis (DBT). In our study, the transfer learning was also used for fine-tuning the model, so the model developed using one dataset can be applied to another. For clinical implementation, the cases were usually acquired in a different setting, and as demonstrated in the present work, re-tuning of the parameters is necessary to improve accuracy. Many companies are developing AI tools, and usually the product can achieve a high accuracy using training datasets. For field implementation in different hospitals, transfer learning based on the specific datasets collected in each hospital is necessary. In the present study, we split the testing cases based on the time of MRI, which represented a realistic clinical scenario. For example, if a commercial AI software is sold to a hospital, it can be re-trained using retrospective datasets, and then applied to prospective cases.

The major limitation was the small case number, particularly for the TN subtype. Unfortunately, this was a common problem for all cancer subtype differentiation studies no matter whether it was based on histology, molecular biomarkers or genetic mutations. Although data augmentation methods were applied, the processed inputs were still similar to each other and highly correlated. For multi-class differentiation to predict breast cancer

molecular subtypes, or to predict different primary tumors in metastasis [32], the overall accuracy was a harsh outcome that often resulted in low accuracy, i.e. each case had to be correctly classified into one of several classes to be counted as accurate. For some clinical applications, combining multi-class into binary classification would be sufficient, e.g. to differentiate lung cancer from other primary cancers in patients with spinal or brain metastasis [177, 226]. The application of machine learning for medical imaging analysis can be designed according to the available case number and the clinical indications, as well as whether there are appropriate datasets that can be used for pre-training.

In conclusion, we have implemented two deep learning networks, conventional CNN and CLSTM, to classify three molecular subtypes that have different treatment strategies. The accuracy in the training dataset could reach 0.8-0.9, but the developed model could not be directly applied to the independent testing dataset acquired in a different hospital using a different scanner. When using part of the testing dataset for re-tuning, the accuracy could be greatly improved by 30%. The results suggest that deep learning can be applied to aid in tumor molecular subtype prediction, and also that transfer learning can be implemented to re-tune the developed model for wide adoption in different clinical settings.

5.3 Differentiation of Benign and Malignant Vertebral Fracture on MR Using ResNet Compared to Radiologist's Reading

5.3.1 Motivation

Imaging plays an important role for evaluation of spinal diseases. Benign and malignant vertebral fracture may be difficult to differentiate due to similar clinical presentations. The correct differentiation and appropriate staging between benign osteoporotic, traumatic and malignant fracture is essential for therapeutic planning, especially in the acute and subacute stages. Benign vertebral lesions occur in approximately one third of cancer patients [227]. Furthermore, fracture resulting from minor trauma is commonly seen in the elderly, which can complicate the evaluation and diagnosis of malignant lesions.

In a clinical setting, images acquired using various modalities are evaluated by radiologists and other clinicians. The diagnostic accuracy is dependent on the medical specialty and the levels of experience [228, 229]. Studies have shown the misdiagnosis rate of vertebral fracture can be as high as 20% [227]. In a study of chest radiography, more than 50% of patients with vertebral fractures are undiagnosed in the original radiology reports when the injury is subtle [230]. The neuroradiologists are in general more sensitive than body radiologist in detection of spinal fractures [231].

MRI is the most helpful imaging modality for characterization of spinal lesions. When the vertebral fat-containing yellow bone marrow is edematous or replaced by enough amount of cancer cells, it shows signal intensity change on T1-weighted (T1W), T2-weighted (T2W), and fat-suppressed images acquired using short tau inversion recovery (STIR) [232, 233]. However, even with the combined information from images acquired

using all sequences, accurate diagnosis of benign and malignant fracture remains challenging in patients with ambiguous features [234, 235].

Recently, artificial intelligence (AI) based imaging analysis has attracted a lot of attention. The methods can provide a comprehensive evaluation of imaging features, which can be used to aid in diagnosis of many diseases. Machine learning methods have been developed to anatomically localize and categorize vertebral compression fractures on CT images [236].

The purpose of this study is to apply an automatic deep learning with residual Network-50 (ResNet50) algorithm to distinguish benign from malignant fractures on MRI. ResNet employs the residual connection in each block which can prevent the gradients vanishing during training, thus all imaging features can be fully utilized [18]. In the training dataset, an experienced radiologist performed visual reading and gave scores for a panel of imaging features and the final diagnostic impression. The diagnostic performance was compared to the results obtained by deep learning. After the model was developed from the training dataset, it was applied to another dataset collected from a different hospital for independent testing to evaluate the applicability, and how the model could be re-tuned to improve accuracy.

5.3.2 Subjects and Image Dataset

Patients

At this initial stage for proof of feasibility, we only included metastatic cancer for malignant fractures, and osteoporosis and minor traumatic injury for benign fractures. The training and testing cases were obtained from two different hospitals. The training dataset

were randomly selected from the radiological reporting system of one hospital in a period of 4 years, using key words of fracture, vertebral collapse, pathological, or metastasis. A total of 190 patients were identified (mean age 66.5, age range 23-95), 140 with benign (mean age 68.8) and 50 with malignant fracture (mean age 61.7). The malignant cases had either biopsy-proven cancer or known history of primary tumor with progressive disease. The most common primary cancer came from lung followed by colon/rectum, breast, and prostate. All benign cases had no known cancer history and have been followed up with stable disease. The images were reviewed by an experienced musculoskeletal radiologist to confirm the lesion(s). The independent testing dataset were obtained from another hospital, consisting of 226 patients (mean age 62.4, age range 14-96), 113 benign (mean age 58.6) and 113 malignant (mean age 66.8). This retrospective study was approved by the Institutional Review Board with waiver of written consent.

MR Imaging protocols

All subjects in the training dataset received MR imaging of the spine on a 1.5T scanner (GE Signa Excite, Milwaukee, Wisconsin, USA). Imaging sequences included axial and sagittal spin-echo T1-weighted non-fat-sat, axial and sagittal fast spin-echo T2-weighted non-fat-sat, and coronal fast-spin echo T2-weighted fat-saturated imaging sequences. The imaging parameters of the two sequences used for analysis in this study were: sagittal spin-echo T1-weighted sequence with repetition time TR 400 ms, echo time TE 15 ms, matrix 320x192, field of view 30cm, and slice thickness 4mm; and sagittal fast-spin echo T2-weighted non-fat-sat image sequence with TR 3200 ms, TE 90 ms, matrix 448x224, field of view 30cm, and slice thickness 4mm. These images were reconstructed into a matrix of

512x512.MRI of the independent testing dataset was performed on two different 3T scanners. One was 3T GE scanner (Milwaukee, Wisconsin, USA) (N=78, 53 benign 25 malignant), and the other was 3T Siemens scanner (Erlangen, Germany) (N=148, 60 benign and 88 malignant). The imaging parameters of the two sequences used for analysis at 3T GE scanner were: sagittal spin-echo T1-weighted sequence with TR 556 ms, TE 8.7 ms, matrix 320x192, field of view 32cm, and slice thickness 4mm; and sagittal fast-spin echo T2-weighted non-fat-sat image sequence with TR 2190 ms, TE 100 ms, matrix 320x256, field of view 32cm, and slice thickness 4mm. These images were reconstructed into a matrix of 512x512. The imaging parameters of the two sequences used for analysis at 3T Siemens were: sagittal spin-echo T1-weighted sequence with TR 469 ms, TE 9.4 ms, matrix 256x180, field of view 30cm, and slice thickness 4mm; and sagittal fast-spin echo T2-weighted non-fat-sat image sequence with TR 2800 ms, TE 97 ms, matrix 384x288, field of view 30cm, and slice thickness 4mm. These images were reconstructed into a matrix of 384x384.

5.3.3 Radiologists' Reading

In the training dataset, a MSK radiologist (LRY, with 28 years of experience) performed reading and gave the binary score for 15 qualitative features, including: 1) absence of collapse, 2) anterior wedge deformity (preserved posterior vertebral height), 3) compression of entire body, 4) central concave deformity, 5) homogeneous marrow signal (no marrow edema or infiltration), 6) band pattern bone marrow edema, 7) intravertebral dark line or band, 8) intravertebral dark patch, 9) fluid or necrotic cleft, 10) diffuse signal change (marrow edema or replacement) of vertebral body >3/4, 11) intravertebral mass-

like or nodular lesion, 12) anterior/posterior protrusion of vertebral body, , 13) epidural/paraspinal soft tissue mass, 14) pedicle/posterior element involvement, and 15) coexisted skipped nodular lesion or mass-like bone marrow replacement in other vertebra. Based on all features for each patient, the radiologist gave a final subjective diagnostic impression of benign vs. malignant fracture. The 15 scores were further combined to develop a classification model using logistic regression. Fisher's exact test was applied to examine the significance of the association (contingency) between the reading scores in benign and malignant groups with confidence interval of 0.95. The p-value was given.

5.3.4 Deep Learning

The deep learning was performed using the most prominent abnormal vertebra for each case as the input, determined by another experienced radiologist (JHC). The abnormal region was identified on the sagittal T2W images. The square bounding box containing the entire abnormal vertebra was generated, and used as the input. The defined box was mapped to T1W images using linear registration. The input of network included both T1W and T2W images of the identified slice with its two neighboring slices that also contained the lesion. Therefore, the total number of input channel was 6. The bounding box was resized to 64x64 by linear interpolation. The intensities of each patch were normalized to mean=0 and standard deviation=1.

The ResNet50 architecture was applied to differentiate the benign and malignant fracture groups, shown in **Figure 5-12**. The convolutional neural network (CNN), such as VGG or AlexNet, learns features using large convolutional network architectures [53]. In contrast, the ResNet can extract residual features, as subtraction of features learned from

input of that layer, using “skip connections” [59]. The ResNet50 architecture contains one 3x3 convolutional layer, one max pooling layer, and 16 residual blocks. Each block contains one 1x1 convolutional layer, one 3x3 convolutional layer and one 1x1 convolutional layer. The residual connection is from the beginning of the block to the end of the block. The output of the last block is connected to a fully-connected layer with sigmoid function to give the prediction. In ResNet, since it is pre-trained with photographs with RGB colors, only 3 sets of images can be used in input channel [59]. Thus, a convolutional layer with 1x1 filter was added to extract interchannel features and transform from 6 channels to 3 channels.

Each individual benign slice was used as independent inputs, and the dataset was further augmented 20 times by using random affine transformations, including translation, scaling and rotation. Since malignant cases have fewer slices, each individual benign slice was used as independent inputs, and the dataset was further augmented 40 times to balance the data. To control the overfitting, L2 regularization term was added to the final loss function and then, during the training process, early stop was applied based on the lowest validation loss to obtain the optimized model [8]. The loss function was cross entropy. The training was implemented using the Adam optimizer [105]. The learning rate was set to 0.0001 with momentum term β to 0.5 to stabilize training. Parameters were initialized using ImageNet [66]. The batch size was set to 32 and the number of epochs was set to 100.

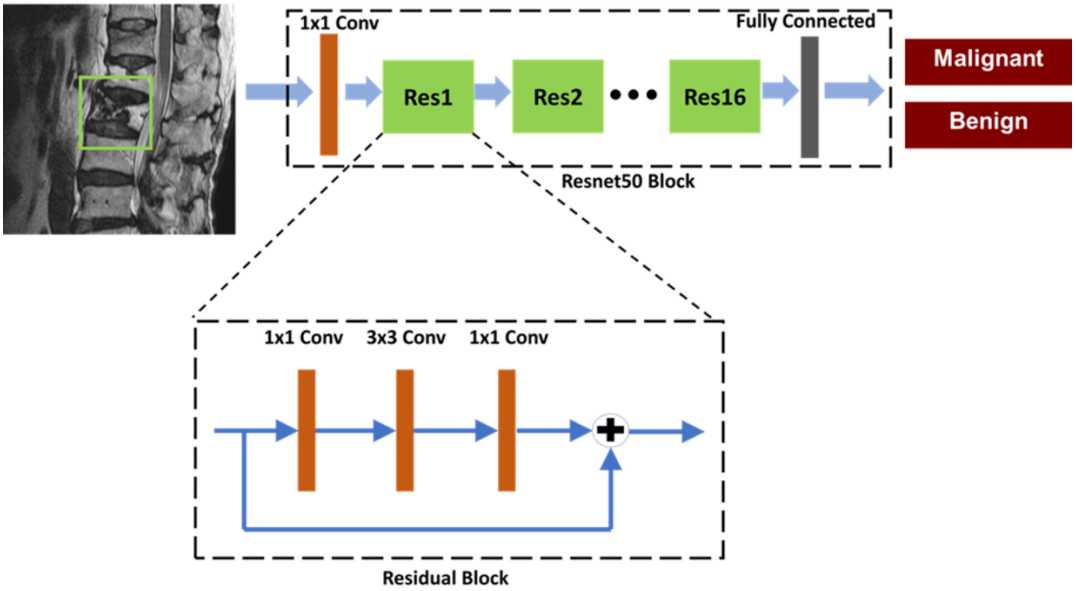

Figure 5-12: Architecture of ResNet50, containing 16 residual blocks. Each residual block begins with one 1x1 convolutional layer, followed by one 3x3 convolutional layer and ends with another 1x1 convolutional layer. The output is then added to the input via a residual connection. The total input number is 6: T1W and T2W of the slice with its two neighboring slices, so one convolutional layer with 1x1 filter is added before ResNet to extract interchannel features and transform from 6 channels to 3 channels as input.

5.3.5 Evaluation in the training and independent testing dataset

In the training dataset, the classification performance was evaluated using 10-fold cross-validation, and each case had only one chance to be included in the validation group. The prediction results based on 2D slices meant each slice had its own diagnostic probability. For per-patient diagnosis, the highest probability among all slices of one patient was considered. Using this definition could increase the false positive rate, and to investigate the difference, the results obtained using per-slice and per-patient basis were compared. From the cross-validation, the best hyper-parameters were determined, and a final model was obtained from the training dataset. The developed model was applied to the second dataset of 113 benign and 113 malignant patients for testing. The malignancy probability for each slice was directly calculated based on the model, and similarly, the

highest malignancy probability among all slices of one patient was used to give a final diagnosis for that patient. To evaluate the diagnostic performance in cases acquired using GE and Siemens scanners, they were separately evaluated.

To adjust the difference between images acquired using GE and Siemens scanners, a resolution fitted model was developed, as shown in **Figure 5-13**. One additional convolutional layer with 3x3 filter size was added for adaptive pre-processing. The input channel number and the output channel number were 6. One third of Siemens patients with matrix size of 384x384 were used to re-tune the trained ResNet50 model, and tested in the remaining two thirds patients for validation.

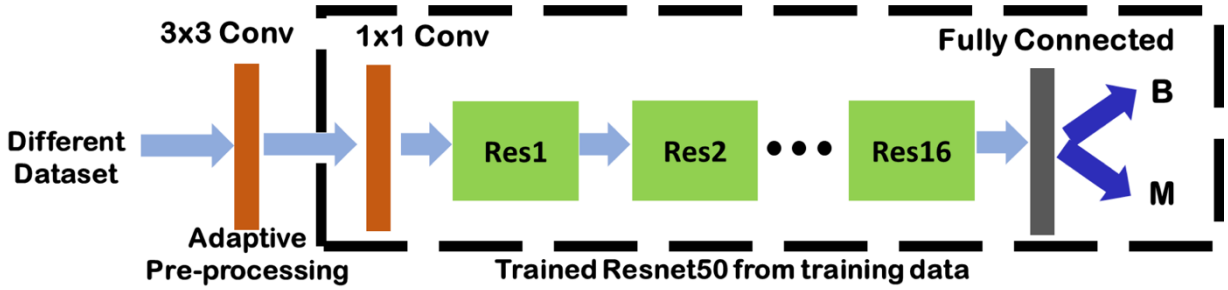

Figure 5-13: Architecture of the resolution fitted model. One convolutional layer with 3x3 filter size is added before the ResNet50 shown in Figure 1 for adaptive pre-processing, to fit the Siemens images reconstructed with 384x384 to training images reconstructed with 512x512. The input channel number and the output channel number are both 6.

5.3.6 Results

Diagnostic performance based on radiologist's reading

The scores of 15 imaging features evaluated by a radiologist in the training dataset is shown in **Table 5.7**. The significance of each imaging feature was evaluated using the Fisher exact test. Ten features showed significant differences between benign and

malignant groups with p<=0.001. The radiologist's diagnostic accuracy based on the final impression was 0.96. When these individual scores were used to build a logistic regression model, the diagnostic accuracy was 0.94. About 34% of malignant vertebral fractures were not associated with apparent collapse or decreased vertebral height. About 41% of benign fractures in our series presented with homogeneous marrow signal (no marrow edema or infiltration), whereas others showed band pattern or diffuse edema. Those without signal change were considered to be old or chronic healed fractures with resolution of the marrow edema. Diffuse signal change occurred more frequently in the malignant group (88%), but still with a considerable percentage in the benign group (22%). Intravertebral dark line or band represented impaction of the bone trabeculae, and was present only in benign fractures (26%). Some fractures showed irregular dark patch in the vertebrae; but, unlike the dark line or band, they were found in both groups with similar incidence (10% vs. 10%). They might represent osteoblastic change, chronic hemorrhage, fibrotic component in tumor, or sclerosis, fibrosis, cement (for vertebroplasty) in benign fracture.

ResNet50 diagnostic performance in training dataset

When deep learning using ResNet50 was applied, the accuracy was 0.84 for per-slice diagnosis, and 0.92 for per-patient diagnosis. There were 3 false negative patients. The mean malignancy probability was 0.25, ranging from 0.03-0.47. There were 12 false positive cases with mean malignancy probability of 0.79, ranging from 0.53-0.97. **Figure 5-14** shows two case examples of malignant fracture correctly diagnosed as true positive. **Figure 5-15** shows two cases of benign fracture correctly diagnosed as true negative. **Figure 5-16** shows two malignant fractures misdiagnosed as benign; and **Figure 5-17** shows two benign fractures misdiagnosed as malignant.

ResNet50 diagnostic performance in testing dataset

In the independent testing dataset from another hospital, the accuracy was different in cases acquired using different scanners. In 78 GE cases with the same matrix size of 512x512 as in the training set, the accuracy was 0.80 for per-slice diagnosis and 0.76 for per-patient diagnosis. In the 148 Siemens cases with a different matrix size of 384x384, the per-slice accuracy was much lower at 0.71, and the per-patient accuracy further reduced to 0.66. After adding the adaptive pre-processing layer in the deep learning architecture to account for different matrix size, the re-tuned model could improve the per-slice accuracy from 0.71 to 0.78, and the per-patient accuracy from 0.66 to 0.74. The diagnostic accuracy is summarized in **Table 5.8**.

Table 5.7: Qualitative Features Evaluated by an Experienced Radiologist

Feature Name	Malignant N=50	Benign N=140	Fisher Test P-value
Absence of collapse	17 (34%)	3 (2%)	<0.001
Anterior wedge deformity (preserved posterior vertebral height)	7 (14%)	77 (55%)	<0.001
Compression of entire body	25 (50%)	54 (39%)	0.18
Central concave deformity	14 (28%)	59 (42%)	0.09
Homogeneous marrow signal (No marrow edema or infiltration)	0 (0%)	48 (41%)	<0.001
Intravertebral dark line, band	0 (0%)	37 (26%)	<0.001
Band pattern bone marrow edema	2 (4%)	44 (31%)	<0.001
Intravertebral dark patch	5 (10%)	14 (10%)	1
Fluid or necrotic cleft	1 (2%)	8 (6%)	0.045
Diffuse signal change (marrow edema or replacement) of vertebral body >3/4	44 (88%)	31 (22%)	<0.001
Intravertebral mass-like or nodular lesion	11 (22%)	0 (0%)	<0.001
Anterior/posterior protrusion of vertebral body	16 (32%)	19 (14%)	0.07
Epidural/paraspinal soft tissue mass	22 (44%)	1 (1%)	<0.001
Pedicle and posterior element involvement	5 (10%)	0 (0%)	0.001
Coexisted skipped nodular lesion or mass-like bone marrow replacement in other vertebra	39 (78%)	8 (6%)	<0.001

The number of patients presenting the feature is reported (percentage)

Table 5.8: Summary of diagnostic accuracy in different datasets using different methods

	Per-slice diagnosis	Per-patient diagnosis
Training Dataset		
Experienced radiologist's diagnosis	N/A	0.96
Logistic model using 15 feature scores	N/A	0.94
Deep learning with ResNet50	0.84	0.92
Independent Testing Dataset		
Directly tested in GE Dataset	0.80	0.76
Directly tested in Siemens Dataset	0.71	0.66
Adaptive processing in Siemens Dataset	0.78	0.74

Figure 5-14: Two true positive malignant cases**.** The image at left panel shows diffuse tumor infiltration at the 7th cervical (C7) vertebral body with posterior cortical destruction and no apparent collapse. The image at right panel shows diffuse tumor infiltration at third thoracic vertebra (T3) with anterior wedge deformity. The fatty change of other cervical vertebrae in the left panel and T2/T4 vertebrae in right panel is post-radiation effect.

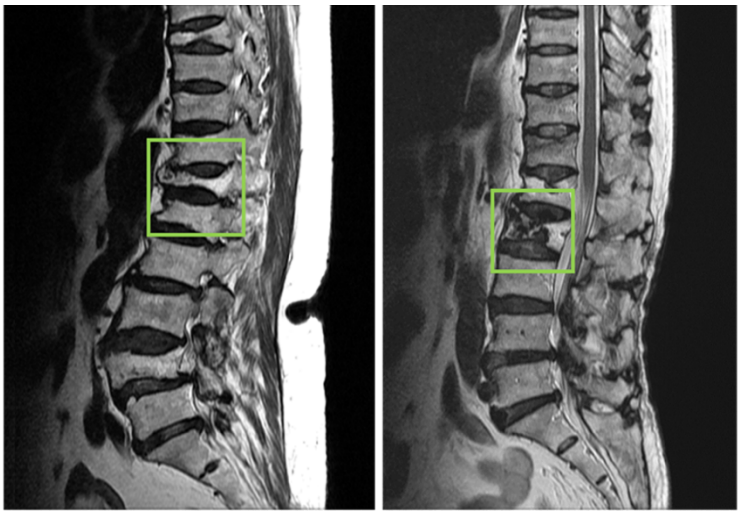

Figure 5-15: Two true negative benign cases. The left case is a chronic benign osteoporotic fracture with resolution of bone marrow edema. Although with severe collapse, the height of posterior vertebral body is still preserved. The right case is a chronic osteoporotic fracture with prior vertebroplasty. The irregular dark patch in the vertebra represents the cement material of vertebroplasty. Both cases show fractures in several other vertebrae.

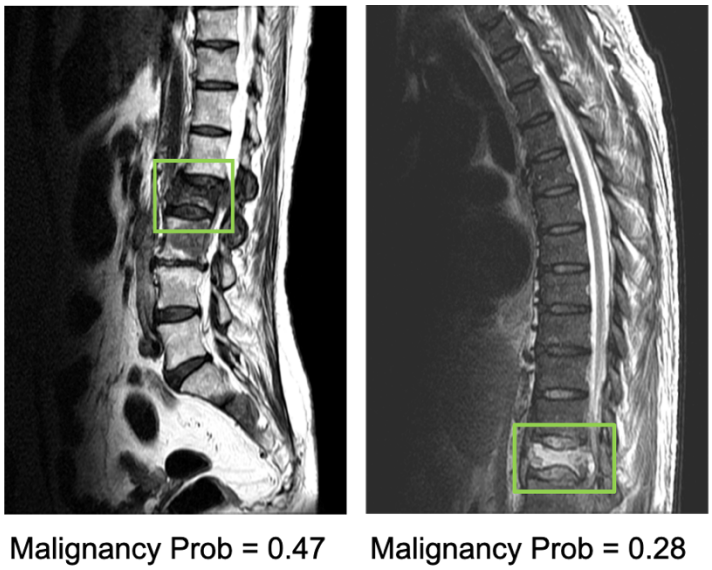

Figure 5-16: Two false negative cases, malignant fracture misdiagnosed as benign. The image at left panel shows diffuse signal change and paravertebral soft tissue mass at L2 vertebra. The coexisted metastatic mass at L3 vertebra is also noted. The right case shows diffuse tumor infiltration, central concave collapse, and paravertebral soft tissue mass.

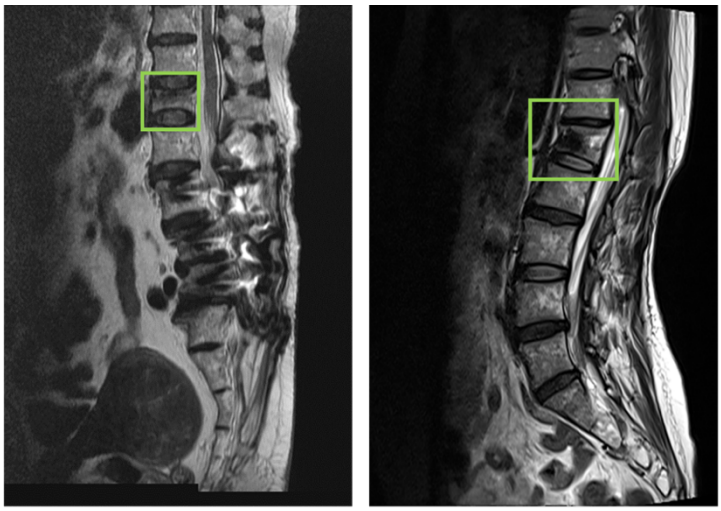

Figure 5-17: Two false positive cases**,** benign fracture misdiagnosed as malignant. The left case is a recent benign fracture with typical band pattern marrow edema. The right case is a benign fracture post cement vertebroplasty.

5.3.7 Summary and Discussion

This study investigated the feasibility of deep learning to differentiate between benign and malignant vertebral fracture on MRI, using T1W and T2W images of three consecutive slices as inputs. The result was compared to the reading of an experienced MSK radiologist. The developed model was further tested in a second dataset obtained from another hospital acquired using two different MR scanners. The results showed that, overall, deep learning using ResNet50 achieved a satisfactory diagnostic accuracy, although inferior to the diagnosis of a senior radiologist who had 28 years of experience. The reasons for a better performance of an experienced radiologist were obvious. First of all, the reading was based on 28 years of diagnostic experience, and no surprise that it could reach 96% accuracy. Secondly, unlike ResNet50 that only considered a small bounding box, the visual assessment included all information derived from the entire images, which also considered

the epidural/paraspinal soft tissue mass, pedicle and posterior element involvement, and coexisted skipped nodular bone marrow replacement in other vertebras, that revealed specific features related to malignancy. Therefore, the head-to-head comparison was not fair. Although the diagnostic performance of deep learning was inferior to that of an experienced radiologist, it may provide a good assistant tool for less experienced radiologists and other physicians. Furthermore, there was a lot of room for improvement to develop a more practical AI model guided by the reading of experienced radiologists, e.g. by considering more inputs from adjacent tissues, more imaging sequences, more imaging planes, etc. The input using only one box covering the most prominent abnormal vertebra in this study was just a starting point to demonstrate the feasibility.

MR Imaging features for the differential diagnosis of benign and malignant vertebral fracture have been well studied [232, 233, 237]. In **Table 5.7,** a total of 15 features were evaluated, and several of them had a good diagnostic implication. Detection of epidural/paraspinal soft tissue mass, pedicle and posterior element involvement, intravertebral mass-like or nodular lesion, and coexisted skipped nodular bone marrow replacement in other vertebra were found to be more specific for malignancy [232, 237]. Band pattern bone marrow lesion and intravertebral dark line or band are more likely benign fracture [232, 237]. The trabeculae of malignant fractures were destroyed before the vertebra collapsed and, theoretically, would have no chance for formation of impacted trabecular band. Whereas malignant vertebral fractures were not always associated with apparent collapse or decreased vertebral height. The cortex destroyed by tumor might not be at weight-bearing part and therefore the vertebral height could still be preserved. Diffuse marrow replacement [233], anterior/posterior protrusion of vertebral body [232],

and non-wedged collapse (central concavity and compression of entire body) are also shared by both benign and malignant fractures. Fracture or collapse in other levels is also not a specific sign since both osteoporotic fracture and malignant fracture can exist in the same patient, especially the elderly [232]. When the fractured vertebra shows equivocal or both features of benign and malignant collapse, the diagnosis may be difficult and challenging, especially for the novice radiologists.

The development of AI-based methods, especially using fully automatic deep learning, not only can assist radiologists to make accurate diagnosis with a higher confidence, also it can be integrated with the clinical workflow and improve the working efficiency [47]. In this study, ResNet50 was used as the architecture of the convolutional neural network. Deep learning with various CNN architectures has been applied to medical images [12]. In our study, each slice was used as individual input, and L2 norm regularization, dropout and data augmentation were applied to control overfitting. In per-slice analysis using 10-fold cross-validation, the AUC's were > 0.90 in all runs, suggesting that the trained model was robust and not over-fitted.

The testing results from the second dataset with images acquired using two different scanners in another hospital clearly showed that the difference in image quality had to be considered. The accuracy in the dataset acquired using GE scanner with 512x512 matrix size was higher compared to the dataset acquired using the Siemens scanner with 384x384 matrix size. However, at this time, it was difficult to pinpoint the most important parameter leading to the different diagnostic results, whether it was the matrix size or the spatial resolution or something else, etc. This was a common problem in implementation of AI tools developed from one setting for another setting. Typically, the re-training, re-tuning, or

transfer learning of the initial model developed from the training dataset by using the second dataset was needed. In the present study, we added one additional layer to adjust the variation of images from different datasets. By using this added layer for re-tuning, the accuracy was improved.

Several studies have applied deep learning for diagnosis of bone fracture on plain radiography [12, 238-240]. Chung et al. used a pre-trained ResNet 152 to classify proximal humerus fractures using plain anteroposterior shoulder radiographs [238]. Olczak at al. analyzed wrist, hand and ankle radiographs using 5 different neural networks [239], and reached accuracies of over 99% on body part, 95% on exam view, 90% on laterality, and 83% on fracture. Kitamura et al. [240] used another 3 different network architectures, including: Inception V3, Resnet, and Xception, to differentiate abnormal from normal radiographs and reached the highest accuracy of 0.8. Also, deep learning has been applied to CT and MR images [241-243]. Raghavendra et al. applied deep learning to distinguish normal cases from thoracolumbar spine injuries [241]. Tomita et al. implemented a sophisticated CNN algorithm using pre-trained ResNet34 Network and Long Short Term Memory (LSTM) to classify the osteoporotic vertebral fractures and normal subjects using CT scans, and achieved 89.2% accuracy [242]. Pedoia et al. employed deep learning using DenseNet for the prediction of osteoarthritis on MRI [243]. All these studies were designed for diagnosis of abnormalities. In the present study, we attempted to differentiate benign from malignant fractures using deep learning, which was much more challenging and rarely reported in the literature, and no results could be compared to.

The training of new radiologists for the interpretation of vertebral fracture takes a very long time and a great effort. The first barrier is the recognition and weighting of features

favoring malignant or benign conditions. Very often, the signs of malignant and benign lesions may coexist in the same patient or even in the same vertebra, and thus, radiologists need to establish their own "weighting" to determine the overall probability and give a final diagnosis. This is a process with a long learning curve that the beginning radiologists usually get frustrated. A similar training strategy can also be implemented in machine learning.

Steps for interpretation of medical images include identification of anatomic structures, detection of lesions, evaluation of image features, and then make a final conclusion based on the data collected. Thus this preliminary study was only the first step to achieve the goal of mimicking the clinical scenario. In this study, the regions of interest were selected by radiologist, and therefore lesion detection was not performed by deep learning. More efforts are required before the AI can be useful clinically, including automatic identification of anatomy and detection of lesions. Refinement of its ability in differential diagnosis is also needed. In the future, a localization strategy including vertebra alignment segmentation and abnormalities search should be established. For spinal segmentation, several literatures have proven different kinds of CNN can obtain satisfactory performance [244-247]. Then applying the presented algorithm in this paper on each segmented vertebral bodies, the malignancy probability of each segmented vertebral bodies can be determined.

This study had several limitations. First, it was a pilot study aiming for demonstrating feasibility, and the case number was relatively small. Second, to limit variations and potential confounding factors, only patients with metastatic cancer were selected in the malignant group, and the results might not be applicable to other primary bone cancers.

Third, although a second dataset from a different hospital was available for validation, they were acquired using two different scanners, and had to be evaluated separately with smaller case numbers. However, images acquired using different scanners allowed us to test the strategy of adaptive pre-processing. The results suggest that for future implementation of AI diagnostic models, re-training is needed for each different clinical setting.

In conclusion, this study investigated the application of deep learning for differential diagnosis of benign and malignant vertebral fracture on MRI. A model using ResNet50 was developed in a training dataset and tested in separate independent datasets. The input used in deep learning was a square box covering a single abnormal vertebral body, without inclusion of the soft tissue, the posterior elements, and the skipped lesions. The differentiation accuracy in the training dataset was 0.92 for per-patient diagnosis, inferior to an experienced radiologist's reading, possibly due to the limited input considered in the deep learning. The testing accuracy in the second dataset acquired from another hospital varied depending on the acquisition protocols or different MR systems. When re-tuning was applied, the accuracy could be improved. The results suggest that deep learning provides a feasible method to consider different T1-weighted and T2-weighted images on MRI to make differential diagnosis. With specific refinement in each clinical setting, the AI-based method has the potential to provide a clinical tool to help less-experienced readers or to improve workflow.

5.4 Differentiation of Spinal Metastases Originated from Lung and Other Cancers using Radiomics and Deep Learning based on DCE-MRI

5.4.1 Motivation and Applications

Patients presenting with pain in the spine are often suspected to have lesions compressing the spinal cord, and MRI is usually performed for diagnosis. The most common malignancy in the spine is metastatic cancer, and approximately 30% of patients present with an unknown primary [248-250]. In these patients, a final diagnosis is needed to proceed with treatment. If the origin of the cancer in the spine can be accurately predicted, this can narrow the search field and help determine the most appropriate imaging method to locate the primary tumor without the need of performing invasive spinal biopsy.

In Western world with established health care systems, PET/ CT is the most commonly used imaging for diagnosis of primary cancer and whole-body staging when the metastatic cancer in the spine is suspected. However, the patient may have to wait for insurance approval and delay the diagnosis. In the developing countries, PET/CT and the 18F-FDG tracer are limited and very expensive, and thus this exam may not be available to many patients. If other cheaper imaging examinations can be used to locate the primary tumor, it will provide a cost-effective management approach to help patients. Among all patients presenting with spinal pain with an unknown primary cancer site, lung metastasis is the most prevalent [250]. If this primary can be accurately predicted by MRI, subsequent workup can be focused to pulmonary imaging, e.g. using CT, which is easily doable and much cheaper.

While conventional MRI can easily detect metastasis in the spine, cancers from different primary appear similar and often indistinguishable. Many studies have shown that dynamic contrast-enhanced MRI (DCE-MRI) can provide additional information for further characterization of the detected spinal lesions [177, 251-257], but only a few tried to differentiate metastasis from different primary cancers [177, 252].

For diagnosis using DCE-MRI, the most common method is to measure the signal intensity time course from a manually-placed region of interest (ROI) to evaluate DCE kinetic parameters. A radiologist can also evaluate the morphological presentation of the tumor, which can be combined with the DCE parameters to make a diagnosis. Additionally, computer-aided or radiomics-based analysis are commonly utilized to extract quantitative parameters from the entire segmented tumor, and that can be used for a thorough evaluation of morphological and DCE kinetic features to aid in diagnosis [20, 22, 258, 259]. Very recently, deep learning has been demonstrated as a feasible, albeit powerful method to automatically evaluate the entire lesion for diagnosis [260-262], or lesion detection [109, 263], without use of pre-defined metrics. All available images can be used as inputs for the algorithm to achieve the best diagnostic accuracy. Each of these three methods has their own pros and cons, which is an active research area for diagnosis.

The purpose of this study is to differentiate metastatic cancer in the spine originated from lung cancer and other non-lung tumors, by using the conventional ROI-based method and the more sophisticated machine-learning based methods, including radiomics and deep learning. The diagnostic results and limitations of these methods were compared.

5.4.2 Subjects and Image Dataset

Patients

This study was approved by the Ethics Committee of our hospital, and the informed consent was waived. In a retrospective review of spinal clinical MRI database in our hospital that included a DCE sequence from 2011 to 2015, a total of 61 patients with confirmed osseous spinal metastases originating from a known primary tumor were identified. The cases were selected by identifying patients who had pain in the supine and came to our hospital for diagnosis using MRI. All of them did not have prior history of any cancer diagnosis. Information regarding primary cancer source was obtained from review of medical records. Distribution of primary cancer sites included: 30 patients confirmed with lung cancer (mean age 56); 9 with breast cancer (mean age 54); 7 with thyroid cancer (mean age 50); 6 with prostate cancer (mean age 72); 6 with liver cancer (mean age 52); 3 with renal cancer (mean age 65). The age and sex distribution between the lung cancer (16 males, 14 females, mean age 56) and other cancer (16 males, 15 females, mean age 57) groups were about the same.

MR Imaging Protocol

MR scans were performed on a 3T Siemens or 3T GE scanner with a consistent protocol. The conventional imaging sequences included transverse T2WI, sagittal T2WI without and with fat suppression, and sagittal T1WI acquired by using the fast spin echo pulse sequence. After the abnormal region was identified on sagittal view, DCE-MRI was performed using the three-dimensional (3D) volume interpolated breath-hold examination (3D VIBE) sequence in the transversal plane to further examine that region. The imaging

parameters were: repetition time TR=4.1 ms, echo time TE=1.5 ms, flip angle=10°, acquisition matrix=256×192 and field of view FOV=250×250 mm. Approximately 30 slices with 3-mm thickness were prescribed to cover the abnormal vertebrae. The temporal resolution varied from 10 to 14 seconds. The contrast agent, 0.1 [mmol/kg] Gd-DTPA, was injected after one set of pre-contrast images were acquired, by using an Ulrich power injector at a rate of 2 ml/s followed by 20 cc saline flush at the same rate. A total of 12 frames were acquired, so the total DCE-MRI acquisition time period ranged from 120 to 168 seconds. When the DCE study was done using the GE scanner, the LAVA (Liver Acceleration Volume Acquisition) pulse sequence with similar spatial and temporal resolution was used. **Figure 5-18** shows two case examples, one from lung and the other from thyroid cancer. The corresponding DCE-MRI, including the pre- and post-contrast images and the subtraction enhancement maps are shown in **Figure 5-19.**

5.4.3 Hot-Spot ROI-based DCE Kinetic Analysis

For each case, an ROI was manually placed on an area that demonstrated avid enhancement and excluded regions with cystic lesions, calcification, necrosis, and hemorrhage, as illustrated in a previous publication [177]. The signal intensity time course was measured and evaluated to find the pre-contrast signal intensity (S0), two adjacent time points that showed the largest difference in their signal intensities (S2 and S1) during the wash-in phase, and the maximum intensity (Smax).

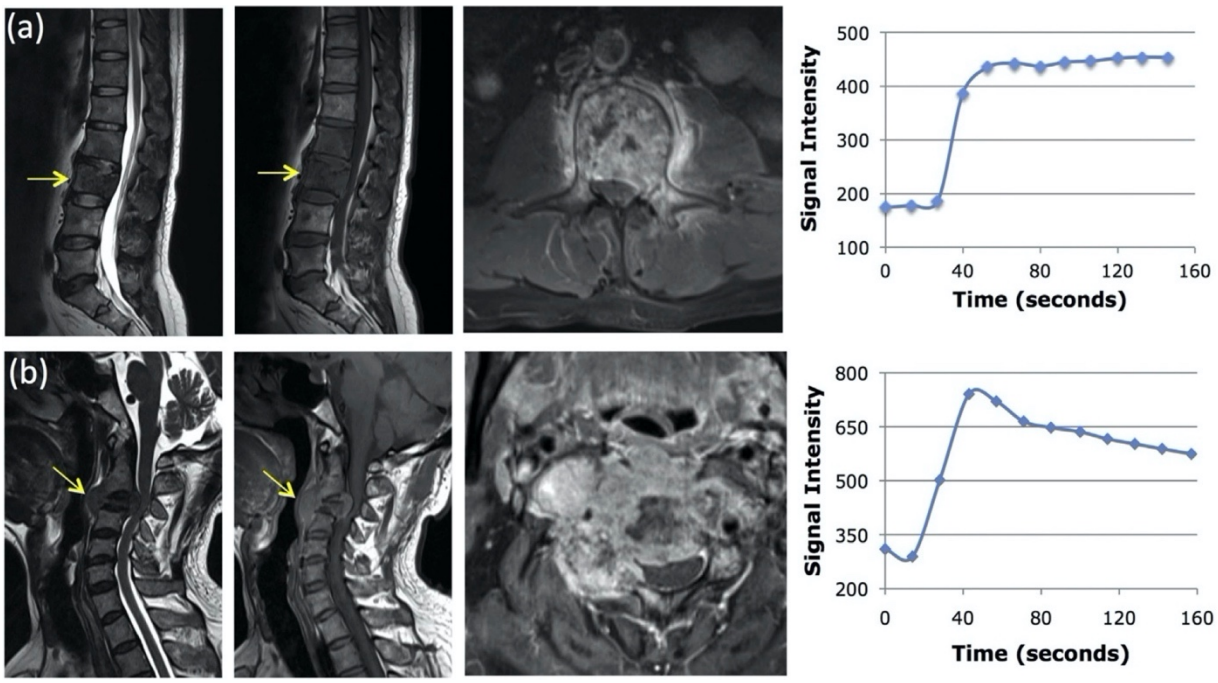

Figure 5-18: Two case examples**.** From left to right: the sagittal T2-w, T1-w, axial contrast-enhanced images, and the signal intensity time course measured from a tumor ROI. Top: (a) A 45-year-old man with metastatic lung cancer, showing the plateau DCE pattern. Bottom: (b) A 55-year-old man with metastatic thyroid cancer, showing the wash-out DCE pattern.

