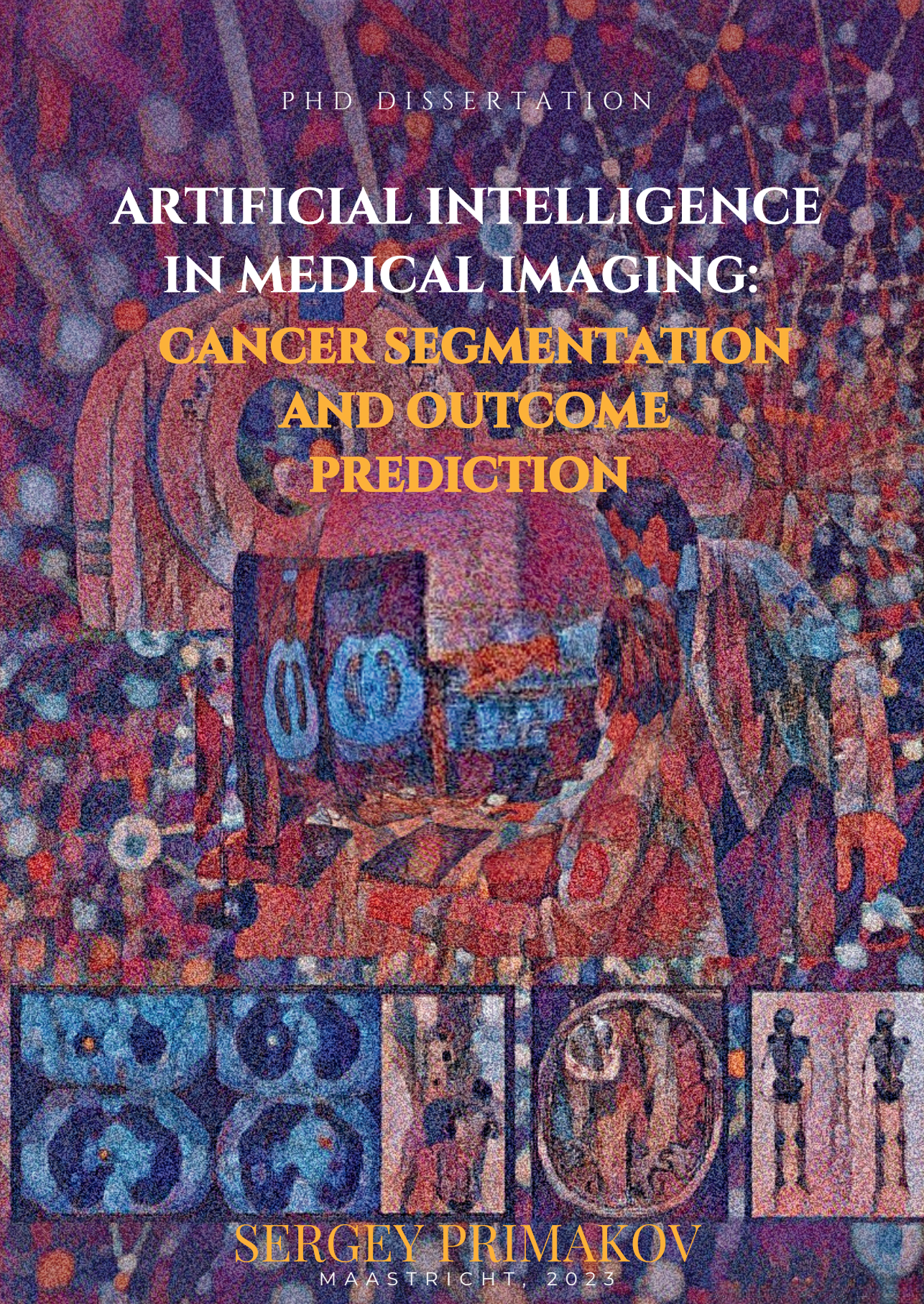
Artificial intelligence in medical imaging



PHD DISSERTATION

ARTIFICIAL INTELLIGENCE IN MEDICAL IMAGING: CANCER SEGMENTATION AND OUTCOME PREDICTION

SERGEY PRIMAKOV

MAASTRICHT, 2023

ARTIFICIAL INTELLIGENCE IN MEDICAL IMAGING: CANCER SEGMENTATION AND OUTCOME PREDICTION

Sergey Primakov

Layout and Cover Design:
Kate Lediakhova || https://www.linkedin.com/in/kate-lediakhova/

Printing: ProefschriftMaken || www.proefschriftmaken.nl

ISBN: 978-94-6469-698-1

The research presented in this thesis was conducted within GROW-School for Oncology and Reproduction, Maastricht University.

The work presented in this book has been financially supported by the European Marie Skladowska-Curie grant, PREDICT-ITN – No. 766276).

ARTIFICIAL INTELLIGENCE IN MEDICAL IMAGING: CANCER SEGMENTATION AND OUTCOME PREDICTION

Dissertation

To obtain the degree of Doctor at Maastricht University,
on the authority of the Rector Magnificus, Prof. dr. Pamela Habibović
in accordance with the decision of the Board of Deans,
to be defended in public
on Tuesday 5th of December at 10:00 hours

by

Sergey Primakov

Promotor:
Prof. dr. Philippe Lambin

Co-promotor:
Dr. Henry C. Woodruff

Assessment committee:
Prof. dr. ir. Andre Dekker (chair)
Prof. dr. Regina Beets-Tan
Dr. Wouter van Elmpt
Prof. dr. Wiro Niessen, University Medical Center Groningen
Dr. Alberto Traverso, San Raffaele Hospital, Milan, Italy

CONTENTS

CHAPTER 1:

GENERAL INTRODUCTION TO THE FIELD AND OUTLINE OF THE THESIS

Cancers

Cancer is a complex and versatile disease that originates from the uncontrolled growth and spread of pathological cells in the body. It stands as one of the primary causes of death globally, affecting individuals of all genders (1). The development of cancer is influenced by a combination of genetic predispositions, environmental factors, and lifestyle choices. These factors contribute to genetic mutations in the cells' DNA, which disrupts their normal growth and division mechanisms, leading to the formation of tumors (2,3).

There are various cancer types, with numerous characteristics and treatment options. Some cancers spread quickly, while others may grow slowly or not at all. Understanding the particular characteristics of each cancer type is crucial for an effective treatment planning.

Diagnosing and treating cancer requires a comprehensive and multidisciplinary approach that utilizes the expertise of various medical specialists, including oncologists, radiologists, and surgeons. The treatment options for cancer are diverse and depend on factors such as the type of cancer and its stage of progression (4,5). Surgery is often employed to remove localized tumors and affected tissues. Radiation therapy utilizes high-energy radiation to target and destroy cancer cells, either as a standalone treatment or in combination with other approaches. Chemotherapy involves the use of drugs that circulate throughout the body to kill cancer cells. Immunotherapy leverages the body's immune system to recognize and eliminate cancer cells. Depending on the specifics of the case, a combination of these treatment options may be recommended to maximize treatment effectiveness (6).

The thesis focuses on the research and development of Artificial Intelligence (AI) based tools to improve the diagnostic and streamline management routines for multiple cancers, including Non-Small Cell Lung Cancer (NSCLC) and Glioblastoma (GBM).

Non-Small Cell Lung Cancer

Non-Small Cell Lung Cancer (NSCLC) is the most common type of lung cancer, accounting for approximately 85% of all lung cancer diagnoses. In 2018, lung cancer claimed the highest number of lives among all cancers affecting both genders. It was responsible for approximately 18.4% of global cancer-related deaths, almost equal to the combined deaths caused by breast and colon cancers (7). Within the NSCLC category,

there are different subtypes, including adenocarcinoma, squamous cell carcinoma, and large cell carcinoma. Adenocarcinoma is the most prevalent subtype of NSCLC, typically originating in the peripheral areas of the lungs. It is more commonly diagnosed in non-smokers and is often associated with genetic mutations, such as mutations in the EGFR (epidermal growth factor receptor) gene. Squamous cell carcinoma, on the other hand, usually arises in the central airways and is strongly linked to tobacco smoking. Large cell carcinoma is a less common subtype and is characterized by the presence of large, abnormal cells (8,9).

The management of NSCLC depends on factors such as cancer stage, molecular characteristics, and the overall health of the patient. Treatment approaches for NSCLC include surgery, radiation therapy, chemotherapy, targeted therapy, and immunotherapy. The specific treatment plan is determined based on individual patient factors and the presence of specific genetic mutations or biomarkers (5). In recent years, significant advancements have been made in the treatment of NSCLC, particularly with the advent of targeted therapies and immunotherapies. Targeted therapies are designed to inhibit specific molecular pathways that drive the growth and survival of cancer cells (10). For example, drugs targeting EGFR mutations or ALK (anaplastic lymphoma kinase) gene rearrangements have shown promising results in patients with NSCLC harboring these mutations (11,12). Immunotherapy, on the other hand, aims to enhance the body's immune response against cancer cells. Immune checkpoint inhibitors, such as PD-1 (programmed cell death protein 1) inhibitors, have demonstrated remarkable efficacy in a subset of NSCLC patients (13,14). Despite these advancements, challenges remain in the treatment of NSCLC. Resistance to targeted therapies, limited treatment options for advanced stages of the disease, and the heterogeneity of NSCLC tumors pose ongoing challenges in achieving favorable outcomes (15). Continued research efforts are focused on identifying new molecular targets, developing combination therapies, and improving patient selection criteria to optimize treatment strategies for NSCLC.

Glioblastoma

Glioblastoma (GBM), commonly known as a grade IV astrocytoma, is a highly malignant brain tumor characterized by its rapid growth and aggressive nature. This type of tumor has the ability to infiltrate the surrounding brain tissue, leading to its classification as an invasive cancer (16). However, GBMs

rarely metastasize or spread to distant organs like other forms of cancer (17). GBMs can arise in the brain de novo, meaning they develop without any prior indication or underlying condition, or they can evolve from lower-grade gliomas. The transformation of a lower-grade glioma into a glioblastoma is often associated with genetic mutations and molecular alterations within the tumor cells (18). These changes lead to a more aggressive phenotype, characterized by rapid cell division, increased blood vessel formation, and resistance to treatment.

Due to their location and aggressive nature, glioblastomas pose significant challenges in terms of treatment and prognosis (19). Standard treatment options for GBM typically involve a combination of surgery, radiation therapy, and chemotherapy. However, complete surgical removal of the tumor is often difficult due to its infiltrative nature and proximity to critical brain regions (20,21). Radiation therapy and chemotherapy are employed to target any remaining tumor cells and slow down their growth. Despite these interventions, glioblastomas have a high rate of recurrence, and the prognosis for patients diagnosed with GBM remains poor (19,22).

Research efforts are underway to better understand the underlying molecular mechanisms of glioblastoma and develop innovative treatment approaches. These include targeted therapies that aim to disrupt specific signaling pathways involved in tumor growth and progression, immunotherapies that harness the immune system to recognize and attack cancer cells, and novel drug delivery methods to improve the effectiveness of treatment (4,23).

Medical Imaging

Medical imaging is an essential tool in the diagnosis, treatment and tumor response to treatment evaluation of various types of cancer (24,25). Computed tomography (CT), magnetic resonance imaging (MRI), nuclear medicine imaging modalities such as positron emission tomography (PET), and bone scintigraphy are among the most commonly used modalities for cancer management. CT combines X-ray technology with an advanced computer processing to produce detailed cross-sectional images of the body. It uses a rotating X-ray beam and detectors to capture multiple X-ray images from different angles, which are then reconstructed by the computer to create a three-dimensional view of the internal structures. Chest CT is an essential imaging tool for the assessment of lung cancer, commonly utilized for both screening and staging purposes(5). It serves as a noninvasive method to examine and characterize lung

lesions, located within the air-filled lung tissue. The size, location, and characteristics of these lesions can be precisely determined through chest CT scans (26).

Magnetic Resonance Imaging (MRI) uses a strong magnetic field and radio waves to generate detailed images of the body's internal structures. It works by aligning the hydrogen protons in the body's tissues and then disrupting their alignment using radio waves. As the protons realign, they emit signals that are captured by specialized detectors and processed by a computer to create highly detailed images. MRI is particularly useful for visualizing soft tissues, such as the brain, spinal cord, muscles, and joints. MRI is the modality of choice for diagnosis and assessment of treatment response in patients with GBM, due to its wide availability, and superior soft tissue visualization over computed tomography (CT) (27).

Nuclear medicine imaging is a branch of medical imaging that utilizes small amounts of radioactive substances, known as radiopharmaceuticals, to visualize and assess the function of organs and tissues within the body. These radiopharmaceuticals are typically injected, swallowed, or inhaled, and they emit gamma rays or positrons that are detected by specialized cameras or scanners. The cameras process the emitted signal and produce images that show the distribution of the radiotracer in the body, highlighting areas of high metabolic activity. Positron Emission Tomography (PET) and Whole Body Bone Bcintigraphy (WBBS) or bone scan are among the ubiquitous nuclear medicine imaging modalities (28). PET is extensively used in oncology for cancer detection, staging, and treatment monitoring (24,25). It helps visualize and assess metabolic activity in tumors, determine the spread of cancer, evaluate treatment response, and detect potential cancer recurrence. WBBS is commonly employed to diagnose and monitor conditions such as bone fractures, infections, tumors, and metastases(29).

With advancements in medical imaging technology, current scanners provide high-resolution images that allow for more accurate and detailed visualizations of various internal body structures. The precise representation of these structures, as well as the high-resolution and volumetric nature of these images, allows for the use of advanced algorithms for quantitative analysis.

Artificial Intelligence

Artificial Intelligence (AI) is a multidisciplinary field of computer science and an umbrella term for a wide range of algorithms and techniques with the goal of enabling machines to perform tasks that

typically require human intelligence (30). AI aims to replicate human-like intelligence and cognitive abilities in machines to automate complex tasks, solve problems, and make informed decisions. Machine learning (ML) is a subfield of AI characterized by a specific approach that typically uses algorithms and statistical models to learn from data and perform tasks like classification, regression, anomaly detection and clustering (31). The unique feature that fueled the recent growth and ubiquitous application of ML algorithms is their ability to learn directly from complex data without being explicitly programmed. ML algorithms can be categorized into various types, such as supervised learning, unsupervised learning, and reinforcement learning. In supervised learning, the ML model learns from labeled data, where inputs and corresponding outputs are provided. The model generalizes from this labeled data to make predictions or classify new, unseen inputs. Regression models can predict continuous values, while classification models assign inputs to predefined categories (30,32). Unsupervised learning, on the other hand, involves analyzing unlabeled data to identify inherent patterns, structures, or relationships within the data. Clustering algorithms group similar data points together, while dimensionality reduction techniques aim to capture essential features of the data while reducing its complexity. Reinforcement learning involves training an agent to make decisions based on interactions with an environment. The agent receives feedback in the form of rewards or penalties, allowing it to learn through trial and error and optimize its decision-making strategy (32).

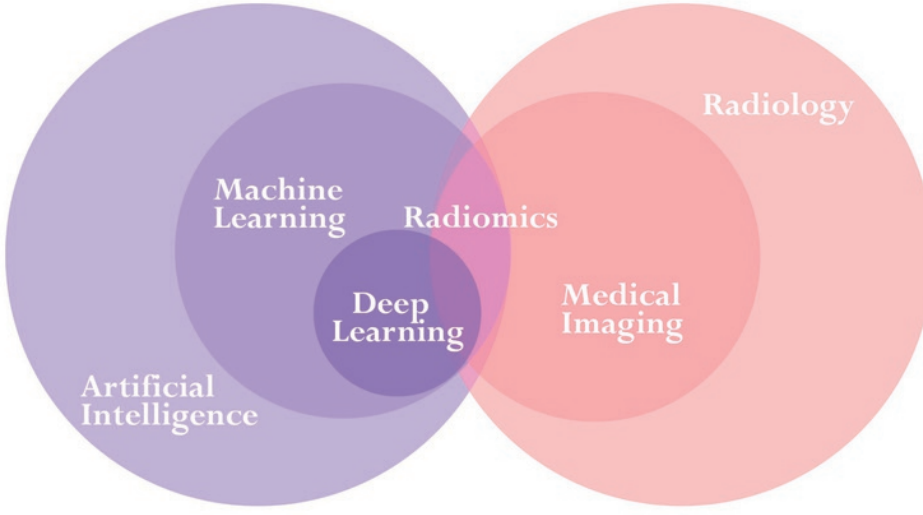


Figure 1: Visual interpretation of Artificial Intelligence algorithms and medical imaging.

Machine learning algorithms quickly found their application in multiple industries and domains, including finance, transportation, marketing and many others. Application of machine learning in radiology has been recently manifested in Radiomics(33).

Radiomics

Radiomics is a rapidly growing field within medical imaging that focuses on the extraction of Handcrafted Radiomics Features (HRFs) from medical images, particularly tumor Region Of Interest, to explore the correlation between the underlying biological characteristics and clinical outcomes (34,35). HRFs are computed using manually designed algorithms that capture various tumor characteristics, including shape, texture, and complex statistical features derived from the voxel intensity data. These features serve as quantitative measurements that aims to characterize the tumor's heterogeneity, spatial patterns, and other distinct properties (36). HRFs coupled with feature selection methods and statistical and machine learning algorithms are used to produce prognostic and predictive models advancing the clinical decision support. As reported in numerous studies HRFs based models has been showing promising performance in developing imaging biomarkers that can help predict patient outcomes, response to treatment and survival time (37–39).
However, it is essential to note that the translation of radiomic models into clinical practice requires rigorous validation, standardization, and integration into existing clinical workflows (35). As the field of radiomics continues to advance, there is a growing need for collaborative efforts, open source data sharing, and robust validation studies to ensure the reliability and generalizability of radiomic findings and their clinical applications.

Deep Learning

Deep learning (DL) is a branch of AI that uses data-driven techniques inspired by the functioning of neurons in the human brain. Unlike radiomics, DL models can automatically identify complex patterns in medical imaging without the need for manual feature engineering (40). The success of DL in the field of computer vision can be largely attributed to the Convolutional Neural Network (CNN) architecture. CNNs have revolutionized image analysis and recognition tasks due to their remarkable performance and ability to automatically extract hierarchical and spatially localized features

from visual inputs. CNNs utilize convolutional layers to apply filters to input data, allowing them to detect patterns and features at different scales. Through repeated application of convolutional layers, pooling layers for downsampling, and fully connected layers for classification, CNNs can learn complex representations and achieve high performance in tasks like image classification, object detection, and image segmentation.

DL has demonstrated remarkable success in various medical imaging tasks, including segmentation, classification, detection, and synthetic data generation (41–43). While DL models have shown superior performance in many medical imaging applications, one challenge associated with their adoption is the interpretability of their decisions. DL models often operate as "black boxes", meaning that the learned features and decision-making processes can be complex and difficult to interpret. Efforts are being made to address the interpretability issue in DL models, including the development of techniques to visualize learned features, identify influential regions within images, and generate heatmaps to highlight regions contributing to predictions (44). Various studies explore methods to provide explanations or justifications for DL model outputs, making them more understandable and transparent to clinicians and patients (45–47).