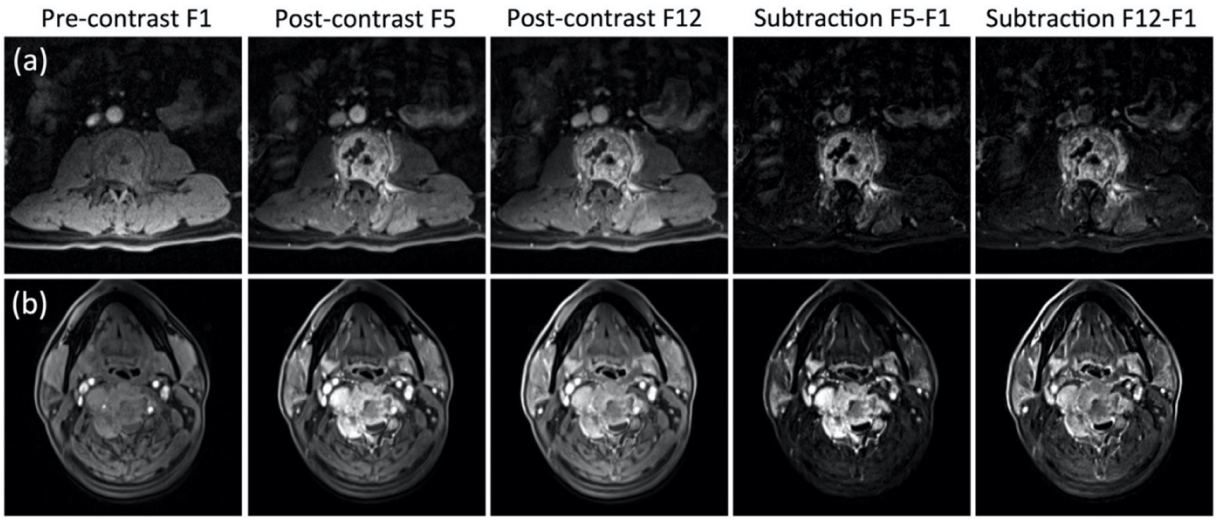

Figure 5-19: The DCE-MRI of two cases shown in **Figure 5-18.** From left to right: the pre-contrast image acquired in Frame-1, the post-contrast images acquired in Frame-5 (50–60 s after injection) and the last Frame-12 (140–150 s after injection), and the subtraction images (F5-F1) and (F12-F1). Top panel: (a) The metastatic lung cancer showing similar enhancements in F5 and F12 and similar subtraction images without a clear wash-out. Bottom panel: (b) The metastatic thyroid cancer showing a stronger enhancement in F5 than in F12, and the subtraction images also clearly demonstrate contrast wash-out.

In the lung metastases case shown in **Figure 5-19**, the S1 and S2 were DCE frames #3 and #4; in the thyroid metastases case, the difference between #3 and #4 was slightly greater than between #2 and #3, and thus used in the calculation. The Smax was DCE frame #12 in the lung metastases case, and #4 in the thyroid metastases case. Two heuristic parameters were calculated as:

Steepest Wash-in Signal Enhancement (SE) Ratio = [(S2-S1) / S0]

Maximum Signal Enhancement (SE) Ratio = [(Smax-S0) / S0]

For cases with a clearly visible peak enhancement occurring approximately 60 seconds after injection, the wash-out slope was calculated using the peak (S_{peak}) and the signal intensity at the last time point (S_{last}). For cases that did not show a peak before 85 seconds, in order to catch the increasing intensities in the DCE time course, the slope between the signal intensities at the 67 second ($S_{67seconds}$) and the last time point was calculated. This method was developed based on the analysis of various lesions in the spine using DCE-MRI, as described in two previous studies [256, 257]. Therefore, the wash-out slope was calculated as:

Wash-out Slope = [($S_{last} - S_{peak}$) / S_{peak}] x 100%

or, [($S_{last} - S_{67seconds}$) / $S_{67seconds}$] x 100%

These three measured parameters were used to differentiate between lung cancer and other cancers by utilizing the logistic regression and Chi-square Automatic Interaction Detector (CHAID) decision tree classification method. In addition to the heuristic analysis, a two-compartmental pharmacokinetic analysis was applied to obtain the in-flux transport constant K^{trans} and the out-flux rate constant k_{ep} ([1/min]), by using the methods reported

previously [257]. The pharmacokinetic parameters were highly correlated with heuristic parameters; thus they were not independent parameters, and not further analyzed.

5.4.4 Normalized Cut and Region Growing

Three-dimensional lesion segmentation was performed for all patients in this study. Since DCE-MRI was acquired in the axial plane based on the abnormal region identified on the sagittal acquisition, an automatic segmentation method was developed following this same approach, the detailed methods are illustrated in **Figure 5-20**. The abnormal area on sagittal T2W images was first manually outlined by a radiologist and then transformed to the axial view DCE-MRI for tumor segmentation using a normalized cut algorithm with region growing [264]. The global coordinates of all voxels outlined on sagittal slices (**Figure 5-20a**) were transformed to axial DCE (**Figure 5-20b**) and used as the initial search area (**Figure 5-20c**). In order to cover the entire lesion, the left and right boundary of the initial search box was expanded by a factor of 5 (**Figure 5-20d**). The normalized cut algorithm was utilized to divide the expanded search area on each slice into partitions, and those overlapping with the initial transformed area were kept (**Figure 5-20e**). Then, the remaining partitions on all DCE slices were combined into a 3D mask (**Figure 5-20f**), and the most strongly enhancing voxel was identified as the seed point for region growing to find the lesion boundary inside this 3D mask (**Figure 5-20g**).

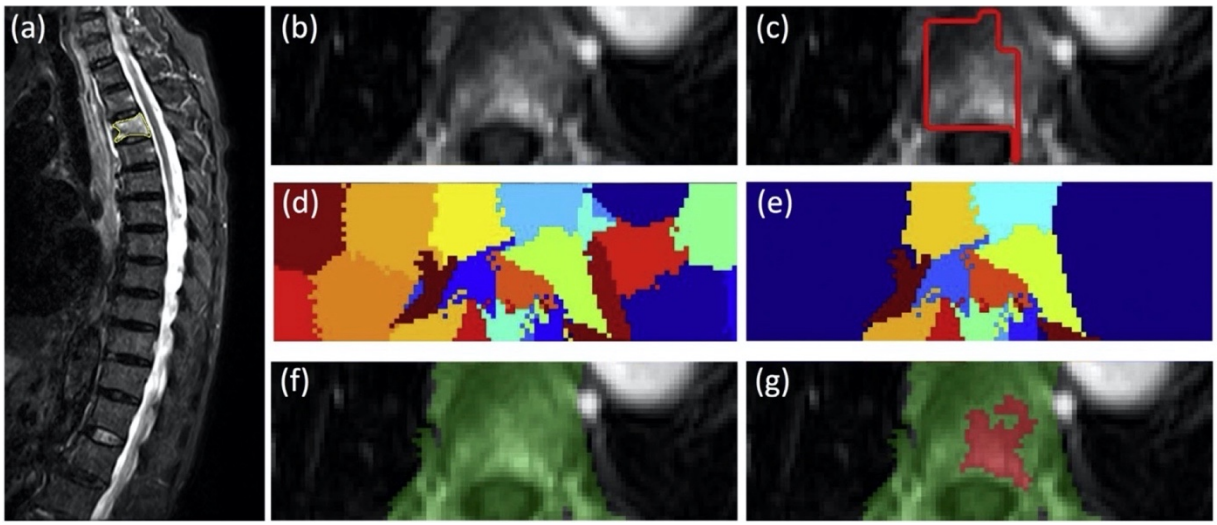

Figure 5-20: Identification and segmentation of the enhanced tumor on Axial DCE images based on the diseased segment drawn on Sagittal T2W images using the normalized cut algorithm. (a) Sagittal T2W image with marked diseased segment. (b) Axial view contrast-enhanced image. (c) The diseased segment in yellow in (a) is transferred to the axial image, shown as red box. (d) The clusters generated using the normalized cut. (d) The clusters containing any part of the red box ROI. (f) The vertebral region combining all remaining clusters. (g) The enhanced tumor lesion (in red) is generated by region growing within the green field.

5.4.5 Radiomics Analysis

Radiomics analysis was performed to extract DCE kinetic parameters and texture features within the segmented lesion based on 3D images. The analysis was done on three computed DCE parametric maps corresponding to the ROI-based analysis, including the steepest wash-in SE map, maximum SE map, and wash-out slope map. These maps were generated on a pixel-by-pixel basis, using the formula described above for hot-spot analysis. In each case, the DCE frames for S1, S2, Smax and S_{peak} were the same as those used in the hot-spot analysis. No apparent patient motion was noted in the short DCE acquisition period of 120-168 seconds, and as such between-frame registration was not needed. The color-coded maps from the lung cancer and thyroid cancer cases are shown in

Figure 5-21. On each map, 20 gray-level co-occurrence matrix (GLCM) texture features were calculated according to Haralick et al. [25], including autocorrelation, cluster prominence, cluster shade, contrast, correlation, dissimilarity, energy, entropy, homogeneity 1, homogeneity 2, maximum probability, sum average, information measure of entropy, sum variance, sum entropy, difference variance, correlation, difference entropy, information measure of correlation 1, and information measure of correlation 2.

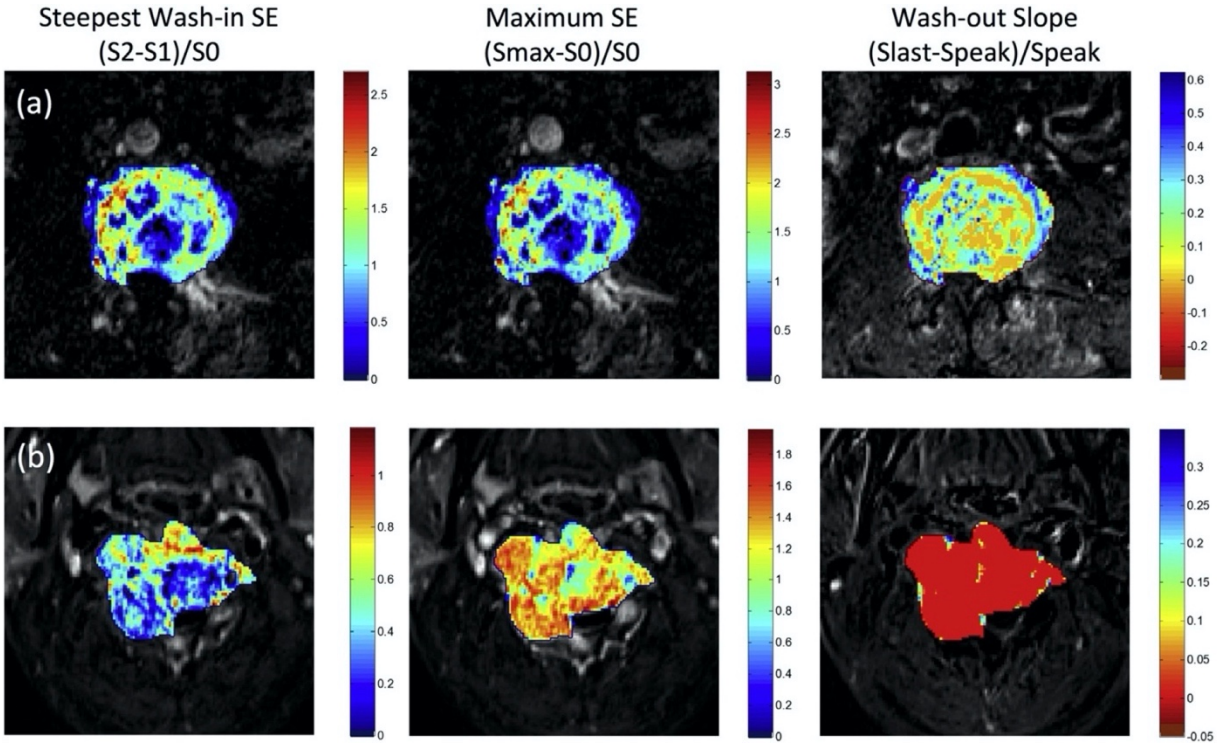

Figure 5-21: The generated DCE maps from the two case examples shown in Figure 5-18 for (a) metastatic lung cancer and (b) metastatic thyroid cancer. The map is generated using the equation for all voxels, but only the voxels within the tumor ROI are shown by color according to the color maps. The entire thyroid cancer shows a clear wash-out pattern in red color; in contrast, the lung cancer mainly shows the plateau pattern in orange to green color.

Furthermore, the histogram or the population distribution curve of all voxels within the tumor ROI on each map was generated, and a total of 13 parameters were obtained,

including the 10%, 20%… 80% to 90% percentile values, mean, standard deviation, kurtosis and skewness. On each map, 20 texture and 13 histogram features were calculated, and a total of 99 quantitative features from three maps were obtained for each patient.

The best radiomics model was generated in three steps: 1) ranking features; 2) selecting combination of features; 3) building a final model based on selected features. The features were first properly normalized. Due to the limited number of cases (total N=61), a random forest algorithm was used to select 3-5 features to form the diagnostic classifier [41]. We first selected parameters with highest significance scores from DCE histograms only, texture features only, and then selected parameters from combined histogram and texture features. The random forest with 500 trees was applied for classification of bootstrap samples randomly selected from 61 patients, and based on these results the discriminating capability of features could be assessed and ranked. Approximately 60% of cases were selected randomly in each run, and the significance of a feature could be assessed as the loss of accuracy after this feature was removed. Then, according to the ranking, the top 3, 4, 5, …. 10 features were selected to build the diagnostic model by using logistic regression. The discrimination accuracy was evaluated by the receiver operating characteristic (ROC) analysis using 10-fold stratified cross-validation. In each fold, 3 cases from lung metastasis (lung metastases) group and 3 cases from the non-lung metastasis (non-lung metastases) group were used as the testing set, and the remaining cases used for training. This process was repeated many times using different combination of selected features (3, 4, 5… 10) and the results were used to find the best model according to the highest AUC. After the features included in the best model were decided, they were used to

build a final diagnostic classifier with logistic regression, and the accuracy was evaluated in the entire dataset of 61 cases.

5.4.6 Deep Learning

Deep learning was used to investigate the diagnostic accuracy that can be achieved using a fully automated approach without manual hand-crafted features. Due to the small case number, the dataset was first augmented by 20 times using a random affine transformation with combination of rotation, translation, scaling and shearing [65]. Detailed augmentation methods have been reported before in Chang et al. [262]. Two separate convolutional neural network (CNN) architectures were applied. First, the three DCE parametric maps were used as independent inputs in a conventional feed-forward CNN. Second, to incorporate time-dependent information from the entire 12 sets of DCE images, we applied used a convolutional long short term memory (CLSTM) network [62, 63], by inputting the 12 DCE datasets into the network in a time order as shown in **Figure 5-22**. Each 2D imaging slice was used as an independent input. For each case, the smallest bounding box containing the segmented tumor was used as input. The segmented tumor ROI's from all slices were projected together, and the smallest bounding box to cover the outer boundary of projected ROI's was used for this case. Since only the tumor ROI was considered, the pixels outside the tumor in the box were set to zero. The 12 sets of DCE images were normalized together to a mean = 0 and standard deviation = 1.

Detailed procedures were described in Chang et al. [262]. For the conventional CNN architecture, the underlying network was composed by a strided convolution in every other layer (i.e. 2^{nd}, 4^{th}, and 6^{th}) to reduce the spatial resolution to 25% of the previous

resolution. Each convolutional operation was followed by a nonlinear rectified linear (ReLU) activation function [48, 60]. This function was chosen because of its well-documented advantages including stable gradients at the extreme values of optimization. Dropout at 50% was applied to all convolutional and fully-connected layers to limit over-fitting and add stochasticity to the training process [61, 265]. The 7th layer output feature maps from all cells were flattened to a one-dimensional vector. The softmax activation function was used for final classification, with a threshold of 0.5. For the CLSTM network, 7 stacked convolutional long short term memory layers were fed into a final fully connected layer before output, as in the architecture shown in **Figure 5-22.**

The algorithm was implemented with a standard cross entropy loss function and the Adam optimizer with an initial learning rate of 0.001, which was kept as a constant throughout the training [105]. The software code was written in Python 3.5 using the open-source TensorFlow 1.0 library (Apache 2.0 license). Experiments were performed on a GPU-optimized workstation with a single NVIDIA GeForce GTX Titan X (12GB, Maxwell architecture). A forward pass for the classification test of a new patient could be achieved in <0.01 second. The results were evaluated using 10-fold cross validation. The range and the mean value with standard deviation were calculated to show the prediction accuracy.

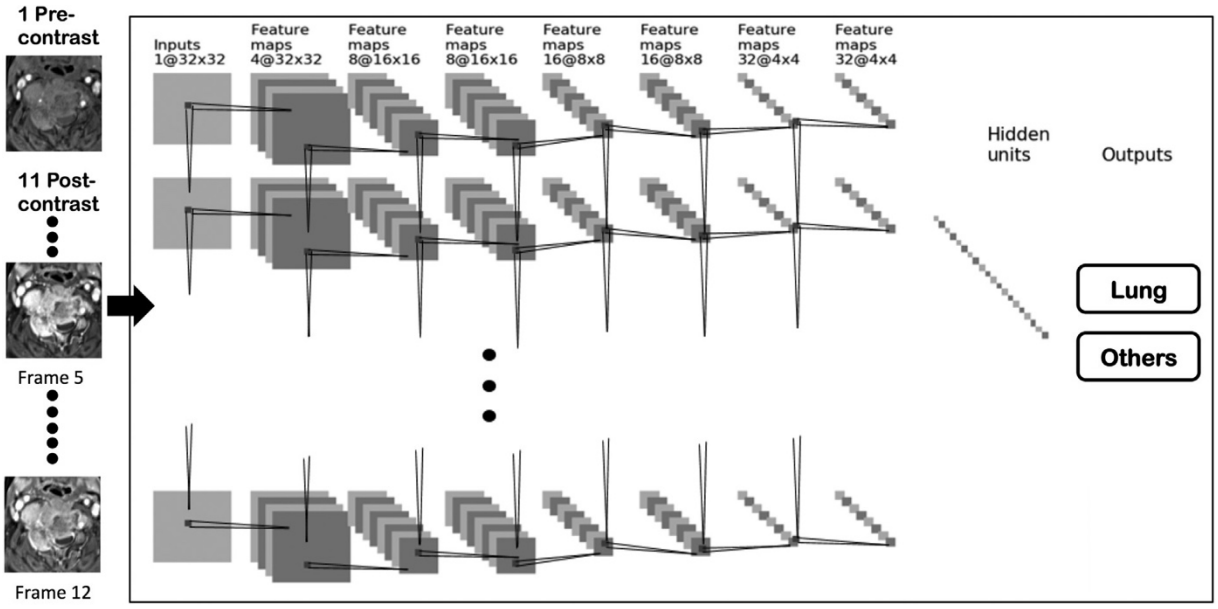

Figure 5-22: Diagram of the recurrent CNN**.** The architecture uses 7 serial convolutional LSTM layers via 3 × 3 filters followed by the ReLU nonlinear activation function. All 12 sets of DCE images are used as inputs, and the signal intensity is normalized using the same scale. Dropout at 50% is applied to all convolutional and fully- connected layers after the second layer. Feature maps are down sampled to 25% of the previous layer by convolutions with a stride length of two. The last dense layer is obtained by flattening the convolutional output feature maps from all states. Softmax is used as the activation function of the last fully connected layer.

5.4.7 Results

Hot-Spot ROI-Based DCE Parameters

Table 5.9 summarizes the mean ± standard deviation of 5 characteristic DCE parameters measured from the manually placed ROI on the hot spot. The wash-out slope and k_{ep} showed a significant difference between the lung cancer and other primary tumors. The mean wash-out slope was 0.25% in lung cancer, indicating most lung cancers showed the plateau DCE kinetic pattern. The mean wash-out slope was -9.8% for other tumors, indicating that many of them showed the wash-out DCE kinetic pattern. In the two examples shown in **Figure 5-18**, the DCE time course shows a plateau pattern for the lung

cancer, and a clear wash-out pattern for the thyroid cancer. In the DCE images shown in **Figure 5-19**, the signal intensity is similar between Frame-5 and Frame-12 for the lung cancer. For the thyroid cancer, the intensity in Frame-12 is clearly lower compared to Frame-5, and the degree of enhancement is lower in the subtraction image of (Frame 12 – Frame 1) compared to that of (Frame 5 – Frame 1). Among all non-lung metastases, the breast (-12.9%) and thyroid (-15.6%) cancers had the most prominent wash-out.

Classification was done by using the logistic regression and CHAID decision tree based on the three heuristic parameters (Steepest wash-in SE, Max SE, Wash-out slope). The accuracy obtained using logistic regression was 0.74, and that by using CHAID with the wash-out slope of -6.6% followed by maximum SE of 98% was 0.79, as shown in **Figure 5-23**. True Positive (TP) for diagnosis of lung cancer = 18/30 cases; False Negative (FN) = 12/30 cases; True Negative (TN) for diagnosis of other tumors = 30/31 cases; and False Positive (FP) = 1/31 case. The Sensitivity = 60%; Specificity = 96.8%; Positive Predicting Value = 94.7%; and Negative Predicting Value = 71.4%.

Table 5.9: The DCE parameters analyzed from the ROI manually placed on the strongly enhanced tissue, data shown is [mean ± standard deviation].

	Tumor origin	**Maximum SE (%)**	**Wash-in SE (%)**	**Wash-out Slope (%)**	**K^{trans} (1/min)**	**k_{ep} (1/min)**
Lung cancer	Lung (N = 30)	243 ± 89	146±60	0.25 ± 10	0.10 ± 0.04	0.39 ± 0.16
Others	Total (N = 31)	220 ± 109	142±76	−9.8 ± 12.9	0.10 ± 0.06	0.58 ± 0.24
	Breast (N = 9)	299 ± 134	197±95	−12.9 ± 8.24	0.14 ± 0.07	0.60 ± 0.16
	Thyroid (N = 7)	210 ± 59	130±47	−15.6 ± 14.4	0.10 ± 0.03	0.68 ± 0.28
	Prostate (N = 6)	165 ± 77	126±74	−7.7 ± 12.9	0.08 ± 0.05	0.56 ± 0.29
	Liver (N = 6)	196 ± 111	118±55	−5.9 ± 16.2	0.08 ± 0.05	0.52 ± 0.27
	Kidney (N = 3)	164 ± 80	87±37	1.3 ± 11.3	0.07 ± 0.04	0.42 ± 0.26
P value		0.199	0.614	0.001	0.634	0.001

* Significant with $P < 0.05$ in the comparison between lung cancer and other tumors.

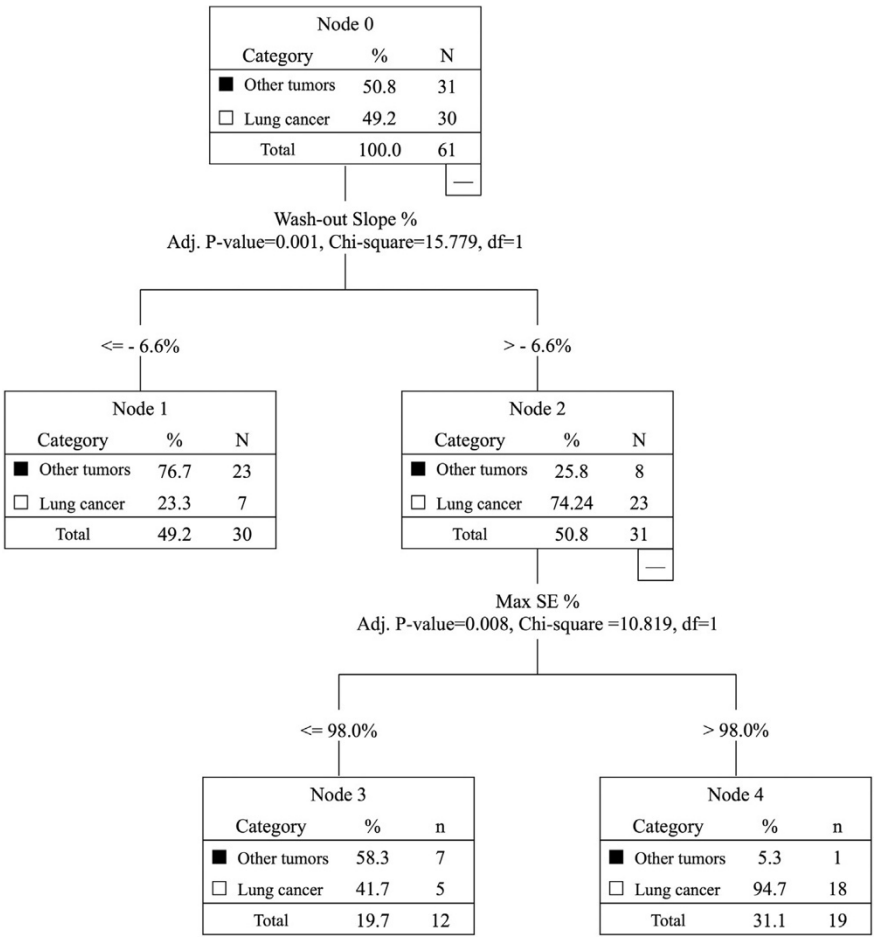

Figure 5-23: The diagnostic results analyzed using the Chi-square Automatic Interaction Detector (CHAID) decision tree classification method. The overall diagnostic accuracy of 0.79 is achieved by using the first threshold of wash-out slope - 6.6%, followed by the second threshold of wash-in SE 98%. True Positive (TP) for diagnosis of lung cancer = 18 cases; False Negative (FN) = 12 cases; True Negative (TN) for diagnosis of other tumors = 30 cases; and False Positive (FP) = 1 case. The Sensitivity = 60%; Specificity = 96.8%; Positive Predicting Value = 94.7%; and Negative Predicting Value = 71.4%.

Radiomics Using DCE Histogram Parameters and Texture

In the color maps shown in **Figure 5-21**, almost all voxels in the entire thyroid cancer show the wash-out pattern (in red color), but the voxels in the lung cancer are more heterogeneous with most of them showing plateau (in orange to green color). Based on the 10-fold cross-validation results, we found that by increasing the number of features from 3 to 4 to 5 the accuracy improved slightly, but then the accuracy did not increase further with

more features; therefore, we only reported the results using 3 and 5 features. The diagnostic accuracy and the selected histogram and texture features are listed in **Table 5.10.** It was noted that the accuracy obtained using the texture features only (0.59-0.62) was lower compared to that analyzed using histogram only (0.67-0.68), or histogram+texture (0.68-0.71). The accuracy of radiomics analysis was worse compared to the hot-spot ROI-based analysis of 0.79.

Table 5.10: Accuracy in differentiating lung metastases from other cancers based on selected features in the radiomics analysis.

Accuracy	Histogram + texture	Histogram features only	Texture features only
3 features	0.68 (90% value and kurtosis from wash-out map, information measure of entropy from max SE map)	0.67 (90% value and kurtosis from wash-out map, mean from max SE map)	0.59 (information measure of entropy from max SE map, entropy and dissimilarity from steepest wash-in map)
5 features	0.71 (90% value, kurtosis and autocorrelation from wash-out map, information measure of entropy from max SE map, entropy from steepest wash-in map)	0.68 (90% value and kurtosis from wash-out map, mean and kurtosis from max SE map, 50% value from steepest wash-in map)	0.62 (information measure of entropy from max SE map, entropy and dissimilarity from steepest wash-in map, dissimilarity and contrast from wash-out map)

Deep Learning Using Convolutional Neural Network

Classification was performed using two different deep learning approaches, evaluated by 10-fold cross-validation. The accuracy achieved using three generated DCE parametric maps as inputs in a conventional CNN was 0.61-0.74, mean 0.71 with standard deviation of 0.043. The accuracy achieved using all 12 sets of DCE images as inputs in a CLSTM network was 0.75-0.84, mean 0.81 with standard deviation of 0.034. The accuracy achieved by CLSTM was significantly higher than that achieved using CNN. The sensitivity for detecting lung metastases was 0.60±0.07 for CNN, 0.75±0.07 for CLSTM; and the specificity was 0.76±0.06 for CNN, 0.83±0.06 for CLSTM. To provide a clear comparison of these three

different analysis approaches, the essential methods, the number of analyzed parameters, the diagnostic evaluation methods, and the obtained results are summarized in **Table 5.11**.

Table 5.11: The comparison of hot-spot, radiomics, and deep learning classification methods and the obtained results.

Method	Number of parameters	Evaluation method	Final cases used in accuracy test	Accuracy
Hot-spot using manually drawn ROI	3	HIAD decision tree	Entire dataset (30 lung mets and 31 others)	0.79
Radiomics using segmented 3D tumor	99	Random forest + logistic regression	Entire dataset (30 lung mets and 31 others)	0.71 (best model from **Table 5.10**)
Deep learning (CNN) using 3 parametric maps			10-Fold cross-validation (6–7 test cases in each run)	0.71 ± 0.043 (range 0.61–0.74)
Deep learning (CLSTM) using 12 DCE images			10-Fold cross-validation (6–7 test cases in each run)	0.81 ± 0.034 (range 0.75–0.84)

5.4.8 Summary and Discussion

In the present study, three different analysis methods based on hot-spot, radiomics and deep learning were applied to differentiate the spinal metastases coming from lung cancer and other tumors. The pros and cons of each method were described, and the achieved accuracy was compared. The results showed that the DCE kinetic measure of wash-out slope from a hot-spot was the best parameter to differentiate primary lung cancer from other tumors. A CHAID decision tree using the wash-out slope followed by maximum SE could achieve an accuracy of 0.79. In comparison, the radiomics analysis performed from the segmented whole tumor in 3D could only achieve the highest accuracy of 0.71, while the CLSTM network using the entire sets of DCE images reached an accuracy of 0.81.

The cause of death in most cancer patients occurs due to metastasis and complications. Thus, early detection of metastasis is critical, as it can be better controlled. The most common metastatic cancer site is the liver, followed by the lung, and then bones, where the

spine is the most vulnerable site to be invaded by metastases in the skeletal system. Patients without a known history of cancer often seek medical attention due to nerve compression and back pain. When metastatic cancer was suspected, finding and confirming the primary lesion became the most important task for treatment planning. While in the Western world PET/CT is the standard of care for such patients, it is expensive; in developing countries the availability of PET/CT system and the [18F]-FDG isotope tracer maybe limited. Since lung metastases are the most common primary, if it is suspected, a CT scan can be performed quickly at a low cost. Even in the Western world, this study also has a good clinical value, to help patients with lung metastases to be diagnosed early without relying on the PET/CT, which needs insurance approval and causes delay. Therefore, in this study we tried to predict origin of spinal metastases that come from lung cancer and other cancers.

The appearance of many spinal lesions was similar on conventional MRI [266-270]. Osteolytic lesion was the most common abnormality seen in the spine, and metastatic lesion was often accompanied with soft tissue mass. The imaging presentation may vary substantially due to many factors, including local myelofibrosis, infarction, edema, pathological compression fracture and infection, adding to the many challenges of the differential diagnosis. DCE-MRI has been proven as a valuable technique for assessing tumor angiogenesis, and it has been widely applied for diagnosis and pre-operative staging for many cancers. For the spine, DCE-MRI has been applied to differentiate various diseases, including primary bone tumors such as myeloma, lymphoma, chordoma [256, 257, 271, 272]; benign lesions such as tuberculosis, giant cell tumor of the bone [255, 256]; as well as metastatic cancers of different origins, e.g. hypervascular renal vs. hypovascular

prostate [254], and cancers of different origins [177, 252]. For surgical planning, the information of blood supply may predict intraoperative blood loss, which can be used to plan for embolization before surgery [273-275].

Many different analysis methods can be applied to extract DCE parameters, either from a hot spot or from the whole tumor. In this study we first measured the signal intensity time course from an ROI manually placed on a strongly enhanced area. The wash-out slope was the best predictor to differentiate the two groups. The CHAID classification accuracy using the wash-out slope of -6.6% followed by maximum SE of 98% was 0.79.

Morphological presentation of spinal lesions could be evaluated using a scoring system based on pre-defined features, as used in a previous study to differentiate chordoma from giant cell tumor of the bone [256]. In recent several years, radiomics analysis has been widely applied to extract thorough information from medical images, for performing many clinical tasks such as differential diagnosis of benign and malignant lesions or subtype cancers [20-22, 258], and predication of therapeutic response or prognosis [214, 276, 277]. The tumor was first segmented, and then the histogram-based parameters and high-level texture features extracted using quantitative algorithms were measured. All these parameters were then combined for feature selection to build an optimal diagnostic/predictive classifier. Radiomics was commonly applied to analyze images acquired in different sequences (e.g. T2, T1 pre- and post- contrast, diffusion weighted imaging, FLAIR, etc.), not multiple sets of post-contrast images acquired using DCE-MRI. Therefore, in this study, we analyzed features on three DCE parametric maps generated corresponding to the hot spot analysis: the wash-in SE map, maximum SE map, and wash-out slope map. To avoid using a “black-box” classification method, we used random forest

algorithm for feature selection, not for the final classification. A similar approach was used in Gallego-Ortiz et al. [184]. The accuracy in the combined histogram+texture analysis was 0.68 by using 3 features and 0.71 by using 5 features. The results showed that the accuracy was inferior to that of hot-spot analysis, and that the texture (i.e. heterogeneity within the tumor) did not add much value for improving differential diagnosis. Although metastatic cancers are highly prevalent in liver, lung, brain and spine, there are few studies trying to predict the origins based on imaging analysis. In a very recent study by Ortiz-Ramón et al, they tried to differentiate brain metastasis coming from 27 lung cancer, 23 melanoma and 17 breast cancer patients [226].

Since DCE-MRI is not a standard procedure for evaluation of spinal lesions, the case number reported in all published studies is small. Furthermore, the heterogeneity from complicated anatomic structures and the vascular bone marrow might limit the prediction accuracy of radiomics analysis. In this study, we also implemented a deep learning network to evaluate the accuracy that could be achieved. Most deep learning methods use images acquired in different sequences as inputs, and currently, there is no established network specifically designed for the full set of pre- and post-contrast images acquired in DCE-MRI [278]. In our protocol we had a total of 12 sets of images. In order to consider the change of signal intensity over time, all DCE images were normalized together. The Long Short Term Memory (LSTM) network, one of the Recurrent Neural Network (RNN), can connect previous information to the present task. The LSTM is explicitly designed to avoid long-term dependency and focus primarily on short-term memory. We used images from different DCE time points as independent inputs to LSTM, but also considered the changes of signal intensity in these different sets of images. Meanwhile, the same hierarchical

features were calculated from each time frame, and thus more information from all DCE frames could be used for prediction of diagnosis. The range of accuracy in 10-fold cross validation was 0.75-0.84, with the mean of 0.81±0.034, slightly better than the 0.79 achieved in hot spot ROI analysis. To compare the results using LSTM with the conventional CNN, we also used the three generated parametric maps as inputs for the CNN, and found that the mean accuracy was only 0.71±0.043, similar to the accuracy of radiomics analysis. The results suggest that LSTM is an appropriate network to consider the entire sets of DCE images and track the change of signal intensity in a time sequence.

The major limitation of this study was the small case number. However, although not optimal for radiomics or deep learning, the results obtained using three different methods could still give insights to their value in solving this very challenging problem. The spine lesion segmentation method and the three different DCE analysis methods presented in this study may be applied to other studies to further investigate their clinical value in predicting diagnosis or further in prognosis.

In conclusion, we have shown that a simple hot-spot ROI analysis could be applied to characterize DCE kinetics of the metastatic cancer in the spine and differentiate the primary from lung cancer and other tumors. We have implemented deep learning and shown the potential in this clinical application. The recurrent neural network using CLSTM could track the change of signal intensity in pre- and post-contrast images in the DCE-MRI, with accuracy comparable to the hot-spot analysis, and better compared to conventional CNN and radiomics. For patients suspected to have metastatic cancer in the spine, DCE-MRI may help to predict the primary cancer from lung, and that may help to reach an early confirmed diagnosis using CT alone without having to wait for the expensive PET/CT.

5.5 Artificial Intelligence Analysis on Prostate DCE-MRI to Distinguish Prostate Cancer and Benign Prostatic Hyperplasia

5.5.1 Motivation

Prostate cancer (PCa) is one of most common malignant tumors in man [279]. The accurate detection of PCa is a challenging task in clinic [280]. The distinction of PCa from benign conditions, including benign prostatic hyperplasia (BPH) and prostatitis, is critical to personalized medicine [281]. Currently, MR images of the prostate are evaluated by radiologists. However, the detection and diagnosis of PCa using MR images varies considerably [282]. Quantitative imaging features may provide additional information for differentiation of the benign and malignant lesions. Furthermore, deep learning using convolutional neural network provides a fully automatic and efficient approach to analyze detailed information in the tumor and the surrounding per-tumor tissue for diagnosis. Several studies have proven that AI has enough potential to diagnose prostate cancer [71, 261, 263]. The goal of this study is to evaluate the accuracy of prediction using the SVM model based on the histogram and texture features extracted from the lesion, as well as deep learning using three different networks. The results to differentiate between prostate cancer and benign prostatic hyperplasia are compared.

5.5.2 Dataset and Method

From September 2014 to September 2018, 67 patients underwent prostate multi-parametric MRI (mpMRI) and were confirmed with PCa by transrectal ultrasonography guided prostate biopsy and followed radical prostatectomy. 37 BPH patients underwent

mpMRI showing PI-RADS v2≤2, and they received biopsy in an interval less than 6 months and were confirmed to have negative findings. MR examinations were carried out on a 3.0 T scanner (Achieve; Philips, The Netherlands) equipped with a sixteen-channel sensitivity-encoding (SENSE) torso coil without an endorectal coil. Four hours of fasting prior to MR examination was required to suppress bowel peristalsis. During the acquisition, a contrast agent (Omniscan, GE, concentration: 0.5 mmol/ml) with a dose of 0.2 ml/kg of body weight at a flow rate of 2 ml/s was injected via a power injector (Spectris Solaris EP, Samedco Pvt Ltd) at the start of the sixth DCE time point followed by a 20 ml saline flush. **Figure 5-24** and **Figure 5-25**show two case examples. Only the DCE images were analyzed in this study. A total of 40 frames were acquired, including 5 pre-contrast (F1-F5) and 35 post-contrast (F6-F40). Two radiologists outlined the whole prostate gland and the index suspicious lesion in consensus on DCE-MRI using imageJ (NIH, USA).

The outlined lesion ROI on all slices were combined to generate a 3D tumor mask, and 13 histogram features and 20 GLCM texture features were extracted on each DCE images [25], with a total of 33x40=1320 features. For differentiation between BPH and PCa using a radiomics method, feature selection was first implemented by using an SVM based sequential feature selection methods to find features with the highest significance[283]. These features were then used to train a final SVM model with Gaussian kernel to serve as the diagnostic classifier. For deep learning, first, a VGG network with 8 convolutional layers were implemented to differentiate between BPH and PCa patients. The 5 pre-contrast frames were averaged as the reference for normalizing post-contrast frames. The last 20 frames were down-sampled to 10 frames, by only selecting every other frame. So, a total of 25 normalized enhancement maps were used.