Objectives & outlines of the thesis

This thesis explores the use of artificial intelligence in medical imaging through application of HRF's based ML and DL models for development of predictive and prognostic models, development and implementation of automatic cancer segmentation pipeline, and creation of open-source tools to facilitate quantitative medical imaging.

More specifically the objectives of this thesis were (1) to evaluate the complementary value of HRFs to clinical features, deep learning based features and qualitative features and investigate its potential in improving patient outcomes by improving the patient prognosis and prediction (Chapters 3 and 4); (2) to investigate the potential of DL in improving current laborious radiotherapy treatment planning routines and decision making through automatic segmentation and better prognosis (Chapters 6 and 7); (3) develop tools to facilitate research needs through open source code for quantitative medical image analysis and in-silico clinical trials, and radiotherapy workflow through software for automatic segmentation of NSCLC (Chapters 8, 9).

The thesis is divided into three sections, comprising a total of ten chapters. The chapters are colorcoded according to the part of the thesis that they reprsent (Figure 2). The following is a brief summary of the content covered in each section.

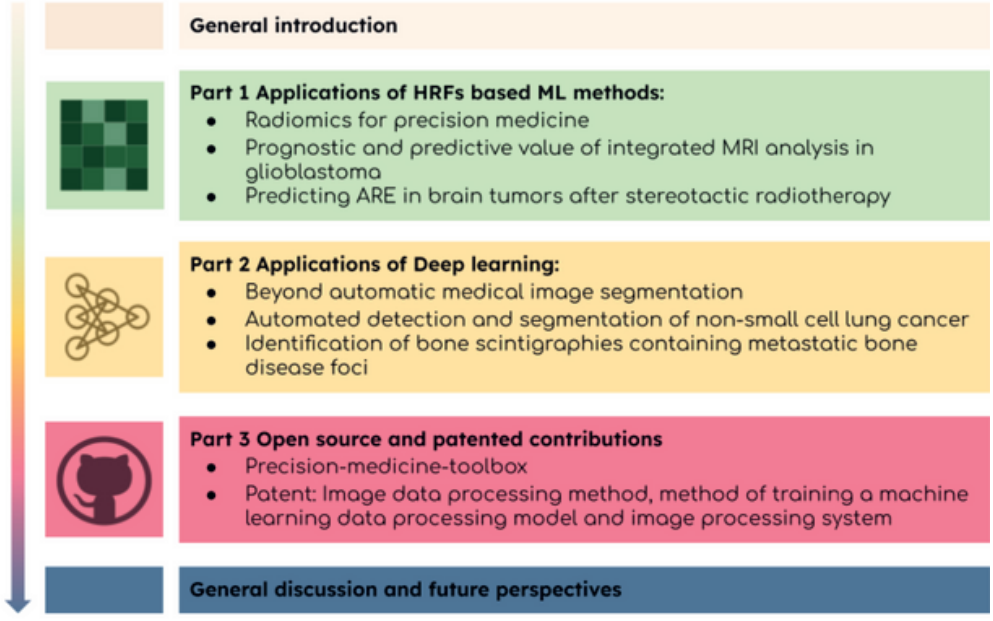


Figure 2: Thesis outline.

General introduction and Outline

Chapters 1 includes a general introduction to the field of medical imaging, presenting the current challenges and technics used to address them, and outline of the thesis.

Part 1 Applications of HRFs based ML methods in medical imaging

Chapter 2 serves as a general introduction to the application of handcrafted radiomics features and deep learning techniques in medical imaging. It encompasses a comprehensive review of the current state of radiomics application, the challenges that it currently confronts, and proposes a novel radiomics framework that focuses on reproducibility of radiomics features.

Chapter 3 investigates the possibility of using the combination of quantitative radiomics features, extracted from magnetic resonance imaging (MRI) with Non-invasive qualitative Visually Accessible Rembrandt Images (VASARI), and clinical features to improve the patient prognosis for the most malignant primary brain tumor - Glioblastoma (GBM). It further explores the potential of quantitative radiomics features in predicting clinically relevant tumor markers for GBM patients that are needed to better guide the clinicians.

Chapter 4 investigates the application of both handcrafted radiomics features and automatically extracted deep learning (DL) features from Gadolinium-enhanced T1-weighted MRIs to predict the likelihood of adverse radiation effects (ARE) in patients with brain metastases (BM) prior to receiving stereotactic radiotherapy (SRT). The study also examines existing techniques for harmonizing the machine learning (ML) and DL pipelines, and proposes an optimal pre-processing method.

Part 2 Applications of Deep learning in medical imaging

Chapters 5 is a comprehensive review on the spectrum of medical image segmentation with the focus on the automatic contouring using deep learning. The review presents the diverse directions being pursued to improve medical image segmentation, by providing a detailed description of various automatic and semi-automatic methodologies for contouring cancers and organs at risk across different medical imaging modalities.

Chapter 6 is a centerpiece of this thesis. The work described in this chapter combines a proposal and comprehensive validation of the DL based method for automatic segmentation of the NSCLC on CT images. Additionally, the chapter explores the uncertainty of manual segmentations, alternative validation through in-silico clinical trial and survival analysis and offers an open-source software for NSCLC segmentation and qualitative assessment.
The work in this chapter were used for the subsequent development of the clinical software for automatic segmentation of NSCLC on CT.

Chapter 7 of this thesis explores the feasibility of utilizing a DL algorithm to identify metastatic bone disease on scintigraphy scans. This study involved multiple centers and included both cancer and non-cancer patients. The performance of the developed software was evaluated against uninformed nuclear medicine physicians in an in-silico clinical trial setting using in-house developed software. The Grad-CAM method was employed to provide improved elucidation of the model's decision-making process through the visualization of the neuron activations.

Part 3 Open source and patented contributions to the field

Chapter 8 presents a precision medicine toolbox, an open-source Python framework that aims to facilitate multiple tasks for researchers, including data curation, image pre-processing, handcrafted radiomics features extraction, and feature exploration. The purpose of this work is to address challenges surrounding data preparation and to enhance the reproducibility of quantitative medical imaging research.

Chapter 9 is a short summary of the patent received for the work on the image data processing method, and method for training a machine learning data processing model.

Part 4 General Discussion and future perspectives

Chapter 10 of this thesis serves as a general discussion of the work presented in this thesis. Furthermore, it elaborates on the potential benefits and limitations of AI-based methods in healthcare, with an emphasis on the use of these methods for the automatic segmentation of medical images, and highlights the key challenges that must be addressed to ensure the safe and effective integration of these methods in healthcare.

References

1. Sung, H. et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J. Clin. 71, 209–249 (2021).
2. Weinberg, R. A. The Biology of Cancer. (Garland Science, 2013).
3. Bertram, J. S. The molecular biology of cancer. Mol. Aspects Med. 21, 167–223 (2000).
4. Tan, A. C. et al. Management of glioblastoma: State of the art and future directions. CA Cancer J. Clin. 70, 299–31 (2020).
5. Herbst, R. S., Morgensztern, D. & Boshoff, C. The biology and management of non-small cell lung cancer. Nature 553, 446–454 (2018).
6. Zappa, C. & Mousa, S. A. Non-small cell lung cancer: current treatment and future advances. Transl Lung Cancer Res 5, 288–300 (2016).
7. Bray, F. et al. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J. Clin. 68, 394–424 (2018).
8. Thomas, A., Liu, S. V., Subramaniam, D. S. & Giaccone, G. Refining the treatment of NSCLC according to histological and molecular subtypes. Nat. Rev. Clin. Oncol. 12, 511–526 (2015).
9. Pikor, L. A., Ramnarine, V. R., Lam, S. & Lam, W. L. Genetic alterations defining NSCLC subtypes and their therapeutic implications. Lung Cancer 82, 179–189 (2013).
10. Sweis, R. F. et al. Concurrent EGFR Mutation and ALK Translocation in Non-Small Cell Lung Cancer. Cureus 8, e513 (2016).
11. Won, J. K. et al. Concomitant ALK translocation and EGFR mutation in lung cancer: a comparison of direct sequencing and sensitive assays and the impact on responsiveness to tyrosine kinase inhibitor. Ann. Oncol. 26, 348–354 (2015).
12. König, D., Savic Prince, S. & Rothschild, S. I. Targeted Therapy in Advanced and Metastatic Non-Small Cell Lung Cancer. An Update on Treatment of the Most Important Actionable Oncogenic Driver Alterations. Cancers 13, (2021).
13. Niu, M., Yi, M., Li, N., Luo, S. & Wu, K. Predictive biomarkers of anti-PD-1/PD-L1 therapy in NSCLC. Exp. Hematol. Oncol. 10, 18 (2021).
14. Dantoing, E., Piton, N., Salaün, M., Thiberville, L. & Guisier, F. Anti-PD1/PD-L1 Immunotherapy for Non-Small Cell Lung Cancer with Actionable Oncogenic Driver Mutations. Int. J. Mol. Sci. 22, (2021).

15. Sun, S.-Y. & Su, C. Challenges and Opportunities of TKIs in the Treatment of NSCLC Patients With Uncommon Mutations. (Frontiers Media SA, 2022).
16. Strepkos, D., Markouli, M., Klonou, A., Piperi, C. & Papavassiliou, A. G. Insights in the immunobiology of glioblastoma. J. Mol. Med. 98, 1–10 (2020).
17. Ghulam Ghous, M., Douglas Miller, M. D., Donald Doll, M. D. & Tolga Tuncer, M. D. A rare case of glioblastoma with extensive liver metastases. Oncology (2021).
18. Jaeckle, K. A. et al. Transformation of low grade glioma and correlation with outcome: an NCCTG database analysis. J. Neurooncol. 104, 253–259 (2011).
19. Noch, E. K., Ramakrishna, R. & Magge, R. Challenges in the Treatment of Glioblastoma: Multisystem Mechanisms of Therapeutic Resistance. World Neurosurg. 116, 505–517 (2018).
20. Davis, M. E. Glioblastoma: Overview of Disease and Treatment. Clin. J. Oncol. Nurs. 20, S2–8 (2016).
21. Gerritsen, J. K. W. et al. Safe surgery for glioblastoma: Recent advances and modern challenges. Neurooncol Pract 9, 364–379 (2022).
22. Delgado-López, P. D. & Corrales-García, E. M. Survival in glioblastoma: a review on the impact of treatment modalities. Clin. Transl. Oncol. 18, 1062–1071 (2016).
23. Bausart, M., Préat, V. & Malfanti, A. Immunotherapy for glioblastoma: the promise of combination strategies. J. Exp. Clin. Cancer Res. 41, 35 (2022).
24. Oliva, M. R. & Saini, S. Liver cancer imaging: role of CT, MRI, US and PET. Cancer Imaging 4 Spec No A, S42–6 (2004).
25. Engbersen, M. P., Van Driel, W., Lambregts, D. & Lahaye, M. The role of CT, PET-CT, and MRI in ovarian cancer. Br. J. Radiol. 94, 20210117 (2021).
26. Park, H. J., Lee, S. H. & Chang, Y. S. Recent advances in diagnostic technologies in lung cancer. Korean J. Intern. Med. 35, 257–268 (2020).
27. Bernstock, J. D. et al. Standard clinical approaches and emerging modalities for glioblastoma imaging. Neurooncol Adv 4, vdac080 (2022).
28. Glaudemans, A. W. J. M. & Signore, A. Nuclear Medicine Imaging Modalities: Bone Scintigraphy, PET-CT, SPECT-CT. in Bone Metastases: A translational and Clinical Approach (eds. Vassiliou, V., Chow, E. & Kardamakis, D.) 71–94 (Springer Netherlands, 2014).
29. Van den Wyngaert, T. et al. The EANM practice guidelines for bone scintigraphy. Eur. J. Nucl. Med. Mol. Imaging 43, 1723–1738 (2016).

30. Ghahramani, Z. Probabilistic machine learning and artificial intelligence. Nature 521, 452–459 (2015).
31. Erickson, B. J., Korfiatis, P., Akkus, Z. & Kline, T. L. Machine Learning for Medical Imaging. Radiographics 37, 505–515 (2017).
32. Raschka, S. Python Machine Learning. (Packt Publishing Ltd, 2015).
33. Aerts, H. J. W. L. et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat. Commun. 5, 4006 (2014).
34. Sanduleanu, S. et al. Tracking tumor biology with radiomics: A systematic review utilizing a radiomics quality score. Radiother. Oncol. 127, 349–360 (2018).
35. Ibrahim, A. et al. Radiomics for precision medicine: Current challenges, future prospects, and the proposal of a new framework. Methods 188, 20–29 (2021).
36. Lambin, P. et al. Radiomics: extracting more information from medical images using advanced feature analysis. Eur. J. Cancer 48, 441–446 (2012).
37. Keek, S. A. et al. A Prospectively Validated Prognostic Model for Patients with Locally Advanced Squamous Cell Carcinoma of the Head and Neck Based on Radiomics of Computed Tomography Images. Cancers 13, (2021).
38. Lambin, P. et al. Radiomics: the bridge between medical imaging and personalized medicine. Nat. Rev. Clin. Oncol. 14, 749–762 (2017).
39. van Timmeren, J. E. et al. Survival prediction of non-small cell lung cancer patients using radiomics analyses of cone-beam CT images. Radiother. Oncol. 123, 363–369 (2017).
40. Chollet, F. Deep Learning with Python. (Manning Publications, 2017).
41. Wei, J. & Fan, Z. Genetic U-Net: Automatically Designed Deep Networks for Retinal Vessel Segmentation Using a Genetic Algorithm. arXiv [eess.IV] (2020).
42. Kalmet, P. H. S. et al. Deep learning in fracture detection: a narrative review. Acta Orthop. 91, 362 (2020).
43. Trimpl, M. J., Primakov, S. & Lambin, P. Beyond automatic medical image segmentation—the spectrum between fully manual and fully automatic delineation. Phys. Med. Biol. (2022).
44. Selvaraju, R. R. et al. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. arXiv [cs.CV] (2016).
45. Zhou, Y., Jiang, F., Cheng, F. & Li, J. Detecting representative characteristics of different genders using intraoral photographs: a deep learning model with interpretation of gradient-weighted class activation mapping. BMC Oral Health 23, 327 (2023).

46. Song, D. et al. A new xAI framework with feature explainability for tumors decision-making in Ultrasound data: comparing with Grad-CAM. Comput. Methods Programs Biomed. 235, 107527 (2023).
47. Tummala, S., Kadry, S., Nadeem, A., Rauf, H. T. & Gul, N. An Explainable Classification Method Based on Complex Scaling in Histopathology Images for Lung and Colon Cancer. Diagnostics (Basel) 13, (2023).

PART 1:

APPLICATIONS OF HANDCRAFTED RADIOMICS FEATURES BASED MACHINE LEARNING METHODS IN MEDICAL IMAGING

CHAPTER 2:

RADIOMICS FOR PRECISION MEDICINE: CURRENT CHALLENGES, FUTURE PROSPECTS, AND THE PROPOSAL OF A NEW FRAMEWORK

Authors: Abdalla Ibrahim, Sergey Primakov[1], Manon Beuque[1], Henry C. Woodruff, Iva Halilaj, Guangyao Wu, Turkey Refaee, Renee Granzier, Yousif Widaatalla, Roland Hustinx, Felix M. Mottaghy, Philippe Lambin

1 These authors have contributed equally.

Adapted from:
Ibrahim A, Primakov S, Beuque M, Woodruff HC, Halilaj I, Wu G, Refaee T, Granzier R, Widaatalla Y, Hustinx R, Mottaghy FM, Lambin P. Radiomics for precision medicine: Current challenges, future prospects, and the proposal of a new framework. Methods 2021; 188:20-29. Doi: https://doi.org/10.1016/j.ymeth.2020.05.022

Acsess link:
https://www.sciencedirect.com/science/article/pii/S1046202320301110

Abstract

The advancement of artificial intelligence concurrent with the development of medical imaging techniques provided a unique opportunity to turn medical imaging from mostly qualitative, to further quantitative and mineable data that can be explored for the development of clinical decision support systems (cDSS). Radiomics, a method for the high throughput extraction of handcrafted features from medical images, and deep learning the data driven modeling techniques based on the principles of simplified brain neuron interactions, are the most researched quantitative imaging techniques. Many studies reported on the potential of such techniques in the context of cDSS. Such techniques could be highly appealing due to the reuse of existing data, automation of clinical workflows, minimal invasiveness, three-dimensional volumetric characterization, and the promise of high accuracy and reproducibility of results and cost-effectiveness. Nevertheless, there are several challenges that quantitative imaging techniques face, and need to be addressed before the translation to clinical use. These challenges include, but are not limited to, the explainability of the models, the reproducibility of the quantitative imaging features, and their sensitivity to variations in image acquisition and reconstruction parameters. In this narrative review, we report on the status of quantitative medical image analysis using radiomics and deep learning, the challenges the field is facing, propose a framework for robust radiomics analysis, and discuss future prospects.

1. Introduction

Advances in artificial intelligence applications, combined with those in medical imaging, have led to the gradual conversion of digital medical images into high-dimensional data appropriate for data mining and data science techniques (1). Meanwhile, computing power and quantitative image analysis (QIA) techniques have made enormous progress, and the application of quantitative imaging techniques on medical imaging gained exponential momentum (2). Currently, radiomics and deep learning are the most researched techniques on medical imaging.

Broadly, radiomics refers to the use of computational or statistical approaches to extract large numbers of quantitative features from a number of medical imaging modalities, such as computed tomography (CT), magnetic resonance imaging (MRI), and positron emission tomography (PET), to develop predictive models

ultimately aiming to enable personalized clinical management (3–5). Radiomics features are quantitative descriptions of the intensity, shape, volume, and texture of the region of interest (ROI), with the recent addition of more abstract features such as radial gradient and radial deviation (6). Radiomics features are broadly divided into histogram-based and texture features. Different statistical methods are used to calculate the radiomics features. The methods include first-order statistics, which depends on the values of single voxels (histogram-based features for e.g. maximum and minimum intensity); second-order statistics, which depends on the relation between two voxels (for e.g. grey-level co-occurrence matrix (GLCM) features), and higher-order statistics (relations among three or more voxels, for e.g. neighborhood grey-tone difference matrices (NGTDM) features) (7,8). The main hypothesis behind radiomics analysis is that radiomics features decode or correlate with the molecular characteristics, phenotype, and genotype of the region of interest (ROI) under study. This information can be used in combination with other patient information to improve patient management. Moreover, as the tumours are of heterogeneous nature (9,10), clinical approaches, such as tissue biopsies, might fail to characterize the entirety of the tumour (11). In contrast, Radiomics takes the whole tumour region (or even the surrounding or healthy tissue) into account, which enables a better characterization (3). Furthermore, frequent clinical imaging can transform radiomics into a non-invasive, easily repeatable, and cost-effective longitudinal approach for cDSS (12).