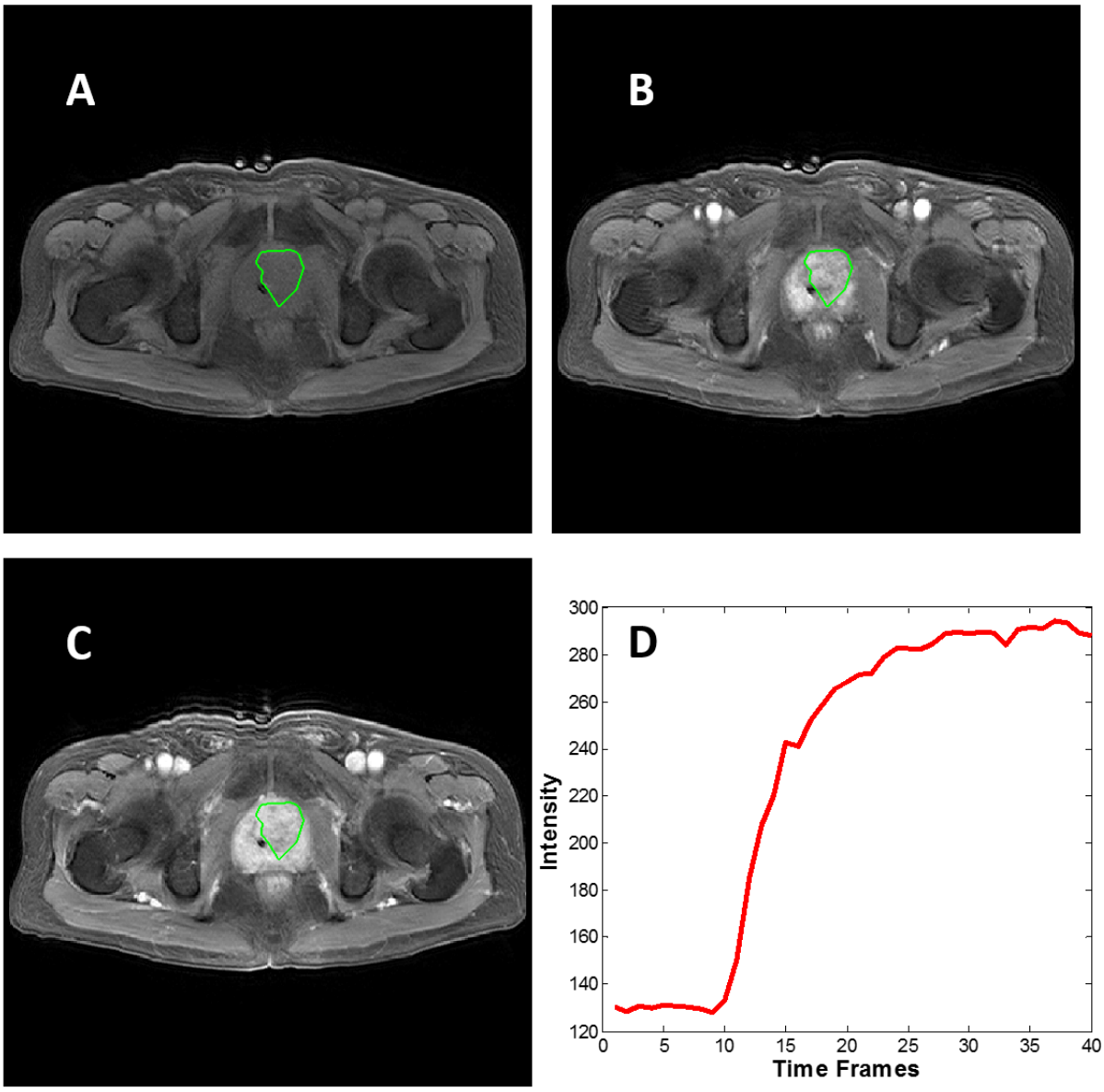

Figure 5-24: A case example from an 80-year-old man with benign prostate hyperplasia (tPSA=10.8 ng/ml). The lesion is manually outlined. (a) The first time frame (pre-contrast image); (b) The 15th time frame (post-contrast image); (c) The 40th time frame (post-contrast image); (d) DCE time intensity curve shows the persistent enhancement pattern.

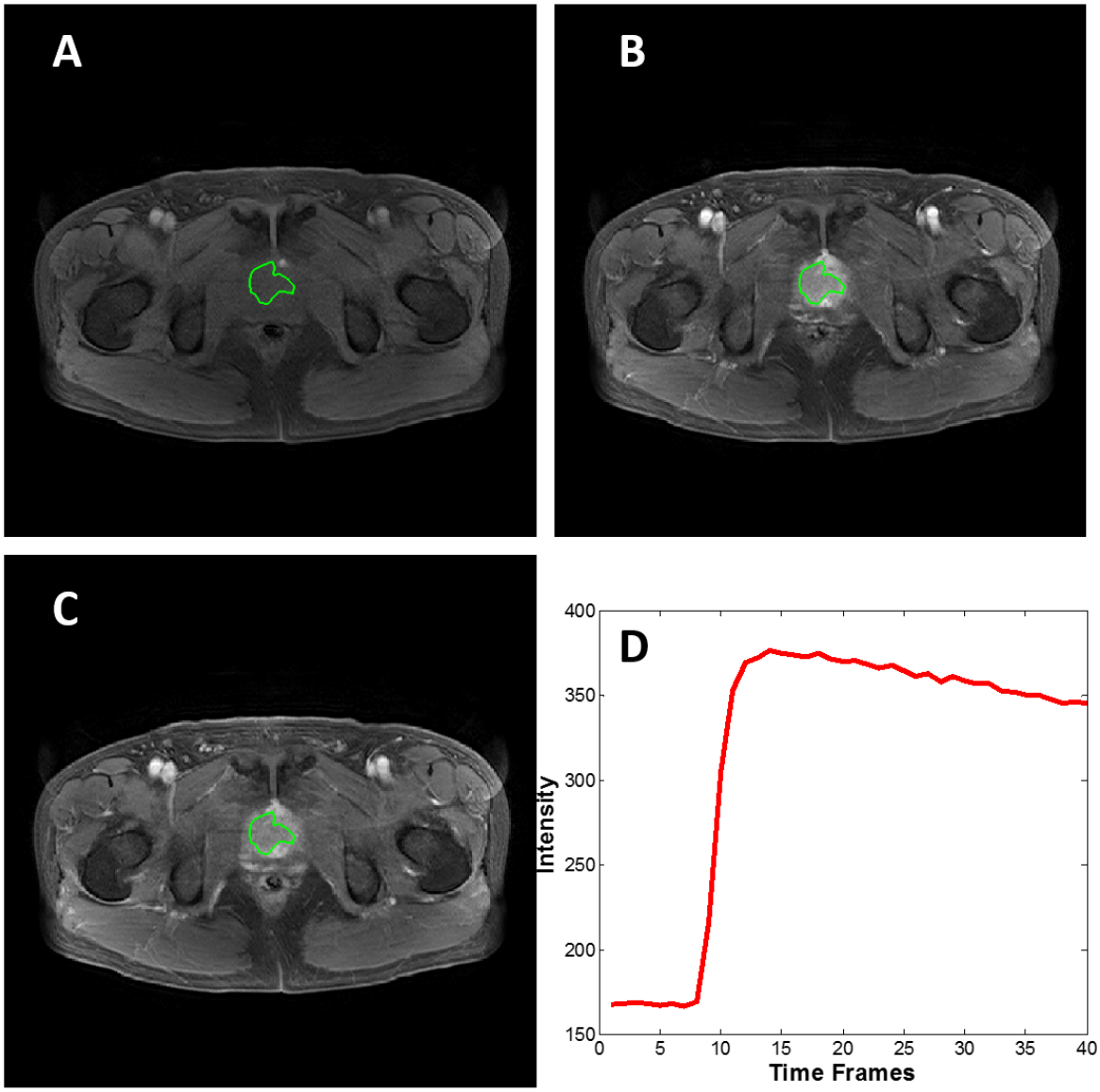

Figure 5-25: A case example from a 65-year-old man with prostate cancer (tPSA=7.13 ng/ml, Gleason Score=4+5). The lesion is manually outlined. (a) The first time frame (pre-contrast image); (b) The 15th time frame (post-contrast image); (c) The 40th time frame (post-contrast image); (d) DCE time intensity curve shows the wash-out kinetic pattern.

Figure 5-26 shows a VGG network architecture which used all 25 sets of images as input without timing information. Then, to consider the change of the signal intensity with time, a convolutional long short term memory (CLSTM) network was applied, shown in **Figure 5-27**[63]. The 25 sets of enhancement maps were added one by one into the

network. However, due to the forgot gate implemented in LSTM, information from early time points contributes less than later time points. To minimize this problem, a bi-directional CLSTM model was applied, shown in **Figure 5-28**. To investigate the contribution from the peri-tumor tissue, region growing was utilized to include connected pixels with the outlined tumor ROI, where the enhancement was > 10% of the mean tumor ROI enhancement on the 10th DCE frame. The results obtained using the expanded ROI and the tumor ROI were compared. To avoid overfitting, the dataset was augmented by random affine transformation [8]. The algorithm was implemented with a cross entropy loss function and Adam optimizer with initial learning rate of 0.001.

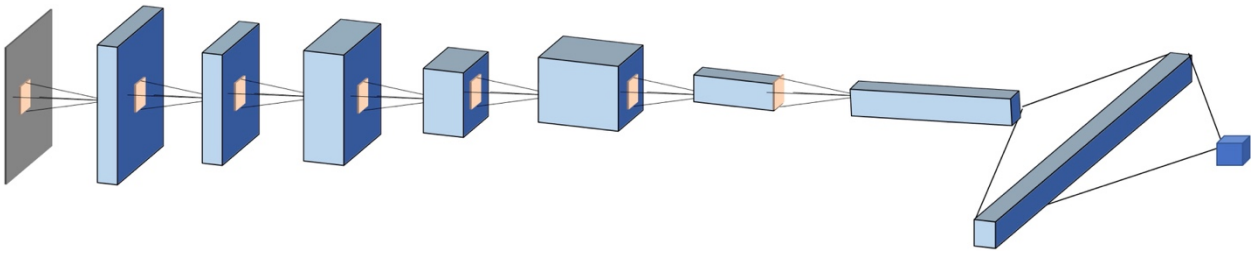

Figure 5-26: Diagram of the VGG convolutional neural network (CNN). The architecture uses 7 serial convolutional 3 x 3 filters followed by the ReLU nonlinear activation function. Dropout at 50% is applied to all convolutional and fully-connected layers after the second layer to avoid over-fitting. Feature maps are down-sampled to 25% of the previous layer by convolutions with a stride length of two. The number of the input channels is 25. The number of activation channels in deeper layers is progressively increased from 8 to 16 to 32 to 64. The activation function of the last layer is Softmax.

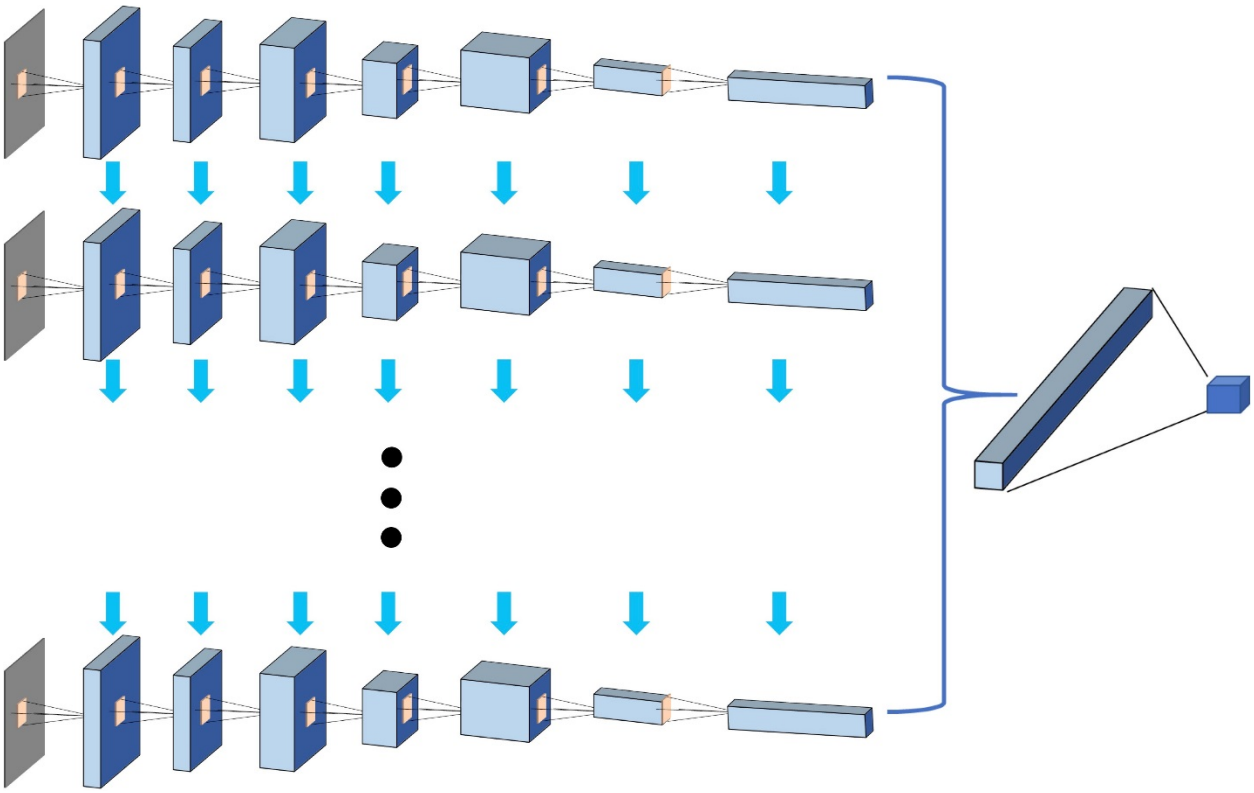

Figure 5-27: Diagram of the Convolutional Long Short Term Memory (CLSTM) network. The architecture uses 7 convolutional LSTM layers via 3x3 filters followed by the ReLU nonlinear activation function. The number of DCE images is reduced from 40 to 25, and used as input by adding them one by one. The number of the input channels is 1 at each time point. The number of activation channels in deeper layers is progressively increased from 4 to 8 to 16 to 32. The last dense layer is obtained by flattening the convolutional output feature maps from all states.

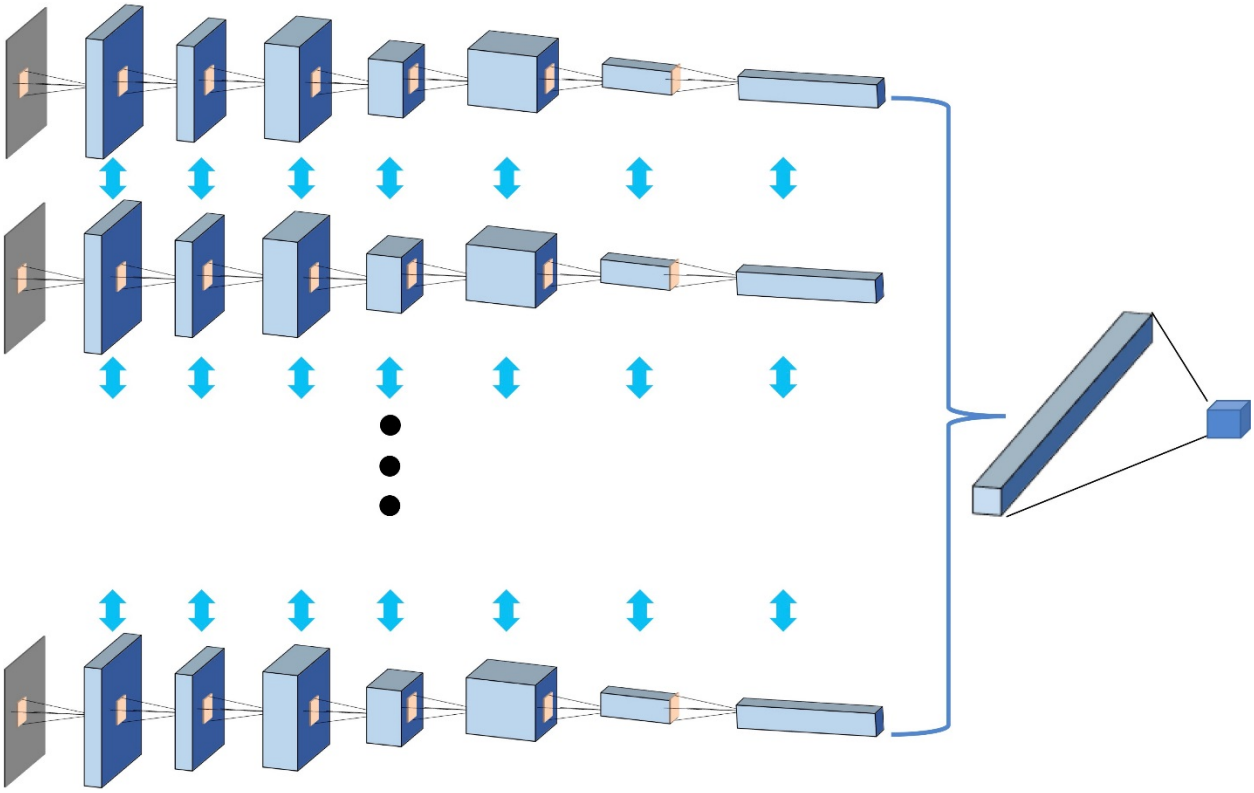

Figure 5-28: Diagram of the bi-directional Convolutional Long Short Term Memory (CLSTM) network. For a series of 25 time points, the train is too long and may lose the early information. Using bi-directional network may minimize this problem. The architecture uses 7 convolutional LSTM layers via 3x3 filters followed by the ReLU nonlinear activation function. The configuration is exactly the same as in Fig. 4, by adding the back direction analysis.

5.5.3 Results

The accuracy for differentiating between BPH and PCa was 0.74 when using the SVM model built based on the histogram and texture parameters. In deep learning using VGG with the manually outlined tumor ROI as inputs, the accuracy in 10-fold cross-validation was 0.60 – 0.81 (mean 0.72). When considering the temporal DCE information using CLSTM, the accuracy was improved to 0.77 – 0.85 (mean 0.82) using one-directional CLSTM architecture, and further to 0.75 – 0.92 (mean 0.87) using the bi-directional CLSTM architecture. When considering the peri-tumor tissues using expanded ROI as inputs, the

accuracy of the bi-directional CLSTM was decreased to 0.60-0.89 (mean 0.80), which was worse compared to the results obtained using the manually drawn tumor ROI as inputs.

5.5.4 Summary and Discussion

In this study we elucidated that prostate DCE-MR images can be analyzed using SVM and deep learning classifiers to differentiate between PCa and BPH patients. The recurrent network using CLSTM could take the change of signal intensity in the DCE series into consideration, and the accuracy was higher compared to the conventional VGG. The train of 40 DCE frames might be too long for CLSTM, so they were down-sampled to 25 by skipping every other frame in the last 20 frames. To further investigate whether the early information, which usually captured the important wash-in phase, was lost in one-directional CLSTM, the bi-directional CLSTM was implemented, and the mean accuracy was improved to 0.84. The results suggest that although the CLSTM is an efficient approach for considering images acquired in a time series, the train length needs to be considered, and novel approaches such as the bi-directional analysis can be considered. When the peritumoral information outside the lesion ROIs was considered, the prediction accuracy was worse, which could be due to the diluted information by including the weakly enhanced tissues into analysis. This study demonstrates that machine learning using radiomics and deep learning, with appropriate consideration of the time series, can be implemented to analyze the DCE-MRI to differentiate between PCa and BPH.

Chapter 6. Improving CBCT Quality for Adaptive Radiation Therapy using Generative Adversarial Network (GAN)

6.1 Motivation and Purpose

Cone-beam CT (CBCT) is widely used in radiotherapy clinics for patient setup and treatment monitoring, and is essential in the context of adaptive radiation therapy. Since it is acquired frequently during the course of radiotherapy, the images may be used for changing the treatment depending on the response, i.e. to facilitate adaptive treatment planning. However, its application for dose calculation or organ delineation is limited by the reduced image quality and inaccurate Hounsfield units (HU) mapping compared to conventional CT used in treatment planning. An alternative way is to deform planning CT to treatment CBCT to account for anatomical and dose changes. However, this strategy relies on deformation algorithm which may not be precise either. It is more beneficial to directly improve CBCT image quality to the CT level and use it for adaptive radiotherapy.

There have been numerous efforts in improving CBCT image quality using scatter correction: such as hardware improvement by adding anti-scatter grid [284] or a lattice-shaped lead beam stopper [285]; or software improvement with iterative filtering [286], raytracing[287], or Monte Carlo (MC) modeling [288, 289]. Especially, raytracing and MC methods have been shown to reproduce HUs to sufficient accuracy for both photon and proton dose calculation. They are, however, limited by the time it takes to perform correction, about 10 minutes for the raytracing-based algorithms, and several hours for the Monte Carlo-based methods. All these strategies are not feasible for on-line real time adaptive radiotherapy due to their lengthy time costs.

Recently, machine-learning based algorithm has been applied to improve image quality and image reconstruction. It has been shown that synthetic CT could be generated from MRI by using convolutional neural network (CNN) for radiotherapy planning without acquiring the actual CT [290, 291]. Similar strategy can also be applied to improve image quality of low-dose CT to match high-resolution CT [292]. The purpose of this study is to develop unsupervised deep-learning model to improve CBCT image quality for adaptive radiotherapy and to further validate the model on different anatomical sites.

6.2 CT and CBCT Datasets

Data from 30 pelvic patients were included. Each patient had one planning CT and five CBCT scans, a total of 150 pairs of CT-CBCT were used for model training and validation purposes. The CBCTs were from the first week of treatment to ensure the closest anatomy to planning CT. Paired CT-CBCT from an additional 15 pelvic image datasets from prostate cancer patients and 10 head-and-neck (HN) datasets from oral cancer patients were used for independent testing purpose. The CBCT scans of the validation set were collected at the first day of treatment on a different Varian TrueBeam.

All treatment planning CT images were collected with a GE LightSpeed16 CT scanner (GE Health Systems, Milwaukee, WI) and the CBCT images of the training set were acquired with an on-board-imager (OBI) equipped Varian TrueBeam STx linear accelerator (Varian Medical Systems, Palo Alto, CA). The original CTs had a resolution of 0.91 × 0.91 × 1.99 mm^3 and dimensions of 512 × 512 x 210. All CBCTs had a resolution of 1.27 × 1.27 × 1.25 mm^3 and dimensions of 512 × 512 x 80. For each patient, the CT images were mapped to each set of CBCT images using Velocity (Varian Medical Systems, Palo Alto, CA) with multi-pass B-

spline based free form deformation to create a reference CT (rCT). All the deep-learning generated synthetic CTs (sCT) were compared to this reference.

6.3 Pix2pix GAN Architecture with Feature Matching

A 2.5 dimensional (2.5D) Pix2pix GAN-based deep-learning model with Feature Matching (FM) was proposed and the architecture is shown in **Figure 6-1** [293]. The Generator was used to generate synthetic CT (sCT) from the original CBCT, and the Discriminator was used to distinguish the synthetic CT (sCT) from the reference CT (rCT). The Generators and Discriminators competed against each other until they reached an optimum.

The Generator was implemented using U-net architecture, in which each Conv-ReLU-BN block consists of either convolution or de-convolution layers with kernel size of 3x3, a batch normalization layer (BN) and a leaky rectified linear unit (ReLU). Concatenate connections were linked between the corresponding layers of the encoder and decoder. The activation function after the last convolutional layer was Elu. Then the synthesize CT (sCT) slices were used as the input of the Discriminator with the reference CT (rCT) slices as ground truth. The discriminator was a classifier that consisted of 8 stages of Conv-ReLU-BN block same as Generator.

The instability during the training of GAN is a critical issue which affects the output image quality from the generator. To address this issue, we implemented feature matching by changing the adversarial loss function [294]. This strategy forced the generator to generate images which could match the expected values of the features on the intermediate layers of the discriminator, besides the output of the discriminator. The loss function for the Discriminator was constructed as:

$$Loss_{D,G} = \sum_{l}\left(\frac{1}{n^{l}}\sum_{n^{l}}\left|f^{l}(D(sCT) - f^{l}(D(rCT))\right|\right) \quad (l = 2, 4, 6, 8)$$

where f^{l} is the output feature map on layer l, and n^{l} represents number of pixels. The sCT and rCT slices were used as input. The corresponding feature maps from the 2nd, 4th, 6th, and 8th layers were obtained with mean absolute error summed together as loss function. To further preserve the HU values between rCT and sCT, the L1 norm distance was added to the loss function:

$$L_1 Loss = \frac{1}{n}\sum_{n}|sCT - rCT|$$

where n is the number of pixels on the images, with the final adversarial loss function as:

$$Loss_{adersarial} = Loss_{D,G} + \alpha L_1 Loss$$

where α is the weight between two different loss functions.

The 2.5 D architecture used a volume set with adjacent three slices as input of the network. This method stacked neighboring three slices together as different channels of the input to provide the network with 2.5D information, providing more morphology information to reconstruct the high-quality images.

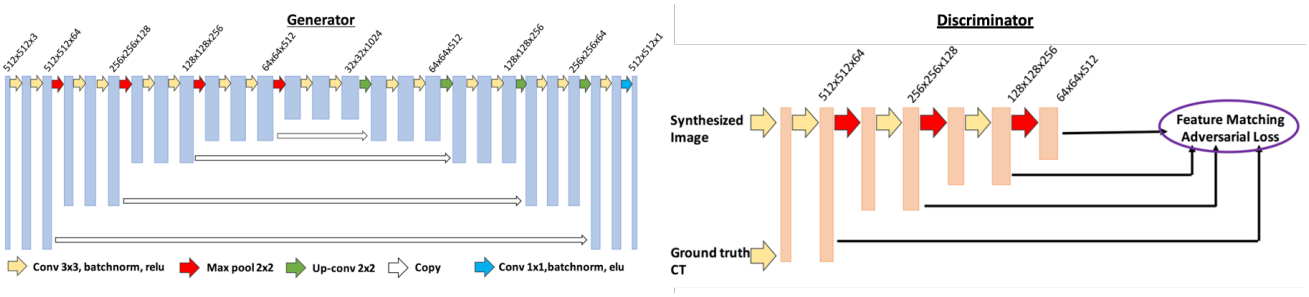

Figure 6-1: The GAN architecture based on the U-Net as the generator. The Generator is used to generate synthetic CT (sCT) from the original CBCT, and the Discriminator is used to distinguish the synthetic CT (sCT) from the reference CT (rCT) which serves as the ground truth. The input data size is 512 × 512 × 3 and the output data size is 512 × 512 × 1; the first two numbers represent resolutions and the third number represents channels. The discriminator is a classifier that consists of 8 stages of Conv-ReLU-BN block.

6.4 Other Network Architectures

Besides feature mapping as mentioned in 2.2, another way to improve the synthesized image quality is to add perceptual loss [295]. The architectures as GAN model with vs. without additional perceptual loss were tested. VGG16 on ImageNet [48, 66, 296] was used to extract the image features for two types of losses: content loss and style loss. The content loss was defined as the Euclidian distance between the feature maps from original and synthesized images of each layer:

$$Loss_{content} = \sum_j \frac{1}{h_j w_j c_j} \left\| \mathrm{f}^j(\mathrm{rCT}) - \mathrm{f}^j(\mathrm{sCT}) \right\|_2^2$$

where $\mathrm{f}^j(\mathrm{CT})$ and $\mathrm{f}^j(\mathrm{sCT})$ stand for the feature maps from the j_{th} layer in the network for the ground-truth and synthesized image, respectively, and h_j, w_j, and c_j stands for the size.

The style loss was used to control the similarity of image styles and was defined as the Euclidian distance between the stylistic feature maps from original and synthesized images of each layer:

$$Loss_{style} = \sum_j \left\| Gram_j(rCT) - Gram_j(sCT) \right\|_2^2$$

where Gram matrix was defined as:

$$Gram_j(y)_{m,n} = \frac{1}{h_j w_j c_j} \sum_{h=1}^{h_j} \sum_{w=1}^{w_j} f^j(y)_{h,w,m} * f^j(y)_{h,w,n}$$

where m and n represent different output channels from the same layer. So the loss function becomes

$$Loss_{percepertual} = Loss_{adversarial} + \beta_1 Loss_{content} + \beta_2 Loss_{style}$$

β_1 and β_2 are the weights.

In addition, we also compared our methods with previously published models as U-net [297, 298] and cycleGAN [299]. U-net is a popular algorithm in image processing field and some investigators have explored its use in this context [135, 297, 298]. In brief, the basic structure consists of convolution and max-pooling layers at the descending part (the left component of U), and convolution and up-sampling layers at ascending part (the right component of U) [135]. In the down-sampling stage, the input image size is divided by the size of the max-pooling kernel size at each max-pooling layer. In the up-sampling stage, the input image size is increased by the operations, which are performed and implemented by convolutions, where kernel weights are learned during training. The arrows between the two components of the U show the incorporation of the information available at the down-sampling stage into the up-sampling stage, by copying the outputs of convolution layers from descending components to the corresponding ascending components. In this way, fine-detailed information captured in descending part of the network is used at the ascending part. The output images share the same size of the input images.

A few work have been done using CycleGAN to obtain synthetic CT from CBCT [299, 300]. In brief, it consisted of two generators as G_A (mapping from CBCT to sCT) and G_B (mapping from CT to sCT). It also had two discriminators as D_A to distinguish rCT from fake CT, and D_B to distinguish real CBCT from fake CBCT. With this bidirectional configurations, cycled CBCT (cycleCBCT) from sCT and cycled CT (cycleCT) from sCBCT could be obtained. Besides adversarial loss from discriminators, cycle loss was added to the final function:

$$Loss_{cycleGAN} = Loss_{adversarial-CT} + Loss_{adversarial-CBCT} + \gamma(Loss_{cycle-CT} + Loss_{cycle-CBCT})$$

where

$$Loss_{cycle-CT} = \frac{1}{n}\sum_{n}|CT - cycleCT|$$

$$Loss_{cycle-CBCT} = \frac{1}{n}\sum_{n}|CBCT - cycleCBCT|$$

and n is the number of pixels on the image and γ is the weight of the cycle loss.

6.5 Model Configuration and Statistical Analysis

The pixel intensities on each slice were normalized to mean=0 and standard deviation=1 for pre-processing. All models were trained with Adam optimization with a mini-batch size of 2 and epoch number of 100 [105]. All weights were initialized from He normal initializer [140]. Batch normalization was used after each convolutional layer [301]. The learning rate was set to 0.0001 with momentum term 0.5 to stabilize training. The generator was trained twice while the discriminator was trained once to keep the balance between the two components. To control the overfitting, three methods were utilized. First, before training, all images were augmented by horizontally flipping, a small angle rotation, as well as adding some background noise. Then L2 regularization term was added to the final loss function. Lastly, during the training process, early stop was applied based on the lowest validation loss to obtain the optimized model.

10-fold cross validation was used to evaluate the performance of the model. Each slice was used as an independent case. The training and validation sets included 150 CBCT-CT pairs, and 90% of cases were used for training while remaining 10% were used for validation purpose. The results from the validation sets were calculated. A separate dataset with additional 15 pelvic patients and 10 head-and-neck patients with paired CT and first-

day CBCT, with CBCTs collected at a different linear accelerator (linac) machine, was used as an independent testing set to evaluate the robustness of proposed algorithm.

Synthetic CT slices (sCT) were firstly generated using the proposed model then rendered into 3D volumes to compare to the reference CT (rCT) images. Two metrics as peak signal-to-noise ratio (PSNR), and mean average error (MAE) were calculated by comparing synthetic CT and reference CT [26]. PSNA measured the maximum possible power of a signal, with higher value indicating better image quality. MAE measured absolute HU differences of every single pixel between target and reconstructed image, with lower value indicating closer similarity to target. A total of 7 models were tested and compared: (1) 2.5D Pix2pix GAN with feature matching – as proposed in this study; (2) 2D Pix2pix GAN without feature matching, using single slice as network input; (3) 2D Pix2pix GAN with feature matching; (4) 2.5D Pix2pix GAN without feature matching; (5) 2.5D Pix2pix GAN with feature matching and perceptual loss; (6) U-net; and (7) cycleGAN.

6.6 Results

Figure 6-2 shows two case examples with reference CT images, raw CBCT images and deep-learning generated synthetic CT (sCT). The intensity differences in Hounsfield Unit (HU) are also displayed. It can be clearly seen that the synthetic CT had much closer HU level to the reference CT compared to the raw CBCT. The group result in the validation dataset is summarized in **Table 6.1**. All deep-learning generated synthetic CTs showed improved image quality with less discrepancies (smaller MAE) to reference CT. The proposed algorithm as 2.5 Pix2pix GAN with feature matching was shown to be the best model among all tested methods with the highest PSNR and the lowest MAE. The mean

MAE improved from 26.1±9.9 HU (CBCT vs. rCT) to 8.0±1.3 HU (sCT vs. rCT). The PSNR also increased significantly from 16.7±10.2 (CBCT vs. rCT) to 24.0±7.5 (sCT vs. rCT) in the validation set. The results showed that changing from 2D to 2.5D input had slight improvement for the PSNR and MAE but not statistically significant, due to only 3 slices information was added into the model. U-net was under-performed than any of GAN networks. As shown in **Figure 6-3**, the U-net generated blurred images and lost detailed information especially at the tissue boundaries. Overall, the deep-learning based CBCT generated through the GAN methods had greatly reduced artifacts compared to the corresponding raw CBCT.

The proposed algorithm was further applied to the independent testing dataset. Due to different linac machine setting, the image discrepancies from raw CBCT to CT was larger compared to the training/validation set. The average MAE was 43.8±6.9 HU for pelvic cases originally, but was improved to 23.6±4.5 with deep-learning. The PSNR was improved from 14.53±6.7 to 20.09±3.4. When extended to head-and-neck regions, though with less extent improvement, the model still produced less MAE discrepancies to 24.1±3.8 from original 32.3±5.7 HU. The PSNR was improved from 20.34±1.6 to 22.79±3.4. An example of the head-and-neck cases is shown in **Figure 6-4**. It shows improved image quality with much closer HU to reference CT.

The network code was written in Python 3.6 and TensorFlow 2.0 and experiments were performed on a GPU-optimized workstation with a single NVIDIA GeForce GTX Titan X (12GB, Maxwell architecture). Once the model was trained, it took 11-12 ms to process one slice and generate a 3D volume of synthetic CT in less than a second.

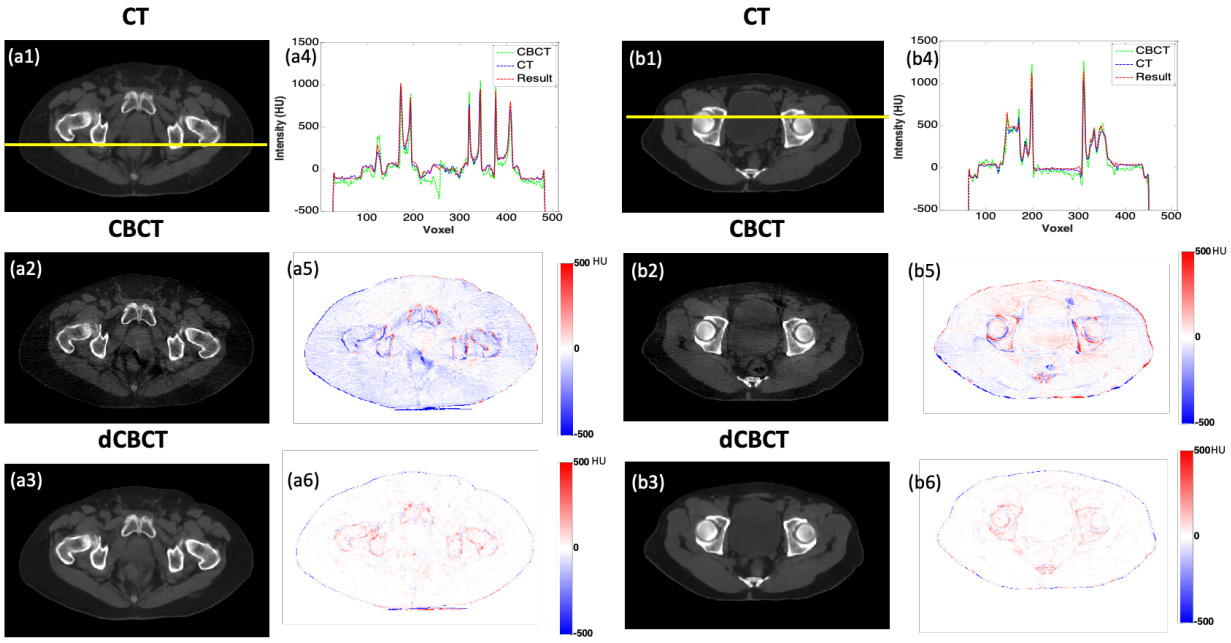

Figure 6-2: Two case examples: (1) CT image, (2) CBCT image, (3) deep-learning based CBCT (dCBCT) predicted using 2.5D GAN with feature matching, (4) line plot showing intensity profile of CT (blue), CBCT (green) and dCBCT (red) in range of [-500, 1500] HU, (5) HU differences between CBCT to CT in range of [-500, 500] HU, (6) HU differences between dCBCT to CT in range of [-500, 500] HU.

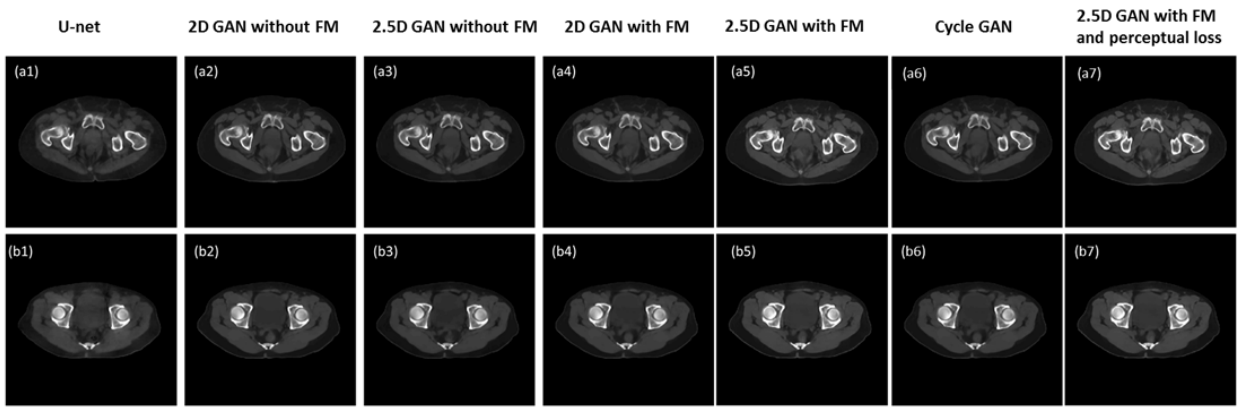

Figure 6-3: Comparison among the presented algorithm and other 4 algorithms using MAE and PSNR, and qualitative assessment using two above examples: (1) Prediction results using U-net. (2) Prediction results using 2D GAN. (3) Prediction results using 2.5D GAN. (4) Prediction results using 2D GAN with feature matching. (5) Deep-learning based CBCT (dCBCT) predicted using 2.5D GAN with feature matching. (6) Prediction results using CycleGAN. (7) Prediction results using 2.5D GAN with feature matching and perceptual loss.

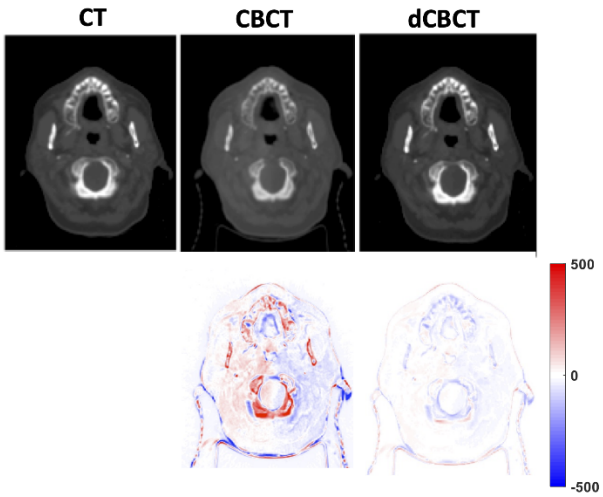

Figure 6-4: One head-and-neck case example from an independent testing dataset. The deep-learning based CBCT showed much closer HU to reference CT.

Table 6.1: The Mean Average Error (MAE) and Peak Signal-to-Noise Ratio (PSNR) of the original CBCT and the synthetic CT generated by using 7 deep learning architectures compared to the original CT.

Network	Mean Average Error (MAE)	Peak Signal-to-Noise Ratio (PSNR)
Original CBCT	26.1±9.9*	16.7±10.2
U-net	19.2±6.4**	18.9±6.7
2D GAN without FM	9.4±1.2	22.4±3.8
2.5D GAN without FM	9.3±2.1	22.7±2.9
2D GAN with FM	8.1±1.4	23.8±1.8
2.5D GAN with FM	**8.1±1.3**	**24±7.5**
2.5D CycleGAN	9.2±1.5	23.2±7.8
2.5D GAN with FM and Perceptual Loss	9.9±3.2	21.3±6.4

* MAE between the original CBCT and CT is significantly higher compared to other methods
** MAE between the U-net generated synthetic CT and original CT is significantly higher compared to other methods

6.7 Summary and Discussion

We have developed a deep-learning based model to generate synthetic CT from routine CBCT images based on pixel-to-pixel (Pix2pix) GAN. The model was built and validated on 30 pelvic patients with 150 paired CT-CBCT images, and further tested with an

independent cohort with 15 additional pelvic cases and 10 head-and-neck cases collected at another linac machine. The image quality of the deep-learning based synthetic CT had been overall improved with much less MAE discrepancies to reference CT in both validation and testing datasets.

The Online CBCT has been widely used for daily positioning and target alignment. It may also allow early assessment of treatment response and be a prognostic factor of treatment outcomes. However, its use in adaptive radiotherapy is limited due to large scattering and inaccurate mapping of HU. Numerous mathematical algorithms have been proposed to improve CBCT image quality, including iterative reconstruction (IR) [286, 287] and compressed sensing (CS) algorithms [302]. However, those algorithms require high computational complexity and are time expensive, thus have not been commonly implemented for clinical use. Alternatively, conventional analytic reconstruction algorithms, such as filtered back-projection, remain as the mainstream due to its fast computation.