Deep learning (DL) is a field of data driven modelling techniques that utilizes the principles of simplified neuron interactions (13). Using artificial neurons started to draw attention decades ago (14), but it only became a major research focus recently (15–17). The artificial neuron model is used as a foundation unit to create complex chains of interactions – DL layers. These layers are used to generate even more complex structures DL architectures (see Figure 1). The neural network (NN) training procedure is typically a cost-function minimization process. The cost function measures the error of predictions based on the ground truth labels (18). Due to the high complexity of the network architectures, computational limitations are reached when trying to solve the optimization task analytically. Henceforth, iterative algorithms are used to overcome this issue. Commonly, these algorithms are variations of the gradient descent (GD). GD iteratively moves in the direction of steepest descent of the cost function, in order to find a local minimum. During the model training process, every image from the training dataset contributes to the cost minimization process. Thereby, a DL network learns how to solve a problem directly from existing data, and apply

it to data it has never seen. These complex models contain the parameters (weights) for millions of neurons, which can be trained for the recognition of problem-related patterns in the data being analyzed. DL has been shown to be efficient in other fields, such as face recognition (19) and autonomous cars (20).

Since the introduction of the field, many studies have reported on the potential of such techniques for predicting patient outcomes (5,21,22). The successful translation of QIA techniques into cDSS will have a significant impact on the clinical workflow and current patient management protocols. Clinicians will be able to non-invasively obtain a more detailed and accurate tumour characterization, in a shorter amount of time. Patients will have to go through less invasive procedures, while having treatment optimized based on their individual characteristics. Furthermore, patient-specific informed decisions can be made with more confidence. However, QIA is still developing in the field of medical imaging and several challenges, including the stability and reproducibility of imaging biomarkers, as well as the interpretability of the developed algorithms, need to be addressed before QIA can be translated to clinical applications.

In this narrative review, we focus on the current status of the potential of radiomics and deep learning to be incorporated in clinical decision support systems (cDSS), their challenges, as well as future prospects for these methods. We further propose a workflow to guide robust radiomics analysis.

2. Quantitative image analysis for precision medicine

The need for personalizing the management of patients has been widely reported (23,24). QIA represents a suitable candidate to be incorporated into the body of personalized medicine due to the non-invasive three-dimensional characterization of the ROIs, the availability of vast amounts of medical images, the longitudinal capabilities, and the cost-effectiveness of the method.

The currently implemented imaging biomarker development workflow is generalizable across different imaging modalities. The workflow can be described as consecutive steps divided into the main categories of data collection, image segmentation, features extraction, development of the signature, and evaluation of the performance (Figure 2), with the segmentation step being optional in the case of deep learning. The workflow has been previously extensively described (22,25).

Many studies have investigated and reported on the added clinical value of radiomics features for predicting various clinical outcomes,

such as overall survival, tumour histology, response to therapy, and genetic profiling, among other endpoints. Furthermore, these studies were performed on various imaging modalities, including CT, MR, and PET.

While the handcrafted radiomics pipeline necessitates the use of machine learning or statistical algorithms after feature extraction for modeling, DL techniques perform feature extraction and modelling internally without the need for further user interaction. DL has its own advantages and drawbacks compared to traditional radiomics. One of the key benefits of using DL is avoiding the contouring problem, the bottleneck of a traditional radiomics pipeline. However, due to the complexity of DL models, it is easier to overfit the model to the training data. As a result, a larger data set is needed for DL compared to handcrafted radiomics. Furthermore, DL is considered a 'black box', i.e the models and features generated are not (or barely) interpretable. This is currently one of the major challenges of the application of artificial intelligence (AI) in medical image analysis. Efforts are being made towards providing explainable AI algorithms, by investigating the correlation of the chosen features with biologic or semantic characteristics. Such correlations would provide an understanding about how the algorithm makes the decision, and ease its incorporation into cDSS.

QIA techniques have a great potential for involvement in developing classification, prognostic and predictive clinical tools. In comparison, classification tasks (for e.g classifying tissue histology) seem to yield a better performance than predictive tasks (for e.g survival prediction). This is in part due to the unaccounted for variables when trying to predict future events. In 2.1 and 2.2, we report on some examples that highlighted the potential of radiomics and deep learning to predict various clinical endpoints, acknowledged or addressed the challenges of QIA techniques used, and/or applied the techniques on a relatively large sample size compared to other studies addressing the same clinical endpoint.

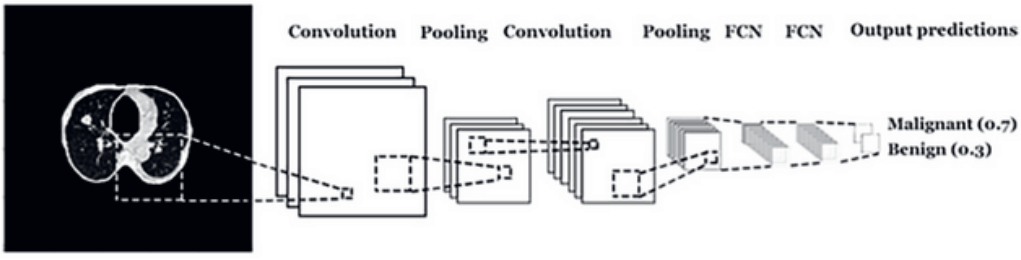


Figure 1. Graphical depiction of DL architectures. * FCN: fully connected network.

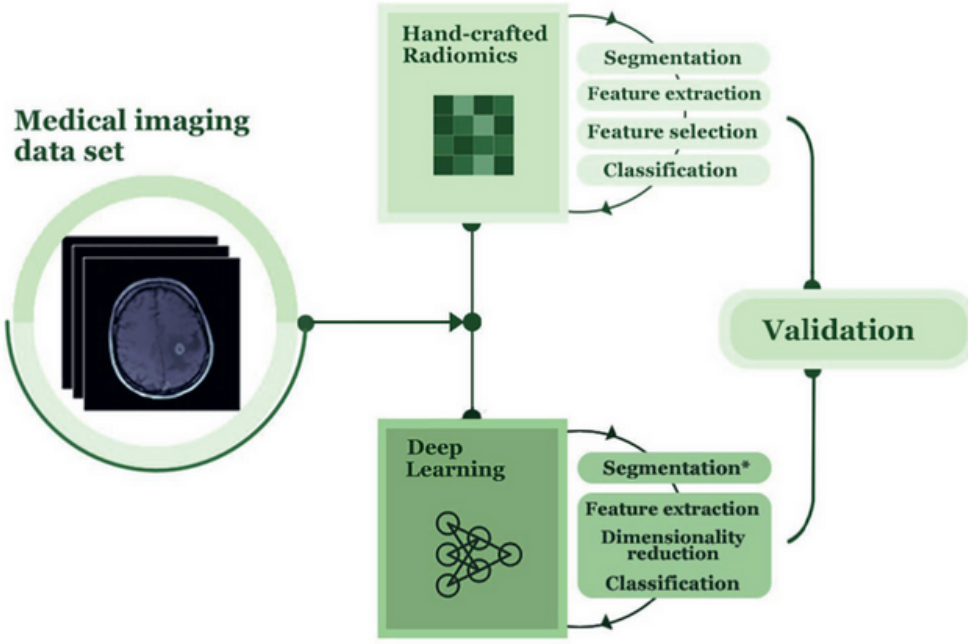


Figure 2. Development of imaging biomarkers using quantitative image analysis.
* Segmentation is not a necessity in the automated radiomics pipeline.

2.1. Handcrafted radiomics

Overall survival

Wang et al. (26) investigated the potential of radiomics signatures to predict overall survival in patients with locally advanced rectal cancer. The authors tried to address the current clinical need for a risk stratification tool for such patients to safely forgo surgical resection, due to the high comorbidities associated. The study included 411 treatment planning CT-scans of patients treated with neoadjuvant chemotherapy followed by surgery. The authors developed a radiomics signature that could stratify patients into low- and high-risk survival groups. The radiomics features included in the signature were found to be independent of the clinical features. Adding radiomics features to the clinical model resulted in an improvement of the predictive power (c- index) of the clinical only model from 0.67 (0.62–0.73) to 0.73 (0.66–0.80) (26). The authors used two investigations to ensure the selection of stable radiomics features, namely test–retest and contour- recontour robustness analysis. The results signifies the added value of properly using radiomics analysis on CT scans in improving patients' risk stratification. Yet, the authors did not externally validate their signature, casting doubt on the generalizability of their signature. It is expected to be of value in cases where the scanning parameters are identical to those used in the study.

Another study by Bae et al. (27) investigated the potential of MR-based radiomics to improve the survival prediction of patients diagnosed with glioblastoma multiforme. The study is an effort to address the unmet clinical need for assessing the survival of the target group following therapy. The authors extracted radiomics features from 217 multiparametric MR scans of patients with glioblastoma. The authors identified 18 radiomics features to build a radiomic signature, and reported that the addition of radiomics features to clinical and genetic profiles of the patients significantly improves the stratification of patients (27). The authors in this study applied a unique approach for the analysis by simultaneously analyzing radiomics features extracted from different co-registered MR sequences. The identified features were independent of the clinical and genetic factors, and the improvement in the survival prediction following their addition, supports the hypothesis of radiomics.

Pitfalls in the study include the lack of assessment of radiomic feature stability before modeling, and as often seen in these studies, a lack of an external validation of the signature. However, their results support the hypothesis that radiomics are

are of great use when applied on scans acquired using identical settings.
Oikonomou et al. (28) reported on the potential of PET/CT-based radiomics to improve the survival stratification of patients with lung cancer treated with stereotactic body radiotherapy. The aim was to identify radiomics features that can improve the prognostication of patients following treatment. The authors extracted radiomics features from 150 PET/CT scans, and built radiomics signatures using 10 radiomics features. The authors reported that the radiomics signature was the sole predictor in the case of overall survival, and provided complementary information for the prediction of regional control (28). The uniqueness in this study is the joint use of radiomics features extracted from the CT-component and PET-component of the PET/CT scans. The authors show how other currently used clinical parameters fail to predict overall survival, while only radiomics could. While the study highlights the potential of radiomics to improve risk stratification, no external validation of the signature was performed.

Progression free survival
Kirienko et al. (29) investigated the role of PET/CT-based radiomics to predict disease free survival in patients with non-small cell lung cancer undergoing surgery. The authors extracted radiomics features from PET, CT, and combined PET/CT images. The authors developed Cox regression models using only CT, only PET, and combined PET/CT radiomics features. They reported that the radiomic signatures they developed improve the current clinical stratification of the targeted patients (29). The authors in this study investigated the reproducibility of radiomics features across the different imaging parameters in their dataset. This ensured selecting the comparable features before proceeding with signature building. The authors also provide evidence of the added value of combining radiomics features extracted from different imaging modalities. Furthermore, the ability to predict disease free survival from the time of diagnosis -which radiomics offer improves physicians and patients decision making. However, the authors in this study did also not perform an external validation of their signature. Further validation of the signature can prompt a prospective validation trial, before incorporation into cDSS.
Another study by Kickingereder et al. (30) investigated the role of MR-based radiomics in predicting survival in patients with glioblastoma multiforme. The authors extracted radiomics features from 119 MR scans, and developed a radiomic signature using 11 features. The developed signature performed significantly better than the radiologic and clinical risk models, and its addition to

those resulted in an overall improvement of progression-free survival stratification (30). The finding that the radiomics signature performed better than the clinical and radiologic models supports the findings reported by Bae at al. (27), and adds more evidence that radiomics features decode complementary biologic information. However, the study did not address the issues of the reproducibility and generalizability sufficiently, leaving a room for improving the performance of radiomics.

Tumour histology

Wu et al. (31) explored the role of radiomics in differentiating between the histologic subtypes of non-small cell lung cancer: adenocarcinoma and squamous cell carcinoma. The study was an effort to address the clinical need for less invasive and easily repeatable methods to determine tumour histology. The authors extracted radiomics features from 350 CT scans of NSCLC patients for whom the tumour histology has been determined from surgical specimens. The developed signature included 5 radiomics features, and they reported an area under the receiver characteristics curve (AUC) of 0.72 (31). This study reflected on the potential of non-invasive radiomic signatures to differentiate between adenocarcinoma and squamous cell carcinoma. They also investigated different machine learning methodologies for building the radiomics signature. While this study generates evidence for the potential of radiomics, the performance of the developed signature is significantly lower than the current gold standard -tissue biopsy. However, there is a great room for improving the development and performance of the signature. The authors did not address the acknowledged challenges in radiomics, nor did they validate their signature on an external dataset. Preselection of reproducible features, external and prospective validation of the signature are necessary steps in the development of radiomics biomarkers.

In another study, Wu et al. (32) investigated the added value of MR-based radiomics features for the prediction of hepatocellular carcinoma (HCC) grade. The authors extracted radiomics features from 170 MRI scans of HCC patients, whose tumour grade was identified through pathological samples. The radiomics-only signature (AUC of 0.74) outperformed the clinical model (AUC of 0.60), and the combination of both significantly improved the prediction (AUC of 0.80) (32). The authors in this study also combined radiomics features extracted from two different MR sequences and analyzed them simultaneously. The significant improvement of the predictions following the combination of clinical and radiomics features supports the independence of radiomics

features from other clinical information. However, external validation of the developed signature is still a necessity before confidently performing prospective validation. Valleries et al. (33) explored the potential of the combination of FDG-PET- and MR-based radiomics features to classify lung nodules. The authors extracted radiomics features from 51 PET and MR scans of histologically confirmed lung lesions in patients with soft-tissue sarcoma. The authors achieved a sensitivity of 0.96 and specificity of 0.93 in diagnosing metastatic nodules using a model with combined radiomics features from both PET and MR modalities. The authors used a novel interesting approach by simultaneously analyzing the features extracted from FDG-PET and MR scans, and were the first to show the potential of this method. The performance of the developed signature makes it a suitable alternative for patients for whom tissue biopsy is contraindicated. Its possible translation to cDSS might significantly improve patient outcomes, as treatment is based on the histologic diagnosis. Yet, further external and prospective validation of the signature is needed.

Response to therapy

Trebeschi et al. (34) explored the role of radiomics in predicting response to anti-PD1 immunotherapy in patients diagnosed with advanced melanoma and NSCLC patients. Immunotherapy has shown promising results. Yet, there is still a need for a tool to determine which patients will benefit from receiving anti-PD-1 antibodies. The authors extracted radiomics features from 1055 ROIs segmented on 203 CT scans. The authors developed a radiomic signature that could predict the response to therapy with an AUC of 0.76; showing the potential of radiomics to predict response to therapy in such patients (34). Interestingly, the authors found correlations between the radiomic biomarker and the genes associated with cell cycle progression and mitosis. Radiomics can become a tool for assisting decision making in immunotherapy, a great unmet clinical need. The study however did not externally validate the signature, and did not sufficiently address the issues of feature stability and reproducibility. Therefore, the application of the developed signature is also limited to the patients who are scanned with the same scanning parameters as used in the training.

In a study by Horvat et al. (35), the authors investigated the role of radiomics in assessing complete clinical response (cCR) after neoadjuvant chemoradiotherapy (CRT) in patients with locally advanced rectal cancer. The guidelines of treating these patients include surgery, but evidence showed recently that a select group of patients can be safely treated with only CRT. The authors extracted

radiomics features from 114 MR scans, and developed a radiomics signature with a sensitivity of 1.00, and a specificity of 0.91, which outperformed qualitative assessment of the response performed by two radiologists. The current clinical standard evaluation of cCR includes digital rectal examination and endoscopy, with an accuracy ranging between 0.71 and 0.88 (35). The developed radiomic signature showed the highest accuracy among the available compared-with tools. Nonetheless, several steps to improve the methodology and performance of the radiomics signature could be made. The sound cCR evaluation following RCT can improve the patient management by eliminating surgical risks, time and money.

2.2. Deep learning

The application of deep learning on medical imaging could potentially fulfil more complicated tasks than handcrafted radiomics, especially when large amounts of data are available. Furthermore, as definition of the ROIs is not a necessity in the automated deep learning workflows, the algorithm will learn patterns from the whole image and possibly make connections with the habitat of the ROIs. The applications of neural networks on medical imaging are also not limited to classification and prediction of clinical end points, but can extend to include other tasks, such as the detection and segmentation of abnormalities or target volumes, which have been investigated for decades (36). Especially the detection and segmentation of lesions can be easily incorporated into the radiomics workflow, further automating the process. In the following paragraphs, we give examples of different applications of DL on medical imaging to perform various tasks on datasets acquired with one of the three main medical images modalities: CT, MRI, and PET.

Automatic segmentation of target structures

Jiang et al. (37) tried to develop a DL model that is able to accurately perform volumetric lung tumour segmentation on CT images. The authors used two versions of multiple resolution residual network models for the delineation of the ROIs. The authors used 377 tumours from the open source dataset available on The Cancer Imaging Archive (TCIA) (https://www.cancerimagingarchive.net) to train the model, and two independent datasets of 304 and 529 lung tumours to validate it. The dice similarity coefficient (DSC), which measures the spatial overlap of the segmentations, was computed to evaluate the

performance of the model. The DSCs of the model on the two validation datasets were 0.75 and 0.68, respectively. The authors reported that there was no significant difference between the DL-generated mask and experts' segmentations (37). The new approach for segmenting medical images used in this study shows to be superior to the traditional use of UNet. The approach generalizes well on external data and overcomes the multiple sizes problem. The major pitfalls is that the authors did not use the 3D geometry of the CTs to compute the results, which would probably increase the performance significantly. The translation of such a tool to clinical practice will significantly reduce the time spent by the clinicians to plan the treatment, or evaluate the response to therapy. Moreover, from a research perspective, it can significantly reduce the time needed for radiomics research, and it will address the issue of inter-observer sensitivity of radiomics features. In the study by Yi et al. (38), the authors developed a DL model for the segmentation of brain tumours based on 274 brain MRIs extracted from the Brain Tumour Image Segmentation Benchmark (BRATS) dataset (39).

Segmentation of brain Glioblastoma on MRI is a time-exhaustive process, and an automated, accurate and reproducible tool for this purpose is considered a clinical need. The model was trained using four different MRIs sequences. The particularity of their convolutional neural network (CNN) model is a fixed difference of Gaussian filters as a first convolution layer, as it was proven to be the most efficient for 3D segmentation. The DSC for the model was 0.89 on the BRATS dataset when compared to ground truth segmentations (38). This article shows the superiority of 3D CNN compared to 2D CNN. The algorithm generated segmentations with a volumetric overlap of 0.89 with the experts' segmentations, which shows the potential of these tools for clinical use. However, the lack of external validation in the study limits the applicability of the algorithm to scanning parameters in the training set. The clinical practice can benefit from such tools, as it significantly reduces the time the clinicians spend, and can provide more accurate evaluation of tumour response than the current clinical routine.