Recently, deep learning based approaches have emerged as a potential solution to overcome computational complexity of prior reconstruction algorithms and inherent poor image quality of CBCT [298, 299, 303, 304]. These approaches have demonstrated promising results for CBCT by applying denoising networks to generate synthetic CT images. Kida *et al.* used a 2D U-net convolutional neural network (CNN) for the pelvic CBCT-to-sCT generation, and reported improvement of MAE from 50 to 31HU using 16 patient cases [303]. Similarly, Li *et al*. used a U-net CNN with residual convolution unit for the head-and-neck CBCT-to-sCT generation with 70 pairs of CBCT-CT [305]. Yuan *et al*. also applied similar technique for head-and-neck patients, but with CBCT collected at fast-scan

low-dose acquisition [298]. Recently, cycleGAN has been proposed to deal with the unpaired training data in multiple applications in medical imaging such as MRI-based sCT generation [290], organ segmentation [306], and CBCT-based sCT generation [299, 307, 308]. CycleGAN incorporates an inverse transformation to better constrain the training model toward one-to-one mapping. In the application of CBCT-to-sCT generation, Liang *et al*. applied cycleGAN to train the CBCT-planning CT dataset without performing deformable registration [299]. The cycleCBCT generated from CT was used to restrain the network. The mean MAE was improved from 32.3±5.7 to 25.0±5 HU for the head-and-neck patient cohort. Similarly, Harms *et al*. published a CBCT-to-sCT generation method using cycleGAN with the incorporation of residual blocks and a novel compound loss in the cycle consistency loss function with improved results [308]. The authors mentioned that although cycleGAN was initially designed for unpaired mapping, in its application in medical imaging, due to the imaging complexity, rigid registration should still be recommended to preserve the qualitative values.

We have compared our proposed deep-learning model with previous reported work. It was found the U-net CNN underperformed than any GAN based methods on our datasets. This might be due to the fact that the algorithm started with multi-layer image smoothing which would in-turn resulted in large signal discrepancies at boundaries. Another tested algorithm as CycleGAN has been widely applied to match unpaired images. Yet, with the co-registration done in the preprocessing step, the input CBCT and the reference CT were matched with similar morphologies. Since the purpose of this study is to generate synthetic CT from CBCT and further to match with reference CT, with this to-match purpose, the cycle loss as used in CycleGAN was not deemed necessary. In addition, we tried to add

perceptual loss into the model. Yet, the initial weights merely captured the features of natural images, and it actually disturbed the training process. By comparing all deep-learning algorithms, 2.5D pix2pix GAN with feature matching was identified as the best model. The model was built on a large pelvic datasets with 150 pairs of CBCT-CT. The pelvic dataset contained enough variation of the anatomy structures, which helped to improve the robustness of the GAN model. The co-registration results contributed to the good correspondence between CT slices and CBCT slices, thus the conversion difficulty was reduced. Notably, the current model not only showed improved results in the validation set, it was further extended to an independent image set with two disease sets collected on a different machine. The improvement was again confirmed by a significant reduction of MAE discrepancies. All these demonstrated its robustness in clinical image sets and potential clinical use.

Despite the promising results, we acknowledge several limitations. Due to technical limitation of the GPU capacity, only three adjacent slices as 2.5D information were used as input. The performance did not show significant improvement compared to 2D single-slice method. The future direction is to include more slices as input or a true 3D information. This may require large data samples with an order of 5-10 times more in high computer power. The second limitation is that signals between tissue boundaries, as body-to-air or bone-to-soft tissue, were not preserved. This may be due to the signal loss during pre-processing as volumetric resizing and image interpolation. To overcome this issue, high-resolution images with original details need to be retained during the pre-processing for which again high computational power is needed. Thirdly, lack of the same day paired CT and CBCT at the same position prevented us to precisely evaluate the exact HU mapping.

Data collection with different disease types is on-going. In addition, not just the mathematical MAE calculation but dosimetric comparison need to be further evaluated.

Overall, CBCT plays a very important role in image-guided radiation therapy (IGRT). Enhancement of its quality can contribute to daily patient setup and adaptive dose delivery, thus enabling higher confidence in patient treatment accuracy. The results of this study demonstrate that the artificial intelligence (AI) technique can improve CBCT image quality without hardware improvement. Once the model is trained, it takes less than a second to process a deep-learning based volumetric CBCT set. The results also show that the improved CBCT can achieve high image quality to be close to the level of conventional CT, thus have the potential to be used for adaptive planning. Overall, the method presented in this study may provide a time-efficient and economic-efficient solution for machines that are coupled with CBCT capability. The output may improve the soft-tissue definition that is necessary for accurate visualization, contouring, deformable image registration, and may enable new applications, such as CBCT-based online adaptive radiotherapy.

Chapter 7. Neoadjuvant Chemoradiation Therapy Response Prediction

7.1 Motivation and Clinical Application

Neoadjuvant chemoradiation therapy (CRT) followed by total mesorectal excision (TME) is the current standard-of-care treatment for locally advanced rectal cancer (LARC). Following CRT, around 15% to 27% of patients can achieve pathologic complete response (pCR) [309, 310]. For these patients without residual invasive cancer remaining, there is a question as to whether they need TME, as this intrusive surgery is associated with significant complications and morbidity [309, 311-313]. Several studies have shown that pCR patients have low rates of local recurrence, and thus less invasive, alternative surgical treatments such as sphincter-saving local excision, or watch-and- wait approaches are gaining popularity [312-315]. However, pCR has to be confirmed after the patient receives surgery, and it is important to identify patients who are likely to be clinical complete responders (CCR) so a less aggressive surgery (not TME) can be performed to confirm pCR.

Medical imaging, especially magnetic resonance imaging (MRI), which can noninvasively evaluate therapeutic response in cancer has shown promise for early predictions of pCR [316-321]. MR imaging done at different times during the course of CRT, including pre-treatment [320, 321], during [317, 319], and after completing CRT [316, 318], can be analyzed separately or in combination to provide anatomic and functional information. A few studies have evaluated the prognostic value of MRI for assessing CRT outcome for LARC [322-326]. The MRI done after completing CRT can be referenced with prior MRI's to assess clinical response and help determine subsequent regimens or select

candidates for an alternative surgical plan.

With the advance of MR imaging technology, several different sequences can be included in the MRI protocol within a reasonable imaging time (< 30 min), and this multi-parametric MRI can provide comprehensive information to facilitate quantitative radiomics analysis for tumor response prediction [276, 327]. Radiomics extracts hundreds of quantitative image features, and then uses sophisticated statistical analysis to classify different groups. A study by Nie et al. showed that radiomics analysis based on pre-treatment multi-parametric MRI performed well in predicting patients who achieved pCR after completion of CRT [276], with a prediction accuracy of 0.8-0.9. Another study by Liu et al., that combined the pre-treatment MRI with post-CRT treatment MRI predicted pCR with an accuracy of 0.97 [327]. These studies indicate the great potential of radiomics analysis based on multi-parametric MRI to predict CRT response. In addition to radiomics, machine learning with convolutional neural network (CNN) provides a new classification strategy based on artificial intelligence pattern recognition of images, without relying on pre-defined metrics. CNN analysis has been employed in the field of oncology for noninvasively profiling tumor heterogeneity to predict neoadjuvant therapy response [328-331].

The purpose of this work was to apply different analysis methods, including whole tumor ROI-based averaged analysis, radiomics and deep learning using CNN, to predict pathological response in LARC patients receiving CRT. The pre-treatment MRI, and the early follow-up MRI performed 3-4 weeks after staring of the radiation therapy, were analyzed to differentiate between pCR and non-pCR patients, and also between good responders (GR) and non-GR patients.

7.2 Subjects and Image Dataset

Patients

A total of 51 patients (mean age 60) with locally advanced rectal cancer**,** based on the American Joint Committee on Cancer (AJCC) TNM system, without distant metastasis were included in this study. Only complete MRI datasets that included all sequences and had high quality for quantitative analysis were analyzed, which included 45 patients with pre-treatment MRI and 41 patients with mid-RT follow-up MRI. Of these, 35 patients had both pre-treatment and mid-RT MRI. **Table 7.1** shows demographic information of these patients. This was a retrospective study approved by the Institutional Ethics committee and the informed consent was waived.

Treatment Protocol

The chemoradiation therapy protocol was done according to the National Comprehensive Cancer Network (NCCN) guidelines. The total radiation dose was 50 Gy, delivered for 25 fractions in 5 weeks using IMRT technique. Patients also received capecitabine 825 mg/m^2 orally, twice daily for 5 consecutive weeks and oxaliplatin 110 mg/m^2 once every 3 weeks. After completing the 5- week CRT, the patients received one additional cycle of chemotherapy using 5-fluorouracil + oxaliplatin or capecitabine + oxaliplatin. After a recovery period of two weeks (6-8 weeks after radiation), TME was performed by either anterior or abdominoperineal resection.

Pathologic Response Evaluation

Following surgery, the specimen was examined by a gastrointestinal pathologist using

the modified tumor regression grade (TRG) based on Ryan's definition [332], to determine the pathologic response. The pathologic complete response (pCR) was defined as the absence of viable adenocarcinoma cells (TRG 0). Additionally, patients were separated into good responders (GR) and non-GR. The GR group included complete response with TRG 0 and those with only a small cluster or isolated cancer cells remaining (TRG 1). The non-GR group included patients with residual cancer remaining but with predominate fibrosis (TRG 2) and patients with poor response with extensive residual cancer (TRG 3). The number of patients in each pathological response group is shown in **Table 7.1**. Among the 45 patients with pre-treatment MRI, 10 (22.2%) were classified as pCR and 35 (77.8%) were non-PCR; and 31 (68.9%) were classified as GR and 14 (31.1%) were non-GR.

MR Imaging Protocol

Patients were scanned with a 3.0 Tesla MR (Signa HDxt, GE Medical Systems) using a phased-array body coil with no special bowel preparation. The imaging protocol consisted of an axial T2-weighted and a T1-weighted image followed by axial diffusion weighted imaging (DWI) acquired with b = 0 and 800 s/mm^2 using a single-shot echo planar imaging sequence. Lastly a multiphase axial T1w DCE-MRI (dynamic-contrast-enhanced) sequence was performed using a spoiled gradient echo sequence LAVA (Liver Acquisition with Volume Acceleration) with 4 frames, one pre-contrast (L1) and three post-contrast at 15 seconds (L2), 60 seconds (L3), and 120 seconds (L4) after the injection of 0.1 mmol/kg body-weight gadolinium contrast agents (Gd-DTPA). The pre-treatment MRI was performed 1-2 weeks prior to CRT, and mid-RT follow-up MRI was performed at 3-4 weeks after the start of CRT. The representative images of one patient are shown in **Figure 7-1**.

Table 7.1: The demographic information, tumor volume and ADC in different response groups

	pCR	Non-pCR	GR	Non-GR
Pre-treatment (N=45)	N=10	N=35	N=31	N=14
Male:Female	5:5	26:9	21:10	10:4
Mean age (SD)	56.3 (11.1)	59.7 (8.0)	58.0 (9.0)	61.1 (8.1)
Mean tumor volume (SD, cm^3)	14.2 (6.0)*	21.5 (15.8)*	15.3 (8.7)‡	28.0 (18.9)‡
Mean ADC (SD, mm^2/s)	0.93 (0.09)	0.95 (0.14)	0.94 (0.14)	0.94 (0.11)
Mid-RT follow-up (N=41)	N=9	N=32	N=27	N=14
Male:Female	5:4	23:9	18:9	10:4
Mean age (SD)	56.4 (11.8)	60.3 (7.9)	58.3 (9.5)	61.9 (7.5)
Mean tumor volume (SD, cm^3)	6.6 (4.5)**	11.7 (12.1)**	6.8 (6.1)‡‡	17.7 (14.5)‡‡
Mean ADC (SD, mm^2/s)	1.33 (0.16)	1.37 (0.18)	1.36 (0.19)	1.33 (0.15)

* The volume is significantly smaller in pCR than in non-pCR (* p=0.009, ** p=0.047)

‡ The volume is significantly smaller in GR than in non-GR (‡ p=0.01, ‡‡ p=0.03)

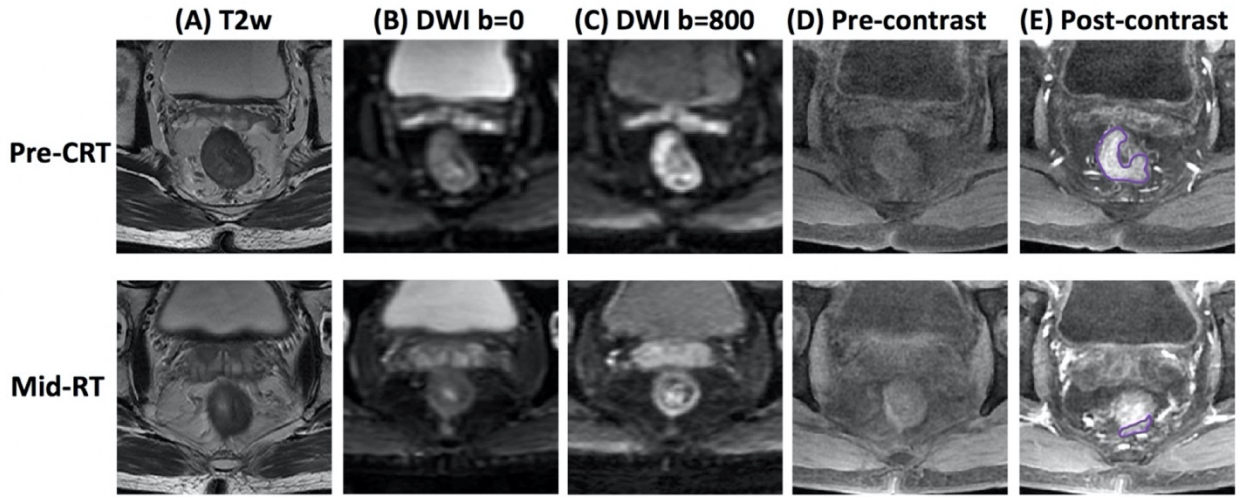

Figure 7-1: MR images of a 51-year-old male with low-rectum cancer at stage of cT3N+M0 taken pre-treatment (top row) and mid-RT (bottom row). (A) T2-weighted image, (B) the diffusion- weighted image with b=0 s/mm2, (C) the diffusion-weighted image with b=800 s/mm2, (D) L1 pre-contrast image, (E) L2 post-contrast image taken at 15 seconds after injection. This patient achieved pCR after completing the entire course of CRT.

7.3 ROI and Radiomics Analysis

All images were reviewed on a MIM Maestro (MIM Software Inc, OH, USA) workstation used for radiotherapy planning, by an MRI radiologist experienced in radiation oncology. The tumor region of interest (ROI) was manually outlined on each slice containing the tumor, excluding the intestinal lumen, on the post-contrast image L2 or L3, while all other sequences were utilized as references. For each patient, the manually drawn ROI was mapped to other images (T2, ADC, other DCE) through co-registration, implemented with a linear rigid transformation algorithm, cubic interpolation, and a mutual information cost function. The transferred ROI was also inspected by a medical physicist, and if necessary, modified. After the ROI is drawn, the total tumor volume was calculated by adding up all tumor areas × slice thickness. The mean apparent diffusion coefficient (ADC) was calculated by averaging the ADC of all tumor pixels. The mean signal intensity on each DCE image, L1, L2, L3 and L4, was also calculated. In addition, the change of intensity (slope) between L3 and L4 was calculated to assess the wash-out DCE pattern.

The radiomics analysis was done performed following the same procedures reported in Nie et al. [276], using two categories: textural features and histogram-based features. The texture analysis was done extracted using the 18 Haralick's Gray Level Co-occurrence Matrix (GLCM) features, including 18 features: autocorrelation, cluster prominence, cluster shade, contrast, correlation, dissimilarity, energy, entropy, homogeneity 1, homogeneity 2, maximum probability, sum average, sum variance, sum entropy, difference variance, difference entropy, information measure of correlation 1, information measure of correlation 2. For the histogram-based analysis, a total of 12 parameters were calculated,

including: 10%, 20% ... 90%, 100% values, kurtosis, and skewness. For each case, a total of 96 parameters were calculated, including 18 texture on T1, 18 texture on T2, 18 texture+12 histogram parameters on the ADC map and 18 texture+12 histogram parameters on the DCE L2 image.

A 3-layer perceptron artificial neural network (ANN) was utilized to select parameters and build the diagnostic model. All parameters from each case were included as input nodes of the ANN, and the output node was either pCR vs. non-pCR or GR vs. non-GR. The number of nodes in the hidden layer was determined by a formula of $m = (n + l)^{1/2} + \alpha$, where m is the number of the hidden nodes, and n is the number of nodes in the input layer, l is the number of nodes in the output layer, and α is a constant from 1 to 10. The forward search strategy was used to search different combinations of predictors by adding predictors one by one to see if the model performance improved. During the training process, the weights were updated by minimizing the error function from the output neuron with mean square error (MSE). The learning process continued until it converged to a predefined value (<0.001) or until the maximum number of iterations, of 10000, was reached. The performance was evaluated using 4-fold cross-validation.

Each case had only one chance to be included in the testing dataset, and after the process was completed, the predicted pCR or GR probability of all cases were used to generate the ROC curve. The ANN analysis was performed in the Matlab Neural Network ToolBox, software version 7.12 (The Mathworks Inc.).

The feature selection was done using an artificial neural network, with 4-fold cross-validation. After a final model was developed, the overall classification performance was evaluated using receiver operating characteristic (ROC) analysis in the entire dataset. The

features analyzed extracted from the T1+T2 images, ADC map, and DCE L2 post-contrast image, were first analyzed separately, and then combined. In addition, the ROI-based parameters including the total tumor volume, mean ADC, and mean signal intensity on the DCE images were added to the radiomics analysis to investigate whether they could further improve the prediction accuracy.

7.4 CNN Configuration

For the deep learning analysis using CNN, the input was the smallest square bounding box covering the tumor ROI. **Figure 7-2** illustrates the determination of the bounding box. The ROI's drawn on all tumor slices were stacked on a projection view, and the smallest square bounding box using the centroid as the center point was determined. The bounding box on each slice was resized to 32×32 pixels as the inputs to CNN. **Figure 7-2A (**top panel) and **Figure 7-2B** (bottom panel) show the generated smallest bounding box for the pre-treatment and mid-RT MRI of one patient. The input box of the T2 and DWI images was were processed using the same method.

The CNN architecture used in this study is shown in **Figure 7-3**. For each patient, the input included 6 sets of images: one T2, two DWI with b= 0 and 800 s/mm^2, and three LAVA frames (L1, L2 and L3). The image intensity was normalized to mean=0, standard deviation=1in each MR sequence was independently normalized using z-score values (μ = 0, σ = 1). The two DWI images were normalized together to consider the intensity changes between b= 0 and 800 s/mm^2 images. Similarly, the LAVA images in the DCE sequence were also normalized together. In order to account for the problem of small case number, each imaging slice was used as independent input, and data augmentation was performed using

Affine transformation, to 20 times. There were The CNN was 7 layers and the size of the convolution kernel was 3×3. For the seven layers, the stride number of the 2nd, 4th, and 6th convolution layers in the output transformation was 2, which reduced the spatial resolution to 1/4 the size of the input feature map.

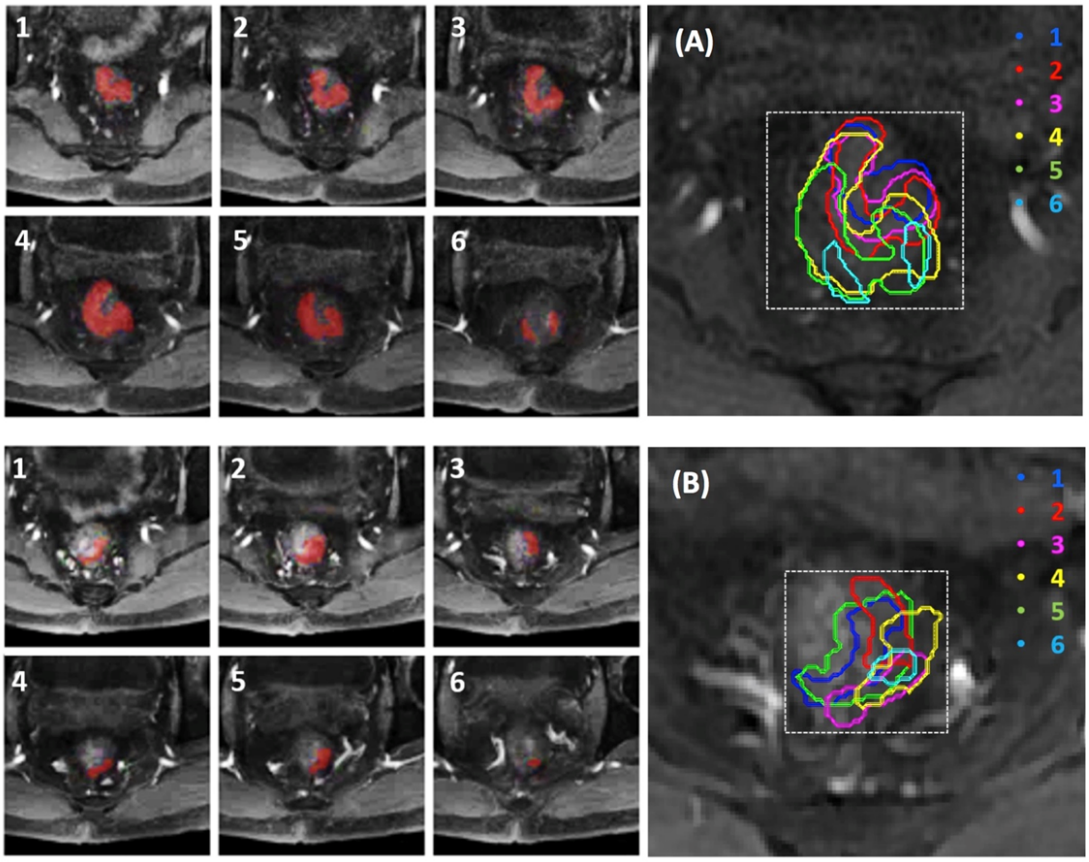

Figure 7-2: Determination of smallest bounding box on pre-treatment MRI (A, top panel) and mid- RT MRI (B, bottom panel) of a 56-year-old male with mid-rectum cancer at stage of cT3N+M0. Tumor ROI (red) outlined on tumor-containing MR slices (1-6) are stacked on a projection view to determine the smallest square bounding box.

Training was implemented using the Adam optimizer, an algorithm for first-order gradient-based optimization of stochastic objective functions, based on adaptive estimates of lower-order moments [105]. Parameters were initialized using the heuristic described by He et al. [140]. L2 regularization was performed to prevent over-fitting of data by

limiting the squared magnitude of the kernel weights. The learning rate was fixed to 0.001. Additionally, a batch normalized gradient algorithm was employed to allow for locally adaptive learning rates that adjust according to changes in the input signal [301]. To control overfitting, dropout layers with 50% preservation rate were added after each convolution layer and the last fully connected layer [61]. The Software code was written in Python 3.5 using the open-source TensorFlow r1.0 library (Apache 2.0 license) [106], on a GPU-optimized workstation with a single NVIDIA GeForce GTX Titan X (12GB, Maxwell architecture).

The classification performance was evaluated by ROC analysis using 10-fold cross-validation, 90% cases for training and the remaining 10% for testing. The CNN was first done using 45 pre-treatment MRI cases and 41 mid-RT MRI cases separately, with the input size of 32×32×6. Then the CNN was done using the 35 patients who had both MRI together, with the input size of 32×32×12. For the combined analysis, in order to consider the change of tumor volume between the pre-treatment and mid-RT, the input bounding box for the pre-treatment and mid-RT of each patient was made the same. The center of the projected tumor ROI shown in **Figure 7-2A and B** was matched, and the smallest square bounding box covering all pre-treatment and mid-RT tumor ROI was used as the inputs in the CNN analysis.

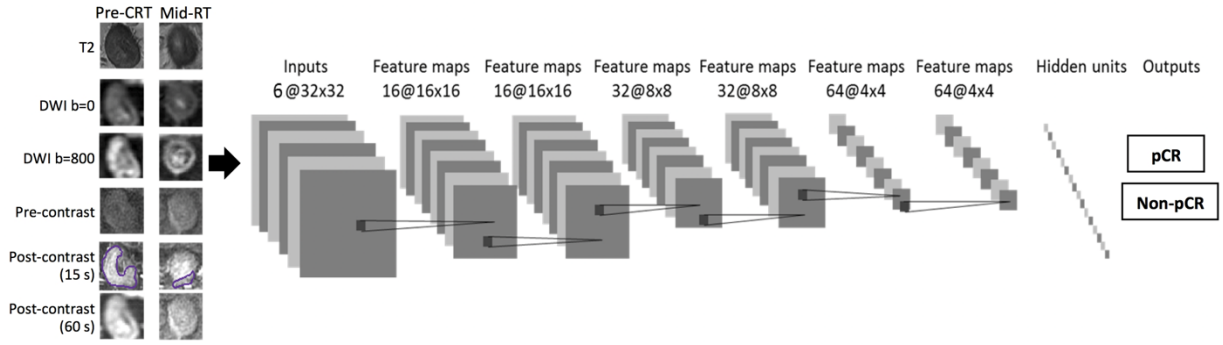

Figure 7-3: Overview of CNN architecture with 7 layers to classify different pathologic response groups: pCR vs. non-pCR, and GR vs. non-GR. Six sets of images are used as inputs: one T2, two DWI with b=0 and 800 s/mm2, and three DCE images (L1, L2 and L3). The analysis is done using pre-treatment MR alone and mid-RT alone (6 input channels), and patients with both pre- treatment and mid-RT together (12 input channels).

7.5 Statistical Evaluation

Statistical analysis was performed using the statistical computing software program R (version 3.5.0). Individual variables were analyzed to evaluate significant differences between groups (pCR vs. non-pCR and GR vs. non-GR) using an independent sample *t-test*. Levene's Test of Equality of Variance was first conducted to test for equal variance. A two-tail *P-value* < 0.05 was considered statistically significant. For radiomics and CNN, the ROC analysis was performed to evaluate the accuracy to differentiate pCR vs. non-pCR and GR vs. non-GR. The difference between two paired ROC curves was compared using the DeLong test.

7.6 Results

Whole Tumor ROI-based Analysis

The tumor volume and the mean ADC and DCE enhancements were calculated from the manually drawn tumor ROI. **Figure 7-4** shows the comparison of the mean tumor volume

and the mean ADC in the 4 different response groups. The tumor volume and ADC value in each group (mean with standard deviation) in the pre-treatment and mid-RT MRI are listed in **Table 7.1.** The tumor volume in the pCR group was significantly smaller than in the non-pCR group (p-value 0.009 and 0.047 for the pre-treatment and mid-RT MRI, respectively, also significantly smaller in the GR compared to the non-GR group (p-value 0.01 and 0.03, respectively). The results suggested that smaller tumors were more likely to achieve a good response either as pCR or GR. Regarding ADC, there was a statistically significant increase after treatment in the mid-RT follow-up MRI compared to the pre-treatment MRI in all 4 groups ($p<0.001$). However, there was no difference among pCR, non-pCR, GR, and non-GR groups for either the pre-treatment or mid-RT MRI. For the signal intensity on the DCE images, there was no significant difference in different groups, or between pre-treatment and mid-RT MRI.

For each patient who had both MRI sets, the percent change in tumor volume in mid-RT compared to pre-treatment was calculated. **Figure 7-5** shows the waterfall plots of the volumetric percent change in patients achieving pCR/non-pCR and GR/non-GR. The mean change was greater in pCR compared to non-pCR groups (-58.1% vs. -45.4%, p=0.28), and greater in GR compared to non-GR groups (-56.0% vs. -32.7%, p=0.03).

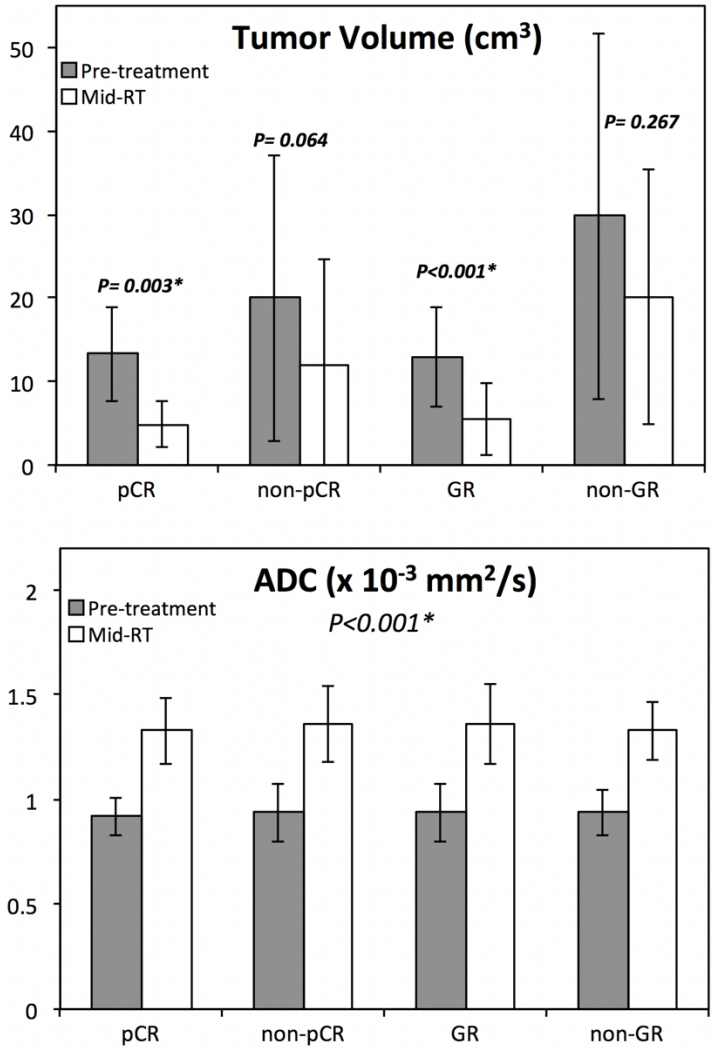

Figure 7-4: Bar plots showing differences of tumor volume and ADC between the pre-treatment (grey) and the mid-RT (white) in 4 response groups. The tumor volume decreases in mid-RT follow-up compared to the pre-treatment MRI is significant for the pCR and GR groups. The ADC increases in the mid-RT MRI compared to the pre-treatment MRI, and significant in all 4 groups.

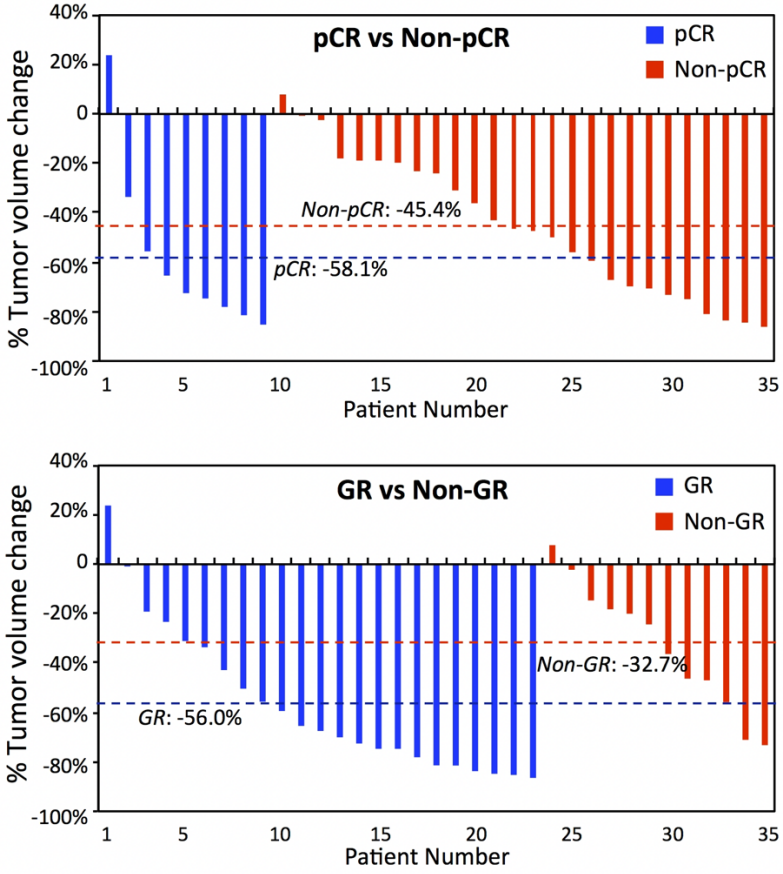

Figure 7-5: Waterfall plots of percent change in tumor volume of 35 patients who have both pre- treatment and mid-RT follow-up MRI. Top: Plot of pCR vs. non-pCR patients with mean change of -58.1% vs. -45.4% (p=0.28). Bottom: Plot of GR vs. non-GR with the mean change of -56.0% vs. -32.7% (p=0.03).

Radiomics

The radiomics prediction model was built from 96 features analyzed from the T1 and T2 images, ADC map, and the L2 post-contrast image using artificial neural network with four-fold cross-validation. The prediction performance of the final model was evaluated using the ROC analysis in the entire dataset. The area under the ROC curve (AUC) based on T1+T2, ADC, DCE post-contrast image, all radiomics, and ROI+radiomics are shown in **Table 7.2**. As expected, the model developed from more features have a better

performance, and the results combining ROI-based parameters and all radiomics features have the highest AUC of 0.80-0.86 (pCR vs. non-pCR) and 0.91-0.93 (GR vs. non-GR). In paired comparison done by the DeLong test, radiomics had a significantly better performance than ROI-based analysis in 3 of 6 response predictions, and combining ROI with radiomics significantly improved the performance only in GR vs. non-GR prediction using mid-RT MRI.

In radiomics analysis, since a forward search strategy was used by adding predictors one by one, we could carefully monitor the trend of change in the training cost and validation cost. Early stopping strategy was applied when the validation cost began to increase. Also, L2 regularization term was added to the cost function to control the overfitting. In most analysis, the AUC achieved by using the first 3-5 parameters are very close to the AUC of the final model, with <0.02 difference. The selected features were also used to build diagnostic models by using the logistic regression and support vector machine (SVM), and the obtained AUC's were very close to the results generated by ANN.

Deep learning using CNN

The prediction performance of the CNN was evaluated using ROC analysis based on ten-fold cross-validation. The range and mean AUC are also listed in **Table 7.2**. Overall, the results of CNN were inferior to radiomics, which was most likely due to the small case number insufficient for training. As shown in **Table 7.2**, when the pre-treatment and mid-RT were used together, the AUC was improved substantially. For pCR vs. non-PCR, the mean AUC was 0.59 for pre-treatment, 0.74 for mid-RT, and increased to 0.83 using both MRI, which was approaching the highest AUC of 0.86 based on ROI +radiomics features.

Table 7.2: The area under the ROC curve in ROI-based parameters, voxelized radiomics analysis and CNN deep learning to differentiate pCR vs. non-pCR and GR vs. non-GR

	ROI	T1+T2	ADC	DCE	Radiomics	ROI + Radiomics	CNN	ROI vs. Radiomics	Radiomics vs. ROI+Radiomics
pCR vs. Non-pCR									
Pre-Treatment	0.75	0.72	0.75	0.76	0.78	0.80	0.51-0.68 (mean 0.59)	Z=1.13 (p=0.31)	Z=1.4 (p=0.43)
Mid-RT Follow-up	0.77	0.69	0.77	0.74	0.80	0.82	0.71-0.75 (mean 0.74)	Z=3.21 (p=0.03)*	Z=1.6 (p=0.15)
Pre-Treatment + mid RT Follow-up	0.84	0.74	0.82	0.78	0.81	0.86	0.71-0.89 (mean 0.83)	Z=1.5 (p=0.22)	Z=1.9 (p=0.07)
GR vs. Non-GR									
Pre-Treatment	0.77	0.74	0.76	0.85	0.88	0.91	0.47-0.55 (mean 0.52)	Z=4.1 (p=0.01)*	Z=1.6 (p=0.14)
Mid-RT Follow-up	0.82	0.72	0.80	0.78	0.81	0.92	0.52-0.58 (mean 0.55)	Z=1.8 (p=0.15)	Z=3.4 (p=0.01)*
Pre-Treatment + mid-RT Follow-up	0.83	0.76	0.83	0.91	0.92	0.93	0.70-0.77 (mean 0.74)	Z=3.1 (p=0.04)*	Z=1.0 (p=0.47)

* Significant between two ROC curves compared by using the DeLong test

7.7 Summary and Discussion

In this study, we applied radiomics and deep learning using CNN based on the pre-treatment and early follow-up MRI after 3-4 weeks of radiation to predict the pathologic response of patients with LARC receiving neoadjuvant CRT. For all methods, the combined information from the pre-treatment and mid-RT follow-up can achieves a higher accuracy in predicting response compared to using either set alone. Using ROI-based averaged tumor volume and mean ADC combined with radiomics features could achieve a high accuracy of 0.86 to differentiate pCR from non-pCR, and 0.93 to differentiate GR from non-GR. Although a CNN with an appropriate normalization scheme could be implemented to predict the response, the range of accuracy was only fair, most likely due to the small number of datasets that were not sufficient for training and cross-validation. However, by combining the pre-treatment and mid-RT MRI together, the CNN could achieve accuracy of

0.83 in the differentiation of pCR and non-pCR, approaching that best radiomics results.

In our study, 22% of patients achieved pCR following CRT. Studies have found significant differences of overall survival (OS) and disease-free survival (DFS) between pCR and non-pCR patients [14]. For pCR patients, since the recurrence rate was very low, intrusive TME surgery probably caused more harm than benefit. Alternative approaches, including watch-and-wait, have been proposed to spare these patients from morbidities associated with TME. Two meta-analyses, including 23 studies of 867 patients and 15 studies of 920 patients, have shown no significant difference between clinical complete response (CCR) patients managed with a watch-and-wait approach or surgery in terms of DFS or OS [333, 334]. Thus, efforts have been devoted in finding reliable clinical or imaging parameters that can accurately identify CCR patients who have a high likelihood of pCR or close to pCR and spare them from surgery. It was recently shown that the accuracy to predict CRT response was increased when the post-CRT MRI information was used in combination with the pre-CRT MRI [276, 327]. Since the post-CRT MRI was done after completing the entire course of CRT, very close to surgery, it should be highly correlated with pathologic response. However, patients who did not respond well have already endured the toxicities of the entire treatment; therefore, using the post-CRT MRI to predict response could not provide much help. In this study, we investigated the value of an early follow-up MRI done 3-4 weeks after the start of CRT. For patients predicted not to responding well to the current regimen, alternative strategies can be considered, such as switching to other drug regimens or going to surgery early without further delay.

The accurate diagnosis of pCR and GR using visual examination on conventional MRI remains challenging in clinical settings. Although methods using multi-modality MRI (e.g.,

combining DWI and conventional MRI [318, 335-337], or PET/CT [338] show promise, further improvement is needed before implementation in clinical practice. Radiomics analysis is an efficient method to extract and integrate many quantitative imaging features, and that has been widely applied for many cancer imaging studies, e.g. diagnosis of benign and malignant lesions, classification of different molecular subtypes, and prediction of response to neoadjuvant chemotherapy, e.g. in breast cancer [214, 339]. Our results showed that the pre-treatment and mid-RT data gave similar prediction accuracies, 0.81 and 0.82 for pCR vs. non-pCR, and 0.91 and 0.92 for GR vs. non-GR, respectively. When the pre-treatment and mid-RT were combined, although the number of patients was smaller, the accuracy was increased to 0.86 for pCR vs. non-pCR, and 0.93 for GR vs. non-GR. The prediction of poor response for non-GR patients at an early time is very important, which could be used to optimize their treatment by changing the planned CRT regimen and to spare them from unnecessary toxicity, or to avoid delayed surgery.

We also analyzed the whole tumor ROI-based parameters, including the total tumor volume, mean ADC, and mean signal intensity on different frames of DCE images. After 3-4 weeks of treatment, there was a significant decrease in tumor volume and increase in ADC in mid-RT compared to pre-treatment MRI. Although these parameters alone were not good predictors for classifying different pathological response groups, they could be added to radiomics analysis to improve accuracy. The studies to investigate the change of tumor volume, ADC, and DCE signal intensity in an early time after starting of neoadjuvant chemotherapy have been reported extensively for breast cancer [340], but not for rectal cancer. Deep learning methods have been applied to evaluate the neoadjuvant therapy responses of different cancers, including bladder [328], esophageal [329], and breast

cancers [330, 331]. In this study, a CNN architecture was implemented to classify pCR vs. non-pCR, and GR vs. non-GR. This CNN model combined T2, DWI, and DCE image datasets as inputs. The results showed that the prediction accuracy of the CNN model was inferior to that of radiomics. This was very likely due to the small case number that was insufficient for training. For most CNN analysis, each 2D image slice was used as independent input, and further, the data augmentation was needed. When the pre-treatment and mid-RT datasets were combined together, the accuracy was greatly improved compared to using either dataset alone. For differentiating pCR vs. non-pCR, the accuracy was 0.59 using pre-treatment, 0.74 using mid-RT, and increased to 0.83 using both together. For differentiating GR vs. non-GR, the accuracy was 0.52 using pre-treatment, 0.55 using mid-RT, and increased to 0.74 using both together.

The major limitation of this study was the small case number, which not only affected the CNN, but also limited the choice of features in the radiomics analysis to predict final pathologic response. For deep learning using CNN, we have shown that it could be implemented by properly considering: 1) the change of signal intensity on the DWI images with different b values, 2) the change of signal intensity on the DCE images before and after injection of Gd contrast agents, and 3) further considering the change of tumor volume between pre-treatment and mid-RT follow-up MRI. These procedures, together with proper data augmentation, were critical to yield reasonable prediction results despite of the small case number. Lastly, the tumor ROI was only contoured once in our study. In radiomics study such as reported in [341], when the segmentation was done twice, it would allow the selection of robust features that had a high intraclass correlation coefficient. Our ROI drawing was carefully done using all MR sequences on an RT treatment

planning workstation, which we believe was valid, and can be implemented in a clinical setting.

In conclusion, we have shown that multi-parametric MRI allows extraction of comprehensive quantitative information to predict pathologic response in LARC patients after completing CRT. Adding an early-treatment follow-up MRI, at 3-4 weeks after starting of therapy, to the pre-treatment MRI could improve the accuracy in predicting final response. In this dataset, the radiomics analysis performed better compared to the deep learning using CNN. Further development of imaging methods is important to improve the care that can be provided to LARC patients. The capability to identify patients who have poor response at an early time is important to change their treatment regimen; and on the other hand, predicting patients who are likely to achieve pCR or close to pCR is important to spare them from morbidities associated with TME surgery.

Chapter 8. Prognostic Prediction for Brain Tumors

In this chapter, 2 projects were presented about brain tumor using brain MRI. The first project is to predict of progression and recurrence in skull based meningioma. In this study, we established a system implementing radiomics to predict P/R in skull based meningioma. Random forest algorithm was applied to evaluate the importance of the extracted features. Another project is to predict of recurrence in nonfunctioning pituitary macroadenomas using brain MRI. In this study, we established a predictive model implementing radiomics to predict P/R in nonfunctioning pituitary macroadenomas Three tumor ROIs including original mask and mask with binary erosions were used. The SVM classifier was applied to evaluate the importance of the extracted features.

8.1 Radiomics Approach for Prediction of Progression and Recurrence in Skull Base Meningioma

8.1.1 Motivation and Application

Meningiomas are the most common primary brain tumors, and 20-30% of them grow in the skull base [342, 343]. Although most meningiomas are classified as benign tumors according to the 2016 WHO classification system [344], a subset may show early progression/recurrence (P/R) after surgical resection [345-347]. Because of the complex neurovascular structures there, complete surgical resection of the skull base meningiomas (SBM) is often difficult to achieve safely [342]. In order to avoid neurological complications of surgery, subtotal tumor resection (STR) or conservative follow up is often used as alternative treatment option [348-351]. When patients with complete resection show

recurrence, or patients with subtotal resection show progression, they are considered to have treatment failure, thus poor prognosis.

In clinical practice, it is important to identify risk factors that correlate with P/R in SBM, so appropriate treatment and follow-up strategies can be chosen for each individual patient. Some MR imaging findings such as tumor size, bone invasion, and proximity to major sinuses are related to P/R in meningiomas [347, 352]; however, quantitative analysis of MRI features for evaluation of clinical outcomes in meningiomas is rarely reported [353]. In recent years, radiomics analysis is emerging as a comprehensive quantitative method to evaluate brain tumors [354], which can extract parameters related to the underlying anatomical microstructure and dynamics of smaller-scale biophysical processes such as gene expression, tumor cell proliferation, and neovascularization [355]. Further, radiomics analysis has been shown capable of providing predictive markers for diagnosis, prognosis, and therapeutic planning in brain tumors [354, 356-359].

In a previous publication[353] we analyzed the preoperative CT and MR imaging features for the prediction of P/R in 73 patients diagnosed with skull base meningiomas, with emphasis on quantitative ADC values. In that study, multiple ROIs were manually placed on the aggressive tumor areas, and an AUC of 0.91 to differentiate between P/R and non-P/R was achieved. It was noted that low ADC values ($< 0.83 \times 10^{-3}$ mm^2/s) and adjacent bone invasion are high-risk factors of P/R. Since subjective ROI placement might vary from operator to operator, in this study we investigated the role of quantitative radiomics analysis based on automatically segmented tumor for the prediction of P/R in SBM.

8.1.2 Subjects and Image Dataset

Patients

From October 2006 to December 2017, 138 patients were diagnosed with SBM (WHO grade I-III) by brain MRI and pathological confirmation. Patients with less than one-year postoperative MRI follow-up (N=34) were excluded. Patients with incomplete preoperative MRI, poor imaging quality, or without preoperative diffusion-weighted imaging (DWI) and apparent diffusion coefficient (ADC) were excluded (N=29). Further, patients with inconsistent imaging sequences compared to the majority of the patients were also excluded (N=15). Finally, 60 patients (14 men, 46 women, median age, 57 years) were included, including 56 benign (WHO grade I), 3 atypical (WHO grade II), and 1 malignant (WHO grade III) SBM. Among the 60 patients, 54 patients were from the dataset of the previous publication [353]. 21 (21/60, 35%) patients had P/R, and the median time to P/R was 27 months (range 2-56 months). The median follow-up time was 52 months (range 12-122 months). Simpson Grade I-III resection (considered gross-total resection, GTR) was performed in 33 patients, and Simpson Grade IV resection (considered subtotal tumor resection, STR) was done in 27 patients.

Postoperative adjuvant RT was usually performed for patients with STR and patients with clinical high-risk features in our hospital. A total of 24 patients (21 benign, 2 atypical, and 1 malignant SBM) received adjuvant RT, including 18 STR and 6 GTR. Of the 6 GTR, 3 were WHO grade I and 3 were WHO grade II or III. The RT was done by using stereotactic radiosurgery (SRS) (N = 15, median dose of 25 Gy, ranging from 18 to 30 Gy; median fraction of 5, ranging from 3 to 5 fractions), or fractionated stereotactic intensity-

modulated radiotherapy (IMRT) (N = 9, dose ranging from 55 to 60 Gy with 30-33 fractions) by linear accelerators.

Determination of progression/recurrence (P/R)

P/R of SBM was evaluated by two experienced neuroradiologists (C.C.K. and T.Y.C.), blinded to the clinical and radiologic findings of the studied patients. In equivocal cases, judgment was made in consensus. Inter-observer reliability with Cohen k value of 0.9 was obtained. P/R is defined as recurrence of tumor in gross-total resection (GTR cases) (Simpson Grade I-III resection), or progression of residual tumor size in STR (Simpson Grade IV resection) on contrast-enhanced T1WI. In STR cases, the threshold of P/R is defined as a 10% increase in tumor volume by comparison with post-operative brain MRI. In patients who received adjuvant RT, P/R was differentiated from post-radiation effect (pseudo-progression) based on progressive tumor growth, not transient increase in tumor volume.

Imaging Acquisition and Tumor Segmentation

The MRI images were acquired using a 1.5T or a 3.0T scanner. The protocol included axial and sagittal spin echo T1-weighted imaging (T1WI), axial and coronal fast spin-echo T2-weighted imaging (T2WI), axial fluid attenuated inversion recovery (FLAIR), axial T2*-weighted gradient-recalled echo (GRE), axial DWI and ADC map, and CE T1WI in axial and coronal sections. **Figure 8-1** shows the flowchart of the analysis process. The lesion was segmented on contrast enhancement maps, by subtracting pre-contrast images from the post-contrast image. For each lesion, the operator placed an initial region of interest (ROI)

indicating the lesion location, and also decided the beginning and ending slices that contained the lesion. Then the outline of the lesion ROI on each imaging slice was automatically obtained using the fuzzy c-means (FCM) clustering-based algorithm [42]. The ROIs from all imaging slices containing this lesion were combined to obtain 3D mask of the whole lesion. Then 3D connected-component labeling was applied to remove scattered voxels not connecting to the main lesion ROI, and hole-filling was applied to include all voxels contained within the main ROI which were labeled as non-lesion. When needed, the operator performed manual correction, and the number of pixels that were changed was recorded. The percentage of corrected pixels was calculated by dividing to the total pixel number of the entire tumor. 28 of 60 cases needed to be corrected, and the corrected pixels were fewer than 5% (mean 3.2 ±2.1 %).

The segmented tumor mask was co-registered to T2W images and ADC maps to transfer the tumor ROI to these images. This process was done by FMRIB's Linear Image *Registration* Tool (FLIRT) [360]. This tool reads the header information of the images which contains the slice locations and the Field of View from T2W images, ADC maps and T1W images. Due to different image resolutions and thickness, the pixels in the tumor masks were mapped to T2W images and ADC maps using affine transformation and linear interpolation.

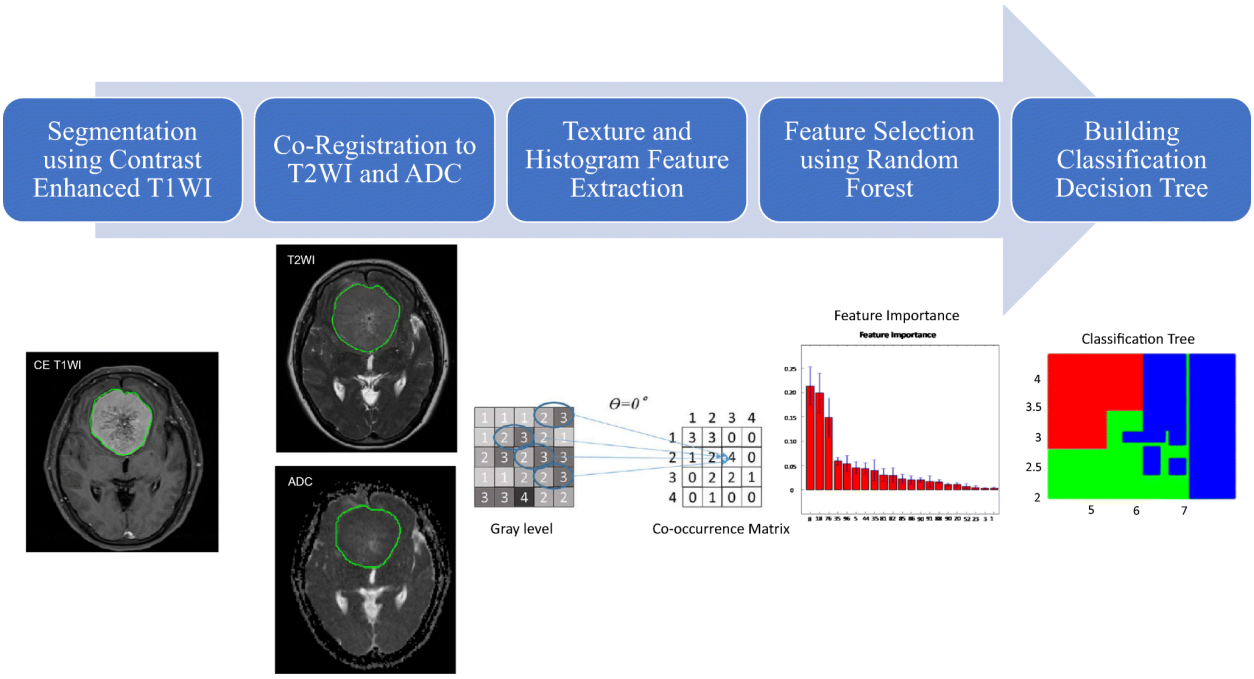

Figure 8-1: Flowchart of the analysis process**.** The tumor is segmented on contrast-enhanced T1WI, and then mapped to T2WI and ADC maps. On each set of images, a total of 33 texture and histogram features are extracted. The random forest algorithm is used to select features for building the classification model by using the decision tree.

8.1.3 Radiomics Analysis

Quantitative Feature Extraction

On each set of the contrast-enhanced T1W images, T2W images and ADC maps, 20 Gray Level Co-occurrence Matrix (GLCM) texture features [25, 26] were calculated from the tumor ROI, including autocorrelation, cluster prominence, cluster shade, contrast, correlation, dissimilarity, energy, entropy, homogeneity 1, homogeneity 2, maximum probability, sum average, sum variance, sum entropy, difference variance, difference entropy, information measure of correlation 1, and information measure of correlation 2, inverse difference normalized and inverse difference moment normalized. In addition, 13 histogram-based parameters were calculated, including 10%, 20%... to 90% percentile

values, mean, standard deviation, kurtosis, and skewness. Thus, a total of 99 parameters were extracted from the three sets of images acquired using different pulse sequences.

Feature Selection and Classification

Random forest algorithms were utilized via Bootstrap-aggregated decision trees to evaluate the importance of these features in differentiating patients in P/R and non-P/R groups [40]. A measure of the feature significance can be assessed as the loss of accuracy after this feature was removed [361]. All features were sorted based on their importance, and then different number of features starting from the top 1, 2, 3… was used to test their classification performance with 10-fold cross-validation. Finally, three features, including T1 Max Probability, T1 Cluster Shade, ADC Correlation, were selected. A decision tree with 5 leaves was used to build the final classification model. This procedure was implemented in Matlab 2018b.

Statistical Analysis

Statistical analyses were performed using statistical package SPSS (V.24.0, IBM, Chicago, Illinois, USA). Mann-Whitney U test was used to compare the obtained 3 parameters for differentiation of P/R. Chi-square or Fisher exact test was used to compare the categorical data. $P < 0.05$ was considered statistically significant.

8.1.4 Results

Clinical Data

The clinical data of SBM with and without P/R were summarized in **Table 8.1**. Twenty-one (21/60, 35%) patients had P/R. Although a higher rate of P/R was observed in patients with STR, there was no statistically significant relationship between the extent of resection and P/R (P=0.17) (**Figure 8-2** and **Figure 8-3**). In 24 patients receiving adjuvant RT, 6 (6/24, 25%) patients still had P/R in the subsequent follow-up. No statistical significance existed between adjuvant RT and P/R (P=0.19). Spheno-orbital region was the most common location amongst SBM with P/R ($P < 0.05$).

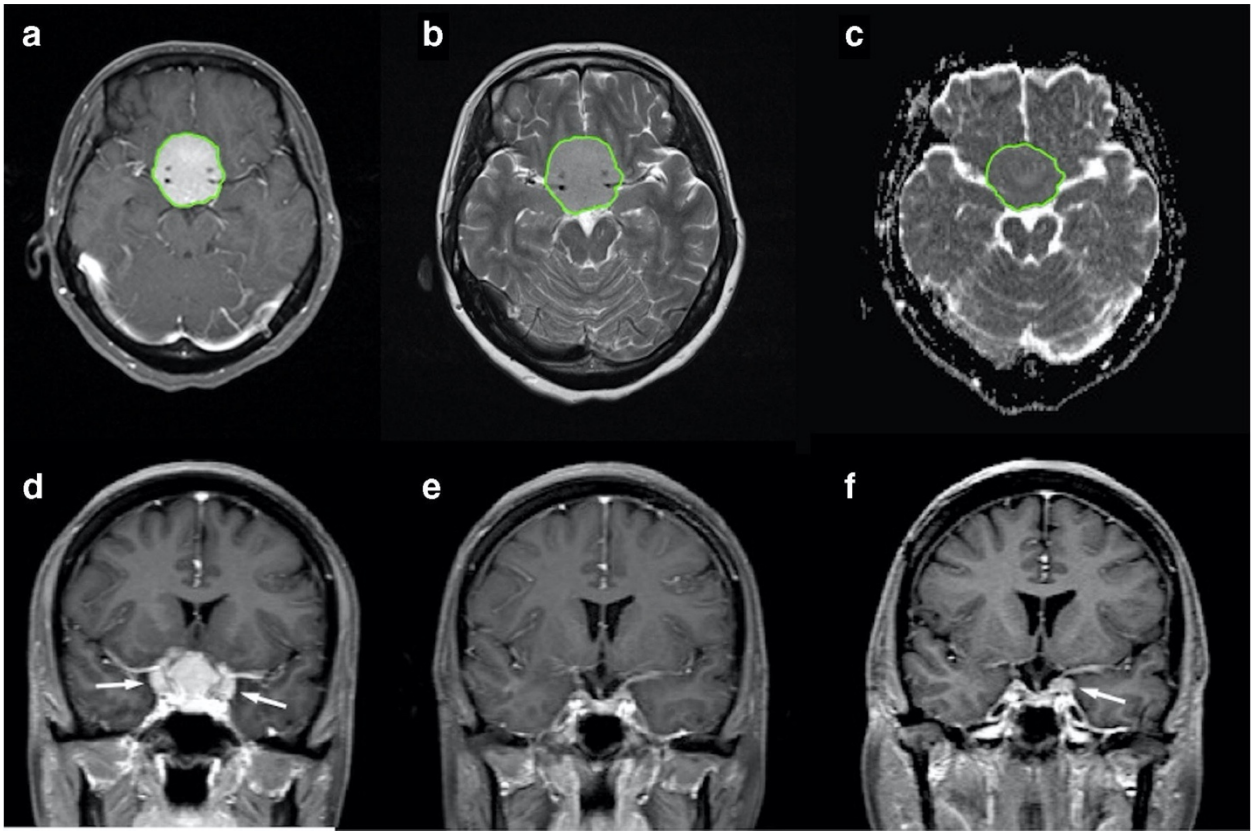

Figure 8-2: A 44-year-old woman with pathologically proven sellar meningioma (WHO grade I). a) Axial contrast-enhanced T1WI showing an enhancing tumor (green outline) involving the sellar/suprasellar region. The tumor (green outline) is segmented on contrast-enhanced T1WI, and then mapped to b) axial T2WI and c) axial ADC maps; d) coronal contrast-enhanced T1WI showing the sellar/suprasellar enhancing tumor (arrows) with bilateral encasement of the proximal internal carotid arteries, middle cerebral arteries, and anterior cerebral arteries; e) gross-total resection was performed, and WHO grade I meningioma was confirmed pathologically; f) recurrent tumor at the left clinoid process (arrow) was observed 36 months after surgical resection.

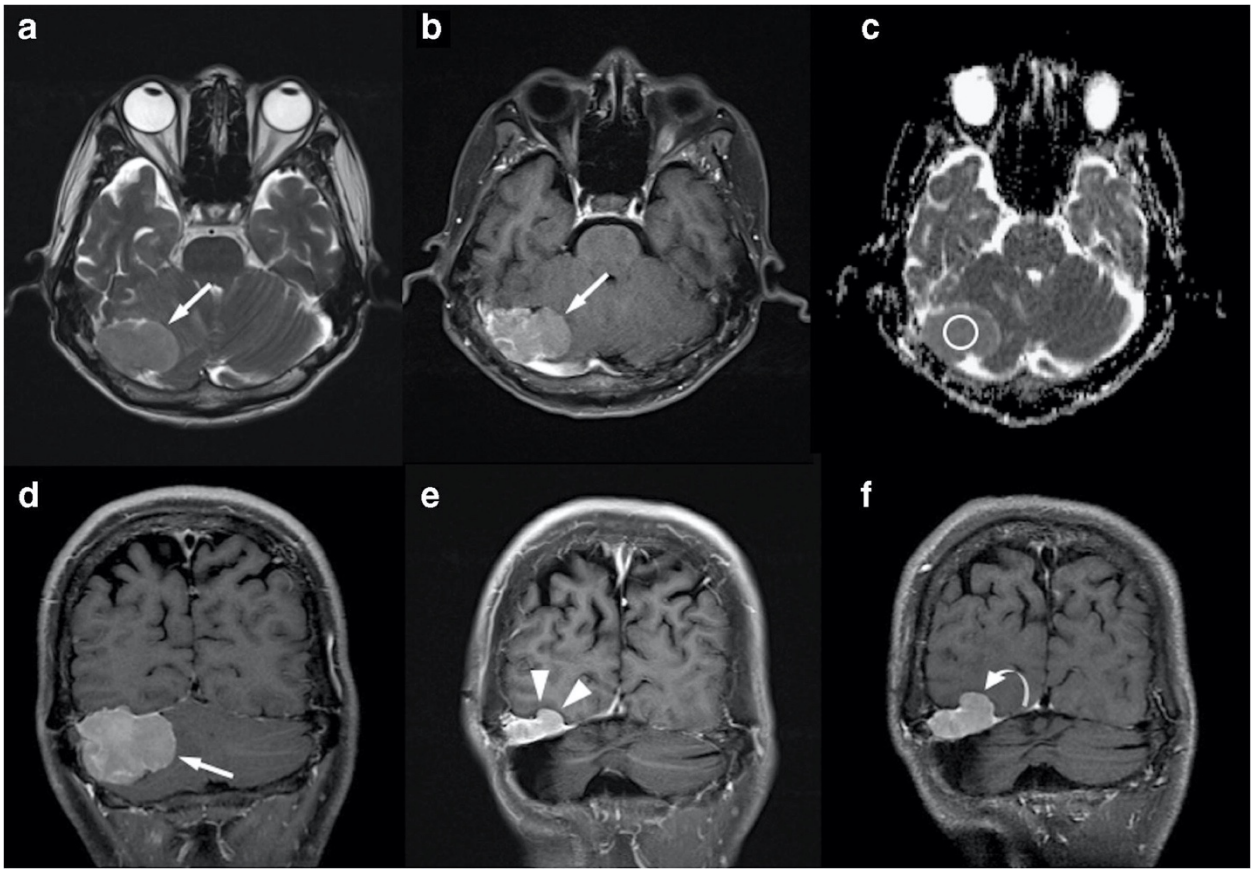

Figure 8-3: A 46-year-old man with pathologically proven right posterior fossa meningioma (WHO grade I). a) Axial T2WI and b) axial contrast-enhanced T1WI showing an enhancing tumor (arrow) in the right posterior fossa with involvement of the right transverse sinus; c) measured ADC value (circular ROI) was 0.823×10^{-3} mm^2/s (b = 1000 s/mm^2); d) coronal contrast-enhanced T1WI showing the enhancing tumor (arrow) arising from the right tentorium with downward extension; e) subtotal resection was performed to preserve the right transverse sinus, with residual tumor (arrowheads) in the right tentorium, and WHO grade I meningioma was confirmed pathologically; f) progression of the residual tumor (curved arrow) was observed 14 months after surgical resection.

Table 8.1: The clinical data of SBM with and without progression/ recurrence (P/R)

	Progression/Recurrence (P/R)	Non-P/R	*p* value
Number	*N* = 21	*N* = 39	
Sex			0.21
Male	7 (33.3%)	7 (17.9%)	
Female	14 (66.7%)	32 (82.1%)	
Age (years)	55 (43.5, 66.5)	58 (50.5, 65.5)	0.42
WHO grade			0.39
Grade I	19 (90.5%)	37 (94.9%)	
Grade II	1 (4.8%)	2 (5.1%)	
Grade III	1 (4.8%)	0	
Histological subtype			0.86
Meningothelial (syncytial)	19 (90.5%)	33 (84.6%)	
Transitional (mixed)	1 (4.8%)	3 (7.7%)	
Fibroblastic (fibrous)	1 (4.8%)	2 (5.1%)	
Psammomatous	0	1 (2.6%)	
Simpson Grade resection			0.17
Grades I, II, and III (gross-total resection, GTR)	9 (42.9%)	24 (61.5%)	
Grades IV and V (subtotal resection, STR)	12 (57.1%)	15 (38.5%)	
Postoperative adjuvant RT			0.19
Yes	6 (28.6%)	18 (46.2%)	
No	15 (71.4%)	21 (53.8%)	
Location			0.03*
Anterior fossa or olfactory groove	1 (4.8%)	13 (33.3%)	
Spheno-orbital	7 (33.3%)	6 (15.4%)	
Temporal floor	5 (23.8%)	5 (12.8%)	
Sellar/cavernous sinus	3 (14.3%)	1 (2.6%)	
Posterior fossa	5 (23.8%)	14 (35.9%)	

Continuous variables were presented as median and interquartile range (IQR)

*Statistical difference ($p < 0.05$)

Radiomics Model to Differentiate between P/R and non-P/R

The importance of all analyzed radiomics features was estimated using the random forest method, and 3 features, including T1 maximum probability, T1 cluster shade and

ADC correlation, were selected to differentiate between P/R and non-P/R groups. The performance could not be improved by adding more features, so these 3 parameters were chosen as the final model. Statistical significance was observed in T1 maximum probability (P=0.004) and T1 cluster shade (P=0.043) between the P/R and non-P/R groups (**Figure 8-4**). The final classification results were generated by using the decision tree (**Figure 8-5**). By using the selected thresholds, there were 18 true positive cases, 36 true negative cases, 3 false positive cases, and 3 false negative cases, with the overall prediction accuracy of 90%. For comparison, the overall accuracy for differentiation of P/R by the mean ADC value obtained from manually placed ROI was 83% (10 false prediction cases).

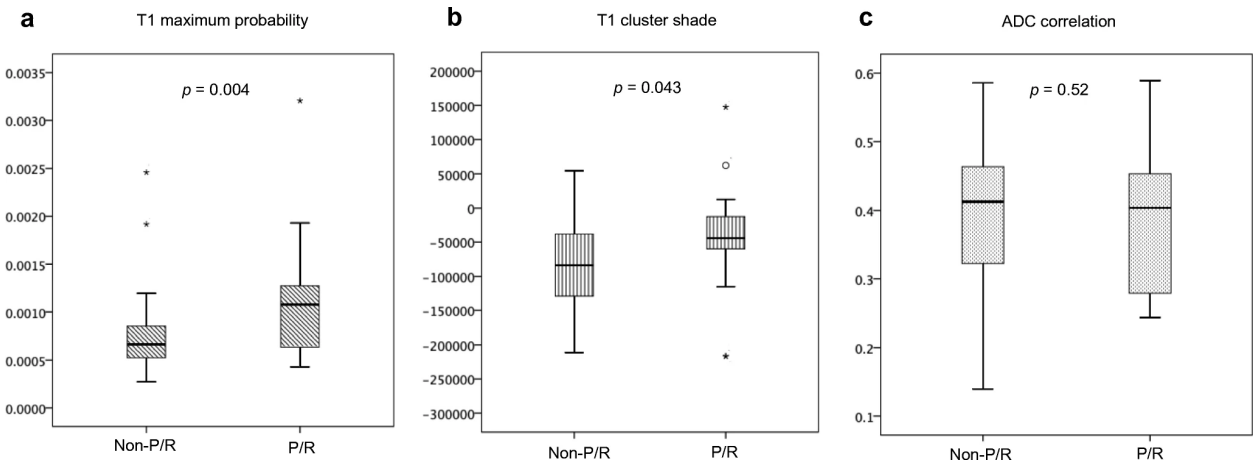

Figure 8-4: Box plot of **a** T1 maximum probability, **b** T1 cluster shade, and **c** ADC correlation in skull base meningiomas with and without progression/recurrence (P/R). Statistical difference ($p < 0.05$) (Mann-Whitney *U* test) in T1 maximum probability and T1 cluster shade was observed. Boxes indicate the interquartile range, and whiskers indicate the range. The horizontal line represents the median in each box. Circles represent outliers, defined as distances greater than 1.5 times the interquartile range above the third quartile. The star represents an extreme value, defined as a distance greater than three times the interquartile range below the first quartile or above the third quartile.

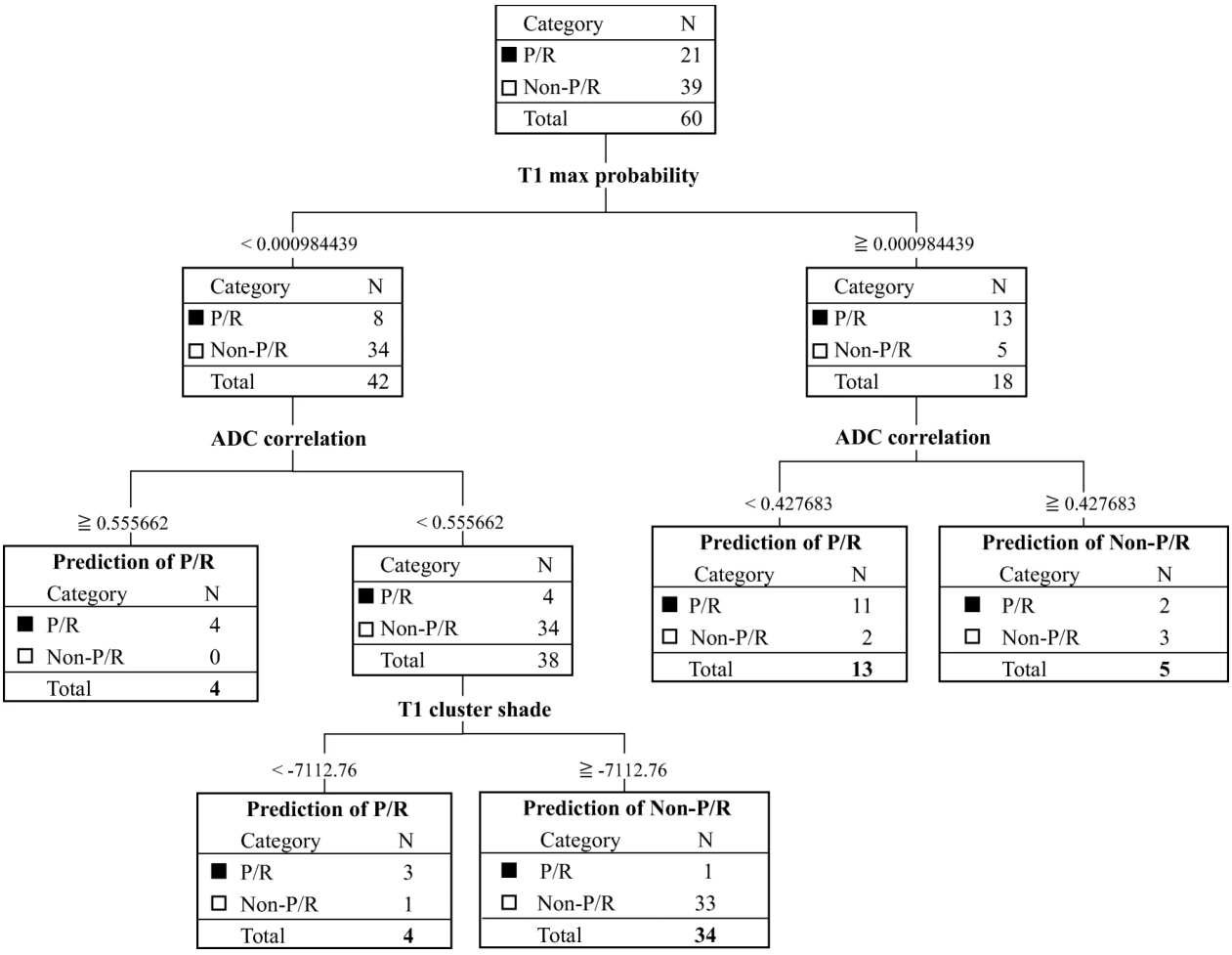

Figure 8-5: The diagnostic decision tree with five leaves to separate patients into P/R and non-P/R groups. The total number of splits is four.

8.1.5 Summary and Discussion

In this study, we established a scheme using radiomics to predict P/R in SBM. Random forest algorithm was applied to evaluate the importance of the extracted features. In the three selected features, two were extracted from contrast-enhanced T1 weighted images and one from ADC maps. The overall accuracy to differentiate between P/R and non-P/R was 90%, with 6 false prediction cases. No histogram parameters were selected in the final model, suggesting that texture provides more important prognostic information.

Although 90% meningiomas are benign (WHO grade I) tumors, about 21% of these tumors recur in 5 years after surgical resection [345, 346]. The risk factors related to progression of SBM were investigated in several studies, and varied recurrence rates from 13.2% to 56% were reported [342, 362-364]. In our study, the relatively high rate of PR (21/60, 35%) may also be caused by small sample size and selection bias. It is known the genetic and pathologic mechanisms between the SBM and non-skull base meningiomas (non-SBM) are different [362, 365, 366]. Further, the recurrence rate and clinical outcomes are also inconsistent between these two entities [342, 367]. Mansouri et al. [342] reported higher recurrence rate in non-SBM. In contrast, Savardekar et al. [367] reported that SBM progressed at a higher rate than non-SBM during the first 10 years' follow-up after surgery. The higher recurrent rate in SBM may be caused by incomplete tumor resection and bone invasion. Since complete surgical resection may result in neurologic complications, prediction of recurrence in SBM is a clinically significant issue for choosing the optimal treatment strategy for each individual patient.

Although conventional MR imaging findings related to recurrence in meningiomas have been reported, most imaging data are reported in qualitative and subjective terms [347, 368, 369]. In contrast, MR radiomics can reproducibly extract objective and quantitative data from different imaging sequences to build diagnostic models to classify different types of lesions [354, 370, 371]. Several authors had reported the application of MR radiomics to provide valuable information for differential diagnosis, tumor staging, prediction of prognosis, and assessment of cancer genetics [356-358]. Recently, MR radiomics and machine learning analyses had been used to differentiate meningioma grade [372, 373]. However, the application of radiomics for predicting clinical outcomes in

meningiomas was rarely reported. Herein, we performed MR radiomic analysis using pre-operative MRI of SBM to predict P/R.

In this study, we used random forest to do feature selection and then used a binary decision tree to build the final classification model. Random forest combines multiple decision trees, with each tree stratifying the feature space into a number of simple non-overlapping regions that can maximize classification accuracy. Compared with other feature selection algorithms, such as LASSO and artificial neural network [10], random forest improves the generalization of the selection process and works better for small dataset. In this study, three features were selected from 99 features. Considering small number of features and cases, a binary decision tree can be constructed and the results can be easily interpreted. Other classification algorithm, such as support vector machine or convolutional neural network, may achieve a very high accuracy, but it requires a huge dataset. Also, these algorithms are considered as 'black-box' classifier and it is difficult to interpret the obtained results [13].

In a previous study we have shown that the mean ADC value measured from manually placed ROI on the aggressive tumor area could be used to predict P/R for SBM [374]. The texture and heterogeneity within the tumor could not be considered using this manual ROI analysis, and missed valuable information that could be extracted. The accuracy for prediction of P/R by using ADC from manual ROI was 83% (10/60 false prediction), which was inferior to the accuracy of 90% (6/60 false prediction) by using the radiomics model.

There were a total of 6 false prediction cases. For the 3 false positive cases, all of them were in right sphenoid ridge, two GTR and one STR. None of them received adjuvant RT. Two had large tumor size (maximal diameter 6.8 cm and 5.6 cm), and showing

heterogeneous contrast enhancement and ADC. For the 3 false negative cases, they were all STR and two occurred in temporal fossa. One patient had adjuvant RT. Relatively homogeneous contrast enhancement and consistent low ADC were seen in all 3 false negative cases. Further investigation in a larger sample size is needed to better understand the reasons leading to false positive and false negative predictions.

Mathiesen et al. [375] reported the recurrence rates of SBM were 3.5-25% in Simpson Grade I-III resection, and 45% in Simpson Grade IV resection. Although it is generally agreed that the extent of surgical resection is an important determining factor in the rate of recurrence [342], recently Voß et al. [376] reported a similar recurrence rate between GTR and STR in 325 SBM [376]. The use of adjuvant radiotherapy may help to decrease the risk of progression in STR patients. In the present study, we also found a lower P/R rate in patients receiving RT. For STR patients, the progression rate was 6/18 (33.3%) in patients with RT, which was lower compared to 6/9 (66.7%) in patients without RT. For GTR patients, only 6 patients received RT, and their recurrence rate was 0/6 (0%). In 27 GTR patients without RT, the recurrence rate was 9/27 (33.3%). Adjuvant RT is known to improve overall survival in high-grade meningiomas, but its role in benign (WHO grade I) meningiomas is still unclear [377]. RT will increase risks of complications such as symptomatic peritumoral edema, cranial nerve deficits, internal carotid artery stenosis, and neurologic deficits, and thus whether it should be given post-surgically without evidence of recurrence is controversial [378, 379]. With advanced imaging analysis, if the risk of P/R can be predicted accurately based on pre-operative imaging, it will help to give RT only to patients who have a high risk of progression. The findings can also be applied to surgical planning. Aggressive tumor resection and close imaging follow-up should be

considered in patients with high likelihood of recurrence. In contrast, for patients with lower possibility of recurrence, the aim of surgery would be relief of mass effect and clinical symptoms, and adjuvant RT may be spared to avoid long-term side effects [351].

Our study has limitations. The retrospective nature of the study may result in bias. All images were acquired at a single site, mostly with a single protocol. Future testing on multi-institutional data and on varying imaging protocols is needed to determine whether the trained classifier is generalizable. The implemented radiomics analysis method is straightforward, but since it is based on pre-defined features, may not fully utilize the information from all images. Due to the small number of cases, only a few features can be selected into the classification model to avoid over-fitting. More cases are expected to improve the model performance. Lastly, as the adjuvant RT and bone invasion may affect the P/R status, they may influence the predictive value of the extracted features, but not considered in the radiomics analysis process. More advanced statistical analysis methods that can take all confounding factors into account may be developed in the future.

In conclusion, this was the first study attempting to apply the MR radiomic analysis to predict P/R in SBM. The results showed that T1 max probability, T1 cluster shade, and ADC correlation were the most important features, with a prediction accuracy of 90%. The results were better compared with our previous analysis approach using ADC measured by operator-defined ROIs. As radiomics can thoroughly evaluate many aspects within the entire tumor, it has a potential to be applied to choose the optimal treatment strategy for each SBM patient, including the choice of surgical types and the use of adjuvant radiotherapy. This will need to be studied when more cases with a long-term follow-up are available.

8.2 Radiomics Approach for Prediction of Recurrence in Nonfunctioning Pituitary Macroadenomas

8.2.1 Motivation and Purpose

Pituitary adenomas constitute 10%-15% of all intracranial tumors [380]. The nonfunctioning pituitary adenoma is the most common type of pituitary adenomas [381, 382]. This nonfunctioning pituitary adenoma often presents as a macroadenoma, defined as tumor size larger than 10 mm. The nonfunctioning pituitary macroadenomas (NFPAs) may cause bitemporal hemianopia due to mass effect by compression of the optic chiasm. Some patients may suffer hypopituitarism due to compression of the normal pituitary gland. According to 2017 WHO classification system, the pituitary tumors are formally classified as adenoma, carcinoma, or blastoma [383]. Although more than 90% of NFPAs are diagnosed as benign tumors, 25-55% of these tumors may show early progression/recurrence (P/R) after surgical resection [384-389]. Gross-total resection (GTR) by a transsphenoidal approach (TSA) is the optimal treatment for NFPAs in clinical practice; however, this aim is often difficult to achieve for the tumors without apoplexy or cystic change [390]. Although adjuvant radiotherapy (RT) is implemented in some institutions to prevent postoperative P/R in NFPAs, this approach may cause progressive pituitary insufficiency and other long-term complications [391]. Conventional MR imaging findings such as cavernous sinus invasion, tumor size, and absence of tumor apoplexy had been reported as the important parameters related to P/R in NFPAs; however, subjective variation may exist in interpretation between each readers [392, 393]. Recently, radiomics analysis is emerging as a comprehensive quantitative method to evaluate brain gliomas, colorectal cancer, and non-small-cell lung cancer [354, 394, 395].

Radiomics is a well-established quantitative approach for image pattern recognition and works by extracting objective information through analysis of the intensity or spatial distribution of intensity variations in images [20, 21, 25]. It extracts a large number of quantitative imaging features from a medical image and then analyses these features by a series of machine learning algorithms [371, 396, 397]. The extracted imaging features are related to the underlying anatomical microstructure and smaller-scale biophysical processes such as genetic expression, tumor proliferation, and neovascularization [355].

Several studies suggest that radiomics is able to provide predictors for diagnosis, prognosis, and therapeutic planning in brain tumors [354, 355, 359, 398-401]. Although radiomics analysis for evaluation of tumor subtypes, consistency, ki-67 proliferation index, and cavernous sinus invasion in NFPAs had been recently reported [359, 402-404], the prediction of clinical outcomes in NFPAs by radiomics approach is rarely reported. In this study, we investigated the role of quantitative radiomics analysis based on automatically segmented tumor for the prediction of P/R in NFPAs.

8.2.2 Subjects and Image Dataset

Patient Selection

From September 2010 to December 2017, 222 patients were diagnosed with benign pituitary macroadenomas (diameter > 10 mm) by pathological confirmation and received preoperative brain MRI studies. Patients with less than 1-year post-operative MRI follow-up were excluded (n = 64), in accordance with previously reported studies [392, 393, 405, 406]. Patients with clinical, biochemical, and histopathological evidence of hormone hypersecretion were also excluded. Therefore, patients diagnosed with prolactinoma (n =

7), acromegaly (n = 6), Cushing's disease (n = 1), thyroid-stimulating hormone (TSH)-secreting pituitary adenoma (n = 1) were excluded. According to studies by Brochier et al. [392] and Hong et al. [407], diagnosis of prolactinoma is considered unlikely if the prolactin level is below 100 mg/L, and this diagnosis was thereafter confirmed by immunocytochemical studies. Patients with incomplete protocol or poor imaging quality on pretreatment MR imaging determined by experienced neuroradiologists (C.C.K. and T.Y.C.) were excluded (n = 41). Patients who received postoperative adjuvant radiotherapy (RT) (n = 52) before P/R were excluded. Finally, 50 patients (29 men, 21 women, age 19 - 80 years; median age, 52 years) diagnosed with benign NFPAs were included in this study. None had previous intracranial radiotherapy. Forty-eight patients received surgery performed by TSA, and 2 patient received craniotomy due to large tumor size. The median follow-up time of all patients was 38 months (range 12 - 115 months). In 28 patients with P/R, the median time to P/R was 20 months (range 6 - 67 months). The clinical and biochemical data were obtained from admission notes.