Chen et al. (40) explored the possibility of developing a DL model that is able to detect and segment cervical tumours on PET imaging. The authors proposed prior information constraint CNN (PIC-CNN), which integrates a CNN with prior information of cervical tumour. The authors reported a DSC of 0.84, which was superior to the other methods tin the comparison, including transfer learning based on fully convolutional neural networks (FCN) (DSC of 0.77), automatic thresholding (DSC of 0.59), and region growing method (DSC of 0.52) (40). The study highlights the potential of deep learning to perform well-defined and robust segmentations on PET imaging.

The novelty of the approach is the use of prior information as input of the model, with delineation of the bladder. This extra information seems to give the traditional model an advantage compared to models that solely segment the tumours. However, the results were not validated on an external dataset. The application of the developed algorithm -after validating it would decrease the need for tissue biopsy, as well as the time spent on segmenting the tumours manually or semi-automatically.

Oncologic classification tasks

Ardila et al. (41) tried to predict the risk of lung cancer using screening low-dose CTs. The algorithm is trained on screening low-dose CT scans of patients who were known to be at risk. The authors trained their DL model on approximately 7000 scans, and validated its performance on 1139 cases. The authors reported that the model achieved the "state-of-the-art" performance (AUC of 0.944). Furthermore, the model outperformed all the radiologists (n = 6) who were asked to give predictions. The model resulted in a significant reduction in the false positive (11%), and false negative rates (5%) (41). While the current low-dose CT screening protocol has substantially improved in terms of consistency, it still faces major limitations represented in the inter-observer variability and incomplete characterization of image findings. The authors in (41) developed an algorithm that achieved significantly better performance than the current protocol, highlighting the potential of DL algorithms to revolutionize the field of lung cancer screening. Other advantages of the algorithm are that it eliminates the current clinical practice limitations.

Ismael et al. (42) investigated the ability of DL algorithms to classify different brain tumours. The algorithm predicts if the lesion is either a meningioma, glioma, or pituitary tumour. The authors developed the algorithm on 3064 T1 MRI images from 233 cancer patients. As input to the algorithm, the 2D images were considered independent from each other, and were split into 80% training and 20% testing, with strictly different patient data. The classifier used is ResNet50, a classic deep learning network, and the resultant balanced accuracy was 0.99 on a slice level and 0.97 at a patient level. This study shows that deep learning can very accurately classify brain tumours based solely on MRI data. However, the data to be used should be acquired using the same scanning parameters, as no external validation was performed in this study. There is a great clinical significance from the development of such a cDSS, as it eliminates the need for risky brain biopsies, while maintaining high accuracy.

In another study by Sibille et al. (43), the authors used the combination of CT, fluorine 18-fluorodeoxyglucose PET, atlas and PET maximum intensity projection (MIP) imaging to classify lung nodules. The study included a set of 629 patients who were diagnosed with either lung cancer or lymphoma. The authors developed models using each of imaging modalities separately, as well as in combination. The recommended algorithm achieved an AUC of 0.98 when both CT and PET were combined (43). This study shows that the combination of CT and PET can achieve an outstanding performance in terms of predictions. The current clinical practice requires the clinician to review and classify all of the increased-uptake foci in a PET/CT scan. The algorithm could help the clinicians to quickly read those images, after highlighting the suspicious areas and their most likely classification using DL.

Non-oncologic classification tasks

Walsh et al. (44) explored the potential of DL to classify fibrotic lung diseases using high resolution CT scans. The current clinical guidelines for classifying fibrotic lung diseases are based on high resolution scans, and diagnoses are made based on the semantic features identified by the radiologists. While these guidelines are the current gold-standard, it suffers greatly from inter-observer variability. The authors tried to address this unmet clinical need using DL approaches. The authors trained their DL model on 929 CT scans, and tested it on 139 scans. The authors reported a performance with human-level accuracy (0.76) (44). Of interest, the algorithm developed had a better agreement with expert radiologists than among them. The ease of application of such methods in clinical settings could benefit clinical practice, especially in centers where such clinical expertise is scarce.

In the study by Ding et al. (45), the authors tried to develop a DL model that is able to diagnose Alzheimer's disease (AD), using 18F-FDG PET scans of the brain. The current clinical guidelines to diagnose AD necessitate the interpretation of scans by an expert, and there is no definitive biomarker. To investigate the potential of DL, the authors collected two datasets: one used for training and testing the model (n = 2109 scans), which was split into 90% training and 10% testing; and an independent dataset (n = 40) for the validation of the model. The authors reported an AUC of 0.98, sensitivity of 1.00 and specificity of 0.82, using scans acquired 75.8 months on average before establishing the diagnosis. The model further outperformed the readers' performance (sensitivity of 0.57 and specificity of 0.91) (45). The significance in this study lies

within the novelty of developing a biomarker for AD that is currently an unmet clinical need. In addition to the significantly better performance compared to human experts, the model can predict that the patient has AD in progression significantly earlier (~6 years). Such an application will revolutionize the clinical management of AD. However, prospective validation of this signature is needed before its translation to clinical practice.

Oh et al. (46) applied a DL based approach in order to classify the neuroimaging data related to AD. Authors used 694 MRI scans (T1-weighted MP-RAGE sequence) for solving several binary classification problems: AD vs. normal control (NC), progressive mild cognitive impairment (pMCI) vs. NC, stable mild cognitive impairment (sMCI) vs. NC and pMCI vs. sMCI. The authors utilized convolutional autoencoder- based unsupervised learning algorithms in order to classify the AD vs. NC. Following that, the authors applied a supervised transfer learning approach to classify the pMCI vs. sMCI. The developed algorithms achieved accuracies of 0.87, 0.77, 0.63, and 0.73 for the AD, pMCI, sMCI and pMCI vs. sMCI classifications, respectively. In comparison to Ding et al. (45), the authors in this study used different DL approaches, and less numbers of patients were available for training and testing the algorithm. Furthermore, the difference in the imaging modality analysed in each study could justify the variation in performance, as AD begins with functional impairment rather than structural changes. Although the model developed by oh et al. (46) was outperformed by human experts, the authors demonstrated the possibility of end-to-end DL algorithms, which could be translated to clinical use after further optimization and prospective validation.

Response to therapy

Lou et al. (47) reported on the potential of DL models to predict response to radiotherapy in patients with lung cancer (primary or metastatic) using CT scans. Currently, all patients are treated similarly, while personalizing radiotherapy remains a desired, but unmet clinical need. The authors in this study collected a total of 849 scans for training the DL algorithm, and 95 scans to validate it. The authors developed a deep learning model (deep profiler) that computes and includes radiomics features in the deep-profiling process. A model combining the deep profiler and clinical variables is then used to calculate a risk score that is used to predict the response to treatment. The algorithm classifies patients into high and low risk groups, with a high performance (c-index of 0.72), which is significantly better compared to the results obtained with solely handcrafted radiomic models (c-index between 0.65 and 0.68) (47). The algorithm developed in this study opens new potentials for individualizing radiotherapy based on patients' sensitivity. Thereby, avoiding over- or under-treatment, and the side-effects of unnecessary treatment. Nevertheless, proper prospective validation of the developed algorithm remains a necessity.

Ypsilantis et al. (48) used convolutional neural networks to develop a model that is capable of predicting response to neo-adjuvant chemotherapy (NAC) in patients with esophageal cancer using PET scans. NAC is considered a standard of care in some cancers. While NAC has favourable outcomes in patients who respond, patients who do not end up with worse outcomes. To investigate the potential of QIA techniques, the authors collected 107 PET scans of patients diagnosed with esophageal cancer, treated with NAC, and followed-up to determine response. The authors compared the performance of handcrafted radiomics with deep learning approaches. The authors reported that the developed deep learning algorithm outperformed the handcrafted radiomics model, and achieved a sensitivity of 0.81 and specificity of 0.82 (48). The algorithm developed in this study highlights the potential of using DL to predict patients' response to therapy at baseline, which is considered a substantial clinical added value.

3. Challenges and future directions

Biomarkers are defined as “objective indications of medical state observed from outside the patient – which can be measured accurately and reproducibly” (49). The core of choosing a biomarker is the ability to measure it objectively. The reproducibility of imaging quantitative features across different imaging parameters is currently the steepest hurdle in QIA. As more research is being performed, other challenges, such as the sensitivity of QIA features to variations in the segmentation of the ROIs; and the lack of feature reproducibility across different implementations of radiomics toolboxes, are becoming increasingly clear.

3.1. The stability and reproducibility of quantitative features

Since the first landmark study in radiomics by Aerts et al. (50), the sensitivity of radiomics features to repeated acquisitions has been acknowledged. The authors performed a test-retest stability investigation and used 100 out of 440 calculated radiomics features based on the stability rank of the features. The authors also acknowledged the sensitivity of features to differences in segmentations, and performed a primary feature selection based on the features’ robustness with regards to differences in both test-retest and segmentations. More recently, several studies reported on the sensitivity of radiomics features to temporal changes in test-retest studies across different modalities, including CT, MRI, and PET.

Anatomical imaging

Anatomical imaging (CT and MRI) is used to explore the underlying anatomical structures. CT imaging is standardized using the hounsfield units (HU) (51). On the other hand, MR imaging has no such standardized intensity measurements (52). Even though CT imaging uses standardized measurements, CT-based radiomics are not necessarily reproducible. Several studies reported that a significant number of CT- based radiomics features are not reproducible in test-retest settings, where the scans are acquired using the same scanning parameters (53–55). Other studies that investigated the reproducibility of CT-based radiomics features across different

imaging acquisition and reconstruction parameters reported that the majority of radiomics features are significantly affected by such differences (53,56,57). Unreproducible radiomics features should be removed before starting the modeling of radiomics signatures. Therefore, it is always necessary to perform preselection of stable radiomics features based on the data under study, before starting the modeling. MR-based radiomics is even more complex and challenging to standardize compared to CT based radiomics, as more factors -in addition to lack of standardized intensity measurements affect MR imaging (58). Some studies reported on the stability of various MR-based features. Fiset et al. (59) investigated the reproducibility of T2- weighted MRI of cervical cancer in three different settings: (i) test–retest; (ii) diagnostic MRI versus simulation MRI; (iii) interobserver variability. The authors reported that 22.6%, 6.2% and 74.4% of 1761 extracted radiomics features were reproducible across test-retest, diagnostic versus simulation MRI, and different observers, respectively. Semi-parametric maps derived from specialized MRI sequences suffer less from the lack of stability: Peerlings et al. (60) reported on the stability of radiomics features extracted from apparent diffusion coefficient (ADC) map in test-retest and across different cancer types, centers, and vendors. The authors reported that out of 1322 extracted radiomics features, 122 features were stable across all cancers, centers, and vendors.
On top of these challenges, using contrast agents for imaging adds another level of complexity to the reproducibility of features, due to the differences in the cardiac function of patients being scanned. Changes in cardiac function can affect the time the distribution of the contrast in the body takes (61). Another factor in contrast-enhanced images is the difference in time between the injection of the contrast and scan acquisition, which might be slightly different across centers and protocols.

Functional imaging

Functional imaging is used to assess the metabolic activity of a region of interest, and includes the injection of radiopharmaceuticals. Some standardized measurements in PET are already being extracted and used in clinical practice, such as the standardized uptake value (SUV), and the metabolically active tumour volume (MTV) (7).
The challenges of radiomics for functional imaging are similar to the challenges of contrast-enhanced anatomical imaging radiomics, where the variability in the injected radiopharmaceutical activity, the time between injection and image acquisition, and acquisition time per bed position have profound

implications on the reproducibility of radiomics features (62). In addition, functional imaging lacks anatomical specificity and suffers from low resolution, which could be addressed by the use of hybrid imaging (22). Tixier et al. (63) investigated the reproducibility of SUV measurements, intensity histogram features, intensity-size zone features, and co-occurrence matrices features. The authors acquired two 18F-FDG PET scans of 16 patients, with a 4-days' time interval. In contrast to further studies, the authors reported that these features were insensitive to the discretization range. Hatt et al. (64) investigated the robustness of PET based heterogeneity textural features with respect to the delineation of functional volumes and partial volume effects correction. The authors reported that these features were significantly affected by the differences in the delineation. The authors further reported that local features, e.g entropy and heterogeneity, were more robust when compared to regional features, e.g intensity variability and size-zone variability. Leijenaar et al. (65) investigated the role of SUV discretization on radiomics features. The authors used two different methods for SUV discretization, and reported that differences in SUV discretization significantly affect the reproducibility of 18F-FDG PET based radiomics features. The authors recommended the standardization of methodology for radiomics analysis. Altazi et al. (66) investigated the reproducibility of PET based radiomics features in cervical cancer patients. The authors investigated the reproducibility in three different scenarios: (i) manual versus computer-aided segmentations, (ii) gray-level discretization, and (iii) reconstruction algorithms. The authors extracted 79 PET radiomics features, and reported that the percentage of stable features in the three scenarios were 13%, 5%, and 1% respectively. Shiri et al. (67) explored the effects of different reconstruction on 18F-FDG PET radiomics. The authors studied the effects of several factors including number of sub- iterations, number of subsets, full width at half maximum (FWHM) of Gaussian filter, and scan time per bed position and matrix size. The authors reported that 47% of the features were found to be robust, and these include shape, 44% of the intensity based features, and 41% of the texture based features. However, with changes in matrix size, the authors reported that only 6% of the features were robust.

The discrepancies in the reported percentages of stable/reproducible features across the reported studies are most likely linked to the variations between the datasets used in each of the studies in the scanners, and scans acquisition and reconstruction parameters combinations. However, these discrepancies are expected because of the different complexity of radiomics features, as well as the interaction between the different scanning parameters. All of the

scanning parameters. All of the above mentioned studies reported that a variable percentage of radiomics features are affected, which highlights the necessity of performing feature stability/ reproducibility studies based on the data under analysis before performing radiomics analysis.

3.2. Sensitivity of quantitative imaging features to variations in the segmentation of the ROIs

In QIA, the medical images are converted to numerical arrays before feature calculation. Consequently, it is intuitive that differences in segmentations affect the quantitative imaging feature values variably, depending on the feature definition. Many studies have identified lists of radiomics features that are robust to variability in segmentations (50,68,69). Furthermore, with the inclusion of deep learning methods in image analysis, efforts are being made to develop reliable and reproducible automatic segmentations of various regions of interest as described in 3.2.1. Deep learning suffers less in this aspect, as the provision of ROIs is not obligatory.

3.3. The different implementations of radiomics feature extraction toolboxes

It is common knowledge in the radiomics community that different radiomics toolboxes use different pre-processing techniques and/or feature definitions, which lead(s) to variations in estimation of radiomics feature values when different software solutions are used. To address this issue, the radiomics community started an initiative – Imaging Biomarkers Standardization Initiative (IBSI) – that aims at standardizing radiomics feature extraction using different toolboxes (70). To date, the IBSI standardized the extraction of 169 radiomics features (71). Limiting the radiomics analysis to the IBSI standardized features can facilitate radiomics features interchangeability across platforms.

3.4. Future directions

To address the issue of radiomics features reproducibility, some harmonization methods have been investigated in the literature. Of the trending methods is Combine Batches (ComBat). ComBat is a statistical method that was developed to remove the batch effects in microarray expressions (72). While several studies have reported on the application of ComBat harmonization in radiomics analysis as a means to remove batch effects (73,74), its direct application on radiomics data is not in concordance with the mathematical definition of ComBat (72), or with the hypothesis that radiomics correlate with biology. This is because ComBat assumes that the differences between batches are attributed to two groups of factors, the first group refers to the biological covariates, which radiomics features are investigated for correlations with. Moreover, adding biologic covariates for ComBat in the training of radiomics signatures will hinder its prospective use, because it will be the outcome the radiomic signature tries to predict. The second group refers to the "non-biologic" factors, such as image acquisition and reconstruction parameters. ComBat was defined to handle one batch effect at a time. In contrast to gene expression arrays for which ComBat was designed, radiomics features have different complexity levels, which are expected to be non-uniformly affected by the variations in imaging parameters. In addition, the differences in image acquisition and reconstruction settings in a given retrospective imaging dataset are usually in more than one imaging parameter. The proper use of ComBat would require the assessment of the reproducibility of radiomics features after applying ComBat on representative objects with no biologic variations, such as phantoms. Then, radiomics features extracted from patients' scans acquired with the same imaging parameters can be transformed based on the location/scale parameters estimated by the application of ComBat on the phantom data. We here propose a framework for performing robust radiomics analysis (Figure 3). Nonetheless, a radiomics-specific harmonization method is still needed to eliminate the need for phantom studies, as the performance of ComBat is expected to be dependent on the variations in scanning parameters in the data. The workflow consists of consecutive steps, and can be used to preselect reproducible and harmonizable radiomics features. The first step in the workflow is the collection of retrospective patient imaging data to be analyzed. In the second step, scan acquisition and reconstruction parameters must be extracted from the collected patient data. The next step includes scanning a phantom with the parameters extracted from the patient imaging data. This allows the assessment of the reproducibility of radiomics

features across the different scan acquisition and reconstruction parameters, and the selection of those features for performing robust radiomics analysis.

Based on our review of existing literature and our own experience, in order to use ComBat in the context of radiomics analysis (steps 5–7), two extra steps are needed. After selecting the features that are insensitive to the variations in the scanning parameters extracted from the patient data, features that are reproducible in test-retest in each of the combinations of those scanning parameters must be identified. ComBat is then applied on the features that are reproducible in test- retest but not across different scanning parameters. The concordance of radiomics features is assessed following the application of ComBat. The location/scale shift parameters estimated by performing ComBat on the phantom data are then applied to the radiomics features extracted from patient data to harmonize them. The combination of the identified stable and harmonizable features can be further used to build the radiomics signature.