Extent of Resection and Progression/Recurrence

The extent of surgical resection was determined by review of operation notes and postoperative MRI by a neuroradiologist (C.C.K.) and a neurosurgeon (S.W.L.). According to published literature [408], GTR is defined as when the percentage of residual tumor volume is less than 10% of its original size; in contrast, subtotal resection (STR) of tumor is defined as when the percentage of residual tumor volume is more than 10% of its original volume. For determining P/R in NFPAs, pretreatment and postoperative MR images were also evaluated by two experienced neuroradiologists (C.C.K, a neuroradiologist with 6

years of experience, and T.Y.C., a neuroradiologist with 18 years of experience), both of whom were blinded to the clinical and imaging outcomes of the studied population. P/R was defined as recurrence of tumor after GTR or enlargement of residual tumor after STR on postoperative coronal and sagittal contrast-enhanced T1WI. The threshold of P/R was defined as a more than 2-mm increase in at least one dimension in comparison with postoperative MRI studies according to published literatures [392, 405]. Interobserver reliability with Cohen k value of 0.9 in determining P/R was obtained. In equivocal cases, judgment was made in consensus. In preoperative MR imaging, cavernous sinus invasion (Knosp classification) [409] and extrasellar extension (Hardy's classification) [410] were determined on coronal T2WI and CE T1WI. Maximum tumor height was measured on coronal CE T1WI. Successful chiasmatic decompression was determined by evidence of the relief of mass effect on the optic chiasm on the postoperative MRI and clinical improvement of associated visual deficit.

MR Imaging Acquisition

Brain MRI images were acquired using a 1.5-T (Siemens, MAGNETOM Avanto) (n = 19), 1.5-T (GE Healthcare, Signa HDxt) (n = 17), or a 3-T (GE Healthcare, Discovery MR750) (n = 14) MR scanner, equipped with 8-channel head coils in each machine. Scanning protocols include axial and sagittal spin echo T1-weighted imaging (T1WI), axial and coronal fast spin echo T2- weighted imaging (T2WI), axial fluid attenuated inversion recovery (FLAIR), axial T2*- weighted gradient-recalled echo (GRE), and axial diffusion-weighted imaging (DWI). Dynamic contrast-enhanced (CE) coronal T1WI images with a small field of view through the pituitary gland, as well as coronal and sagittal CE T1WI with fat saturation,

were performed after intravenous administration of 0.1 mmol/kg of body weight of gadobutrol or gadoterate meglumine.

8.2.3 Tumor Segmentation and Radiomics Analysis

Tumor Segmentation

Because radiomics in T2WI and CE T1WI were associated with cavernous sinus invasion, histopathologic subtypes, consistency, and therapeutic response in pituitary tumors [359, 401, 403, 411-413], the two sequences (slice thickness/spacing, 3 mm/3 mm) were selected for analysis in our study. Figure 1 showed the flowchart of the analysis process. Since NFPAs in general enhance very well, tumor segmentation was performed from the coronal CE T1WI by a volunteer physician who knows the anatomy well. For each lesion, the operator placed an initial rectangle region of interest (ROI) on coronal CE T1WI which can locate the lesion roughly, and also decided the beginning and ending slices that contained the lesion. Then the fuzzy c-mean (FCM) clustering based algorithm was developed to calculate the outline of the lesion ROI on each imaging slice [10]. An experienced radiologist (J.H.C) familiar with brain MRI checked the accuracy of tumor segmentation slice by slice. In cases of under- or over-segmentation, manual correction by inclusion of more tumor tissue or exclusion of unnecessary normal tissue was adopted. After segmentation/correction, the ROIs from all imaging slices containing this lesion were combined to obtain 3D information of the whole lesion. Then 3D connected-component labeling was applied to remove scattered voxels not connecting to the main lesion, and hole-filling algorithm was applied to include all voxels contained within the main ROI which were labeled as non-lesion. The segmented tumor mask was co-registered to coronal

T2WI to localize the tumor location on corresponding images using affine transformation and linear interpolation. This process was done by FLIRT [360], which could read the header information of the images that contained the slice locations and the field of view from CE T1WI and T2WI.

Texture Feature Extraction and Selection

Within segmented tumor on coronal enhanced T1W images and T2W images , 107 imaging features, including 32 first order features and 75 textural features were extracted on each modality (**Figure 8-6**). Considering some NFPAs are small in volumes or the tumors are often inseparable with the normal pituitary tissue, the boundary pixels of the tumor masks on each slice were removed by binary erosion for accurate results [414]. We used 2 lengths, 0.5 cm and 0.25 cm, to determine the outer shells of the boundary pixels to be removed. Then we have 3 types of tumor ROIs, including original masks, original mask with 0.25cm erosion, and original mask with 0.5cm erosion. From each of them, we totally obtained 214 descriptors from coronal CE T1WI and T2WI. To evaluate the importance of these features in differentiate patients with and without P/R, sequential feature selection process was utilized via constructing multiple support vector machine (SVM) classifiers. Sequential feature selection process is to measure the characteristics of data to select candidate features for classification using a reasonable criterion [283]. In this process, we used SVM with Gaussian kernel as the objective function to test the potential of a subset of the features [15, 16]. In this project, a subset of features was employed to train SVM models and to test the performance of models. At the start of the selection process, an empty candidate set was presented, and features were sequentially added to it until the addition

of further features does not decrease the criterion. 10-fold cross validation method was applied to test the model performance [415]. In each iteration, the training process was repeated 1000 times to explore the robustness of each feature. After each iteration, the feature which led to the best performance was added into the candidate set. Once the addition of features does not meet the criterion, the selection process stopped then a selected feature set containing the optimal features was obtained. Here, we used 10^{-6} as termination tolerance for the objective function value. Three features, including T1 surface-to-volume ratio, T1 GLCM Informational Measure of Correlation, and T2 NGTDM Coarseness with the highest importance were selected to build the final SVM classification model. This procedure was implemented in MATLAB 2019b.

Statistical Analysis

Statistical analyses were performed using statistical package SPSS for Windows (V.24.0, IBM, Chicago, IL, USA). For evaluation of the clinical parameters and conventional MR imaging, chi-square (or Fisher exact test) and Mann-Whitney U tests were performed for categorical and continuous data, respectively. The true positive (TP), true negative (TN), false positive (FP), false negative (FN), accuracy, and area under the receiver operating characteristic curve (ROC) curve (AUC) in prediction models of different tumor masks were calculated. P-value < 0.05 was considered statistically significant.

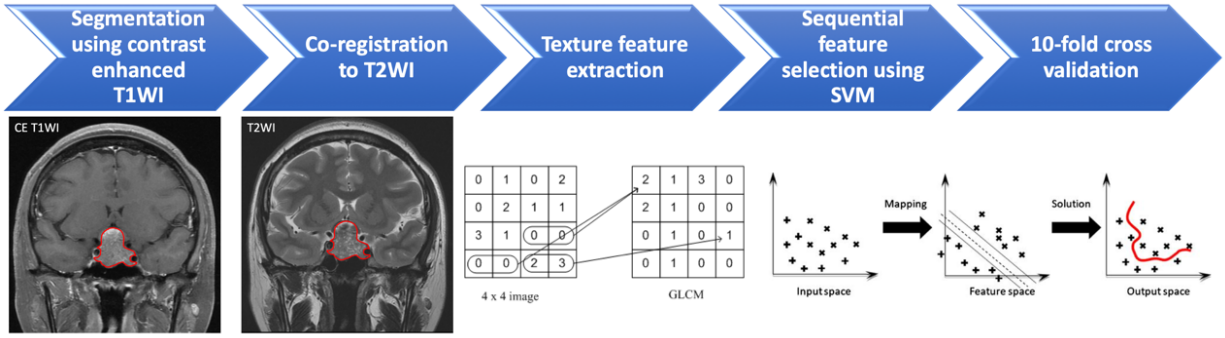

Figure 8-6: Flowchart of the analysis process [modified from reference 62]. The nonfunctioning pituitary macroadenoma (NFPA) (red outline) is segmented on coronal contrast-enhanced T1WI, and then mapped to coronal T2WI. On each set of images, a total of 107 imaging features including 32 first order features and 75 textural features were extracted. The most important three features were selected by sequential feature selection and support vector machine (SVM) classifiers to build the prediction model. 10-fold cross validation method was applied to test the model performance.

8.2.4 Results

Clinical Data and Conventional MRI Findings

The clinical data and conventional MRI findings of the included 50 NFPA patients are summarized in **Table 8.2**. Twenty-eight (28/50, 56%) patients are diagnosed with P/R. Although a higher rate of P/R was observed in patients receiving STR, no statistical significance was found between the extent of resection and P/R ($p = 0.157$). Visual disturbance, hypopituitarism, extrasellar extension (Hardy's classification grade 3 or 4), compression of the 3rd ventricle, larger tumor height and volume occurred more frequently in the P/R group ($p < 0.05$) (**Figure 8-7**). Besides, more successful chiasmatic decompression were observed in the non-P/R group ($p < 0.05$).

Table 8.2: The clinical data and conventional MR imaging of nonfunctioning pituitary macroadenomas (NFPAs) with and without progression/recurrence (P/R)

	P/R	Non-P/R	p value
Number of patients	28	22	
Sex			0.111
Male	19 (67.9%)	10 (45.5%)	
Female	9 (32.1%)	12 (54.5%)	
Age (y)	53.5 (44, 63)	42 (23.5, 60.5)	0.089
Clinical symptoms			
Visual disturbance	26 (92.9%)	13 (59.1%)	0.006*
Headache	8 (28.6%)	11 (50%)	0.121
Decreased libido, sexual dysfunction, and/or amenorrhea/oligomenorrhea	5 (17.9%)	1 (4.5%)	0.211
Incidental	2 (7.1%)	4 (18.2%)	0.385
Hypopituitarism			0.047*
No	12 (42.9%)	17 (77.3%)	
Single	8 (28.6%)	3 (13.6%)	
Multiple	8 (28.6%)	2 (9.1%)	
Hyperprolactinemia	10 (35.7%)	6 (27.3%)	0.525
Extent of surgical resection			0.157
Gross-total resection (GTR)	3 (10.7%)	6 (27.3%)	
Gross-total resection (STR)	25 (89.3%)	16 (72.7%)	
Successful chiasmatic decompression	9 (32.1%)	17 (77.3%)	0.002*
Cavernous sinus invasion (Knosp classification)			0.077
Grade 1-2	18 (64.3%)	19 (86.4%)	
Grade 3-4	10 (35.7%)	3 (13.6%)	
Extrasellar extension (Hardy's classification)			0.045*
Grade 1-2	17 (60.7%)	19 (86.4%)	
Grade 3-4	11 (39.3%)	3 (13.6%)	
Compression of optic chiasm	27 (96.4%)	17 (77.3%)	0.075

Compression of the 3rd ventricle	21 (75%)	9 (40.9%)	0.015*
Hydrocephalus	2 (7.1%)	1 (4.5%)	1
Giant (> 40 mm)	9 (32.1%)	2 (9.1%)	0.085
Maximum tumor height (mm)	35.5 (27.5, 43.5)	18 (10, 26)	< 0.001*
Tumor volume (cm^3)	12.3 (4.4, 20.1)	2.7 (1.2, 8)	< 0.001*

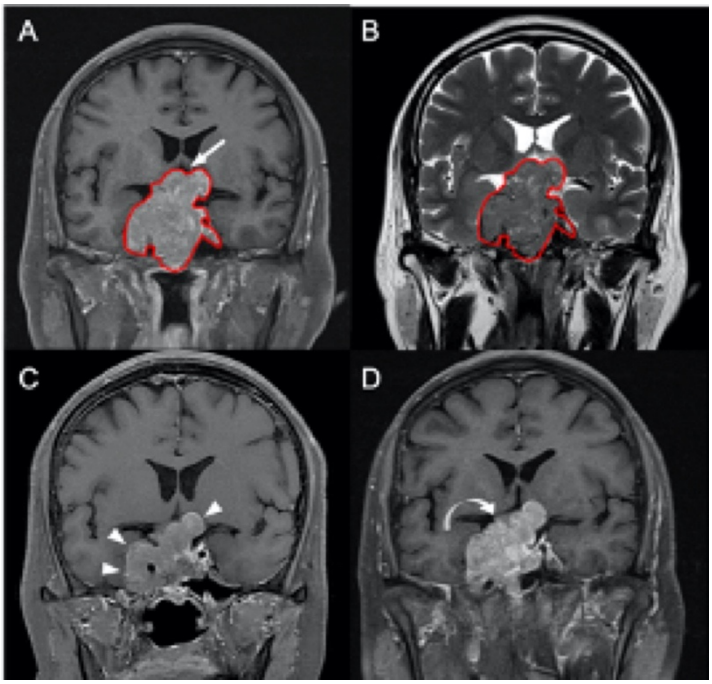

Figure 8-7: A 55-year-old male patient with left hemianopia and pathologically proven NFPA. Coronal contrast-enhanced (CE) T1WI shows an enhancing sellar tumor (red outline) with upward suprasellar extension and invasion into bilateral cavernous sinuses, causing compression of the optic chiasm and the third ventricle (arrow indicates area of optic chiasm and third ventricle). The tumor (red outline) is segmented on coronal CE T1WI (A), and then mapped to coronal T2WI (B). Improvement of blurred vision after subtotal tumor resection by transsphenoidal approach was clinically documented, and the maximum height of the residual tumor (arrowheads) measured from coronal CE T1WI is 38 mm (C). Recurrent visual deterioration with enlargement of the residual tumor (curved arrow) (maximum height up to 48 mm) occurred 19 months after surgical resection (D).

Radiomics Approach for Prediction of P/R

In radiomics analysis, the most significant three parameters selected by the final SVM model for prediction of P/R were T1 surface-to-volume ratio, T1 GLCM-informational measure of correlation, and T2 NGTDM-coarseness (**Figure 8-8**). Significantly statistical difference (Mann-Whitney U test) in T1 surface-to-volume ratio ($p < 0.001$), T1 GLCM-informational measure of correlation ($p = 0.037$), and T2 NGTDM-coarseness ($p = 0.001$) between the P/R and non-P/R groups were observed (**Figure 8-8**). The SVM classification results by original mask showed 16 TP cases, 25 TN cases, 3 FP cases, and 6 FN cases (**Figure 8-9**). The overall prediction accuracy is 82% and the AUC of the prediction model is 0.78. Similar accuracy with values of 80% and 82%, and AUC of 0.8 and 0.79 are observed in mask with 0.25cm and 0.5cm erosions respectively (**Table 8.3**).The detailed MR imaging features of the 9 false prediction cases in original mask are listed in **Table 8.4**.

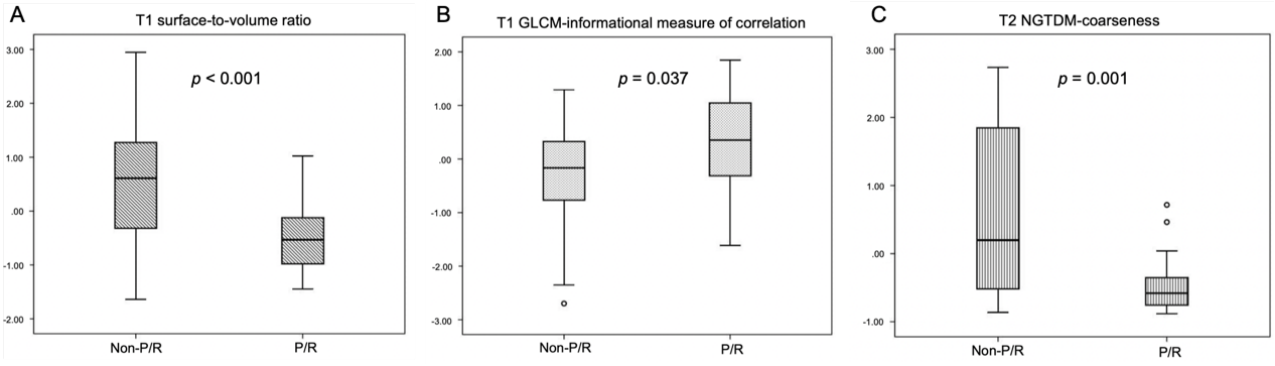

Figure 8-8: Box plot of T1 surface-to-volume ratio (A), T1 GLCM-informational measure of correlation (B), and T2 NGTDM-coarseness (C) in NFPAs with and without progression/recurrence (P/R). Significantly statistical difference ($p < 0.05$) (Mann-Whitney U test) in the three selected features was observed. Boxes indicate the interquartile range (IQR), and whiskers indicate the range. The horizontal line represents the median in each box. Circles represent outliers, defined as distances greater than 1.5 times the IQR above the third quartile.

Table 8.3: The accuracy and AUC in prediction models without and with binary erosions.

	True Positive (TP)	True Negative (TN)	False Positive (FP)	False Negative (FN)	Accuracy (%)	AUC
Original mask	16	25	3	6	82%	0.78
With 0.25 cm erosion	16	24	4	6	80%	0.80
With 0.5 cm erosion	17	24	4	5	82%	0.79

Table 8.4: The MR imaging features of the 3 false positive (FP) and 6 false negative (FN) NFPAs

False prediction	GTR	Cavernous sinus invasion (Knosp grade 3-4)	Extrasellar extension (Hardy's grade 3-4)	Apoplexy/ cystic change	Heterogeneous enhancement	Maximum tumor height (mm)	Tumor volume (cm^3)
FP	-	-	-	+	+	32	11.6
FP	-	+	-	+	+	33	10.2
FP	+	+	-	+	+	36	18.7
FN	-	-	-	+	+	62	43.5
FN	+	-	-	-	-	19	2.7
FN	-	-	-	-	-	12	1.2
FN	-	+	+	-	-	41	24.1
FN	-	+	-	-	-	22	8.2
FN	+	-	-	-	-	13	1.8

+ / -: Yes / No

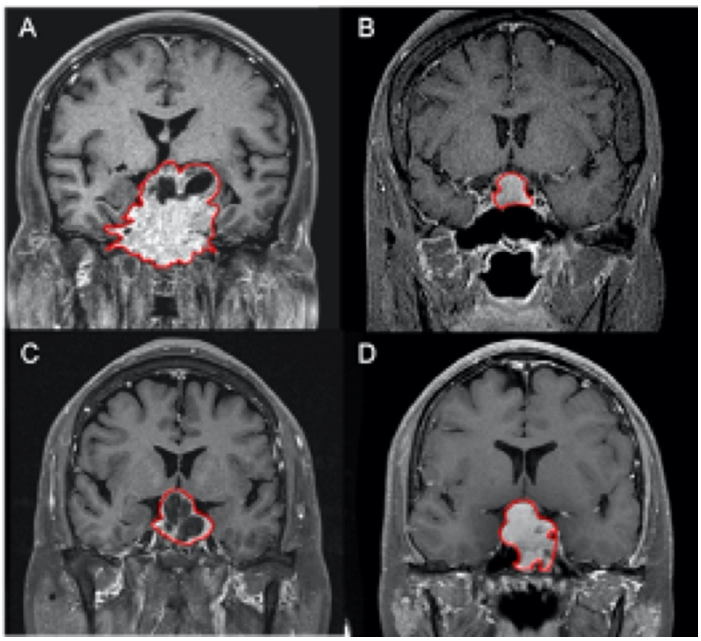

Figure 8-9: Examples of NFPAs (red outline) on coronal CE T1WI showing true positive (TP) (A), true negative (TN) (B), false positive (FP) (C), and false negative (FN) (D) in prediction model. In TP group (A), large tumor size (median tumor height of 36mm) with heterogeneous contrast enhancement due to focal cystic or hemorrhagic change were observed in most cases. In contrast, smaller tumor size (median tumor height of 16.5mm) with homogeneous contrast enhancement were found in most TN cases (B). 2 of the 3 FP cases (C) showed macrocystic component (presence of a dominant cyst exceeding 50% of the tumor volume) or macro hemorrhage (apoplexy or presence of dominant blood products exceeding 50% of the tumor volume). Although 5 of the 6 FN cases (D) also showed relatively homogeneous contrast enhancement as in TN cases, relatively large tumor height (median of 20.5mm) was found in FN cases (D) as compared with TN group.

8.2.5 Summary and Discussion

In this study, we established a predictive model implementing radiomics to predict P/R in NFPAs. Three tumor ROIs including original mask and mask with binary erosions were used. The SVM classifier was applied to evaluate the importance of the extracted features. In the three selected features, two were extracted from CE T1WI and one from the T2WI. The overall accuracy of 80 % to 82% with AUC of 0.78 to 0.8 were obtained in three tumor

ROIs. Obviously, the removal of the boundary pixels of the tumor masks on each slice by binary erosion didn't affect the results much.

Although more than 90% of NFPAs are benign pituitary adenomas according to the 2017 WHO classification system [383], 25-55% of benign NFPAs may show early P/R within 5 years after surgical resection [384-387, 416]. According to the 2017 WHO classification, Ki-67 index, mitotic count, and tumor invasion are associated with aggressive clinical behavior [383]. However, the invasive growth of NFPAs was not clearly defined in the WHO criteria, and it is usually underestimated if no corresponding information from MR imaging is taken into consideration [387, 417]. Furthermore, in a meta-analysis including 143 studies by Roelfsema et al. [387], it is known that postoperative hormone concentration was an important predictor for P/R in functioning adenomas; in contrast, no specific factor could be used to predict P/R in NFPAs.

On conventional MR imaging, invasion of the skull base bone and larger tumor size had been reported as important imaging features related to P/R in NFPAs [392, 393], and our study revealed similar results. Although conventional MRI findings associated with recurrence in NFPAs had been reported, most imaging data were presented in subjective and qualitative terms. Recently, low apparent diffusion coefficient (ADC) values on diffusion-weight MR imaging (DWI) were reported to be associated with tumor progression in NFPAs [405, 418]. However, the ADC values could only be measured for solid NFPAs because incorrect information may exist at hemorrhagic or cystic part of NFPAs due to susceptibility artifact [390, 405, 419]. Because of limited reports so far, the clinical value of ADC thus needs to be further investigated. In contrast, radiomics analysis based on whole tumor segmentation is able to reproducibly extract objective and

quantitative data from different imaging sequences to build diagnostic and predictive models classifying different lesion types [354, 394, 395, 420].

Radiomics is a relatively new field in radiology, meaning the extraction of a high number of quantitative features from medical images. Artificial intelligence (AI) is a broad concept that covers many machine learning techniques such as support vector machines, decision trees, and neural networks, that basically learn the patterns in the provided data to make predictions for unseen data sets. Radiomics can be combined with AI because it is superior in managing a massive amount of data compared with the traditional statistical methods. The primary purpose of these fields is to analyze as much and meaningful hidden quantitative data as possible to be used in medical decision and prediction [420]. A general pipeline of radiomics analysis including feature extraction, feature selection, and prediction [420, 421]. Feature extraction can quantitatively define the imaging parameters from the specified areas of the images. Feature selection can evaluate the feature importance based on the objectives. Then the prediction model will be established by selected features. Radiomics in texture and shape analysis had been widely used to evaluate medical images with promising results [20, 21, 25]. Spatial and temporal texture features of radiomics are based on the compression and destruction of normal brain structures by tumor mass, tumor cellularity, and perifocal edema [399]. Some of that cannot be detected by human visual reading [399, 400, 422]. Some authors had reported that texture analysis can reveal visually imperceptible information extends beyond radiology to histopathology, and it could be a potentially useful approach for estimating grades and molecular status in brain tumors [399, 400, 422].

For application of radiomics and machine learning (ML) in pituitary tumors, Saha et al. [421] reported a review article included 16 studies from the past 10 years (2009-2019). Of these studies, 10 appeared in 2018 to 2019, and most of the studies utilized single-centered, retrospective data, semi-automatic ML pipeline, and binary classification as in our study. Due to absence of standardized procedure, the ML algorithms vary significantly as different types of classifiers were applied and only few of the models were validated on an external set. All of the studies indicated the need of further validation before their models can be translated to clinical practice. Some authors had performed MR radiomics analyses in the differentiation of subtypes, consistency, cavernous invasion, and radiotherapeutic response in pituitary adenomas. Zhang et al. [359] reported preoperative radiomics analysis on T1WI could differentiate null cell adenomas and other subtypes in NFPAs, with AUC values of 0.8 to 0.83. Rui et al. [401] and Zeynalova et al. [413] reported preoperative radiomics texture and histogram analysis could predict tumor consistency in pituitary macroadenomas, with an AUC of 0.836 and 0.71 respectively. Fan et al. [402, 411] and Kocak et al. [412] used ML-based radiomics to predict response of radiotherapy and somatostatin analogues in acromegaly, with AUC of 0.96 and 0.845 respectively. Niu et al. [403] used radiomics analysis for prediction of cavernous sinus invasion in NFPAs, with AUC values of 0.826 to 0.852. Therefore, it is postulated that radiomics features may play a potential role in prediction of recurrence in NFPAs. However, the application of radiomics for predicting outcomes in NFPAs is rare. To the best of our knowledge, we have thus undertaken the first MR radiomics analysis for preoperative prediction of P/R in NFPAs.

In the proposed method, SVM algorithm was utilized for feature selection and classification. SVMs are among the best (and many believe are indeed the best) “off-the-

shelf" supervised learning algorithms [10]. SVM utilizes kernel method which gives a way to apply inputs efficiently in very high dimensional (such as infinite-dimensional) feature spaces. At the same time, this algorithm can guarantee an optimal margin between different data classes. Therefore, the variance of the classification results can be reduced to a reasonable level. Moreover, with SVM as objective function, we used sequential forward selection method to add important features to an empty candidate set to limit the number of selected features to control overfitting [283]. Compared with other method, such as least absolute shrinkage and selection operator (LASSO) and artificial neural network [276], this method improves the generalization of the selection process as well as guarantees the classification performance [13]. Although some algorithms such as random forest and LASSO are suitable for small dataset [398], the sequential selection method can deal with the overfitting issue properly.

The overall prediction accuracy in this study was 82% and the AUC of the prediction model was 0.78. The results were based on the three selected features, T1 surface-to-volume ratio, T1 GLCM-informational measure of correlation, and T2 NGTDM-coarseness, used for the prediction model. The surface-to-volume ratio is the ratio of surface area to volume. It compares the size of the outside of an object and the amount inside. For example, small or thin objects have a large surface area compared to the volume. T1 GLCM-informational measure of correlation is the informational measurement of the joint probability occurrence of the pixel pairs entropy on T1 weighted images. If the distribution of the intensities are more homogeneous, the value of this feature can be higher. T2 NGTDM-coarseness is an inverse measure of the level of the spatial rate of change in

intensity. A higher value indicates a lower spatial change rate and a locally more uniform texture [25, 423].

In this study, we used three ROIs methods, including the original tumor mask, and two masks with different erosion of the boundary pixels. The goal was to evaluate whether the potential inclusion of normal pituitary gland and other surrounding non-tumor tissue would affect the prediction. Our results showed that erosion of the boundary pixels didn't improve the prediction accuracy of PR, with only minimal improvement of AUC. One possible reason for the results was that the eroded pixels was minimal compared to the whole tumor mask, thus would not affect the results much.

There were 41 true and 9 false prediction cases in predictive model with original tumor mask. Larger tumor sizes (median tumor height of 36 mm) were observed in the TP cases as compared with TN cases. 9 (9/16, 56.3%) TP cases were giant NFPAs (> 40 mm). In addition, 8 (8/16, 50%) and 12 (12/16, 75%) TP cases showed cavernous sinus invasion (Knops grade 3 to 4) and extrasellar sphenoid bone extension (Hardy's grade 3 to 4) respectively. Heterogeneous T2WI signal with heterogeneous contrast enhancement on T1WI due to focal cystic or hemorrhagic change (cyst or hemorrhage smaller than 50% of the tumor volume) were observed in most TP cases. GTR was performed in 2 (2/16, 12.5%) TP cases. In contrast, smaller tumor sizes (median tumor height of 16.5mm) with less cavernous sinus invasion (1/25, 4%) and extrasellar extension (1/25, 4%) were found in TN cases. Further, homogeneous isointense T2WI signal with homogeneous contrast enhancement were found in most TN cases. GTR was achieved in 4 (4/25, 16%) TN cases. All 3 FP cases had tumor height between 3 to 4 cm, and 2 of the 3 cases showed macrocystic component (presence of a dominant cyst exceeding 50% of the tumor volume)

or macro hemorrhage (apoplexy or presence of dominant blood products exceeding 50% of the tumor volume). Although 5 of the 6 FN cases also showed homogeneous T2WI signal and homogeneous contrast enhancement as in TN cases, relatively large tumor heights (median tumor height of 20.5mm) were found in FN group. Further study involving a larger sample size is necessary for further understanding factors related to true and false predictions.

It is known that the extent of tumor resection is an important determining factor in recurrence rate of NFPAs [392, 406]. Although no statistical difference existed between GTR and P/R in our study, this result may be explained by the small sample size. On the other hand, significant correlation between the number of surgical resections and complication rates in NFPAs was reported [424]. Anterior pituitary insufficiency and diabetes insipidus are the most common post-operative complications in NFPAs, with occurrence rates of 19.4% and 17.8% respectively [424]. In our study, 3 patients still had tumor recurrence after receiving GTR; in contrast, 16 patients had stable disease even if receiving STR only. Since most NFPAs are benign tumors, preoperative prediction of P/R in NFPAs offers clinically valuable information on treatment choices. For patients with high possibility of tumor recurrence, aggressive resection combined with postoperative adjuvant RT and close MR imaging follow up should be considered; in contrast, for patients with lower possibility of recurrence, the aim of surgery would be relief of clinical symptoms by decreasing tumor mass effect. The optimal surgical planning for the low risk patients will reduce the potential complications of endocrine disorders.

It is known that postoperative adjuvant RT offers excellent tumor control with rate up to 96% in non-secreting adenomas [425]. However, whether postoperative RT is beneficial

for patients with low possibility of recurrence is controversial because RT may increase risks of complications such as visual deterioration, hypopituitarism, cerebrovascular accident, and dementia in NFPAs [425, 426]. Since adjuvant RT may affect the independent predictive value of the preoperative MR radiomics analysis for P/R, patients with adjuvant RT before P/R were excluded from our study.

Although we performed the first radiomics model for preoperative prediction of P/R in NFPAs, our study still had several limitations. Selection bias may exist due to the retrospective nature in our study. All images were acquired at a single site, and mostly with a single protocol. Future testing with multi-institutional data and varying imaging protocols is necessary to determine whether the trained classifier is generalizable. The implemented radiomics method is straightforward, and it may not utilize the information from all images completely since it is based on pre-defined features. Because the sample size is small, only a few imaging features can be selected into the classification model to avoid over-fitting. More cases are expected to improve the model performance. Besides, more advanced statistical analysis methods that can take all clinical and imaging factors into account need to be developed in the future. Moreover, convolutional neural network can be taken into consideration when an increasing number of cases is available. Convolutional neural network is a machine learning strategy which is designed for computer vision and obtained some satisfactory results in the radiology field. The future work should consider using CNN to improve the prediction performance.

To the best of our knowledge, this was the first study attempting to apply the MR radiomics approach to predict P/R in NFPAs. With the analysis of CE T1WI and T2WI, the overall accuracy of 82% and AUC of 0.78 were obtained in SVM predictive model. Although

the results of this study were preliminary, due to the objective and quantitative measures of radiomics, it may likely offer valuable information for the preoperative and postoperative planning in the management of NFPAs, such as the extent of surgical resection, implementation of postoperative adjuvant RT, and the time interval of MR imaging follow-up. Nevertheless, this approach still needs to be validated when studies with more cases and a long-term follow-up are conducted.

Chapter 9. Conclusions and Future Plans

In this dissertation, several machine learning methods, including radiomics and deep learning, were proposed and implemented. These methods were applied to 6 different clinical scenarios: i) lesion detection; ii) organ/tissue segmentation; iii) diagnosis and subtype classification; iv) treatment planning; v) treatment response prediction; and vi) prognosis prediction. Different deep learning and radiomics algorithms were implemented according to the collected datasets in these applications, that covered various diseases in different organ sites, including breast, brain, spine, lung, and prostate. The satisfactory prediction results suggest there is a potential for these methods to contribute in solving clinical problems, and also that the methods may be further developed into commercial products for wider clinical adoption to benefit many patients and improve their care and management. My intellectual contributions in this dissertation are in four main areas:

1. Design Convolutional Neural Networks According to the Image Data Structures:

CLSTM: For the breast DCE-MRI studies, the pattern of the DCE kinetics (or, signal intensity time curve) is known to provide important information for lesion diagnosis, which can be taken into consideration in deep learning architecture. To consider the full spectrum of this time-dependent intensity information in DCE MRI, CLSTM architecture was developed to process the DCE images set by set. By using this architecture, the temporal features contained in the DCE sequences can be fully utilized.

Bi-directional CLSTM: For the prostate MRI project, the DCE train has a total of 40 images in the time series, and it is too long for conventional LSTM. In the implementation of LSTM, the “forget gate” will cause the information from the early dataset to contribute less than

the later dataset. To minimize this problem, I developed a bi-directional CLSTM model, which treats the later and early datasets equally.

Mask R-CNN: For the breast lesion detection project in MRI, I implemented a mask R-CNN architecture with modified criteria to improve lesion detection accuracy, which utilized the enhancement information from the contralateral side breast based on symmetry.

Weakly Supervised Learning using Class Activation Map: I have implemented another detection algorithms to localize the malignant fractures on MR images, by generating the intermediated feature maps from CNN.

In this dissertation, there are 6 projects using convolution classification neural network. In the project for diagnosis of benign and malignant breast lesions on DCE-MRI and the project for differentiation of benign and malignant vertebral fracture, the ResNet50 architecture was chosen. This was decided after an exploration using different algorithms. The deep learning was also performed by using 4 different convolutional neural networks, including, VGG16, VGG19, Xception, and InceptionV3. From the cross validation of the training set, VGG16 and VGG19 resulted in poorer performance. The prediction accuracies of Xception and InceptionV3 were comparable to that of ResNet50. However, the corresponding prediction accuracies on the independent testing set was much lower compared to ResNet50, which meant the generalization of Xception and InceptionV3 was not satisfactory. Therefore, ResNet50 was the final best choice among these pre-defined large scale networks.

When the input case number was limited, ResNet50 might not work well, and the 7-layer customized CNN was implemented for the other 4 projects, including prediction of breast cancer molecular subtypes, differentiation of spinal metastases, differentiation of

prostate cancer and benign prostatic hyperplasia, and neoadjuvant chemoradiation therapy response prediction in rectal cancer. It is known that the larger scale of the network leads to more powerful computational capability, but it requires more input cases. Some pre-defined large scale networks, such ResNet50, Inceptionv3, VGG16, VGG19, and Xception, contain more layers and trainable parameters than the customized CNN with 7 layers. Although techniques to control overfitting have been implemented, the output performance in studies with small training cases cannot be improved to acceptable levels. During the exploration of the CNN architecture, I have tried different settings, such as layer numbers and feature map numbers. I tried to use different number of convolutional layers with 5, 6, … to 15. With regularization term and augmented inputs, the performance of the network with 7 to 12 layers could achieve comparable performance. When the number of layers was lower than 5, the number of features extracted from the network became lower which led to less information being decoded. When the number of layers was higher than 12, the validation loss was obviously higher than the training loss. To keep the generalization of the whole system, I selected the architecture with the fewest trainable parameters, thus the 7-layer CNN was implemented in these 4 projects.

Based on these experiences, in order to get better performance in future deep learning studies, the experiment can start from the pre-defined large scale network, such as ResNet50 if the input case number is sufficient. By implementing the proper methods to avoid overfitting, good prediction results may be obtained. If the performance is not satisfactory, customized CNN with smaller number of layers may be considered. From my experience, we can start with CNN with 7 layers, and then fine-tune the architecture to get the best performance.

2. Develop Transfer Learning Strategies for Improving the Classification Performance in Independent Testing Datasets:

The independent testing cases were usually acquired using a different setting, and re-tuning of the parameters is necessary to improve accuracy. In the *breast cancer molecular subtype classification study*, we split the testing cases based on the time of MRI, which represented a realistic clinical scenario. When an AI product developed by a company is implemented in a hospital, the old retrospective cases can be used for re-training, and then the obtained specific model can be used in analyzing the new, or prospective, cases. For the *segmentation of the breast and fibroglandular tissue*, I designed a transfer learning strategy to apply the obtained non-fat-sat model for training of fat-sat images. The results showed that transfer learning could be applied to improve the segmentation accuracy compared to the direct training, and also that the training efficiency could be improved, thus not requiring a large number of input data to obtain satisfactory performance. Also, in the *classification of benign and malignant fracture on MRI*, in order to improve the accuracy in the dataset acquired using a different scanner with different matrix size, re-training was implemented with one additional pre-processing layer. The results showed that by using this added layer for re-tuning, the accuracy was improved.

3. Develop Data-Specific Segmentation and Registration Algorithms According to the Lesion Contrast and Surrounding Tissues:

There are several image segmentation methods implemented in this dissertation. I developed the *normalized-cut algorithm to segment spinal metastasis* on the post-contrast

MR images. Also, a *non-mass breast lesion segmentation algorithm on MRI* was developed based on the region growing method. The threshold of region growing was calculated from the intensity distribution of the lesion areas and normal tissue areas. I also developed a new method based on the *registration of baseline and follow-up CT images to enable the segmentation and evaluation of COVID-19 lesions* on the corresponding image space. A two-step registration method was developed, first by applying the Affine registration based on the whole images, and then followed by the non-rigid registration algorithm focusing on the segmented lung tissues. In a project to *match the rectal cancers between baseline and follow-up MRI*, I developed a method to segment the rectum on B/L and F/U images, so they can be registered to evaluate the change of lesions after receiving neoadjuvant chemoradiation therapy.

4. Implement Various Machine Learning Algorithms for Feature Selection and Model Building in Radiomics:

There are 4 projects using radiomics models in this dissertation, and I implemented several supervised learning algorithms for feature selection and classification, including random forest (RF), support vector machine (SVM), artificial neural network (ANN), etc. In radiomics, choosing proper machine learning algorithms is very important to optimize the final performance. In the project for diagnosis of benign and malignant breast lesions and the project for prediction of breast cancer molecular subtypes, random forest (RF) was employed. In the project for diagnosis of benign and malignant prostate tumors, I utilized SVM. In the project for neoadjuvant chemoradiation therapy response prediction in rectal

cancer, I implemented an artificial neural network (ANN) to do feature selection and classification prediction.

Among these algorithms, SVM is the most flexible, and the most likely to give a high accuracy for a high dimensional dataset [10]. SVM is a special kind of linear model with a specific kernel [15, 16]. The kernel in SVM works as a transformation, which maps input parameters into a different feature space where the transformed data can be divided more obviously. Also, the proper choice of the cost function allows a wide margin between different classes. This can improve the robustness. However, SVM algorithm is sensitive to noise. Thus, a careful outlining of ROI on the original images is essential to obtain good results. Also, the complicity of SVM model usually requires a large dataset, and overfitting is a common problem in small datasets.

Compared to SVM, random forest can tolerate noise, and can deal with unbalanced data. Meanwhile, random forest can avoid overfitting during the training of the model. This is the reason why I chose random forest instead of SVM in some projects. Random forest algorithms can estimate all feature importance, and then we can determine the number of features as desired, usually 1/5-1/3 of number of the input case number. But due to the simplicity of the decision trees, the performance may not reach that of SVM when the volume and quality of the input images can meet the requirement for SVM.

In the project for neoadjuvant chemoradiation therapy response prediction in rectal cancer, I implemented a 3-layer artificial neural network (ANN) to do feature selection and classification prediction. Artificial neural network can perform various kinds of complicated nonlinear mapping. But, the training of the neural network is challenging, and may get stuck in local minima problem. Also, the determination of optimal network nodes

is difficult. In this project, since both the number of cases and features are small, the artificial neural network works well. But for other projects, it may not be a proper choice.

For the future radiomics studies, our experience suggests that the data can be processed using the SVM first. If the case number is not sufficient, or the input image quality is bad, then the random forest can be applied to perform the feature selection and establish the prediction model. If the case number is far larger than the feature number, some dimension reduction algorithms, such as Principle Component Analysis (PCA), Independent Component Analysis (ICA) can be applied to reduce the input dimensions. If the number of selected features from SVM or random forest is low, the logistic regression is another option which can be applied to build a robust model with satisfactory performance.

Artificial Intelligence (AI) using machine learning has been proven as a feasible approach for object recognition and computer vision tasks that can achieve a satisfactory performance comparable to human observers. The success that has been demonstrated so far, combined with the rapid technological advancement, makes it reasonable for people to believe that these methods will further revolutionize the clinical workflow, not only in radiology but also in general medical practices. When we consider the use of AI in medical imaging, we anticipate this technological innovation to serve as a collaborative platform, aiming to decrease the burden and distraction of human observers in performing many repetitive and humdrum tasks, rather than to replace physicians. The use of deep learning and AI in radiology is currently in the stage of infancy, but it is progressing very rapidly. One of the key factors for the development and its proper clinical adoption in medicine would be a good mutual understanding of the AI technology, and its most appropriate form of clinical practice and workflow, by both clinicians and computer scientists/engineers.

AI companies are working on commercialization of their developed products. Many AI software for image analysis and interpretation has been cleared by the Food and Drug Administration (FDA), and the list is expanding very rapidly. These products only need to have specific intended use, e.g. they can focus on image analysis, visualization display, abnormality detection and segmentation, disease diagnosis and assessment, etc. The American College of Radiology is keeping track of FDA cleared products on this website:

https://www.acrdsi.org/DSI-Services/FDA-Cleared-AI-Algorithms

So far, the most common applications are for several organs: brain, breast, chest (heart and lung), pelvic, and musculoskeletal systems. For these FDA cleared AI algorithms in radiology, CT is the most popular imaging modality. Most of the head and neck, lung, liver, and MSK products are designed based on CT images or CT angiography. MRI is another popular modality due to its great soft-tissue contrast, which provides better capability for imaging of cancer in the whole body, and imaging of vascular function in the heart and the brain. Also, a few products are designed for X-ray, ultrasound, mammography, and digital breast tomosynthesis (DBT). In these AI products, the deep learning algorithms for the application in the brain and the lung are the most mature. Several companies have developed products for automatic detection of intracerebral hemorrhage (ICH), large vessel occlusion (LVO), and pulmonary embolism; however, instead of marketing them as diagnostic tools, the product is mainly designed to flag/prioritize or triage cases based on suspicious findings. The most significant impact is the clinical application in emergency medicine, when the patient presents with significant findings and needs urgent care. The AI tool can trigger an immediate alert, which can be very helpful especially at night times when experienced attending physicians are not on-site to take care of patients.

Another major area of FDA approved AI products is for detection and characterization of suspicious malignant cancer, with the major focus for breast cancer on mammography, ultrasound, and DBT. Several other approved AI products are for detection of bone fracture on x-ray or CT images. Based on the probability, the system can identify the area with different markers or colors, and with specified malignancy score. The requirement to clear products aiming to flag cases with clinically significant findings or to provide suspected abnormality is much lower compared to products aiming to give a final diagnosis of malignant vs. benign findings. Perhaps the liability issue is the main reason for the current AI product to be marketed as an assisting tool for decision-making support, not for guiding clinical decisions. Therefore, there still exists a huge room for improvements in the detection and diagnosis of cancers and bone fractures. In this dissertation, we applied radiomics and deep learning to investigate these problems, and obtained encouraging results. The presented studies may provide a solid foundation for further refinement of the AI methods towards future clinical adoption to help solving more clinical problems. For example, several projects were set out to investigate the application of AI in cancer diagnosis and density segmentation on breast MRI, which can be further developed as FDA approved products.

In this dissertation, the main application is in oncology, and the majority of the projects is focusing on MR images. In addition to deep learning, we also performed diagnosis using radiologists' reading and radiomics methods, so the results can be compared to understand the value of the developed AI algorithms. Although deep learning is a very powerful method, it will require a large dataset for training and validation. In contrast, radiomics can be performed in a small dataset to successfully train a model. Another advantage of

radiomics is the capability to further analyze the selected features, so the link with visual reading descriptors can be established, and the developed model can be explained and accepted by the radiology society more easily.

Although the AI techniques have demonstrated promising applications in medical field, there are still some major obstacles for them to reach a high impact. One issue is the explainability of deep learning algorithms. Many reports in the literature described it as a 'black box'. For clinicians, if the performance of the algorithms cannot be clearly explained, it will lead to hesitation for its adoption for patient care. However, many other aspects of clinical practice in medicine are also unexplained, and the most important requirement is for the method to demonstrate its clinical value, e.g. correctly identify hemorrhage, cancer, fracture, etc. [427]. Another obstacle is the requirement of large data volume for training and validation. The lack of data is a crucially important issue for deep leaning, especially in medical applications. Due to the fast development of machine learning algorithms, the computation scales have led to high computational workload and large number of trainable parameters. Thus, this requires more training cases as well as upgraded computer hardware, such as GPU. The shortage of data will lead to overfitting, and the low performance hardware will limit the memory and result in extremely long training time. Meanwhile, if the distribution of training samples is very different from that of testing cases, the generalization of the network will be low. Considering the great distinction between the high-quality images which were often used in the research work and the actual image quality in the real clinical world, this is a major issue when implementing the commercial products developed by deep learning algorithms [49]. Even with regularization or transfer learning methods described in previous chapters, current performance of the

applications in processing medical images is worse compared to natural image studies. In the future, when larger clinical datasets gradually become available, deep learning algorithms are expected to achieve better performance.

With the fast development of deep learning, there are a lot of new techniques being published for natural image applications, for example dilated convolution [428] and PSPnet [429]. These algorithms showed very good performance. The homogeneity among different images in the same class for the natural images and the medical images are not similar. The noise levels of the natural images and medical images vary a lot as well. Therefore, the capability of these advanced algorithms should be explored in the future. Appropriate modifications tailored to the image dataset should be investigated during the application.

In conclusion, my PhD research includes 6 clinical tasks commonly performed in radiology, including detection, segmentation, differential diagnosis, therapy planning, response monitoring, and prognosis prediction. In these projects, the MR and CT images are processed, and several novel deep learning algorithms are developed according to different types of images included in the dataset. The clinical applications in breast cancer, brain cancer, rectal cancer, spinal cancer, prostate cancer and spine fracture are demonstrated. In the future, we will apply more advanced and novel machine learning algorithms to further improve the performance in these tasks, and possibly, extend to other diseases. Overall, I am very grateful to many people who have helped me in many different ways in this PhD research. I sincerely hope that what I have achieved in this dissertation can be further improved and extended, and contribute in the precision care that can be provided to each individual patient.

References

1. Bhargavan M, Kaye AH, Forman HP, Sunshine JH: Workload of radiologists in United States in 2006–2007 and trends since 1991–1992. Radiology 252:458-467, 2009
2. Wang S, Summers RM: Machine learning and radiology. Medical image analysis 16:933-951, 2012
3. Davis PL, et al.: Breast cancer measurements with magnetic resonance imaging, ultrasonography, and mammography. Breast cancer research and treatment 37:1-9, 1996
4. Agrawal G, Su MY, Nalcioglu O, Feig SA, Chen JH: Significance of breast lesion descriptors in the ACR BI-RADS MRI lexicon. Cancer 115:1363-1380, 2009
5. Radiology ESo: The future role of radiology in healthcare. Insights into imaging 1:2-11, 2010
6. Fenton JJ, et al.: Influence of computer-aided detection on performance of screening mammography. New England Journal of Medicine 356:1399-1409, 2007
7. Lehman CD, Wellman RD, Buist DS, Kerlikowske K, Tosteson AN, Miglioretti DL: Diagnostic accuracy of digital screening mammography with and without computer-aided detection. JAMA internal medicine 175:1828-1837, 2015
8. Goodfellow I, Bengio Y, Courville A: Deep learning: MIT press, 2016
9. Paiva OA, Prevedello LM: The potential impact of artificial intelligence in radiology. Radiologia brasileira 50:V-VI, 2017
10. Nasrabadi NM: Pattern recognition and machine learning. Journal of electronic imaging 16:049901, 2007
11. Chollet F: Deep Learning with Python. 978-1617294433. Shelter Island: Manning Publications:384, 2017
12. Lee J-G, et al.: Deep Learning in Medical Imaging: General Overview. Korean Journal of Radiology 18:570-584, 2017
13. Byvatov E, Fechner U, Sadowski J, Schneider G: Comparison of support vector machine and artificial neural network systems for drug/nondrug classification. Journal of chemical information and computer sciences 43:1882-1889, 2003
14. Cortes C, Vapnik V: Support-vector networks. Machine learning 20:273-297, 1995
15. Drucker H, Burges CJ, Kaufman L, Smola AJ, Vapnik V: Support vector regression machines. Proc. Advances in neural information processing systems: City
16. Tong S, Chang E: Support vector machine active learning for image retrieval. Proc. Proceedings of the ninth ACM international conference on Multimedia: City
17. David B: Bayesian Reasoning and Machine Learning. London: Cambridge, 2012
18. Bengio Y: Learning deep architectures for AI. Foundations and trends® in Machine Learning 2:1-127, 2009
19. Shin H-C, et al.: Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE transactions on medical imaging 35:1285-1298, 2016
20. Kumar V, et al.: Radiomics: the process and the challenges. Magnetic resonance imaging 30:1234-1248, 2012
21. Lambin P, et al.: Radiomics: extracting more information from medical images using advanced feature analysis. European journal of cancer 48:441-446, 2012

22. Gillies RJ, Kinahan PE, Hricak H: Radiomics: images are more than pictures, they are data. Radiology 278:563-577, 2016
23. Fernández-Delgado M, Cernadas E, Barro S, Amorim D: Do we need hundreds of classifiers to solve real world classification problems? The journal of machine learning research 15:3133-3181, 2014
24. Fusco R, et al.: Pattern recognition approaches for breast cancer DCE-MRI classification: a systematic review. Journal of medical and biological engineering 36:449-459, 2016
25. Haralick RM, Shanmugam K: Textural features for image classification. IEEE Transactions on systems, man, and cybernetics:610-621, 1973
26. Haralick RM, Shapiro LG: Computer and robot vision: Addison-wesley, 1992
27. Degenhard A, Tanner C, Hayes C, Hawkes D, Leach M: Comparison between radiological and artificial neural network diagnosis in clinical screening. Physiological measurement 23:727, 2002
28. Fusco R, Sansone M, Maffei S, Raiano N, Petrillo A: Dynamic contrast-enhanced MRI in breast cancer: A comparison between distributed and compartmental tracer kinetic models. Journal of Biomedical Graphics and Computing 2:23, 2012
29. Brix G, et al.: Microcirculation and microvasculature in breast tumors: pharmacokinetic analysis of dynamic MR image series. Magnetic resonance in medicine 52:420-429, 2004
30. Sansone M, Fusco R, Petrillo A, Petrillo M, Bracale M: An expectation-maximisation approach for simultaneous pixel classification and tracer kinetic modelling in dynamic contrast enhanced-magnetic resonance imaging. Medical & biological engineering & computing 49:485-495, 2011
31. Fusco R, Sansone M, Petrillo M, Petrillo A: Influence of parameterization on tracer kinetic modeling in DCE-MRI. Journal of Medical and Biological Engineering 34:157-163, 2014
32. McLaren CE, Chen W-P, Nie K, Su M-Y: Prediction of malignant breast lesions from MRI features: a comparison of artificial neural network and logistic regression techniques. Academic radiology 16:842-851, 2009
33. Zheng Y, Baloch S, Englander S, Schnall MD, Shen D: Segmentation and classification of breast tumor using dynamic contrast-enhanced MR images. Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention: City
34. Ikeda DM, et al.: Development, standardization, and testing of a lexicon for reporting contrast-enhanced breast magnetic resonance imaging studies. Journal of Magnetic Resonance Imaging 13:889-895, 2001
35. Castellano G, Bonilha L, Li L, Cendes F: Texture analysis of medical images. Clinical radiology 59:1061-1069, 2004
36. Chu A, Sehgal CM, Greenleaf JF: Use of gray value distribution of run lengths for texture analysis. Pattern Recognition Letters 11:415-419, 1990
37. Galloway MM: Texture analysis using grey level run lengths. NASA STI/Recon Technical Report N 75, 1974
38. Thibault G, et al.: Shape and texture indexes application to cell nuclei classification. International Journal of Pattern Recognition and Artificial Intelligence 27:1357002, 2013

39. Vomweg T, et al.: Improved artificial neural networks in prediction of malignancy of lesions in contrast-enhanced MR-mammography. Medical physics 30:2350-2359, 2003
40. Segal MR: Machine learning benchmarks and random forest regression. Center for Bioinformatics & Molecular Biostatistics, 2004
41. Ho TK: Random decision forests. Proc. Document analysis and recognition, 1995, proceedings of the third international conference on: City
42. Bezdek JC: Objective Function Clustering: Springer, 1981
43. Clendenen TV, Zeleniuch-Jacquotte A, Moy L, Pike MC, Rusinek H, Kim S: Comparison of 3-point dixon imaging and fuzzy C-means clustering methods for breast density measurement. Journal of Magnetic Resonance Imaging 38:474-481, 2013
44. Schmidhuber J: Deep learning in neural networks: An overview. Neural networks 61:85-117, 2015
45. LeCun Y, Bengio Y, Hinton G: Deep learning. nature 521:436, 2015
46. Anwar SM, Majid M, Qayyum A, Awais M, Alnowami M, Khan MK: Medical image analysis using convolutional neural networks: a review. Journal of medical systems 42:226, 2018
47. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJ: Artificial intelligence in radiology. Nature Reviews Cancer 18:500-510, 2018
48. Krizhevsky A, Sutskever I, Hinton GE: Imagenet classification with deep convolutional neural networks. Proc. Advances in neural information processing systems: City
49. Lundervold AS, Lundervold A: An overview of deep learning in medical imaging focusing on MRI. Zeitschrift für Medizinische Physik 29:102-127, 2019
50. Shen D, Wu G, Suk H-I: Deep learning in medical image analysis. Annual review of biomedical engineering 19:221-248, 2017
51. Suzuki K: Overview of deep learning in medical imaging. Radiological physics and technology 10:257-273, 2017
52. Csáji BC: Approximation with artificial neural networks. Faculty of Sciences, Etvs Lornd University, Hungary 24:7, 2001
53. LeCun Y, Bengio Y: Convolutional networks for images, speech, and time series. The handbook of brain theory and neural networks 3361:1995, 1995
54. Hubel DH, Wiesel TN: Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. The Journal of physiology 160:106-154, 1962
55. Gelsema ES, Kanal LN: Pattern recognition in practice: proceedings of an international workshop held in Amsterdam, May 21-23, 1980: North-Holland, 1980
56. LeCun Y, Bottou L, Bengio Y, Haffner P: Gradient-based learning applied to document recognition. Proceedings of the IEEE 86:2278-2324, 1998
57. Serre T, Wolf L, Bileschi S, Riesenhuber M, Poggio T: Robust object recognition with cortex-like mechanisms. IEEE transactions on pattern analysis and machine intelligence 29:411-426, 2007
58. Szegedy C, et al.: Going deeper with convolutions. Proc. Proceedings of the IEEE conference on computer vision and pattern recognition: City
59. He K, Zhang X, Ren S, Sun J: Deep residual learning for image recognition. Proc. Proceedings of the IEEE conference on computer vision and pattern recognition: City

60. Nair V, Hinton GE: Rectified linear units improve restricted boltzmann machines. Proc. Proceedings of the 27th international conference on machine learning (ICML-10): City
61. Srivastava N, Hinton GE, Krizhevsky A, Sutskever I, Salakhutdinov R: Dropout: a simple way to prevent neural networks from overfitting. Journal of machine learning research 15:1929-1958, 2014
62. Hochreiter S, Schmidhuber J: Long short-term memory. Neural computation 9:1735-1780, 1997
63. Xingjian S, Chen Z, Wang H, Yeung D-Y, Wong W-K, Woo W-c: Convolutional LSTM network: A machine learning approach for precipitation nowcasting. Proc. Advances in neural information processing systems: City
64. Szegedy C, Ioffe S, Vanhoucke V, Alemi AA: Inception-v4, inception-resnet and the impact of residual connections on learning. Proc. AAAI: City
65. Hazewinkel M: Affine transformation. Encyclopedia of Mathematics, Springer, 2001
66. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L: Imagenet: A large-scale hierarchical image database. Proc. Computer Vision and Pattern Recognition, 2009 CVPR 2009 IEEE Conference on: City
67. Haarburger C, et al.: Transfer Learning for Breast Cancer Malignancy Classification based on Dynamic Contrast-Enhanced MR Images: Springer, 2018
68. Tan C, Sun F, Kong T, Zhang W, Yang C, Liu C: A survey on deep transfer learning. Proc. International Conference on Artificial Neural Networks: City
69. Samala RK, Chan H-P, Hadjiiski L, Helvie MA, Richter CD, Cha KH: Breast cancer diagnosis in digital breast tomosynthesis: effects of training sample size on multi-stage transfer learning using deep neural nets. IEEE Transactions on Medical Imaging 38:686-696, 2018
70. Sevakula RK, Singh V, Verma NK, Kumar C, Cui Y: Transfer learning for molecular cancer classification using deep neural networks. IEEE/ACM transactions on computational biology and bioinformatics, 2018
71. Yuan Y, et al.: Prostate cancer classification with multiparametric MRI transfer learning model. Medical physics 46:756-765, 2019
72. Sites A: SEER cancer statistics review 1975-2011. Bethesda, MD: National Cancer Institute, 2014
73. Chakraborti K, Bahl P, Sahoo M, Ganguly S, Oberoi C: Magentic resonance imaging of breast masses: Comparison with mammography. Indian Journal of Radiology and Imaging 15:381, 2005
74. Montemurro F, et al.: Relationship between DCE-MRI morphological and functional features and histopathological characteristics of breast cancer. European radiology 17:1490-1497, 2007
75. Kuhl C: The current status of breast MR imaging part I. Choice of technique, image interpretation, diagnostic accuracy, and transfer to clinical practice. Radiology 244:356-378, 2007
76. Kuhl CK: Current status of breast MR imaging part 2. Clinical applications. Radiology 244:672-691, 2007
77. Chen W, Giger ML, Lan L, Bick U: Computerized interpretation of breast MRI: Investigation of enhancement-variance dynamics. Medical physics 31:1076-1082, 2004

78. Renz D, et al.: Clinical value of computer-assisted analysis in MR mammography. A comparison between two systems and three observers with different levels of experience. RoFo: Fortschritte auf dem Gebiete der Rontgenstrahlen und der Nuklearmedizin 180:968-976, 2008
79. Lehman CD, Blume JD, DeMartini WB, Hylton NM, Herman B, Schnall MD: Accuracy and interpretation time of computer-aided detection among novice and experienced breast MRI readers. American Journal of Roentgenology 200:W683-W689, 2013
80. Djilas-Ivanovic D, et al.: Breast MRI: intraindividual comparative study at 1.5 and 3.0 T; initial experience. Journal of BU ON: official journal of the Balkan Union of Oncology 17:65-72, 2012
81. Pediconi F, et al.: Breast lesion detection and characterization at contrast-enhanced MR mammography: gadobenate dimeglumine versus gadopentetate dimeglumine. Radiology 237:45-56, 2005
82. Pediconi F, et al.: Contrast-enhanced MR mammography: improved lesion detection and differentiation with gadobenate dimeglumine. American Journal of Roentgenology 191:1339-1346, 2008
83. Martincich L, et al.: Multicenter, double-blind, randomized, intraindividual crossover comparison of gadobenate dimeglumine and gadopentetate dimeglumine for breast MR imaging (DETECT Trial). Radiology 258:396-408, 2011
84. Gubern-Mérida A, et al.: Automated detection of breast cancer in false-negative screening MRI studies from women at increased risk. European journal of radiology 85:472-479, 2016
85. Chang Y-C, Huang Y-H, Huang C-S, Chen J-H, Chang R-F: Computerized breast lesions detection using kinetic and morphologic analysis for dynamic contrast-enhanced MRI. Magnetic resonance imaging 32:514-522, 2014
86. Dorrius MD, Jansen-van der Weide MC, van Ooijen PM, Pijnappel RM, Oudkerk M: Computer-aided detection in breast MRI: a systematic review and meta-analysis. European radiology 21:1600-1608, 2011
87. Renz DM, et al.: Detection and classification of contrast-enhancing masses by a fully automatic computer-assisted diagnosis system for breast MRI. Journal of Magnetic Resonance Imaging 35:1077-1088, 2012
88. Vignati A, et al.: Performance of a fully automatic lesion detection system for breast DCE-MRI. Journal of Magnetic Resonance Imaging 34:1341-1351, 2011
89. Codari M, Schiaffino S, Sardanelli F, Trimboli RM: Artificial Intelligence for Breast MRI in 2008–2018: A Systematic Mapping Review. American Journal of Roentgenology 212:280-292, 2019
90. Al-masni MA, et al.: Simultaneous detection and classification of breast masses in digital mammograms via a deep learning YOLO-based CAD system. Computer methods and programs in biomedicine 157:85-94, 2018
91. Dalmış MU, Vreemann S, Kooi T, Mann RM, Karssemeijer N, Gubern-Mérida A: Fully automated detection of breast cancer in screening MRI using convolutional neural networks. Journal of Medical Imaging 5:014502, 2018
92. Sheth D, Giger ML: Artificial intelligence in the interpretation of breast cancer on MRI. Journal of Magnetic Resonance Imaging, 2019

93. Yap MH, et al.: Automated breast ultrasound lesions detection using convolutional neural networks. IEEE journal of biomedical and health informatics 22:1218-1226, 2017
94. Kooi T, et al.: Large scale deep learning for computer aided detection of mammographic lesions. Medical image analysis 35:303-312, 2017
95. Samala RK, Chan HP, Hadjiiski L, Helvie MA, Wei J, Cha K: Mass detection in digital breast tomosynthesis: Deep convolutional neural network with transfer learning from mammography. Medical physics 43:6654-6666, 2016
96. Kim E-K, et al.: Applying data-driven imaging biomarker in mammography for breast cancer screening: preliminary study. Scientific reports 8:2762, 2018
97. Zhou J, et al.: Weakly supervised 3D deep learning for breast cancer classification and localization of the lesions in MR images. Journal of Magnetic Resonance Imaging, 2019
98. Ribli D, Horváth A, Unger Z, Pollner P, Csabai I: Detecting and classifying lesions in mammograms with deep learning. Scientific reports 8:4165, 2018
99. He K, Gkioxari G, Dollár P, Girshick R: Mask r-cnn. Proc. Computer Vision (ICCV), 2017 IEEE International Conference on: City
100. Chang P, et al.: Hybrid 3D/2D convolutional neural network for hemorrhage evaluation on head CT. American Journal of Neuroradiology 39:1609-1616, 2018
101. Rohit Malhotra K, Davoudi A, Siegel S, Bihorac A, Rashidi P: Autonomous detection of disruptions in the intensive care unit using deep mask R-CNN. Proc. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops: City
102. Couteaux V, et al.: Automatic knee meniscus tear detection and orientation classification with Mask-RCNN. Diagnostic and interventional imaging 100:235-242, 2019
103. Nie K, Chen J-H, Hon JY, Chu Y, Nalcioglu O, Su M-Y: Quantitative analysis of lesion morphology and texture features for diagnostic prediction in breast MRI. Academic radiology 15:1513-1525, 2008
104. Lin T-Y, Goyal P, Girshick R, He K, Dollár P: Focal loss for dense object detection. IEEE transactions on pattern analysis and machine intelligence, 2018
105. Kingma D, Ba J: Adam: A method for stochastic optimization. arXiv preprint arXiv:14126980, 2014
106. Abadi M, et al.: TensorFlow: A System for Large-Scale Machine Learning. Proc. OSDI: City
107. Bejnordi BE, et al.: Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. Jama 318:2199-2210, 2017
108. Rajpurkar P, et al.: CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. arXiv preprint arXiv:171105225, 2017
109. Wang J, Fang Z, Lang N, Yuan H, Su M-Y, Baldi P: A multi-resolution approach for spinal metastasis detection using deep Siamese neural networks. Computers in Biology and Medicine 84:137-146, 2017
110. Bahl M, Baker JA, Bhargavan-Chatfield M, Brandt EK, Ghate SV: Impact of breast density notification legislation on radiologists' practices of reporting breast density: a multi-state study. Radiology 280:701-706, 2016
111. Wernli KJ, et al.: Patterns of breast magnetic resonance imaging use in community practice. JAMA internal medicine 174:125-132, 2014

112. Brandt KR, et al.: Comparison of clinical and automated breast density measurements: implications for risk prediction and supplemental screening. Radiology 279:710-719, 2015
113. Ram S, Sarma N, López JE, Liu Y, Li C-S, Aminololama-Shakeri S: Impact of the California Breast Density Law on Screening Breast MR Utilization, Provider Ordering Practices, and Patient Demographics. Journal of the American College of Radiology 15:594-600, 2018
114. Kuhl CK, Schrading S, Strobel K, Schild HH, Hilgers R-D, Bieling HB: Abbreviated breast magnetic resonance imaging (MRI): first postcontrast subtracted images and maximum-intensity projection—a novel approach to breast cancer screening with MRI. Journal of Clinical Oncology 32:2304-2310, 2014
115. Kerlikowske K, et al.: Combining quantitative and qualitative breast density measures to assess breast cancer risk. Breast Cancer Research 19:97, 2017
116. Lundberg FE, et al.: Association of infertility and fertility treatment with mammographic density in a large screening-based cohort of women: a cross-sectional study. Breast Cancer Research 18:36, 2016
117. Chen J-H, et al.: Reduction of breast density following tamoxifen treatment evaluated by 3-D MRI: preliminary study. Magnetic resonance imaging 29:91-98, 2011
118. Lin M, Chen JH, Wang X, Chan S, Chen S, Su MY: Template-based automatic breast segmentation on MRI by excluding the chest region. Medical physics 40, 2013
119. Petridou E, et al.: Breast fat volume measurement using wide-bore 3 T MRI: comparison of traditional mammographic density evaluation with MRI density measurements using automatic segmentation. Clinical radiology 72:565-572, 2017
120. Ribes S, et al.: Automatic segmentation of breast MR images through a Markov random field statistical model. IEEE transactions on medical imaging 33:1986-1996, 2014
121. Doran SJ, et al.: Breast MRI segmentation for density estimation: Do different methods give the same results and how much do differences matter? Medical physics 44:4573-4592, 2017
122. Chen H, Dou Q, Yu L, Qin J, Heng P-A: VoxResNet: Deep voxelwise residual networks for brain segmentation from 3D MR images. NeuroImage, 2017
123. Moeskops P, et al.: Evaluation of a deep learning approach for the segmentation of brain tissues and white matter hyperintensities of presumed vascular origin in MRI. NeuroImage: Clinical 17:251-262, 2018
124. Tong N, Gou S, Yang S, Ruan D, Sheng K: Fully Automatic Multi-Organ Segmentation for Head and Neck Cancer Radiotherapy Using Shape Representation Model Constrained Fully Convolutional Neural Networks. Medical physics, 2018
125. Commandeur F, et al.: Deep learning for quantification of epicardial and thoracic adipose tissue from non-contrast CT. IEEE Transactions on Medical Imaging, 2018
126. Oktay O, et al.: Anatomically constrained neural networks (ACNNs): application to cardiac image enhancement and segmentation. IEEE transactions on medical imaging 37:384-395, 2018
127. He K, Cao X, Shi Y, Nie D, Gao Y, Shen D: Pelvic Organ Segmentation Using Distinctive Curve Guided Fully Convolutional Networks. IEEE transactions on medical imaging, 2018

128. Gibson E, et al.: Automatic multi-organ segmentation on abdominal CT with dense v-networks. IEEE Transactions on Medical Imaging, 2018
129. Lu F, Wu F, Hu P, Peng Z, Kong D: Automatic 3D liver location and segmentation via convolutional neural network and graph cut. International journal of computer assisted radiology and surgery 12:171-182, 2017
130. Ha R, et al.: Fully Automated Convolutional Neural Network Method for Quantification of Breast MRI Fibroglandular Tissue and Background Parenchymal Enhancement. Journal of digital imaging:1-7, 2018
131. Dalmış MU, et al.: Using deep learning to segment breast and fibroglandular tissue in MRI volumes. Medical physics 44:533-546, 2017
132. Kallenberg M, et al.: Unsupervised deep learning applied to breast density segmentation and mammographic risk scoring. IEEE transactions on medical imaging 35:1322-1331, 2016
133. Zhou Z, Zhao G, Kijowski R, Liu F: Deep convolutional neural network for segmentation of knee joint anatomy. Magnetic resonance in medicine, 2018
134. Trebeschi S, et al.: Deep learning for fully-automated localization and segmentation of rectal cancer on multiparametric MR. Scientific reports 7:5301, 2017
135. Ronneberger O, Fischer P, Brox T: U-net: Convolutional networks for biomedical image segmentation. Proc. International Conference on Medical image computing and computer-assisted intervention: City
136. Shin GW, et al.: Role of dynamic contrast-enhanced MRI in evaluating the association between contralateral parenchymal enhancement and survival outcome in ER-positive, HER2-negative, node-negative invasive breast cancer. Journal of Magnetic Resonance Imaging, 2018
137. Chen JH, et al.: Consistency of breast density measured from the same women in four different MR scanners. Medical physics 39:4886-4895, 2012
138. Lin M, et al.: A new bias field correction method combining N3 and FCM for improved segmentation of breast density on MRI. Medical physics 38:5-14, 2011
139. Nie K, et al.: Development of a quantitative method for analysis of breast density based on three-dimensional breast MRI. Medical physics 35:5253-5262, 2008
140. He K, Zhang X, Ren S, Sun J: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. Proc. Proceedings of the IEEE international conference on computer vision: City
141. Zou KH, et al.: Statistical validation of image segmentation quality based on a spatial overlap index1: scientific reports. Academic radiology 11:178-189, 2004
142. Zhang Y, et al.: Automatic Breast and Fibroglandular Tissue Segmentation in Breast MRI Using Deep Learning by a Fully-Convolutional Residual Neural Network U-Net. Academic radiology, 2019
143. Chen J-H, et al.: Decrease in breast density in the contralateral normal breast of patients receiving neoadjuvant chemotherapy: MR imaging evaluation 1. Radiology 255:44-52, 2010
144. Hennessey S, et al.: Bilateral symmetry of breast tissue composition by magnetic resonance in young women and adults. Cancer Causes & Control 25:491-497, 2014
145. Chen J-H, Zhang Y, Chan S, Chang R-F, Su M-Y: Quantitative analysis of peri-tumor fat in different molecular subtypes of breast cancer. Magnetic resonance imaging 53:34-39, 2018

146. Pujara AC, et al.: Comparison between qualitative and quantitative assessment of background parenchymal enhancement on breast MRI. Journal of Magnetic Resonance Imaging 47:1685-1691, 2018
147. Jung Y, Jeong SK, Kang DK, Moon Y, Kim TH: Quantitative analysis of background parenchymal enhancement in whole breast on MRI: Influence of menstrual cycle and comparison with a qualitative analysis. European journal of radiology 103:84-89, 2018
148. Hu X, Jiang L, Li Q, Gu Y: Quantitative assessment of background parenchymal enhancement in breast magnetic resonance images predicts the risk of breast cancer. Oncotarget 8:10620, 2017
149. King V, Brooks JD, Bernstein JL, Reiner AS, Pike MC, Morris EA: Background parenchymal enhancement at breast MR imaging and breast cancer risk. Radiology 260:50-60, 2011
150. Dontchos BN, et al.: Are qualitative assessments of background parenchymal enhancement, amount of fibroglandular tissue on MR images, and mammographic density associated with breast cancer risk? Radiology 276:371-380, 2015
151. Wang Y, Morrell G, Heibrun ME, Payne A, Parker DL: 3D multi-parametric breast MRI segmentation using hierarchical support vector machine with coil sensitivity correction. Academic radiology 20:137-147, 2013
152. Fashandi H, Kuling G, Lu Y, Wu H, Martel AL: An investigation of the effect of fat suppression and dimensionality on the accuracy of breast MRI segmentation using U-nets. Medical physics 46:1230-1244, 2019
153. Ivanovska T, Jentschke TG, Daboul A, Hegenscheid K, Völzke H, Wörgötter F: A deep learning framework for efficient analysis of breast volume and fibroglandular tissue using MR data with strong artifacts. International journal of computer assisted radiology and surgery:1-7, 2019
154. Chang DHE, et al.: Comparison of breast density measured on MR images acquired using fat-suppressed versus nonfat-suppressed sequences. Medical physics 38:5961-5968, 2011
155. Harvey JA, Hendrick RE, Coll JM, Nicholson BT, Burkholder BT, Cohen MA: Breast MR imaging artifacts: how to recognize and fix them. Radiographics 27:S131-S145, 2007
156. Clauser P, et al.: A survey by the European Society of Breast Imaging on the utilisation of breast MRI in clinical practice. European radiology 28:1909-1918, 2018
157. McCormack VA, dos Santos Silva I: Breast density and parenchymal patterns as markers of breast cancer risk: a meta-analysis. Cancer Epidemiology and Prevention Biomarkers 15:1159-1169, 2006
158. Boyd N, et al.: Breast-tissue composition and other risk factors for breast cancer in young women: a cross-sectional study. The lancet oncology 10:569-580, 2009
159. Chen J-H, et al.: Decrease in breast density in the contralateral normal breast of patients receiving neoadjuvant chemotherapy: MR imaging evaluation. Radiology 255:44-52, 2010
160. Vovk U, Pernus F, Likar B: A review of methods for correction of intensity inhomogeneity in MRI. IEEE transactions on medical imaging 26:405-421, 2007
161. Zhou L, et al.: A method of radio-frequency inhomogeneity correction for brain tissue segmentation in MRI. Computerized Medical Imaging and Graphics 25:379-389, 2001

162. Committee WE: Statement on the second meeting of the International Health Regulations (2005) Emergency Committee regarding the outbreak of novel coronavirus (COVID-19). Geneva: WHO, 2020, 2005
163. Organization WH: WHO Director-General's opening remarks at the media briefing on COVID-19-11 March 2020, 2020
164. Huang C, et al.: Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. The lancet 395:497-506, 2020
165. Xu Z, et al.: Pathological findings of COVID-19 associated with acute respiratory distress syndrome. The Lancet respiratory medicine 8:420-422, 2020
166. MacMahon H, et al.: Guidelines for management of incidental pulmonary nodules detected on CT images: from the Fleischner Society 2017. Radiology 284:228-243, 2017
167. Dong E, Du H, Gardner L: An interactive web-based dashboard to track COVID-19 in real time. The Lancet infectious diseases 20:533-534, 2020
168. Huang L, et al.: Serial quantitative chest ct assessment of covid-19: Deep-learning approach. Radiology: Cardiothoracic Imaging 2:e200075, 2020
169. Shi F, et al.: Review of artificial intelligence techniques in imaging data acquisition, segmentation and diagnosis for covid-19. IEEE reviews in biomedical engineering, 2020
170. Rajinikanth V, Dey N, Raj ANJ, Hassanien AE, Santosh K, Raja N: Harmony-search and otsu based system for coronavirus disease (COVID-19) detection using lung CT scan images. arXiv preprint arXiv:200403431, 2020
171. Fan D-P, et al.: Inf-Net: Automatic COVID-19 Lung Infection Segmentation from CT Images. IEEE Transactions on Medical Imaging, 2020
172. Chan TF, Vese LA: Active contours without edges. IEEE Transactions on image processing 10:266-277, 2001
173. Kroon D-J, Slump CH: MRI modalitiy transformation in demon registration. Proc. Biomedical Imaging: From Nano to Macro, 2009 ISBI'09 IEEE International Symposium on: City
174. Thirion J-P: Non-rigid matching using demons. Proc. Computer Vision and Pattern Recognition, 1996 Proceedings CVPR'96, 1996 IEEE Computer Society Conference on: City
175. Prokop M, et al.: CO-RADS–A categorical CT assessment scheme for patients with suspected COVID-19: definition and evaluation. Radiology, 2020
176. Zhou J, et al.: Diagnosis of benign and malignant breast lesions on DCE-MRI by using radiomics and deep learning with consideration of peritumor tissue. Journal of Magnetic Resonance Imaging, 2019
177. Lang N, Su M-Y, Hon JY, Lin M, Hamamura MJ, Yuan H: Differentiation of myeloma and metastatic cancer in the spine using dynamic contrast-enhanced MRI. Magnetic resonance imaging 31:1285-1291, 2013
178. Marino MA, Helbich T, Baltzer P, Pinker-Domenig K: Multiparametric MRI of the breast: A review. Journal of Magnetic Resonance Imaging 47:301-315, 2018
179. Mann RM, Kuhl CK, Moy L: Contrast-enhanced MRI for breast cancer screening. Journal of Magnetic Resonance Imaging 50:377-390, 2019
180. Hill DA, et al.: Utilization of breast cancer screening with magnetic resonance imaging in community practice. Journal of general internal medicine 33:275-283, 2018

181. Gilhuijs KG, Giger ML, Bick U: Computerized analysis of breast lesions in three dimensions using dynamic magnetic-resonance imaging. Medical physics 25:1647-1654, 1998
182. Gweon HM, Cho N, Seo M, Chu AJ, Moon WK: Computer-aided evaluation as an adjunct to revised BI-RADS Atlas: improvement in positive predictive value at screening breast MRI. European radiology 24:1800-1807, 2014
183. Newell D, et al.: Selection of diagnostic features on breast MRI to differentiate between malignant and benign lesions using computer-aided diagnosis: differences in lesions presenting as mass and non-mass-like enhancement. European radiology 20:771-781, 2010
184. Gallego-Ortiz C, Martel AL: Improving the accuracy of computer-aided diagnosis for breast MR imaging by differentiating between mass and nonmass lesions. Radiology 278:679-688, 2015
185. Tsougos I, Vamvakas A, Kappas C, Fezoulidis I, Vassiou K: Application of radiomics and decision support systems for breast MR differential diagnosis. Computational and mathematical methods in medicine 2018, 2018
186. Chitalia RD, Kontos D: Role of texture analysis in breast MRI as a cancer biomarker: A review. Journal of Magnetic Resonance Imaging 49:927-938, 2019
187. Li H, et al.: Quantitative MRI radiomics in the prediction of molecular classifications of breast cancer subtypes in the TCGA/TCIA data set. NPJ breast cancer 2:16012, 2016
188. Liang C, et al.: An MRI-based radiomics classifier for preoperative prediction of Ki-67 status in breast cancer. Academic radiology 25:1111-1117, 2018
189. Arevalo J, González FA, Ramos-Pollán R, Oliveira JL, Lopez MAG: Representation learning for mammography mass lesion classification with convolutional neural networks. Computer methods and programs in biomedicine 127:248-257, 2016
190. Becker AS, Marcon M, Ghafoor S, Wurnig MC, Frauenfelder T, Boss A: Deep learning in mammography: diagnostic accuracy of a multipurpose image analysis software in the detection of breast cancer. Investigative radiology 52:434-440, 2017
191. Chougrad H, Zouaki H, Alheyane O: Deep Convolutional Neural Networks for Breast Cancer Screening. Computer Methods and Programs in Biomedicine, 2018
192. Diniz JOB, Diniz PHB, Valente TLA, Silva AC, de Paiva AC, Gattass M: Detection of mass regions in mammograms by bilateral analysis adapted to breast density using similarity indexes and convolutional neural networks. Computer Methods and Programs in Biomedicine, 2018
193. Antropova N, Abe H, Giger ML: Use of clinical MRI maximum intensity projections for improved breast lesion classification with deep convolutional neural networks. Journal of Medical Imaging 5:014503, 2018
194. Truhn D, Schrading S, Haarburger C, Schneider H, Merhof D, Kuhl C: Radiomic versus Convolutional Neural Networks Analysis for Classification of Contrast-enhancing Lesions at Multiparametric Breast MRI. Radiology:181352, 2018
195. Kim Y, Stolarska MA, Othmer HG: The role of the microenvironment in tumor growth and invasion. Progress in biophysics and molecular biology 106:353-379, 2011
196. Wu J-s, Sheng S-r, Liang X-h, Tang Y-l: The role of tumor microenvironment in collective tumor cell invasion. Future Oncology 13:991-1002, 2017

197. Lee A, DeLellis RA, Silverman ML, Heatley GJ, Wolfe HJ: Prognostic significance of peritumoral lymphatic and blood vessel invasion in node-negative carcinoma of the breast. Journal of Clinical Oncology 8:1457-1465, 1990
198. Mohammed ZM, et al.: The relationship between lymphovascular invasion and angiogenesis, hormone receptors, cell proliferation and survival in patients with primary operable invasive ductal breast cancer. BMC clinical pathology 13:31, 2013
199. Freed M, et al.: Evaluation of breast lipid composition in patients with benign tissue and cancer by using multiple gradient-echo MR imaging. Radiology 281:43-53, 2016
200. Shin HJ, et al.: Characterization of tumor and adjacent peritumoral stroma in patients with breast cancer using high-resolution diffusion-weighted imaging: Correlation with pathologic biomarkers. European journal of radiology 85:1004-1011, 2016
201. Cheon H, et al.: Invasive breast cancer: Prognostic value of peritumoral edema identified at preoperative MR imaging. Radiology 287:68-75, 2018
202. Chai H, Brown RE: Field effect in cancer–an update. Annals of Clinical & Laboratory Science 39:331-337, 2009
203. Heaphy CM, Griffith JK, Bisoffi M: Mammary field cancerization: molecular evidence and clinical importance. Breast cancer research and treatment 118:229-239, 2009
204. Fan M, He T, Zhang P, Zhang J, Li L: Heterogeneity of diffusion-weighted imaging in tumours and the surrounding stroma for prediction of Ki-67 proliferation status in breast cancer. Scientific reports 7:2875, 2017
205. Whitney HM, et al.: Additive benefit of radiomics over size alone in the distinction between benign lesions and luminal A cancers on a large clinical breast MRI dataset. Academic radiology 26:202-209, 2019
206. Antropova N, Huynh B, Li H, Giger ML: Breast lesion classification based on dynamic contrast-enhanced magnetic resonance images sequences with long short-term memory networks. Journal of Medical Imaging 6:011002, 2018
207. Reig B, Heacock L, Geras KJ, Moy L: Machine learning in breast MRI. Journal of Magnetic Resonance Imaging, 2019
208. Sandhu R, Parker JS, Jones WD, Livasy CA, Coleman WB: Microarray-based gene expression profiling for molecular classification of breast cancer and identification of new targets for therapy. Laboratory Medicine 41:364-372, 2010
209. Houssami N, Turner RM, Morrow M: Meta-analysis of pre-operative magnetic resonance imaging (MRI) and surgical treatment for breast cancer. Breast cancer research and treatment 165:273-283, 2017
210. Agner SC, et al.: Computerized image analysis for identifying triple-negative breast cancers and differentiating them from other molecular subtypes of breast cancer on dynamic contrast-enhanced MR images: a feasibility study. Radiology 272:91-99, 2014
211. Chang R-F, Chen H-H, Chang Y-C, Huang C-S, Chen J-H, Lo C-M: Quantification of breast tumor heterogeneity for ER status, HER2 status, and TN molecular subtype evaluation on DCE-MRI. Magnetic resonance imaging 34:809-819, 2016
212. Sutton EJ, et al.: Breast cancer molecular subtype classifier that incorporates MRI features. Journal of Magnetic Resonance Imaging 44:122-129, 2016
213. Fan M, Li H, Wang S, Zheng B, Zhang J, Li L: Radiomic analysis reveals DCE-MRI features for prediction of molecular subtypes of breast cancer. PloS one 12:e0171683, 2017