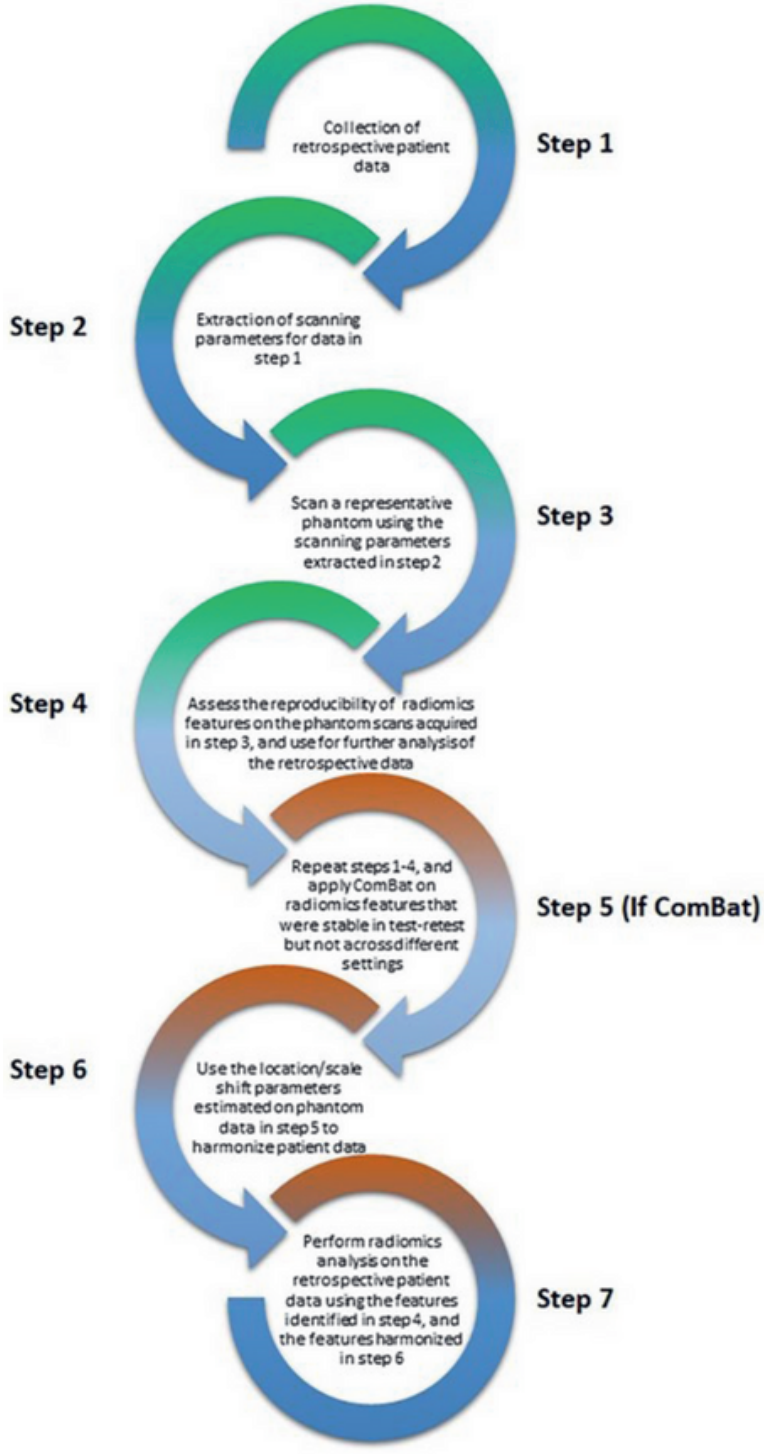


Figure 3. Proposed workflow for robust radiomics analysis.

The challenges discussed above raise questions about the future applications of radiomics, and the development of radiomic signatures as clinical biomarkers. To begin with, how to approach the concept of external validation in radiomics studies. Do radiomic signatures need to be externally validated as is the case with other biomarkers, given all the challenges of reproducibility across different imaging settings? Or would the observatory prospective validation of a given signature in a specific image setting suffice? Does the development of radiomic signatures need to be specific for a scanner model and imaging settings? The ultimate solution will be the development of specific quantitative imaging parameters, as there is currently a clinical direction to personalize imaging settings per patient, which will have its toll on radiomics analysis. The direct application of radiomics analysis on data acquired heterogeneously could lead to spurious results, and inability of translating the results in a meaningful manner.

4. Conclusion

Quantitative imaging techniques (radiomics and deep learning) present a perfect candidate for personalizing patients' management. Applying these techniques in a sound manner can provide highly accurate and reproducible tools that minimize costs and time loss. However, to incorporate QIA in cDSS, the quantitative features should fulfil the definition of a biomarker, namely the stability and reproducibility. The future of quantitative image analysis in general lies within harmonizing the imaging protocols across centers and scanners, or within the development of a unique global protocol for quantitative analysis scans. Hence, the development of radiomics-specific tools to harmonize medical images and facilitate meaningful quantitative image analysis of the currently available retrospective data remains a necessity. Our proposed framework is expected to improve the robustness of radiomics analysis. Nevertheless, the benefits of the proper application and translation of QIA on medical imaging are undoubted. QIA techniques will be a valuable asset for both: the clinicians and patients. QIA can become an efficient means for aiding clinicians in risk stratification, early diagnosis, and improved management of patients.

Competing interests

Dr. Philippe Lambin reports, within and outside the submitted

work, grants/sponsored research agreements from Varian medical, Oncoradiomics, ptTheragnostic, Health Innovation Ventures and DualTpharma. He received an advisor/presenter fee and/or reimbursement of travel costs/external grant writing fee and/or in kind manpower contribution from Oncoradiomics, BHV, Merck and Convert pharmaceuticals. Dr. Lambin has shares in the company Oncoradiomics SA and Convert pharmaceuticals SA and is co-inventor of two issued patents with royalties on radiomics (PCT/NL2014/050248, PCT/ NL2014/050728) licensed to Oncoradiomics and one issue patent on mtDNA (PCT/EP2014/059089) licensed to ptTheragnostic/DNAmito, three non-patentable invention (softwares) licensed to ptTheragnostic/ DNAmito, Oncoradiomics and Health Innovation Ventures. Dr. Woodruff has (minority) shares in the company Oncoradiomics.

CRediT authorship contribution statement

A. Ibrahim: Conceptualization, Methodology, Formal analysis, Data curation, Writing - original draft, Project administration. S. Primakov: Formal analysis, Data curation, Writing - original draft, Visualization.
M. Beuque: Formal analysis, Data curation, Writing - original draft.
H.C. Woodruff: Supervision, Writing - review & editing. I. Halilaj: Visualization. G. Wu: Resources, Data curation. T. Refaee: Resources.
R. Granzier: Resources. Y. Widaatalla: Resources. R. Hustinx: Supervision. F.M. Mottaghy: Supervision, Writing - review & editing.
P. Lambin: Conceptualization, Methodology, Writing - review & editing, Project administration, Supervision.

Acknowledgements

Authors acknowledge financial support from ERC advanced grant (ERC-ADG-2015, n° 694812 - Hypoximmuno), ERC-2018-PoC (n° 81320– CL-IO). We further acknowledge the financial support from Maastricht-Liege imaging valley grant. This research is also supported by the Dutch technology Foundation STW (grant n° P14-19 Radiomics STRaTegy), which is the applied science division of NWO, and the Technology Programme of the Ministry of Economic Affairs. Authors also acknowledge financial support from SME Phase 2 (RAIL – n°673780), EUROSTARS (DART, DECIDE), the European

Program H2020-2015-17 (BD2Decide - PHC30-689715, ImmunoSABR – n° 733008, PREDICT – ITN – n° 766276), TRANSCAN Joint Transnational Call 2016 (JTC2016 'CLEARLY'- n° UM 2017-8295) and Interreg V-A Euregio Meuse-Rhine ('Euradiomics'). Authors further acknowledge financial support by the Dutch Cancer Society (KWF Kankerbestrijding), Project number 12085/2018-2.

References

1. S. Walsh, E.E.C. de Jong, J.E. van Timmeren, A. Ibrahim, I. Compter, J. Peerlings, S. Sanduleanu, T. Refaee, S. Keek, R.T.H.M. Larue, Y. van Wijk, A.J.G. Even, A. Jochems, M.S. Barakat, R.T.H. Leijenaar, P. Lambin, Decision support systems in oncology, JCO Clin. Cancer Inform. 3 (2019) 1–9, https://doi.org/10.1200/CCI.18. 00001.
2. P. Lambin, E. Rios-Velazquez, R. Leijenaar, S. Carvalho, R.G.P.M. van Stiphout, P. Granton, C.M.L. Zegers, R. Gillies, R. Boellard, A. Dekker, H.J.W.L. Aerts, Radiomics: extracting more information from medical images using advanced feature analysis, Eur. J. Cancer 48 (2012) 441–446, https://doi.org/10.1016/j.ejca. 2011.11.036.
3. R.J. Gillies, P.E. Kinahan, H. Hricak, Radiomics: images are more than pictures, they are data, Radiology 278 (2016) 563–577, https://doi.org/10.1148/radiol. 2015151169.
4. P. Lambin, R.T.H. Leijenaar, T.M. Deist, J. Peerlings, E.E.C. de Jong, J. van Timmeren, S. Sanduleanu, R.T.H.M. Larue, A.J.G. Even, A. Jochems, Y. van Wijk, H. Woodruff, J. van Soest, T. Lustberg, E. Roelofs, W. van Elmpt, A. Dekker, F.M. Mottaghy, J.E. Wildberger, S. Walsh, Radiomics: the bridge between medical imaging and personalized medicine, Nat. Rev. Clin. Oncol. 14 (2017) 749–762, https://doi.org/10.1038/nrclinonc.2017.141.
5. T. Refaee, G. Wu, A. Ibrahim, I. Halilaj, R.T.H. Leijenaar, W. Rogers, H.A. Gietema, L.E.L. Hendriks, P. Lambin, H.C. Woodruff, The emerging role of radiomics in COPD and lung cancer, Respiration 99 (2020) 99–107, https://doi.org/10.1159/ 000505429.
6. R.C. Hardie, S.K. Rogers, T. Wilson, A. Rogers, Performance analysis of a new computer aided detection system for identifying lung nodules on chest radiographs, Med. Image Anal. 12 (2008) 240–258, https://doi.org/10.1016/j.media.2007.10. 004.
7. G.J.R. Cook, M. Siddique, B.P. Taylor, C. Yip, S. Chicklore,V. Goh, Radiomics in PET: principles and applications, Clin. Transl. Imaging

2 (2014) 269–276, https:// doi.org/10.1007/s40336-014-0064-0.
8. S. Chicklore, V. Goh, M. Siddique, A. Roy, P.K. Marsden, G.J.R. Cook, Quantifying tumour heterogeneity in 18F-FDG PET/CT imaging by texture analysis, Eur. J. Nucl. Med. Mol. Imaging 40 (2013) 133–140, https://doi.org/10.1007/s00259-012- 2247-0.
9. C. Swanton, Intratumor heterogeneity: evolution through space and time, Cancer Res. 72 (2012) 4875–4882, https://doi.org/10.1158/0008-5472.CAN-12-2217.
10. M. Gerlinger, A.J. Rowan, S. Horswell, M. Math, J. Larkin, D. Endesfelder, E. Gronroos, P. Martinez, N. Matthews, A. Stewart, P. Tarpey, I. Varela, B. Phillimore, S. Begum, N.Q. McDonald, A. Butler, D. Jones, K. Raine, C. Latimer, C.R. Santos, M. Nohadani, A.C. Eklund, B. Spencer-Dene, G. Clark, L. Pickering, G. Stamp, M. Gore, Z. Szallasi, J. Downward, P.A. Futreal, C. Swanton, Intratumour heterogeneity and branched evolution revealed by multiregion sequencing, N. Engl. J. Med. 366 (2012) 883–892, https://doi.org/10.1056/NEJMoa1113205. 33
11. T.M. Soo, M. Bernstein, J. Provias, R. Tasker, A. Lozano, A. Guha, Failed stereotactic biopsy in a series of 518 cases, Stereotact. Funct. Neurosurg. 64 (1995) 183–196, https://doi.org/10.1159/000098747.
12. S.S.F. Yip, H.J.W.L. Aerts, Applications and limitations of radiomics, Phys. Med. Biol. 61 (2016) R150–66, https://doi.org/10.1088/0031-9155/61/13/R150.
13. Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (2015) 436–444, https://doi.org/10.1038/nature14539.
14. W.S. McCulloch, W. Pitts, A logical calculus of the ideas immanent in nervous activity, Bull. Math. Biophys. 5 (1943) 115–133, https://doi.org/10.1007/ BF02478259.
15. L. Hongtao, Z. Qinchuan, Applications of Deep Convolutional Neural Network in Computer Vision, J. Data Acquisition Process. (2016). http://en.cnki.com.cn/ Article_en/CJFDTotal-SJCJ201601001.htm.
16. H. Shirani-Mehr, Applications of deep learning to sentiment analysis of movie reviews, Tech. Rep. NAVTRADEVCEN (2014) 1-8.
17. L. Deng, A tutorial survey of architectures, algorithms, and applications for deep learning, APSIPA Trans. Signal InformationProcessing 3 (2014), https://doi.org/ 10.1017/atsip.2013.9.
18. K. Janocha, W.M. Czarnecki, On loss functions for deep neural networks in classification, Schedae Informaticae. 1/2016 (2017). doi: 10.4467/20838476si.16.004. 6185.
19. A. Voulodimos, N. Doulamis, A. Doulamis, E. Protopapadakis, Deep learning for computer vision: a brief review, Comput. Intell. Neurosci. 2018 (2018) 7068349, https://doi.org/10.1155/2018/7068349.

20. R. Simhambhatla, K. Okiah, S. Kuchkula, R. Slater, Self-Driving Cars: Evaluation of Deep Learning Techniques for Object Detection in Different Driving Conditions, SMU Data Science Review. 2 (2019) 23. https://scholar.smu.edu/datasciencereview/vol2/iss1/23/ (accessed May 14, 2020).
21. D. Shen, G. Wu, H.-I. Suk, Deep learning in medical image analysis, Annu. Rev. Biomed. Eng. 19 (2017) 221–248, https://doi.org/10.1146/annurev-bioeng- 071516-044442.
22. A. Ibrahim, M. Vallières, H. Woodruff, S. Primakov, M. Beheshti, S. Keek, T. Refaee,
S. Sanduleanu, S. Walsh, O. Morin, P. Lambin, R. Hustinx, F.M. Mottaghy, Radiomics analysis for clinical decision support in nuclear medicine, Semin. Nucl. Med. 49 (2019) 438–449, https://doi.org/10.1053/j.semnuclmed.2019.06.005.
23. L.R. Cardon, H. Watkins, Waiting for the working draft from the human genome project. A huge achievement, but not of immediate medical use, BMJ 320 (2000) 1223–1224, https://doi.org/10.1136/bmj.320.7244.1223.
24. N.J. Schork, Personalized medicine: time for one-person trials, Nature 520 (2015) 609–611, https://doi.org/10.1038/520609a.
25. C. Parmar, J.D. Barry, A. Hosny, J. Quackenbush, H.J.W.L. Aerts, Data analysis strategies in medical imaging, Clin. Cancer Res. 24 (2018) 3492–3499, https://doi. org/10.1158/1078-0432.CCR-18-0385.
26. J. Wang, L. Shen, H. Zhong, Z. Zhou, P. Hu, J. Gan, R. Luo, W. Hu, Z. Zhang, Radiomics features on radiotherapy treatment planning CT can predict patient survival in locally advanced rectal cancer patients, Sci. Rep. 9 (2019) 15346, https://doi.org/10.1038/s41598-019-51629-4.
27. S. Bae, Y.S. Choi, S.S. Ahn, J.H. Chang, S.-G. Kang, E.H. Kim, S.H. Kim, S.-K. Lee, Radiomic MRI Phenotyping of Glioblastoma: Improving Survival Prediction, Radiology 289 (2018) 797–806, https://doi.org/10.1148/radiol.2018180200. 34
28. A. Oikonomou, F. Khalvati, P.N. Tyrrell, M.A. Haider, U. Tarique, L. Jimenez-Juan, M.C. Tjong, I. Poon, A. Eilaghi, L. Ehrlich, P. Cheung, Radiomics analysis at PET/CT contributes to prognosis of recurrence and survival in lung cancer treated with stereotactic body radiotherapy, Sci. Rep. 8 (2018) 4003, https://doi.org/10.1038/ s41598-018-22357-y.
29. M. Kirienko, L. Cozzi, L. Antunovic, L. Lozza, A. Fogliata, E. Voulaz, A. Rossi, A. Chiti, M. Sollini, Prediction of disease-free survival by the PET/CT radiomic signature in non-small cell
lung cancer patients undergoing surgery, Eur. J. Nucl. Med. Mol. Imaging 45 (2018) 207–217, https://doi.org/10.1007/s00259-017-3837-7.

30. P. Kickingereder, S. Burth, A. Wick, M. Götz, O. Eidel, H.-P. Schlemmer, K.H. Maier- Hein, W. Wick, M. Bendszus, A. Radbruch, D. Bonekamp, Radiomic profiling of glioblastoma: identifying an imaging predictor of patient survival with improved performance over established clinical and radiologic risk models, Radiology 280 (2016) 880–889, https://doi.org/10.1148/radiol.2016160845.
31. W. Wu, C. Parmar, P. Grossmann, J. Quackenbush, P. Lambin, J. Bussink, R. Mak, H.J.W.L. Aerts, Exploratory study to identify radiomics classifiers for lung cancer histology, Front. Oncol. 6 (2016) 71, https://doi.org/10.3389/fonc.2016.00071.
32. M. Wu, H. Tan, F. Gao, J. Hai, P. Ning, J. Chen, S. Zhu, M. Wang, S. Dou, D. Shi, Predicting the grade of hepatocellular carcinoma based on non-contrast-enhanced MRI radiomics signature, Eur. Radiol. 29 (2019) 2802–2811, https://doi.org/10.1007/s00330-018-5787-2.
33. M. Vallières, C.R. Freeman, S.R. Skamene, I. El Naqa, A radiomics model from joint FDG-PET and MRI texture features for the prediction of lung metastases in soft- tissue sarcomas of the extremities, Phys. Med. Biol. 60 (2015) 5471–5496, https://doi.org/10.1088/0031-9155/60/14/5471.
34. S. Trebeschi, S.G. Drago, N.J. Birkbak, I. Kurilova, A.M. Calin, A. Delli Pizzi, F. Lalezari, D.M.J. Lambregts, M. W. Rohaan, C. Parmar, E.A. Rozeman, K.J. Hartemink, C. Swanton, J.B.A.G. Haanen, C.U. Blank, E.F. Smit, R.G.H. Beets-Tan, H.J.W.L. Aerts, Predicting response to cancer immunotherapy using noninvasive radiomic biomarkers, Annals of. (2019). https://academic.oup.com/annonc/article-abstract/30/6/998/5416144.
35. N. Horvat, H. Veeraraghavan, M. Khan, I. Blazic, J. Zheng, M. Capanu, E. Sala, J. Garcia-Aguilar, M.J. Gollub, I. Petkovska, MR imaging of rectal cancer: radiomics analysis to assess treatment response after neoadjuvant therapy, Radiology 287 (2018) 833–843, https://doi.org/10.1148/radiol.2018172300.
36. J. Alirezaie, M.E. Jernigan, C. Nahmias, Automatic segmentation of cerebral MR images using artificial neural networks, IEEE Trans. Nucl. Sci. 45 (1998) 2174–2182, https://doi.org/10.1109/23.708336.
37. J. Jiang, Y.-C. Hu, C.-J. Liu, D. Halpenny, M.D. Hellmann, J.O. Deasy, G. Mageras, H. Veeraraghavan, Multiple resolution residually connected feature streams for automatic lung tumor segmentation from CT images, IEEE Trans. Med. Imaging 38 (2019) 134–144, https://doi.org/10.1109/TMI.2018.2857800.
38. D. Yi, M. Zhou, Z. Chen, O. Gevaert, 3-D Convolutional Neural Networks for Glioblastoma Segmentation, arXiv [cs.CV]. (2016). http://arxiv.org/abs/1611. 04534.

39. B.H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest, L. Lanczi, E. Gerstner, M.-A. Weber, T. Arbel, B.B. Avants, N. Ayache, P. Buendia, D.L. Collins, N. Cordier, J.J. Corso, A. Criminisi, T. Das, H. Delingette, Ç. Demiralp, C.R. Durst, M. Dojat, S. Doyle, J. Festa, F. Forbes, E. Geremia, B. Glocker, P. Golland, X. Guo, A. Hamamci, K.M. Iftekharuddin, R. Jena, N.M. John, E. Konukoglu, D. Lashkari, J.A. Mariz, R. Meier, S. Pereira, D. Precup, S.J. Price, T.R. Raviv, S.M.S. Reza, M. Ryan, D. Sarikaya, L. Schwartz, H.-C. Shin, J. Shotton, C.A. Silva, N. Sousa, N.K. 35
Subbanna, G. Szekely, T.J. Taylor, O.M. Thomas, N.J. Tustison, G. Unal, F. Vasseur, M. Wintermark, D.H. Ye, L. Zhao, B. Zhao, D. Zikic, M. Prastawa, M. Reyes, K. Van Leemput, The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS), IEEE Trans. Med. Imaging 34 (2015) 1993–2024, https://doi.org/10.1109/TMI.2014.2377694.
40. L. Chen, C. Shen, S. Li, G. Maquilan, K. Albuquerque, M.R. Folkert, J. Wang, Automatic PET cervical tumor segmentation by deep learning with prior information, in: Medical Imaging 2018: Image Processing, International Society for Optics and Photonics, 2018: p. 1057436. doi: 10.1117/12.2293926.
41. D. Ardila, A.P. Kiraly, S. Bharadwaj, B. Choi, J.J. Reicher, L. Peng, D. Tse, M. Etemadi, W. Ye, G. Corrado, D.P. Naidich, S. Shetty, Author Correction: End-to- end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography, Nat. Med. 25 (2019) 1319, https://doi.org/10.1038/s41591-019-0536-x.
42. S.A. Abdelaziz Ismael, A. Mohammed, H. Hefny, An enhanced deep learning approach for brain cancer MRI images classification using residual networks, Artif. Intell. Med. 102 (2020) 101779, https://doi.org/10.1016/j.artmed.2019.101779.
43. L. Sibille, R. Seifert, N. Avramovic, T. Vehren, B. Spottiswoode, S. Zuehlsdorff, M. Schäfers, 18F-FDG PET/CT uptake classification in lymphoma and lung cancer by using deep convolutional neural networks, Radiology 294 (2020) 445–452, https://doi.org/10.1148/radiol.2019191114.
44. S.L.F. Walsh, L. Calandriello, M. Silva, N. Sverzellati, Deep learning for classifying fibrotic lung disease on high-resolution computed tomography: a case-cohort study, Lancet Respir. Med. 6 (2018) 837–845, 30286-8.
45. Y. Ding, J.H. Sohn, M.G. Kawczynski, H. Trivedi, R. Harnish, N.W. Jenkins, D. Lituiev, T.P. Copeland, M.S. Aboian, C. Mari Aparici, S.C. Behr, R.R. Flavell, S.-Y. Huang, K.A. Zalocusky, L. Nardo, Y. Seo, R.A. Hawkins, M. Hernandez Pampaloni, D. Hadley, B.L. Franc, A deep learning model to predict a diagnosis of

Alzheimer disease by using 18F-FDG PET of the brain, Radiology 290 (2019) 456–464, https://doi.org/10.1148/radiol.2018180958.
46. K. Oh, Y.-C. Chung, K.W. Kim, W.-S. Kim, I.-S. Oh, Author Correction: Classification and visualization of Alzheimer's disease using volumetric convolutional neural network and transfer learning, Sci. Rep. 10 (2020) 5663, https://doi.org/10.1038/s41598-020-62490-1.
47. B. Lou, S. Doken, T. Zhuang, D. Wingerter, M. Gidwani, N. Mistry, L. Ladic, A. Kamen, M.E. Abazeed, An image-based deep learning framework for individualising radiotherapy dose: a retrospective analysis of outcome prediction, The Lancet Digital Health 1 (2019) e136–e147, https://doi.org/10.1016/s2589-7500(19)30058-5.
48. P.-P. Ypsilantis, M. Siddique, H.-M. Sohn, A. Davies, G. Cook, V. Goh, G. Montana, Predicting response to neoadjuvant chemotherapy with PET imaging using convolutional neural networks, PLoS ONE 10 (2015) e0137036, , https://doi.org/10.1371/journal.pone.0137036.
49. K. Strimbu, J.A. Tavel, What are biomarkers? Curr. Opin. HIV AIDS 5 (2010) 463 https://www.ncbi.nlm.nih.gov/pmc/articles/pmc3078627/.
50. H.J.W.L. Aerts, E.R. Velazquez, R.T.H. Leijenaar, C. Parmar, P. Grossmann, S. Carvalho, J. Bussink, R. Monshouwer, B. Haibe-Kains, D. Rietveld, F. Hoebers, M.M. Rietbergen, C.R. Leemans, A. Dekker, J. Quackenbush, R.J. Gillies, P. Lambin, Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach, Nat. Commun. 5 (2014) 4006, https://doi.org/10.1038/ncomms5006. 36
51. U. Schneider, E. Pedroni, A. Lomax, The calibration of CT Hounsfield units for radiotherapy treatment planning, Phys. Med. Biol. 41 (1996) 111–124, https://doi. org/10.1088/0031-9155/41/1/009.
52. L.G. Nyúl, J.K. Udupa, On standardizing the MR image intensity scale, Magn. Reson. Med. 42 (1999) 1072–1081. https://doi.org/3.0.co;2-m.“ > 10.1002/(sici)1522-2594(199912)42:6 < 1072::aid-mrm11 > 3.0.co;2-m.
53. R. Berenguer, M.D.R. Pastor-Juan, J. Canales-Vázquez, M. Castro-García, M.V. Villas, F. Mansilla Legorburo, S. Sabater, Radiomics of CT Features May Be Nonreproducible and Redundant: Influence of CT Acquisition Parameters, Radiology. 288 (2018) 407–415. doi: 10.1148/radiol.2018172361.
54. J.E. van Timmeren, R.T.H. Leijenaar, W. van Elmpt, J. Wang, Z. Zhang, A. Dekker, P. Lambin, Test-retest data for radiomics feature stability analysis: generalizable or study-specific? Tomography. 2 (2016) 361–365, https://doi.org/10.18383/j.tom. 2016.00208.

55. L. Lu, R.C. Ehmke, L.H. Schwartz, B. Zhao, Assessing agreement between radiomic features computed for multiple CT imaging settings, PLoS ONE 11 (2016) e0166550, https://doi.org/10.1371/journal.pone.0166550.
56. D. Mackin, X. Fave, L. Zhang, D. Fried, J. Yang, B. Taylor, E. Rodriguez-Rivera, C. Dodge, A.K. Jones, L. Court, Measuring computed tomography scanner variability of radiomics features, Invest. Radiol. 50 (2015) 757–765, https://doi.org/10.1097/RLI.0000000000000180.
57. I. Zhovannik, J. Bussink, A. Traverso, Z. Shi, P. Kalendralis, L. Wee, A. Dekker, R. Fijten, R. Monshouwer, Learning from scanners: bias reduction and feature correction in radiomics, Clin. Transl. Radiat. Oncol. 19 (2019) 33–38, https://doi.org/10.1016/j.ctro.2019.07.003.
58. A. Webb, G.C. Kagadis, Introduction to Biomedical Imaging, Med. Phys. 30 (2003) 2267–2267. doi: 10.1118/1.1589017.
59. S. Fiset, M.L. Welch, J. Weiss, M. Pintilie, J.L. Conway, M. Milosevic, A. Fyles, A. Traverso, D. Jaffray, U. Metser, J. Xie, K. Han, Repeatability and reproducibility of MRI-based radiomic features in cervical cancer, Radiother. Oncol. 135 (2019) 107–114, https://doi.org/10.1016/j.radonc.2019.03.001.
60. J. Peerlings, H.C. Woodruff, J.M. Winfield, A. Ibrahim, B.E. Van Beers, A. Heerschap, A. Jackson, J.E. Wildberger, F.M. Mottaghy, N.M. DeSouza, P. Lambin, Stability of radiomics features in apparent diffusion coefficient maps from a multi-centre test-retest trial, Sci. Rep. 9 (2019) 4800, https://doi.org/10. 1038/s41598-019-41344-5.
61. K.T. Bae, Intravenous contrast medium administration and scan timing at CT: considerations and approaches, Radiology 256 (2010) 32–61, https://doi.org/10. 1148/radiol.10090908.
62. G.J.R. Cook, G. Azad, K. Owczarczyk, M. Siddique, V. Goh, Challenges and promises of PET radiomics, Int. J. Radiat. Oncol. Biol. Phys. 102 (2018) 1083–1089, https://doi.org/10.1016/j.ijrobp.2017.12.268.
63. F. Tixier, M. Hatt, C.C. Le Rest, A. Le Pogam, L. Corcos, D. Visvikis, Reproducibility of tumor uptake heterogeneity characterization through textural feature analysis in 18F-FDG PET, J. Nucl. Med. 53 (2012) 693–700, https://doi.org/10.2967/jnumed.111.099127.
64. M. Hatt, F. Tixier, C.C. Le Rest, O. Pradier, Robustness of intratumour 18F-FDG PET uptake heterogeneity quantification for therapy response prediction in oesophageal carcinoma, European Journal of. (2013). https://link.springer.com/article/10.1007/s00259-013-2486-8.

65. R.T.H. Leijenaar, G. Nalbantov, S. Carvalho, W.J.C. van Elmpt, E.G.C. Troost, R. Boellaard, H.J.W.L. Aerts, R.J. Gillies, P. Lambin, The effect of SUV discretization in quantitative FDG-PET Radiomics: the need for standardized methodology in tumor texture analysis, Sci. Rep. 5 (2015) 11075, https://doi.org/10.1038/srep11075. 37
66. B.A. Altazi, G.G. Zhang, D.C. Fernandez, M.E. Montejo, D. Hunt, J. Werner, M.C. Biagioli, E.G. Moros, Reproducibility of F18-FDG PET radiomic features for different cervical tumor segmentation methods, gray-level discretization, and reconstruction algorithms, J. Appl. Clin. Med. Phys. 18 (2017) 32–48 https://aapm.onlinelibrary.wiley.com/doi/abs/10.1002/acm2.12170.
67. I. Shiri, A. Rahmim, P. Ghaffarian, P. Geramifar, H. Abdollahi, A. Bitarafan-Rajabi, The impact of image reconstruction settings on 18F-FDG PET radiomic features: multi-scanner phantom and patient studies, Eur. Radiol. 27 (2017) 4498–4509, https://doi.org/10.1007/s00330-017-4859-z.
68. R.T.H. Leijenaar, S. Carvalho, E.R. Velazquez, W.J.C. van Elmpt, C. Parmar, O.S. Hoekstra, C.J. Hoekstra, R. Boellaard, A.L.A.J. Dekker, R.J. Gillies, H.J.W.L. Aerts, P. Lambin, Stability of FDG-PET Radiomics features: an integrated analysis of test-retest and inter-observer variability, Acta Oncol. 52 (2013) 1391–1397, https://doi.org/10.3109/0284186X.2013.812798.
69. M. Pavic, M. Bogowicz, X. Würms, S. Glatz, T. Finazzi, O. Riesterer, J. Roesch, L. Rudofsky, M. Friess, P. Veit-Haibach, M. Huellner, I. Opitz, W. Weder, T. Frauenfelder, M. Guckenberger, S. Tanadini-Lang, Influence of inter-observer delineation variability on radiomics stability in different tumor sites, Acta Oncol. 57 (2018) 1070–1074, https://doi.org/10.1080/0284186X.2018.1445283.
70. M. Hatt, M. Vallieres, D. Visvikis, A. Zwanenburg, IBSI: an international community radiomics standardization initiative, J. Nucl. Med. 59 (2018) 287–287. http://jnm.snmjournals.org/content/59/supplement_1/287.abstract.
71. A. Zwanenburg, M. Vallières, M.A. Abdalah, H.J.W.L. Aerts, V. Andrearczyk, A. Apte, S. Ashrafinia, S. Bakas, R.J. Beukinga, R. Boellaard, M. Bogowicz, L. Boldrini, I. Buvat, G.J.R. Cook, C. Davatzikos, A. Depeursinge, M.-C. Desseroit, N. Dinapoli, C.V. Dinh, S. Echegaray, I. El Naqa, A.Y. Fedorov, R. Gatta, R.J. Gillies, V. Goh, M. Götz, M. Guckenberger, S.M. Ha, M. Hatt, F. Isensee, P. Lambin, S. Leger, R.T.H. Leijenaar, J. Lenkowicz, F. Lippert, A. Losnegård, K.H. Maier-Hein, O. Morin, H. Müller, S. Napel, C. Nioche, F. Orlhac, S. Pati, E.A.G. Pfaehler, A. Rahmim, A.U.K. Rao, J. Scherer, M.M. Siddique, N.M. Sijtsema, J. Socarras Fernandez, E. Spezi, R.J. H.M. Steenbakkers, S. Tanadini-Lang, D. Thorwarth, E.G.C. Troost, T. Upadhaya, V. Valentini, L.V. van Dijk, J. van

Griethuysen, F.H.P. van Velden, P. Whybra, C. Richter, S. Löck, The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping, Radiology. (2020) 191145. doi: 10.1148/radiol.2020191145.
72. W.E. Johnson, C. Li, A. Rabinovic, Adjusting batch effects in microarray expression data using empirical Bayes methods, Biostatistics 8 (2007) 118–127, https://doi.org/10.1093/biostatistics/kxj037.
73. F. Orlhac, F. Frouin, C. Nioche, N. Ayache, I. Buvat, Validation of A method to compensate multicenter effects affecting CT radiomics, Radiology 291 (2019) 53–59, https://doi.org/10.1148/radiol.2019182023.
74. F. Orlhac, S. Boughdad, C. Philippe, H. Stalla-Bourdillon, C. Nioche, L. Champion, M. Soussan, F. Frouin, V. Frouin, I. Buvat, A postreconstruction harmonization method for multicenter radiomic studies in PET, J. Nucl. Med. 59 (2018) 1321–1328, https://doi.org/10.2967/jnumed.117.199935.

CHAPTER 3:

PROGNOSTIC AND PREDICTIVE VALUE OF INTEGRATED QUALITATIVE AND QUANTITATIVE MAGNETIC RESONANCE IMAGING ANALYSIS IN GLIOBLASTOMA

Authors: Maikel Verduin,[1] Sergey Primakov,[1] Inge Compter, Henry C. Woodruff, Sander M. J. van Kuijk, Bram L. T. Ramaekers, Maarten te Dorsthorst, Elles G. M. Revenich, Mark ter Laan, Sjoert A. H. Pegge, Frederick J. A. Meijer, Jan Beckervordersandforth, Ernst Jan Speel, Benno Kusters, Wendy W. J. de Leng, Monique M. Anten, Martijn P. G. Broen, Linda Ackermans, Olaf E. M. G. Schijns, Onno Teernstra , Koos Hovinga, Marc A. Vooijs, Vivianne C. G. Tjan-Heijnen, Danielle B. P. Eekers, Alida A. Postma, Philippe Lambin,[2] Ann Hoeben[2]

1 These authors have contributed equally.
2 Share senior authorship.

Adapted from:
Maikel Verduin, Sergey Primakov, Inge Compter, Henry C. Woodruff, Sander M. J. van Kuijk, Bram L. T. Ramaekers, Maarten te Dorsthorst, Elles G. M. Revenich, Mark ter Laan, Sjoert A. H. Pegge, Frederick J. A. Meijer, Jan Beckervordersandforth, Ernst Jan Speel, Benno Kusters, Wendy W. J. de Leng, Monique M. Anten, Martijn P. G. Broen, Linda Ackermans, Olaf E. M. G. Schijns, Onno Teernstra , Koos Hovinga, Marc A. Vooijs, Vivianne C. G. Tjan-Heijnen, Danielle B. P. Eekers, Alida A. Postma, Philippe Lambin, Ann Hoeben, Prognostic and Predictive Value of Integrated Qualitative and Quantitative Magnetic Resonance Imaging Analysis in Glioblastoma, Cancers 2021, 13(4), 722; doi: https://doi.org/10.3390/cancers13040722

Acsess link:
https://www.mdpi.com/2072-6694/13/4/722

Simple Summary

Glioblastoma (GBM) is the most malignant primary brain tumor, for which improving patient outcome is limited by a substantial amount of tumor heterogeneity. Magnetic resonance imaging (MRI) in combination with machine learning offers the possibility to collect qualitative and quantitative imaging features which can be used to predict patient prognosis and relevant tumor markers which can aid in selecting the right treatment. This study showed that combining these MRI features with clinical features has the highest prognostic value for GBM patients; this model performed similarly in an independent GBM

Abstract

Glioblastoma (GBM) is the most malignant primary brain tumor for which no curative treatment options exist. Non-invasive qualitative (Visually Accessible Rembrandt Images (VASARI)) and quantitative (radiomics) imaging features to predict prognosis and clinically relevant markers for GBM patients are needed to guide clinicians. A retrospective analysis of GBM patients in two neuro-oncology centers was conducted. The multimodal Cox-regression model to predict overall survival (OS) was developed using clinical features with VASARI and radiomics features in isocitrate dehydrogenase (IDH)-wild type GBM. Predictive models for IDH-mutation, 06-methylguanine-DNA-methyltransferase (MGMT)-methylation and epidermal growth factor receptor (EGFR) amplification using imaging features were developed using machine learning. The performance of the prognostic model improved upon addition of clinical, VASARI and radiomics features, for which the combined model performed best. This could be reproduced after external validation (C-index 0.711 95% CI 0.64–0.78) and used to stratify Kaplan–Meijer curves in two survival groups (p-value < 0.001). The predictive models performed significantly in the external validation for EGFR amplification (area-under-the-curve (AUC) 0.707, 95% CI 0.582–8.25) and MGMT-methylation (AUC 0.667, 95% CI 0.522–0.82) but not for IDH-mutation (AUC 0.695, 95% CI 0.436–0.927). The integrated clinical and imaging prognostic model was shown to be robust and of potential clinical relevance. The prediction of molecular markers showed promising results in the training set but could not be validated after external validation in a clinically relevant manner. Overall, these results show the potential of combining clinical features with imaging features for prognostic

and predictive models in GBM, but further optimization and larger prospective studies are warranted.
Keywords: glioblastoma; radiomics; MRI; prognosis; prediction; machine learning; survival

1. Introduction

Glioblastoma (GBM) is the most malignant type of primary brain cancer with an incidence of 2–3 cases per 100,000 (1). Currently, a median survival of fifteen months is achieved with multimodal treatment (2) with a five-year overall relative survival of only 6.8% (3). However, despite this intensive treatment by neurosurgical intervention, concurrent chemoradiation and adjuvant temozolomide (TMZ) (2), GBM is still considered incurable and recurrence is inevitable. Although major improvements in the treatment of cancer have been made, the current standard-of-care for GBM has largely remained unchanged over the past decade.