214. Braman NM, et al.: Intratumoral and peritumoral radiomics for the pretreatment prediction of pathological complete response to neoadjuvant chemotherapy based on breast DCE-MRI. Breast Cancer Research 19:57, 2017
215. Ma W, et al.: Breast Cancer Molecular Subtype Prediction by Mammographic Radiomic Features. Academic radiology, 2018
216. Shi L, et al.: Machine learning for prediction of chemoradiation therapy response in rectal cancer using pre-treatment and mid-radiation multi-parametric MRI. Magnetic resonance imaging 61:33-40, 2019
217. Zhu Z, Albadawy E, Saha A, Zhang J, Harowicz MR, Mazurowski MA: Deep learning for identifying radiogenomic associations in breast cancer. Computers in biology and medicine 109:85-90, 2019
218. Xie T, et al.: Machine learning-based analysis of MR multiparametric radiomics for the subtype classification of breast cancer. Frontiers in oncology 9:505, 2019
219. Ha R, et al.: Predicting breast cancer molecular subtype with MRI dataset utilizing convolutional neural network algorithm. Journal of digital imaging 32:276-282, 2019
220. Chang P, et al.: Deep-Learning Convolutional Neural Networks Accurately Classify Genetic Mutations in Gliomas. American Journal of Neuroradiology, 2018
221. Michael KY, Ma J, Fisher J, Kreisberg JF, Raphael BJ, Ideker T: Visible Machine Learning for Biomedicine. Cell 173:1562-1565, 2018
222. Couture HD, et al.: Image analysis with deep learning to predict breast cancer grade, ER status, histologic subtype, and intrinsic subtype. NPJ breast cancer 4:1-8, 2018
223. Jaber MI, et al.: A deep learning image-based intrinsic molecular subtype classifier of breast tumors reveals tumor heterogeneity that may affect survival. Breast Cancer Research 22:12, 2020
224. Nishio M, et al.: Computer-aided diagnosis of lung nodule classification between benign nodule, primary lung cancer, and metastatic lung cancer at different image size using deep convolutional neural network with transfer learning. PloS one 13, 2018
225. Byra M, et al.: Breast mass classification in sonography with transfer learning using a deep convolutional neural network and color conversion. Medical physics 46:746-755, 2019
226. Ortiz-Ramón R, Larroza A, Ruiz-España S, Arana E, Moratal D: Classifying brain metastases by their primary site of origin using a radiomics approach based on texture analysis: a feasibility study. European radiology 28:4514-4523, 2018
227. Avellino AM, et al.: The misdiagnosis of acute cervical spine injuries and fractures in infants and children: the 12-year experience of a level I pediatric and adult trauma center. Child's Nervous System 21:122-127, 2005
228. Casez P, Uebelhart B, Gaspoz J-M, Ferrari S, Louis-Simonet M, Rizzoli R: Targeted education improves the very low recognition of vertebral fractures and osteoporosis management by general internists. Osteoporosis international 17:965-970, 2006
229. Goradia D, Blackmore CC, Talner LB, Bittles M, Meshberg E: Predicting Radiology Resident Errors in Diagnosis of Cervical Spine Fractures1. Academic radiology 12:888-893, 2005
230. Li Y, Yan L, Cai S, Wang P, Zhuang H, Yu H: The prevalence and under-diagnosis of vertebral fractures on chest radiograph. BMC musculoskeletal disorders 19:235, 2018

231. Zhou AL, Bonham LW, Verde F: Comparative analysis of body radiologist to neuroradiologist evaluation of the spine in trauma settings. Journal of the American College of Radiology 15:1687-1691, 2018
232. Schwaiger BJ, Gersing AS, Baum T, Krestan CR, Kirschke JS: Distinguishing benign and malignant vertebral fractures using CT and MRI. Proc. Seminars in musculoskeletal radiology: City
233. Baker LL, Goodman SB, Perkash I, Lane B, Enzmann DR: Benign versus pathologic compression fractures of vertebral bodies: assessment with conventional spin-echo, chemical-shift, and STIR MR imaging. Radiology 174:495-502, 1990
234. Diacinti D, et al.: Misdiagnosis of vertebral fractures on local radiographic readings of the multicentre POINT (Prevalence of Osteoporosis in INTernal medicine) study. Bone 101:230-235, 2017
235. Genant H, Jergas M: Assessment of prevalent and incident vertebral fractures in osteoporosis research. Osteoporosis International 14:43-55, 2003
236. Burns JE, Yao J, Summers RM: Vertebral body compression fractures and bone density: automated detection and classification on CT images. Radiology 284:788-797, 2017
237. Takigawa T, Tanaka M, Sugimoto Y, Tetsunaga T, Nishida K, Ozaki T: Discrimination between malignant and benign vertebral fractures using magnetic resonance imaging. Asian spine journal 11:478, 2017
238. Chung SW, et al.: Automated detection and classification of the proximal humerus fracture by using deep learning algorithm. Acta orthopaedica 89:468-473, 2018
239. Olczak J, et al.: Artificial intelligence for analyzing orthopedic trauma radiographs: deep learning algorithms—are they on par with humans for diagnosing fractures? Acta orthopaedica 88:581-586, 2017
240. Kitamura G, Chung CY, Moore BE: Ankle fracture detection utilizing a convolutional neural network ensemble implemented with a small sample, de novo training, and multiview incorporation. Journal of digital imaging 32:672-677, 2019
241. Raghavendra U, Bhat NS, Gudigar A, Acharya UR: Automated system for the detection of thoracolumbar fractures using a CNN architecture. Future Generation Computer Systems 85:184-189, 2018
242. Tomita N, Cheung YY, Hassanpour S: Deep neural networks for automatic detection of osteoporotic vertebral fractures on CT scans. Computers in biology and medicine 98:8-15, 2018
243. Pedoia V, Lee J, Norman B, Link T, Majumdar S: Diagnosing osteoarthritis from T2 maps using deep learning: an analysis of the entire Osteoarthritis Initiative baseline cohort. Osteoarthritis and cartilage 27:1002-1010, 2019
244. Chen H, et al.: Automatic localization and identification of vertebrae in spine CT via a joint learning model with deep neural networks. Proc. International conference on medical image computing and computer-assisted intervention: City
245. Whitehead W, Moran S, Gaonkar B, Macyszyn L, Iyer S: A deep learning approach to spine segmentation using a feed-forward chain of pixel-wise convolutional networks. Proc. 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018): City
246. Sekuboyina A, Kukačka J, Kirschke JS, Menze BH, Valentinitsch A: Attention-driven deep learning for pathological spine segmentation. Proc. International Workshop and

Challenge on Computational Methods and Clinical Applications in Musculoskeletal Imaging: City
247. Lu J-T, et al.: Deepspine: Automated lumbar vertebral segmentation, disc-level designation, and spinal stenosis grading using deep learning. arXiv preprint arXiv:180710215, 2018
248. Sciubba DM, Gokaslan ZL: Diagnosis and management of metastatic spine disease. Surgical oncology 15:141-151, 2006
249. Robson P: Metastatic spinal cord compression: a rare but important complication of cancer. Clinical medicine 14:542, 2014
250. Piccioli A, Maccauro G, Spinelli MS, Biagini R, Rossi B: Bone metastases of unknown origin: epidemiology and principles of management. Journal of Orthopaedics and Traumatology 16:81-86, 2015
251. Zha Y, Li M, Yang J: Dynamic contrast enhanced magnetic resonance imaging of diffuse spinal bone marrow infiltration in patients with hematological malignancies. Korean journal of radiology 11:187-194, 2010
252. Khadem N, et al.: Characterizing hypervascular and hypovascular metastases and normal bone marrow of the spine using dynamic contrast-enhanced MR imaging. American journal of neuroradiology 33:2178-2185, 2012
253. Dutoit JC, Vanderkerken MA, Verstraete KL: Value of whole body MRI and dynamic contrast enhanced MRI in the diagnosis, follow-up and evaluation of disease activity and extent in multiple myeloma. European journal of radiology 82:1444-1452, 2013
254. Saha A, Peck KK, Lis E, Holodny AI, Yamada Y, Karimi S: Magnetic resonance perfusion characteristics of hypervascular renal and hypovascular prostate spinal metastases: clinical utilities and implications. Spine 39:E1433, 2014
255. Lang N, Su M-Y, Hon JY, Yuan H: Differentiation of tuberculosis and metastatic cancer in the spine using dynamic contrast-enhanced MRI. European Spine Journal 24:1729-1737, 2015
256. Lang N, Su MY, Xing X, Yu HJ, Yuan H: Morphological and dynamic contrast enhanced MR imaging features for the differentiation of chordoma and giant cell tumors in the axial skeleton. Journal of Magnetic Resonance Imaging 45:1068-1075, 2017
257. Lang N, Yuan H, Hon JY, Su M-Y: Diagnosis of spinal lesions using heuristic and pharmacokinetic parameters measured by dynamic contrast-enhanced MRI. Academic radiology 24:867-875, 2017
258. Aerts HJ, et al.: Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nature communications 5:1-9, 2014
259. Lambin P, et al.: Radiomics: the bridge between medical imaging and personalized medicine. Nature reviews Clinical oncology 14:749, 2017
260. Antropova N, Huynh BQ, Giger ML: A deep feature fusion methodology for breast cancer diagnosis demonstrated on three imaging modality datasets. Medical physics 44:5162-5171, 2017
261. Le MH, et al.: Automated diagnosis of prostate cancer in multi-parametric MRI based on multimodal convolutional neural networks. Physics in Medicine & Biology 62:6497, 2017
262. Chang PD, Chow DS, Yang PH, Filippi CG, Lignelli A: Predicting glioblastoma recurrence by early changes in the apparent diffusion coefficient value and signal intensity on FLAIR images. American Journal of Roentgenology 208:57-65, 2017

263. Zhu Y, et al.: MRI-based prostate cancer detection with high-level representation and hierarchical classification. Medical physics 44:1028-1039, 2017
264. Huang S-H, Chu Y-H, Lai S-H, Novak CL: Learning-based vertebra detection and iterative normalized-cut segmentation for spinal MRI. IEEE transactions on medical imaging 28:1595-1605, 2009
265. Baldi P, Sadowski P: The dropout learning algorithm. Artificial intelligence 210:78-122, 2014
266. Erlemann R, et al.: Musculoskeletal neoplasms: static and dynamic Gd-DTPA--enhanced MR imaging. Radiology 171:767-773, 1989
267. Hermann G, Abdelwahab IF, Miller TT, Klein MJ, Lewis MM: Tumour and tumour-like conditions of the soft tissue: magnetic resonance imaging features differentiating benign from malignant masses. The British journal of radiology 65:14-20, 1992
268. Moulton J, Blebea J, Dunco D, Braley S, Bisset 3rd G, Emery K: MR imaging of soft-tissue masses: diagnostic efficacy and value of distinguishing between benign and malignant lesions. AJR American journal of roentgenology 164:1191-1199, 1995
269. May DA, Good RB, Smith D, Parsons TW: MR imaging of musculoskeletal tumors and tumor mimickers with intravenous gadolinium: experience with 242 patients. Skeletal radiology 26:2-15, 1997
270. Kim HJ, Ryu KN, Choi WS, Choi BK, Choi JM, Yoon Y: Spinal involvement of hematopoietic malignancies and metastasis: differentiation using MR imaging. Clinical imaging 23:125-133, 1999
271. Lang P, et al.: Musculoskeletal neoplasm: perineoplastic edema versus tumor on dynamic postcontrast MR images with spatial mapping of instantaneous enhancement rates. Radiology 197:831-839, 1995
272. Moulopoulos L, Maris T, Papanikolaou N, Panagi G, Vlahos L, Dimopoulos M: Detection of malignant bone marrow involvement with dynamic contrast-enhanced magnetic resonance imaging. Annals of oncology 14:152-158, 2003
273. Kato S, Hozumi T, Takaki Y, Yamakawa K, Goto T, Kondo T: Optimal schedule of preoperative embolization for spinal metastasis surgery. Spine 38:1964-1969, 2013
274. Qiao Z, Jia N, He Q: Does preoperative transarterial embolization decrease blood loss during spine tumor surgery? Interventional neuroradiology 21:129-135, 2015
275. Jiang L, et al.: Surgical treatment options for aggressive osteoblastoma in the mobile spine. European Spine Journal 24:1778-1785, 2015
276. Nie K, et al.: Rectal cancer: assessment of neoadjuvant chemoradiation outcome based on radiomics of multiparametric MRI. Clinical cancer research 22:5256-5264, 2016
277. Wang G, He L, Yuan C, Huang Y, Liu Z, Liang C: Pretreatment MR imaging radiomics signatures for response prediction to induction chemotherapy in patients with nasopharyngeal carcinoma. European journal of radiology 98:100-106, 2018
278. Antropova NO, Abe H, Giger ML: Use of clinical MRI maximum intensity projections for improved breast lesion classification with deep convolutional neural networks. Journal of Medical Imaging 5:014503, 2018
279. Siegel RL, Miller KD, Jemal A: Cancer statistics, 2018. CA: a cancer journal for clinicians 68:7-30, 2018
280. Weinreb JC, et al.: PI-RADS prostate imaging–reporting and data system: 2015, version 2. European urology 69:16-40, 2016

281. Herold CJ, et al.: Imaging in the age of precision medicine: summary of the proceedings of the 10th Biannual Symposium of the International Society for Strategic Studies in Radiology. Radiology 279:226-238, 2016
282. Venderbos LD, Roobol MJ: PSA-based prostate cancer screening: the role of active surveillance and informed and shared decision making. Asian journal of andrology 13:219, 2011
283. Guyon I, Elisseeff A: An introduction to variable and feature selection. Journal of machine learning research 3:1157-1182, 2003
284. Siewerdsen JH, Moseley D, Bakhtiar B, Richard S, Jaffray DA: The influence of antiscatter grids on soft-tissue detectability in cone-beam computed tomography with flat-panel detectors: Antiscatter grids in cone-beam CT. Medical physics 31:3506-3520, 2004
285. Cai W, Ning R, Conover D: Scatter correction using beam stop array algorithm for cone-beam CT breast imaging: SPIE, 2006
286. Wang J, Li T, Xing L: Iterative image reconstruction for CBCT using edge-preserving prior. Medical Physics 36:252-260, 2009
287. Jia X, Yan H Fau - Cervino L, Cervino L Fau - Folkerts M, Folkerts M Fau - Jiang SB, Jiang SB: A GPU tool for efficient, accurate, and realistic simulation of cone beam CT projections. Med Phys 39(12):7368-7378, 2012
288. Zbijewski W, Beekman FJ: Efficient Monte Carlo based scatter artifact reduction in cone-beam micro-CT. IEEE transactions on medical imaging 25:817-827, 2006
289. Xu Y, et al.: A practical cone-beam CT scatter correction method with optimized Monte Carlo simulations for image-guided radiation therapy. Physics in Medicine & Biology 60:3567, 2015
290. Maspero M, et al.: Dose evaluation of fast synthetic-CT generation using a generative adversarial network for general pelvis MR-only radiotherapy. Phys Med Biol 63:185001, 2018
291. Liu Y, et al.: MRI-based treatment planning for liver stereotactic body radiotherapy: validation of a deep learning-based synthetic CT generation method. Br J Radiol 92:20190067, 2019
292. Isola P, Zhu J-Y, Zhou T, Efros AA: Image-to-image translation with conditional adversarial networks. Proc. Proceedings of the IEEE conference on computer vision and pattern recognition: City
293. Mroueh Y, Sercu T, Goel V: Mcgan: Mean and covariance feature matching gan. arXiv preprint arXiv:170208398, 2017
294. Salimans T, Goodfellow I, Zaremba W, Cheung V, Radford A, Chen X: Improved techniques for training gans. Proc. Advances in neural information processing systems: City
295. Johnson J, Alahi A, Fei-Fei L: Perceptual losses for real-time style transfer and super-resolution. Proc. European conference on computer vision: City
296. Simonyan K, Zisserman A: Very Deep Convolutional Networks for Large-Scale Image Recognition. CoRR abs/1409.1556, 2015
297. Hansen DC, et al.: ScatterNet: A convolutional neural network for cone-beam CT intensity correction. Med Phys 45:4916-4926, 2018
298. Yuan N, et al.: Convolutional neural network enhancement of fast-scan low-dose cone-beam CT images for head and neck radiotherapy. Phys Med Biol 65:035003, 2020

299. Liang X, et al.: Generating synthesized computed tomography (CT) from cone-beam computed tomography (CBCT) using CycleGAN for adaptive radiation therapy. Phys Med Biol 64:125002, 2019
300. Zhu J-Y, Park T, Isola P, Efros AA: Unpaired image-to-image translation using cycle-consistent adversarial networks. Proc. Proceedings of the IEEE international conference on computer vision: City
301. Ioffe S, Szegedy C: Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:150203167, 2015
302. Yu H, Wang G: A soft-threshold filtering approach for reconstruction from a limited number of projections. Phys Med Biol 55:3905-3916, 2010
303. Kida S, et al.: Cone Beam Computed Tomography Image Quality Improvement Using a Deep Convolutional Neural Network. Cureus 10:e2548, 2018
304. Chen L, Liang X, Shen C, Jiang S, Wang J: Synthetic CT generation from CBCT images via deep learning. Med Phys 47:1115-1125, 2020
305. Li Y, et al.: A preliminary study of using a deep convolution neural network to generate synthesized CT images based on CBCT for adaptive radiotherapy of nasopharyngeal carcinoma. Phys Med Biol 64:145010, 2019
306. Lei Y, et al.: Male pelvic multi-organ segmentation aided by CBCT-based synthetic MRI. Phys Med Biol 65:035013, 2020
307. Liu Y, et al.: CBCT-based synthetic CT generation using deep-attention cycleGAN for pancreatic adaptive radiotherapy. Med Phys, 2020
308. Harms J, et al.: Paired cycle-GAN-based image correction for quantitative cone-beam computed tomography. Med Phys 46:3998-4009, 2019
309. Maas M, et al.: Long-term outcome in patients with a pathological complete response after chemoradiation for rectal cancer: a pooled analysis of individual patient data. The lancet oncology 11:835-844, 2010
310. Sanghera P, Wong D, McConkey CC, Geh J, Hartley A: Chemoradiotherapy for rectal cancer: an updated analysis of factors affecting pathological response. Clinical oncology 20:176-183, 2008
311. Habr-Gama A, et al.: Patterns of failure and survival for nonoperative treatment of stage c0 distal rectal cancer following neoadjuvant chemoradiation therapy. Journal of gastrointestinal surgery 10:1319-1329, 2006
312. Borschitz T, Wachtlin D, Möhler M, Schmidberger H, Junginger T: Neoadjuvant chemoradiation and local excision for T2-3 rectal cancer. Annals of surgical oncology 15:712-720, 2008
313. Renehan AG, et al.: Watch-and-wait approach versus surgical resection after chemoradiotherapy for patients with rectal cancer (the OnCoRe project): a propensity-score matched cohort analysis. The Lancet Oncology 17:174-183, 2016
314. Ludwig KA: Sphincter-sparing resection for rectal cancer. Clinics in colon and rectal surgery 20:203, 2007
315. Marijnen CA: Organ preservation in rectal cancer: have all questions been answered? The lancet oncology 16:e13-e22, 2015
316. Maas M, et al.: Assessment of clinical complete response after chemoradiation for rectal cancer with digital rectal examination, endoscopy, and MRI: selection for organ-saving treatment. Annals of surgical oncology 22:3873-3880, 2015

317. Lambrecht M, et al.: Value of diffusion-weighted magnetic resonance imaging for prediction and early assessment of response to neoadjuvant radiochemotherapy in rectal cancer: preliminary results. International Journal of Radiation Oncology* Biology* Physics 82:863-870, 2012
318. Lambregts DM, et al.: Diffusion-weighted MRI for selection of complete responders after chemoradiation for locally advanced rectal cancer: a multicenter study. Annals of surgical oncology 18:2224-2231, 2011
319. de Lussanet QG, et al.: Dynamic contrast-enhanced magnetic resonance imaging of radiation therapy-induced microcirculation changes in rectal cancer. International Journal of Radiation Oncology* Biology* Physics 63:1309-1315, 2005
320. Oberholzer K, et al.: Rectal cancer: Assessment of response to neoadjuvant chemoradiation by dynamic contrast-enhanced MRI. Journal of Magnetic Resonance Imaging 38:119-126, 2013
321. DeVries AF, et al.: Pretreatment evaluation of microcirculation by dynamic contrast-enhanced magnetic resonance imaging predicts survival in primary rectal cancer patients. International Journal of Radiation Oncology* Biology* Physics 90:1161-1167, 2014
322. Rengo M, et al.: Magnetic resonance tumor regression grade (MR-TRG) to assess pathological complete response following neoadjuvant radiochemotherapy in locally advanced rectal cancer. Oncotarget 8:114746, 2017
323. Aker M, Boone D, Chandramohan A, Sizer B, Motson R, Arulampalam T: Diagnostic accuracy of MRI in assessing tumor regression and identifying complete response in patients with locally advanced rectal cancer after neoadjuvant treatment. Abdominal Radiology 43:3213-3219, 2018
324. Zhang C, Ye F, Liu Y, Ouyang H, Zhao X, Zhang H: Morphologic predictors of pathological complete response to neoadjuvant chemoradiotherapy in locally advanced rectal cancer. Oncotarget 9:4862, 2018
325. Xu Q, et al.: Quantitative intravoxel incoherent motion parameters derived from whole-tumor volume for assessing pathological complete response to neoadjuvant chemotherapy in locally advanced rectal cancer. Journal of Magnetic Resonance Imaging 48:248-258, 2018
326. Cusumano D, et al.: Fractal-based radiomic approach to predict complete pathological response after chemo-radiotherapy in rectal cancer. La radiologia medica 123:286-295, 2018
327. Liu Z, et al.: Radiomics Analysis for Evaluation of Pathological Complete Response to Neoadjuvant Chemoradiotherapy in Locally Advanced Rectal Cancer. Clinical Cancer Research:clincanres. 1038.2017, 2017
328. Cha KH, et al.: Bladder cancer treatment response assessment in CT urography using two-channel deep-learning network. Proc. Medical Imaging 2018: Computer-Aided Diagnosis: City
329. Ypsilantis P-P, et al.: Predicting response to neoadjuvant chemotherapy with PET imaging using convolutional neural networks. PloS one 10:e0137036, 2015
330. Huynh BQ, Antropova N, Giger ML: Comparison of breast DCE-MRI contrast time points for predicting response to neoadjuvant chemotherapy using deep convolutional neural network features with transfer learning. Proc. Medical Imaging 2017: Computer-Aided Diagnosis: City

331. Ravichandran K, Braman N, Janowczyk A, Madabhushi A: A deep learning classifier for prediction of pathological complete response to neoadjuvant chemotherapy from baseline breast DCE-MRI. Proc. Medical Imaging 2018: Computer-Aided Diagnosis: City
332. Ryan R, et al.: Pathological response following long-course neoadjuvant chemoradiotherapy for locally advanced rectal cancer. Histopathology 47:141-146, 2005
333. Dossa F, Chesney TR, Acuna SA, Baxter NN: A watch-and-wait approach for locally advanced rectal cancer after a clinical complete response following neoadjuvant chemoradiation: a systematic review and meta-analysis. The lancet Gastroenterology & hepatology 2:501-513, 2017
334. Sammour T, Price BA, Krause KJ, Chang GJ: Nonoperative management or 'watch and wait'for rectal cancer with complete clinical response after neoadjuvant chemoradiotherapy: a critical appraisal. Annals of surgical oncology 24:1904-1915, 2017
335. van der Paardt MP, Zagers MB, Beets-Tan RG, Stoker J, Bipat S: Patients who undergo preoperative chemoradiotherapy for locally advanced rectal cancer restaged by using diagnostic MR imaging: a systematic review and meta-analysis. Radiology 269:101-112, 2013
336. Kim SH, et al.: Locally advanced rectal cancer: added value of diffusion-weighted MR imaging in the evaluation of tumor response to neoadjuvant chemo-and radiation therapy. Radiology 253:116-125, 2009
337. Curvo-Semedo L, et al.: Rectal cancer: assessment of complete response to preoperative combined radiation therapy with chemotherapy—conventional MR volumetry versus diffusion-weighted MR imaging. Radiology 260:734-743, 2011
338. Song I, Kim SH, Lee S, Choi J, Kim M, Rhim H: Value of diffusion-weighted imaging in the detection of viable tumour after neoadjuvant chemoradiation therapy in patients with locally advanced rectal cancer: comparison with T 2 weighted and PET/CT imaging. The British journal of radiology 85:577-586, 2012
339. Aghaei F, Tan M, Hollingsworth AB, Zheng B: Applying a new quantitative global breast MRI feature analysis scheme to assess tumor response to chemotherapy. Journal of Magnetic Resonance Imaging 44:1099-1106, 2016
340. Marinovich M, et al.: Early prediction of pathologic response to neoadjuvant therapy in breast cancer: systematic review of the accuracy of MRI. The Breast 21:669-677, 2012
341. Bibault J-E, et al.: Deep Learning and Radiomics predict complete response after neo-adjuvant chemoradiation for locally advanced rectal cancer. Scientific reports 8:1-8, 2018
342. Mansouri A, et al.: Surgically resected skull base meningiomas demonstrate a divergent postoperative recurrence pattern compared with non-skull base meningiomas. Journal of neurosurgery 125:431-440, 2016
343. Wiemels J, Wrensch M, Claus EB: Epidemiology and etiology of meningioma. J Neurooncol 99:307-314, 2010
344. Louis DN, et al.: The 2016 World Health Organization classification of tumors of the central nervous system: a summary. Acta neuropathologica 131:803-820, 2016

345. Perry A, Stafford SL, Scheithauer BW, Suman VJ, Lohse CM: Meningioma grading: an analysis of histologic parameters. The American journal of surgical pathology 21:1455-1465, 1997
346. Maillo A, et al.: Early recurrences in histologically benign/grade I meningiomas are associated with large tumors and coexistence of monosomy 14 and del(1p36) in the ancestral tumor cell clone. Neuro Oncol 9:438-446, 2007
347. Ildan F, et al.: Predicting the probability of meningioma recurrence in the preoperative and early postoperative period: a multivariate analysis in the midterm follow-up. Skull base : official journal of North American Skull Base Society [et al] 17:157-171, 2007
348. Nanda A, Vannemreddy P: Recurrence and outcome in skull base meningiomas: do they differ from other intracranial meningiomas? Skull base : official journal of North American Skull Base Society [et al] 18:243-252, 2008
349. Black PM, Villavicencio AT, Rhouddou C, Loeffler JS: Aggressive surgery and focal radiation in the management of meningiomas of the skull base: preservation of function with maintenance of local control. Acta neurochirurgica 143:555-562, 2001
350. Kreil W, Luggin J, Fuchs I, Weigl V, Eustacchio S, Papaefthymiou G: Long term experience of gamma knife radiosurgery for benign skull base meningiomas. Journal of neurology, neurosurgery, and psychiatry 76:1425-1430, 2005
351. Sekhar LN, Juric-Sekhar G, Brito da Silva H, Pridgeon JS: Skull Base Meningiomas: Aggressive Resection. Neurosurgery 62 Suppl 1:30-49, 2015
352. Escribano Mesa JA, et al.: Risk of Recurrence in Operated Parasagittal Meningiomas: A Logistic Binary Regression Model. World neurosurgery 110:e112-e118, 2018
353. Ko C-C: Applications of Diffusion-Weighted MR Imaging in Brain Tumors, 2018
354. Zhou M, et al.: Radiomics in Brain Tumor: Image Assessment, Quantitative Feature Descriptors, and Machine-Learning Approaches. AJNR Am J Neuroradiol 39:208-216, 2018
355. Gatenby RA, Grove O, Gillies RJ: Quantitative imaging in cancer evolution and ecology. Radiology 269:8-15, 2013
356. Kickingereder P, et al.: Radiomic Profiling of Glioblastoma: Identifying an Imaging Predictor of Patient Survival with Improved Performance over Established Clinical and Radiologic Risk Models. Radiology 280:880-889, 2016
357. Chaddad A, Sabri S, Niazi T, Abdulkarim B: Prediction of survival with multi-scale radiomic analysis in glioblastoma patients. Medical & biological engineering & computing, 2018
358. Rathore S, et al.: Radiomic MRI signature reveals three distinct subtypes of glioblastoma with different clinical and molecular characteristics, offering prognostic value beyond IDH1. Scientific reports 8:5087, 2018
359. Zhang S, et al.: Non-invasive radiomics approach potentially predicts non-functioning pituitary adenomas subtypes before surgery. European radiology 28:3692-3701, 2018
360. Jenkinson M, Bannister P, Brady M, Smith S: Improved optimization for the robust and accurate linear registration and motion correction of brain images. Neuroimage 17:825-841, 2002
361. Friedman J, Hastie T, Tibshirani R: The elements of statistical learning: Springer series in statistics New York, NY, USA:, 2001

362. McGovern SL, Aldape KD, Munsell MF, Mahajan A, DeMonte F, Woo SY: A comparison of World Health Organization tumor grades at recurrence in patients with non-skull base and skull base meningiomas. Journal of neurosurgery 112:925-933, 2010
363. Savardekar A, et al.: Differential tumor progression patterns in skull base versus non-skull base meningiomas: A critical analysis from a long-term follow-up study and review of literature. World neurosurgery, 2017
364. Hashimoto N, et al.: Slower growth of skull base meningiomas compared with non-skull base meningiomas based on volumetric and biological studies. Journal of neurosurgery 116:574-580, 2012
365. Clark VE, et al.: Genomic analysis of non-NF2 meningiomas reveals mutations in TRAF7, KLF4, AKT1, and SMO. Science (New York, NY) 339:1077-1080, 2013
366. Sade B, Chahlavi A, Krishnaney A, Nagel S, Choi E, Lee JH: World Health Organization Grades II and III meningiomas are rare in the cranial base and spine. Neurosurgery 61:1194-1198; discussion 1198, 2007
367. Savardekar AR, et al.: Differential Tumor Progression Patterns in Skull Base Versus Non-Skull Base Meningiomas: A Critical Analysis from a Long-Term Follow-Up Study and Review of Literature. World neurosurgery 112:e74-e83, 2018
368. Hwang WL, et al.: Imaging and extent of surgical resection predict risk of meningioma recurrence better than WHO histopathological grade. Neuro-oncology 18:863-872, 2015
369. Ildan F, et al.: Predicting the probability of meningioma recurrence in the preoperative and early postoperative period: a multivariate analysis in the midterm follow-up. Skull Base 17:157, 2007
370. Gillies RJ, Kinahan PE, Hricak H: Radiomics: Images Are More than Pictures, They Are Data. Radiology 278:563-577, 2016
371. Kumar V, et al.: Radiomics: the process and the challenges. Magn Reson Imaging 30:1234-1248, 2012
372. Hale AT, Stonko DP, Wang L, Strother MK, Chambless LB: Machine learning analyses can differentiate meningioma grade by features on magnetic resonance imaging. Neurosurgical Focus 45:E4, 2018
373. Coroller TP, et al.: Radiographic prediction of meningioma grade by semantic and radiomic features. PLoS One 12:e0187908, 2017
374. Ko CC, Lim SW, Chen TY, Chen JH, Li CF, Shiue YL: Prediction of progression in skull base meningiomas: additional benefits of apparent diffusion coefficient value. J Neurooncol, 2018
375. Mathiesen T, Lindquist C, Kihlstrom L, Karlsson B: Recurrence of cranial base meningiomas. Neurosurgery 39:2-7; discussion 8-9, 1996
376. Voss KM, et al.: The Simpson grading in meningioma surgery: does the tumor location influence the prognostic value? J Neurooncol 133:641-651, 2017
377. Maclean J, Fersht N, Short S: Controversies in radiotherapy for meningioma. Clinical oncology (Royal College of Radiologists (Great Britain)) 26:51-64, 2014
378. Stafford SL, et al.: Meningioma radiosurgery: tumor control, outcomes, and complications among 190 consecutive patients. Neurosurgery 49:1029-1037; discussion 1037-1028, 2001

379. Mathiesen T, Kihlstrom L, Karlsson B, Lindquist C: Potential complications following radiotherapy for meningiomas. Surgical neurology 60:193-198; discussion 199-200, 2003
380. Sivakumar W, Chamoun R, Nguyen V, Couldwell WT: Incidental pituitary adenomas. Neurosurgical focus 31:E18, 2011
381. Greenman Y, Stern N: Non-functioning pituitary adenomas. Best practice & research Clinical endocrinology & metabolism 23:625-638, 2009
382. Molitch ME: Nonfunctioning pituitary tumors and pituitary incidentalomas. Endocrinology and metabolism clinics of North America 37:151-171, xi, 2008
383. Lloyd RV, Osamura RY, Klöppel G, Rosai J, Cancer IAfRo: WHO Classification of Tumours of Endocrine Organs: International Agency for Research on Cancer, 2017
384. Dekkers OM, et al.: The natural course of non-functioning pituitary macroadenomas. European journal of endocrinology 156:217-224, 2007
385. Ferrante E, et al.: Non-functioning pituitary adenoma database: a useful resource to improve the clinical management of pituitary tumors. European journal of endocrinology 155:823-829, 2006
386. O'Sullivan EP, et al.: The natural history of surgically treated but radiotherapy-naive nonfunctioning pituitary adenomas. Clinical endocrinology 71:709-714, 2009
387. Roelfsema F, Biermasz NR, Pereira AM: Clinical factors involved in the recurrence of pituitary adenomas after surgical remission: a structured review and meta-analysis. Pituitary 15:71-83, 2012
388. Di Ieva A, Rotondo F, Syro LV, Cusimano MD, Kovacs K: Aggressive pituitary adenomas--diagnosis and emerging treatments. Nat Rev Endocrinol 10:423-435, 2014
389. Meij BP, Lopes MB, Ellegala DB, Alden TD, Laws ER, Jr.: The long-term significance of microscopic dural invasion in 354 patients with pituitary adenomas treated with transsphenoidal surgery. Journal of neurosurgery 96:195-208, 2002
390. Boxerman JL, Rogg JM, Donahue JE, Machan JT, Goldman MA, Doberstein CE: Preoperative MRI evaluation of pituitary macroadenoma: imaging features predictive of successful transsphenoidal surgery. AJR American journal of roentgenology 195:720-728, 2010
391. Snead FE, Amdur RJ, Morris CG, Mendenhall WM: Long-term outcomes of radiotherapy for pituitary adenomas. International journal of radiation oncology, biology, physics 71:994-998, 2008
392. Brochier S, et al.: Factors predicting relapse of nonfunctioning pituitary macroadenomas after neurosurgery: a study of 142 patients. European journal of endocrinology 163:193-200, 2010
393. Losa M, et al.: Early results of surgery in patients with nonfunctioning pituitary adenoma and analysis of the risk of tumor recurrence. Journal of neurosurgery 108:525-532, 2008
394. Huang YQ, et al.: Development and Validation of a Radiomics Nomogram for Preoperative Prediction of Lymph Node Metastasis in Colorectal Cancer. J Clin Oncol 34:2157-2164, 2016
395. Wu J, et al.: Early-Stage Non-Small Cell Lung Cancer: Quantitative Imaging Characteristics of (18)F Fluorodeoxyglucose PET/CT Allow Prediction of Distant Metastasis. Radiology 281:270-278, 2016

396. Lambin P, et al.: Radiomics: extracting more information from medical images using advanced feature analysis. Eur J Cancer 48:441-446, 2012
397. Aerts HJ, et al.: Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat Commun 5:4006, 2014
398. Zhang Y, et al.: Radiomics approach for prediction of recurrence in skull base meningiomas. Neuroradiology 61:1355-1364, 2019
399. Park YW, et al.: Whole-Tumor Histogram and Texture Analyses of DTI for Evaluation of IDH1-Mutation and 1p/19q-Codeletion Status in World Health Organization Grade II Gliomas. AJNR Am J Neuroradiol 39:693-698, 2018
400. Park YW, et al.: Radiomics and machine learning may accurately predict the grade and histological subtype in meningiomas using conventional and diffusion tensor imaging. European radiology, 2018
401. Rui W, et al.: MR textural analysis on contrast enhanced 3D-SPACE images in assessment of consistency of pituitary macroadenoma. Eur J Radiol 110:219-224, 2019
402. Fan Y, et al.: Preoperative Noninvasive Radiomics Approach Predicts Tumor Consistency in Patients With Acromegaly: Development and Multicenter Prospective Validation. Front Endocrinol (Lausanne) 10:403, 2019
403. Niu J, et al.: Preoperative prediction of cavernous sinus invasion by pituitary adenomas using a radiomics method based on magnetic resonance images. European radiology 29:1625-1634, 2019
404. Ugga L, et al.: Prediction of high proliferative index in pituitary macroadenomas using MRI-based radiomics and machine learning. Neuroradiology, 2019
405. Ko CC, Chen TY, Lim SW, Kuo YT, Wu TC, Chen JH: Prediction of recurrence in solid nonfunctioning pituitary macroadenomas: additional benefits of diffusion-weighted MR imaging. Journal of neurosurgery:1-9, 2019
406. Lee MH, et al.: Clinical Concerns about Recurrence of Non-Functioning Pituitary Adenoma. Brain tumor research and treatment 4:1-7, 2016
407. Hong JW, Lee MK, Kim SH, Lee EJ: Discrimination of prolactinoma from hyperprolactinemic non-functioning adenoma. Endocrine 37:140-147, 2010
408. Wang S, Lin S, Wei L, Zhao L, Huang Y: Analysis of operative efficacy for giant pituitary adenoma. BMC surgery 14:59, 2014
409. Knosp E, Steiner E, Kitz K, Matula C: Pituitary adenomas with invasion of the cavernous sinus space: a magnetic resonance imaging classification compared with surgical findings. Neurosurgery 33:610-617; discussion 617-618, 1993
410. Hardy J: Acromegaly : Surgical treatment by transsphenoidal microsurgical removal of the pituitary adenoma. Clinical Management of Pituitary Disorders, 1979
411. Fan Y, Jiang S, Hua M, Feng S, Feng M, Wang R: Machine Learning-Based Radiomics Predicts Radiotherapeutic Response in Patients With Acromegaly. Front Endocrinol (Lausanne) 10:588, 2019
412. Kocak B, et al.: Predicting response to somatostatin analogues in acromegaly: machine learning-based high-dimensional quantitative texture analysis on T2-weighted MRI. European radiology 29:2731-2739, 2019
413. Zeynalova A, et al.: Preoperative evaluation of tumour consistency in pituitary macroadenomas: a machine learning-based histogram analysis on conventional T2-weighted MRI. Neuroradiology 61:767-774, 2019

414. Gonzalez RC, Woods RE, Eddins SL: Digital Image processing using MATLAB®, [United States]: Gatesmark Publishing, 2009
415. Hastie T, Tibshirani R, Friedman J: The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition: Springer New York, 2009
416. Chatzellis E, Alexandraki KI, Androulakis, II, Kaltsas G: Aggressive pituitary tumors. Neuroendocrinology 101:87-104, 2015
417. Trouillas J, et al.: A new prognostic clinicopathological classification of pituitary adenomas: a multicentric case-control study of 410 patients with 8 years post-operative follow-up. Acta neuropathologica 126:123-135, 2013
418. Tamrazi B, Pekmezci M, Aboian M, Tihan T, Glastonbury CM: Apparent diffusion coefficient and pituitary macroadenomas: pre-operative assessment of tumor atypia. Pituitary 20:195-200, 2017
419. Bradley WG, Jr.: MR appearance of hemorrhage in the brain. Radiology 189:15-26, 1993
420. Kocak B, Durmaz ES, Ates E, Kilickesmez O: Radiomics with artificial intelligence: a practical guide for beginners. Diagn Interv Radiol 25:485-495, 2019
421. Saha A, Tso S, Rabski J, Sadeghian A, Cusimano MD: Machine learning applications in imaging analysis for patients with pituitary tumors: a review of the current literature and future directions. Pituitary, 2020
422. Yan PF, et al.: The Potential Value of Preoperative MRI Texture and Shape Analysis in Grading Meningiomas: A Preliminary Investigation. Transl Oncol 10:570-577, 2017
423. Amadasun M, King R: Textural features corresponding to textural properties. IEEE Transactions on systems, man, and Cybernetics 19:1264-1274, 1989
424. Ciric I, Ragin A, Baumgartner C, Pierce D: Complications of transsphenoidal surgery: results of a national survey, review of the literature, and personal experience. Neurosurgery 40:225-236; discussion 236-227, 1997
425. Rim CH, Yang DS, Park YJ, Yoon WS, Lee JA, Kim CY: Radiotherapy for pituitary adenomas: long-term outcome and complications. Radiation oncology journal 29:156-163, 2011
426. Sebastian P, Balakrishnan R, Yadav B, John S: Outcome of radiotherapy for pituitary adenomas. Rep Pract Oncol Radiother 21:466-472, 2016
427. Topol EJ: High-performance medicine: the convergence of human and artificial intelligence. Nature medicine 25:44-56, 2019
428. Yu F, Koltun V: Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:151107122, 2015
429. Zhao H, Shi J, Qi X, Wang X, Jia J: Pyramid scene parsing network. Proc. IEEE Conf on Computer Vision and Pattern Recognition (CVPR): City