GBM is diagnosed using gadolinium contrast-enhanced magnetic resonance imaging (MRI) followed by histopathological examination of tumor tissue specimen obtained after either biopsy or resection. Further characterization of GBM has led to the introduction of the 2016 updated world health organization (WHO) classification of central nervous system tumors (4). This classification integrates histopathological and morphological examination of the tumor with molecular markers (5). Thus far, the only predictive marker that has been established into clinical practice is the 06-methylguanine-DNA-methyltransferase (MGMT) methylation status, which is predictive of an improved response to alkylating chemotherapy such as TMZ (6). However, a substantial “grey zone” between MGMT methylated and unmethylated patients still exists for which the efficacy of TMZ is still to be determined (7). Additionally, the presence of a mutation in the isocitrate dehydrogenase (IDH) genes —which has been identified as a positive prognostic marker—is linked to dedifferentiated low-grade gliomas which have a distinctly different clinical behavior compared to IDH wild-type (WT) GBM (8). Epidermal growth factor receptor (EGFR) amplification is one of the most common genetic alterations (±50%) in GBM (9). This oncogenic molecular alteration poses a potential therapeutic target but also identifies a biological different subtype of GBM which responds differently to established treatments (10,11). However, the role of EGFR amplification as a prognostic factor still remains controversial (11,12,13) and studies using targeted agents for EGFR have so far been unsuccessful but are still ongoing (3). Additionally,

multiple other molecular targets (genetic mutations, amplifications and protein fusion products) have been identified which have either failed in previous clinical trials to improve patient survival or are currently still under investigation(3). All in all, the integration of molecular markers has led to an improvement in prediction of prognosis and treatment response but a substantial variety remains and no improvement in treatment outcome has been made, which is thought to be due to extensive inter- and intratumor heterogeneity (14).

Intratumor heterogeneity complicates treatment efficacy as different regions within the same tumor may contain cells having distinct genetic compositions, transcriptional subtypes and/or proliferation kinetics (3). Furthermore, temporal heterogeneity has been observed in which changes in the expression of molecular targets occur over time which limits efficacy of targeted approaches (15,16). In clinical practice and currently used diagnostic techniques and available prognostic models intratumor heterogeneity is not accounted for, since single-cell sequencing is not routinely used. Additionally, it is not clear if molecular GBM heterogeneity can be captured by qualitative and/or quantitative analysis of imaging features.

Imaging techniques have the advantage over standard pathological examination to also analyze the invasive, non-resected, components of GBM and thus capture and analyze the tumor as a whole. Especially temporal heterogeneity of expression of molecular targets cannot be evaluated using routine clinical diagnostics, as re-resection of tumors is not always feasible, making non-invasive imaging an interesting alternative. In order to make a standardized analysis of qualitative MR imaging features, the Visually Accessible Rembrandt Images (VASARI) features were previously developed (17). VASARI features include tumor size, location and morphology and have previously been shown to be reproducible and of prognostic value (17). Quantitative imaging analysis using radiomics is an approach to extract imaging features by high-throughput data mining on textures, shapes and intensities (18). Radiomics has shown prognostic and predictive potential in multiple solid tumors (19,20) including GBM (21). Furthermore, radiomics features have the potential to analyze the entire tumor and to identify intratumor molecular heterogeneity and underlying biological processes (22,23). In glioma, radiomics models have been developed to predict tumor grade (24), overall survival (OS) (25) and in GBM trying to predict molecular subtypes (26). Although IDH-mutation status is established as the best prognostic marker in GBM (27), defining different IDH wild-type GBM prognostic subgroups is still warranted due to their heterogeneous prognosis and clinical behavior.

The main challenge in developing prognostic and predictive

imaging-based models is their generalizability towards all GBM patients treated at different centers. Differences in diagnostic techniques (i.e., scanner vendors and protocols) and treatment and population variety can greatly influence model performances [28]. Due to these challenges, this study utilizes two multi-center datasets to train and validate the developed models.
The objective of this study was to investigate the additive value of qualitative and quantitative imaging heterogeneity analysis to established prognostic clinical features. These data were used to develop a prognostic model for OS in a real-world multi-center GBM population for IDH1/2 wild-type (IDH-WT) GBM. Furthermore, the value of imaging features as predictor for clinically relevant molecular markers for GBM was explored.

2. Results

2.1. Patient and Tumor Characteristics

In total, 142 patients were included in the training cohort and 46 patients in the validation cohort. Median OS was 12.0 months (range, 0–142 months) in the training cohort and 7.3 months (range, 0–30 months) in the validation cohort (log rank p-value, 0.001). Patients in the validation cohort more frequently received no adjuvant treatment, but these data were not available for all patients. Patient demographics, received treatment schedules and tumor characteristics are listed in Table 1. Molecular data for a subset of patients are reported as missing due to insufficient formalin-fixed paraffin-embedded (FFPE) material or poor quality or quantity of extracted DNA. VASARI features were available for all patients in both cohorts. For radiomics analysis, T1+Gadolinium and T2- weighted images were available for 105 patients in the training cohort and 44 patients in the validation cohort. MRI characteristics such as types and manufacturers of scanners and imaging protocols are reported in Figures S1 and S2. The numbers of patients that were eligible in the two cohorts for the different models are reported in Table S1.

Table 1. Overview of patient, treatment and tumor characteristics in the training and validation cohort.

Table 1. Overview of patient, treatment and tumor characteristics in the training and validation cohort.

Demographics	Training Cohort (n = 142)	Validation Cohort (n = 46)	p-Value
Median age at diagnosis (range)	61.4 years (15–85)	61.7 years (18–81)	0.991
Sex (%)	Male: 85 (59.9%) Female: 57 (40.1%)	Male: 29 (63.0%) Female: 17 (37.0%)	0.258
Treatment characteristics			
Surgical treatment (%)	Biopsy: 54 (38.0%) Debulking: 88 (62.0%)	Biopsy: 17 (37.0%) Debulking: 29 (63.0%)	0.112
Adjuvant treatment (%)	STUPP completed: 67 (47.2%) STUPP not completed or Non-STUPP regimen: 75 (52.8%)	STUPP completed: 17 (37.0%) STUPP not completed or Non-STUPP regimen: 16 (34.8%) Missing: 13 (28.2%)	0.288
Tumor characteristics			
Isocitrate dehydrogenase (IDH1) (R132H) mutation status (%)	*IDH1/2*-WT: 129 (91.5%) *IDH1*-mutation: 12 (8.5%) Missing: 1	*IDH1/2*-WT: 39 (84.8%) *IDH1*-mutation: 5 (10.9%) Missing: 2	1.000
Methylguanine methyltransferase (MGMT)-methylation status (%)	*MGMT*-methylated: 37 (26.2%) *MGMT* non-methylated: 104 (73.8%) Missing: 1	*MGMT*-methylated: 18 (39.1%) *MGMT* non-methylated: 26 (56.5%) Missing: 2	0.045
Epidermal growth factor receptor (EGFR) amplification status (%)	*EGFR* amplified: 47 (37.3%) *EGFR* non-amplified: 79 (62.7%) Missing: 16	*EGFR* amplified: 20 (43.5%) *EGFR* non-amplified: 26 (56.5%) Missing: 0	0.738

2.2. Prognostic Value of Integrative MRI Imaging Analysis in IDH-Wild Type GBM Population

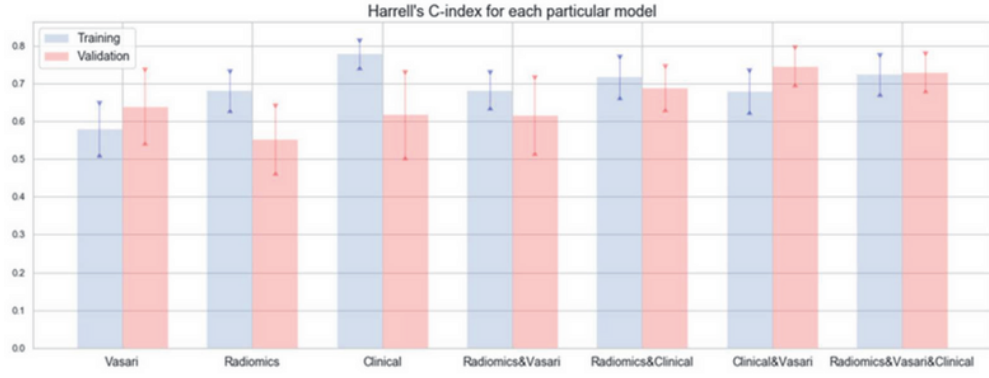
Median OS was 11.2 months (1.2–132.80 months) in the training cohort and 7.0 months (0.4–29.4 months) in the validation cohort in the IDH-WT GBM population. Univariate Cox-regression analysis of VASARI features for OS in the training cohort resulted in 13 features selected for inclusion in multivariable analysis (Table S2). The multivariable Cox-regression model consisted of five VASARI features (Model 1). For radiomics, five radiomics features were selected to predict OS (Model 2) (Table 2). In this study, none of the radiomics features showed evidence of a significant correlation with tumor volume (Figure S3). Additionally, no significant correlation were found between VASARI, radiomics and clinical features (Figure S4). An elaborate explanation of these radiomics features can be found on the Pyradiomics website (29) and in a previous study (30).

Table 2. Multivariate Cox-regression model using Visually Accessible Rembrandt Images (VASARI), radiomics and/or clinical features for overall survival (OS) prediction in isocitrate dehydrogenase wild type (IDH-WT) glioblastoma (GBM) patients in different prognostic models based on the training cohort (n = numbers of patients used for model development).

Prognostic Model Variables	Hazard Ratio (95% CI)	*p*-Value
Model 1: VASARI features model ($n = 129$)		
Involvement of eloquent cortex	1.28 (0.88–1.87)	0.198
Multifocality	1.72 (0.97–3.05)	0.064
Subependymal extension	1.75 (1.21–2.53)	0.003
Low proportion of edema	Reference	Reference
Medium proportion of edema	1.09 (0.75–1.61)	0.653
High proportion of edema	0.45 (0.24–0.83)	0.011
Increased T1FLAIR-ratio	0.59 (0.37–0.94)	0.026
Model 2: Radiomics features model ($n = 95$)		
T1_wavelet.HHH_firstorder_Median	1.04 (0.81–1.3)	0.754
T2_log.sigma.2.0.mm.3D_glszm_LargeAreaLowGrayLevelEmphasis	1.00 (0.83–1.2)	0.958
T2_log.sigma.3.0.mm.3D_glszm_LargeAreaLowGrayLevelEmphasis	1.33 (1.07–1.6)	0.009
T2_wavelet.LLH_firstorder_Mean	1.70 (1.32–2.2)	0.001
T2_wavelet.HHL_glszm_LargeAreaLowGrayLevelEmphasis	0.92 (0.75–1.1)	0.404
Model 3: Clinical features model ($n = 95$)		
Sex (male vs. female)	1.12 (0.70–1.77)	0.644
Type of surgery (resection vs. biopsy)	0.48 (0.31–0.76))	0.002
Age at diagnosis (>70 vs. <70)	1.10 (0.60–2.02)	0.749
Adjuvant treatment (non-STUPP vs. STUPP)	4.92 (2.79–8.67)	0.001
Methylguanine methyltransferase (*MGMT*)-methylation	0.61 (0.35–1.06)	0.082
Model 4: Integrated imaging model (VASARI + radiomics) ($n = 95$)		
VASARI prognostic score	2.2 (1.4–3.4)	<0.001
Radiomics prognostic score	2.92 (1.9–4.5)	<0.001
Model 5: Integrated VASARI and clinical model ($n = 95$)		
VASARI prognostic score	2.0 (1.3–3.2)	0.003
Clinical prognostic score	2.7 (1.8–3.9)	<0.001
Model 6: Integrated Radiomics and clinical model ($n = 95$)		
Radiomics prognostic score	2.6 (1.7–4.0)	<0.001
Clinical prognostic score	2.8 (1.9–4.1)	<0.001
Model 7: Integrated imaging and clinical model ($n = 95$)		
VASARI prognostic score	2.1 (1.4–3.3)	<0.001
Radiomics prognostic score	3.0 (1.9–4.7)	<0.001
Clinical prognostic score	2.1 (1.4–3.3)	<0.001

Clinical features that were selected in the clinical model were chosen based on previous studies (31) and clinical expertise (Model 3). Next, VASARI features, radiomics features and clinical features multivariable Cox-regression models were combined in different combinations. Model 4 was developed by combining VASARI prognostic index (PI) and Radiomics PI, Model 5 by combining VASARI PI and Clinical PI and Model 6 by combining Radiomics PI and Clinical PI (Model 4–6). Finally, clinical features were combined with the integrated VASARI and radiomics prognostic score to develop an integrated clinical and imaging prognostic model (Model 7) (Table 2). The calibration slope of the PI of Model 7 on the validation set was 0.79 (log-rank test p-value 0.27), indicating there is no certainty for the slope in the validation set being different from 1. The joint test of all predictors with the offsetting of the predicted PI results in the p-value of 0.23, indicating that there is no evidence of a lack of fit on the validation.

To assess the reproducibility performance of the prognostic models, all models were tested on the external validation set (n = 38) and the discriminative prognostic value in both cohorts was analyzed using Harrell's C-index (Figure 1A). Model 1 achieved a C-index of 0.61 (95% CI 0.55–0.68) when tested on the whole training cohort (n = 129). In order to make a comparison between the different models, the C-index for the VASARI-only model was also calculated using only the patients available in all other models (n = 95). In order to visualize the prognostic potential of the integrated imaging and clinical model (Model 7), the data-set was split in a low- and high-risk group at a set cut-off value (75th percentile) of the prognostic index in the training cohort. This same cut-off value was applied to the external validation cohort. Two survival groups could be identified (p-value < 0.0001) in both the training and validation cohort (Figure 1B,C).



A

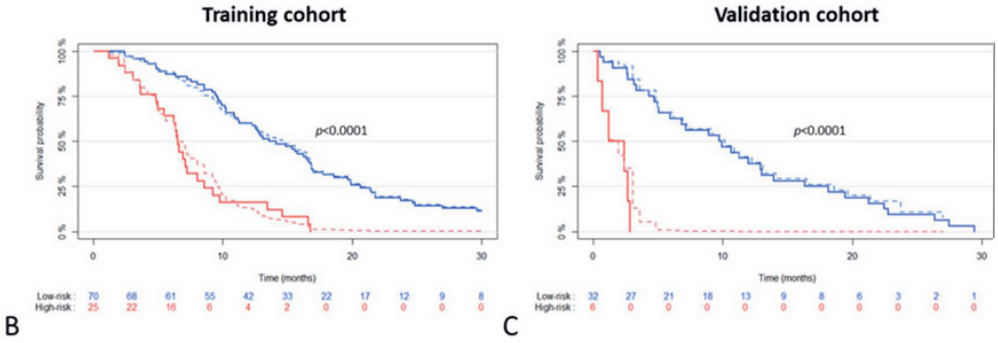
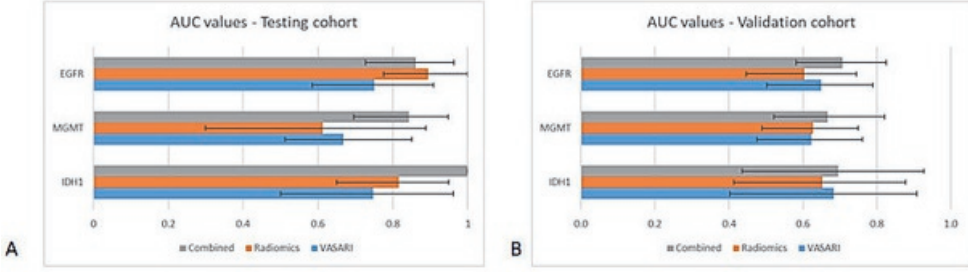
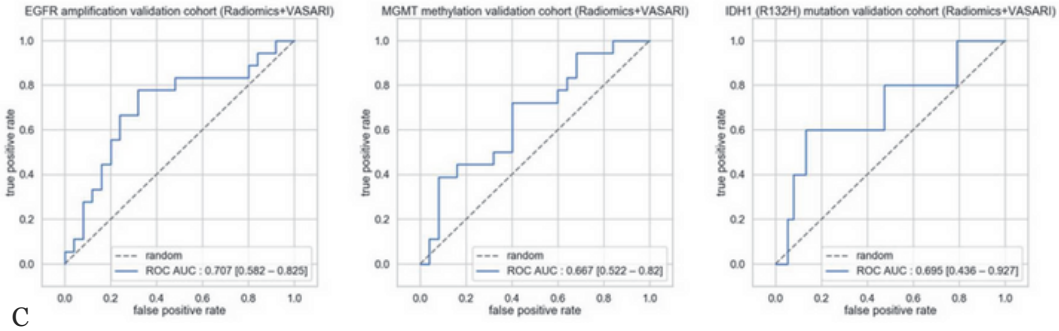


Figure 1. Performance of prognostic models: (A) visualization of C-index for all prognostic models (including 95% CI) in the training (n = 95) and validation cohort (n = 38); (B) Kaplan–Meier curve of integrated radiomics, Visually Accessible Rembrandt Images (VASARI) and clinical model (Model 7) in the training cohort and (C) validation cohort. Low- and high-risk groups (blue and red lines, respectively) cut-off values were determined by set cut-off (75th percentile) in the training cohort. The solid lines represent the observed survival curves, the dashed the corresponding predicted survival curves.

2.3. Predictive Value of Integrative Imaging Analysis

In order to develop the predictive models for molecular markers (EGFR amplification, MGMT-methylation and IDH1 mutation), the Maastricht University Medical Center+ (MUMC+) cohort was split into a training (70%) and test (30%) cohort. For the prediction of EGFR amplification, in total eleven VASARI features and four radiomics features were selected in the predictive models using the XGBoost machine learning algorithm (Table 3). Both VASARI and radiomics models alone were able to significantly predict EGFR amplification in the test dataset (Figure 2A). In the external validation set, both VASARI and radiomics features reached similar results to each other; however, an increased predictive value was observed when both models were combined (area-under-the-curve (AUC) 0.707 (95% CI 0.582–0.825); Figure 2B,C).





C

Figure 2. Performance of predictive models: (A) Area-under-the-curve (AUC) values and corresponding 95% confidence intervals of different predictive models in the testing cohort and (B) in the validation cohort; (C) receiver operating characteristic (ROC)-curves of combined VASARI and radiomics model predictive performance in external validation set.

Table 3. Selected VASARI and radiomics features in predictive models for epidermal growth factor receptor (EGFR) amplification, methylguanine methyltransferase(MGMT)-methylation and isocitrate dehydrogenase 1(IDH1) mutation in GBM patients in the training cohort.

VASARI Features	**Radiomics Features**
EGFR amplification (n = 64)	
Size: Major Axis, Minor Axis, Mean Minor Axis, Median Major Axis Location: Eloquent location, Midline cross of enhancing tumor Morphology: Proportion necrosis Tumor characteristics: Hemorrhage, Subependymal Extension, Pial invasion, Definition enhancing margin	T1+Gado: wavelet-HLH_glcm_Correlation T2: log-sigma-2-0-mm-3D_gldm_LargeDependenceLowGrayLevelEmphasis; wavelet-LLH_glcm_ClusterShade; wavelet-LLH_firstorder_Skewness
MGMT-methylation (n = 74)	
Size: Major Axis, Minor Axis, Median Major Axis, Mean Major Axis Morphology: Proportion non-enhancing tumor Tumor characteristics: Deep white matter invasion, Subependymal extension	T1+Gado: no features selected T2: wavelet-HLL_gldm_LargeDependenceHighGrayLevelEmphasis; log-sigma-5-0-mm-3D_glrlm_HighGrayLevelRunEmphasis; log-sigma-5-0-mm-3D_glszm_SmallAreaHighGrayLevelEmphasis
IDH1 mutation (n = 72)	
Size: Minor Axis, Major Axis Location: Tumor side, Eloquent location Morphology: Proportion non-enhancing tumor, Proportion Edema Tumor characteristics: Pial invasion, Thickness Enhancing Margin, Definition enhancing margin, T1-FLAIR-ratio	T1+Gado: wavelet-HLL_glcm_Contrast; wavelet-HLL_glcm_DifferenceAverage T2: log-sigma-2-0-mm-3D_firstorder_90Percentile; Original_glrlm_LongRunHighGrayLevelEmphasis; log-sigma-3-0-mm-3D_firstorder_90Percentile; original_firstorder_10Percentile; log-sigma-4-0-mm-3D_firstorder_Uniformity; wavelet-HLL_gldm_DependenceEntropy; wavelet-HLL_glcm_Correlation

The predictive models developed for MGMT-methylation status consisted of seven VASARI features (logistic regression analysis) and three radiomics features (XGBoost algorithm) (Table 3). VASARI features alone reached similar predictive values in the test and validation dataset with an AUC of 0.668 (95% CI 0.513–0.850) and 0.622 (95% CI 0.475–0.761) respectively. Radiomics features alone could not predict MGMT-methylation in both datasets. An

increased predictive value was observed when VASARI features and radiomics features were combined in one predictive model, with an AUC of 0.843 (95% CI 0.696–0.948) in the test dataset but did not perform as well in the external validation dataset (AUC 0.667 (95% CI 0.522–0.820); Figure 2B,C).

For the prediction of the IDH1 mutation ten VASARI features were included in the multivariate VASARI model and nine radiomics features in the radiomics prediction model developed using the XGBoost machine learning algorithm (Table 3). In the test dataset, only radiomics features reached statistical significance with an ROC AUC of 0.816 (95% CI 0.650–0.950), which improved upon combining with VASARI features (Figure 2A). In the external validation set, neither VASARI nor radiomics features or the combination were able to predict the IDH1 status (Figure 2B,C). ROC curves for all predictive models in the training and validation cohort are reported in Figures S5 and S6, respectively. Next, histogram heterogeneity was assessed to identify whether radiomics features demonstrate significant differences between the outcome groups in a univariate manner. Only for IDH1 mutation was a significant difference found for two features that could explain the heterogeneity in the outcome. The histograms of heterogeneity for each predictive model and significance values for IDH1-mutation are reported in Figures S7 and S8, respectively.

2.4. Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) Statement and Radiomics Quality Score

The TRIPOD statement adherences were calculated at 77% for this study. The radiomics quality score (RQS) score calculated for this study was 47%. An overview of point allocation towards the TRIPOD statement and RQS score can be found in Tables S3 and S4, respectively.

3. Discussion

Increasing curation rates by optimizing treatment strategies is being hampered by the highly invasive nature and GBM specific inter- and intratumoral molecular heterogeneity. MR imaging is currently the preferred diagnostic imaging technique for GBM. However, integrated standardized qualitative and quantitative analysis of different MR sequences has not yet been introduced into prognostic and predictive GBM models. This study retrospectively analyzed two multi-center GBM patient cohorts to develop integrated clinical and imaging prognostic models and predictive models for clinically relevant molecular markers.

Combining clinical features with quantitative and qualitative imaging features resulted in the most optimal prognostic model which could be reproduced in the external validation cohort (C-index 0.72 in training cohort and 0.73 in validation cohort). Despite promising results for predicting EGFR amplification and IDH1-mutation in the test cohort, none of the predictive models for molecular markers were able to predict these markers in a clinically relevant manner in the external validation set.

The prognostic model described in this study is developed for IDH-WT GBM patients as this patient group makes up the majority of GBM and exhibits large variation in prognosis and treatment response. This variance is also reflected in statistically significant differences in baseline characteristics for OS and MGMT-methylation. However, these differences are also known to exist between centers, in which different treatment decisions and strategies are being implemented. The aim of this study was to investigate the performance of prognostic models in such heterogeneous GBM cohorts. To predict OS, five VASARI features were identified to be of most prognostic relevance. Three of these features are well known prognostic factors and were also previously identified to be negatively associated with OS (involvement of eloquent cortex, multifocality and subependymal extension) and can be attributed to a more invasive growth of the tumor (32,33). The other selected features, proportion of edema and T1-FLAIR-ratio showed opposite prognostic value in this study when compared to previous studies (33,34,35,36,37). However, other studies reported no prognostic value for these features and therefore this still remains controversial(32,38).

Radiomics features that were identified to have prognostic value were mainly derived from T2-weighted imaging. This is in line with the hypothesis that the T2-weighted signal corresponds with intratumor heterogeneity and infiltrative tumor growth (39) and this area is accountable for the majority of local recurrences (40). Therefore, radiomics features from this area are expected to be of importance for survival prediction as was also shown in previous studies (41,42). The radiomics signature for OS consists of five features, from which two features are the first order Mean (T2-weighted) and Median (T1-weighted) describing the mean and median intensity values after the LLH and HHH wavelet decomposition of the original MR images. The remaining three features quantify gray level zones in an T2-weighted image, more precisely measuring the proportion in the image of the joint distribution of larger size zones with lower gray-level values after image transformation (Laplacian of Gaussian) which is

useful for edge detection. These gray level zone features can potentially be associated with the measure of intratumor heterogeneity (43).
In this study, VASARI features alone or radiomics features alone were not able to predict OS in the external validation dataset in a clinically relevant manner. Interestingly, the performance of the prognostic model improved upon combining VASARI, radiomics and clinical features (C-index 0.723 in training cohort and 0.730 in validation cohort) and became clinically relevant. The robustness of this combined model also improved as the model performed similarly in the training- and validation cohort and the uncertainty decreased as represented by a smaller confidence interval of the C-index. Model 5 and 6 report similar performances when compared to the model combining all features. However, the final combined model seems to remain mostly stable between both cohorts, though the actual additive value should be further validated in larger patient cohorts.
The combined model was also able to accurately split the two cohorts in a high- and low-risk group (p-value < 0.001) (Figure 1B,C). Previous studies also observed that combining clinical features with imaging features improves the prognostic value of the model (42,44,45,46,47). The model developed in this study performed similar or better compared to previous findings, even after external validation in a heterogeneous patient cohort. This highlights the clinical relevant potential of combining these features into a multimodal prognostic model which can potentially be applied in clinical practice.
As a proof-of-concept study, this study investigated the capability of VASARI and radiomics features to link phenotype to genotype and predict clinically relevant molecular markers, IDH1-mutation, MGMT-methylation and EGFR amplification, by machine learning approaches. Overall, the predictive models had promising performance on the test set, especially when VASARI and radiomics features were combined (Figure 2A). Unfortunately, none of the developed models were able to predict in the external validation set in a clinically relevant manner with a wide spread in confidence intervals of the AUC values (Figure 2B,C). In order for a model predicting molecular markers to be clinically relevant, much higher AUC values are desired. Since the presence of the molecular markers has biological consequences on tumor growth and development, specific imaging techniques that reflect biological processes have shown more promising results in the prediction of these markers and should therefore be used for further research. Perfusion-weighted and/or diffusion-weighted MRI features have been used to predict EGFR amplification (48,49,50) and MGMT-methylation (51),

whereas MR spectroscopy (52) and amino acid tracer PET imaging (FET–PET) (53) can predict IDH1 mutation status due to its effects on tumor metabolism.

In addition, by analyzing the heterogeneity histogram for EGFR amplification based on the validation cohort, we can notice that none of the radiomics features has demonstrated significant difference between the outcome groups in the univariate manner. Heterogeneity histogram for MGMT-methylation also did not demonstrate the significant difference between the outcome groups. For IDH1 mutation, however, we can point out a significant difference ($p < 0.05$) for T2_original_firstorder_10Percentile, T1_wavelet_HLL_glcm_DifferenceAverage features, which indicates the ability of these features to reflect the heterogeneity in the outcome (Figure S8). These findings also highlight the value of multivariate predictive analysis.

The overall RQS of 47% achieved in this study is higher than generally reported in neuro-oncology radiomics studies (54).

The main strength of this study includes the usage of two independent multicenter datasets. Though the performance of previous prognostic models based on VASARI or radiomics features is generally better, most of these studies only use internal validation methods and lack validation in an independent external dataset (34,55). The same applies to the performance of predictive models for molecular markers. However, the fact that the promising results for the predictive models in this study in the testing cohort could not be replicated in the external validation cohort stresses the importance of external validation.

Additionally, most studies use a more homogeneous patient cohort, for example, with regards to treatment characteristics, whereas this present study comprises two heterogeneous cohorts which more reflects daily clinical practice. For example, corticosteroid usage is known to decrease the amount of edema, therefore altering the T2-weighted signal, which can influence both VASARI and radiomics features. Previous studies either do not mention corticosteroid usage or exclude patients using corticosteroids (36,37) even though a significant amount of GBM patients are known to use corticosteroids. Furthermore, multiple studies only use single-institute data in which real-life heterogeneity between MRI acquisition is not represented (56) which is important for the generalizability of radiomics models.

Several limitations should be taken into account when considering the results of this study. The main limitation of this study is the number of patients that were included. Though for the OS models the number of patients is in accordance with the majority of

previous studies, especially the limited available molecular data in the external validation set limits the validation capacity of the predictive models. Especially IDH1/2 mutations rarely occur in both cohorts, which is to be expected in GBM, leading to wide confidence intervals and complications in the validation of the model. Future studies using a larger IDH-mutated cohort are needed to accurately test the models developed in this study. Next, the fact that this study is a retrospective study poses a potential selection bias. Additionally, the Karnofsky Performance Score (KPS) is an established prognostic feature which could not be included in this study due to lack of reporting of the KPS in patient files during the time period used for this study. Furthermore, it could be stated that a limitation of this study was the lack of advanced MRI sequences such as diffusion- and perfusion-weighted imaging and PET-MRI. However, this study specifically chose to focus on the relevance of conventional MRI images as these are widely available in clinical centers. Furthermore, MRI radiomics features are known to be dependent on differences in MRI scanners and scanning protocols. The images used in this study were collected from more than ten different hospitals over a ten-year time-period resulting in large differences in technical MRI characteristics. Again, even though this limits the performance of radiomics, an ideal prognostic and predictive model should not be dependent on homogeneous data. These differences in MRI acquisition methods are present in the real-life multicenter setting and should be accounted for in order to provide a relevant, clinical applicable model.

In order to further improve the prognostic and predictive potential of non-invasive imaging models, several steps need to be taken. First of all, larger (big data) datasets and preferably prospective studies are warranted to develop more accurate and generalizable models. This could pose a challenge, especially in less common types of cancer such as GBM. Next, the first studies on radiomics have been conducted on computed tomography (CT) imaging, which can be quantified using standardized Hounsfield units. For MRI radiomics, such a unit does not exist which poses problems due to inter- and intra-scanner variability. Multiple pre-processing methods have been developed, though not all radiomics features were shown to be robust between different pre-processing approaches (57,58,59). This calls for a generalized pre-processing pipeline and focus on features that are shown to be robust. Robust features and normalization methods can be achieved by applying phantom studies to account for differences between MRI acquisition protocols (60). Tumor delineation poses another important aspect of radiomics feature extraction. Manual delineation is still generally seen as the golden standard, though a substantial

inter-observer variability exists, despite international guidelines on tumor delineation (61) and it is a time consuming process. It has been shown that this inter-observer variation influences the radiomics analysis in multiple tumors (62). Automatic segmentation methods using a deep learning neural network approach are widely developed and can be beneficial in future radiomics studies and its clinical applicability by decreasing workload on clinicians and inter-observer variability (63,64). This is expected to lead to more robust radiomics features due to standardization of the delineation method.
Parallel to the establishment of MR signatures that are able to predict clinically significant expression of specific biomarkers, there is a need for imaging signatures that capture the level of intratumoral heterogeneity. However, it needs to be emphasized that is not yet clarified how to quantify GBM MR imaging heterogeneity and moreover how to non-invasively analyze the level of intratumoral heterogeneous expression of predictive markers, since the golden standard, single cell RNA sequencing, is missing in standard of care. By extracting radiomics features from the whole tumor and the surrounding area of edema we identified several features that are associated with intratumor heterogeneity. However, different steps could be taken to include more aspects of tumor heterogeneity. Improved performance of radiomics has been reported when features are extracted from distinct tumor areas (active tumor, necrosis and edema) separately (65,66), though this is a more labor-intensive approach which might limit its clinical applicability. In this aspect, automatic segmentation algorithms have shown to be useful for prognostic radiomics modelling (47).
Additionally, more biologically relevant MRI sequences such as diffusion- or perfusion-weighted MRI have been shown to outperform radiomics models based on conventional MRI (25). These approaches should be taken into account in future studies as they will be able to encompass more features concerning intratumor heterogeneity (67) and have shown improved performance with regards to predicting prognosis and molecular markers. Ultimately, studies correlation pathological and genetic examination of multiregional biopsies towards imaging features are needed to study the value of imaging features for tumor heterogeneity.

4. Materials and Methods

4.1. Patient Population

All patients treated by the neuro-oncology team of the Maastricht University Medical Centre (Maastricht UMC+, Maastricht, the Netherlands) between January 2004 and August 2014 for a glioblastoma (WHO grade IV) were considered for inclusion in the retrospective training cohort. Patients were excluded if no diagnostic, pre-operative MRI-images were available (minimum T1+Gadolinium and T2-weighed imaging), if survival data were unknown or no histological diagnosis was available. All patient records were reviewed considering patient and tumor characteristics, received treatments and survival data. The external validation cohort was constructed using the same criteria on an independent dataset of patients treated in Radboud University Medical Center (Radboudumc, Nijmegen, The Netherlands) in the same time period. Both Maastricht UMC+ and Radboudumc are academic reference centers for GBM patients in the Netherlands, implying MRI-images were also obtained in hospitals that refer their patients to these academic centers. Numbers of patients used for each analysis are reported in Table S1. The requirement for informed consent for this retrospective study was waived by the medical ethics committee of the MUMC+ (METC 16-4-022).

4.2. Image Acquisition and Qualitative Imaging Feature Assessment

Pre-operative MRI images were collected, pseudonymized and pooled in a database combining MRI images from different types and manufacturers of scanners using different imaging protocols to reflect the real-life inter-center heterogeneity (Figures S1 and S2). A quantitative and qualitative imaging analysis pipeline was set-up (Figure 3). All diagnostic MRI-scans were analyzed by dedicated neuro-radiologists (SP, AJ, AP), blinded for outcome and scored using the VASARI Imaging Features. A previous study conducted by the VASARI research project group showed a strong overall inter-observer agreement among six readers for the VASARI features (29). When needed, multi-categorical and continuous VASARI features were recoded into different groups based on their clinical relevance prior to the start of analysis (Table S5).

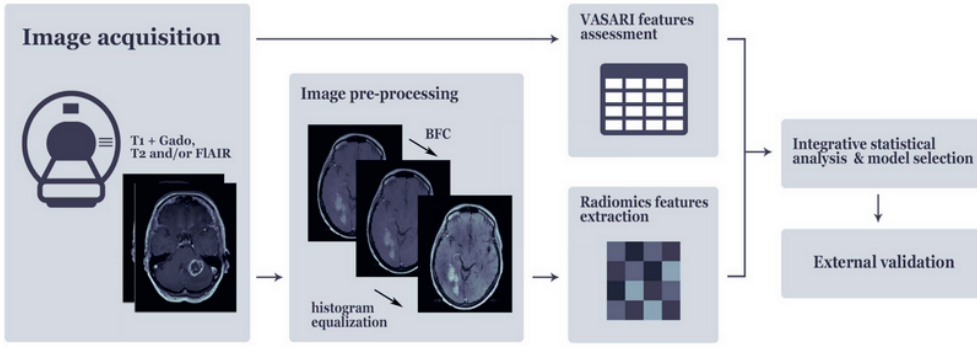


Figure 3. Quantitative and qualitative imaging analysis pipeline.

4.3. Tumor Delineation, Image Pre-Processing and Extraction of Radiomics Features

Using Osirix Lite (Pixmeo SARL, Bernex, Switzerland) and MiM software (version 7.0.4, MIM Software Inc., Cleveland, OH, USA), regions of interests (enhancing tumor on T1+Gadolinium images and combined tumor/edema portion on T2-weighted images) were manually delineated on all diagnostic MRI-images of the training and validation cohort, supervised by two experienced neuro-radiation oncologists (DE, IC).

Using Python 3.7 and the dedicated packages (cv2 version 4.1.0, https://pypi.org/project/opencv-python/, (accessed on 23 December 2020)), SimpleITK version 1.2.0 (https://simpleitk.org/, (accessed on 23 December 2020)) and scikit-image version 0.14.2, (https://scikit-image.org/, (accessed on 23 December 2020)), an image pre-processing routine was developed to handle the broad variability of image acquisition and reconstruction parameters.

At first, spatial resolution of the images was normalized with respect to the image sequence (final pixels are: 0.449 mm2 and slice thickness of: 5.5 mm). The mode of the pixel spacing and slice thickness distributions from the Maastricht UMC+ cohort were used as reference values for the resampling procedure to minimize the number of resampled images. A bicubic interpolation over 4 × 4 pixel neighborhood was used for both upsampling and downsampling. In order to correct the low frequency intensity non-uniformity, which is intrinsic for MRI images, the N4 bias field correction algorithm was used (68).

Furthermore, the histogram equalization method implemented in the scikit-image 0.15.0 package (69) was used to enhance the contrast of MRI images (70). As the last step of the pre- processing routine, image intensities were normalized using Z-score standardization method (71). A pre-processing routine was applied to both cohorts,

where parameters (mu, sigma) for the Z-score transformation were evaluated on the training cohort and transferred to the validation cohort. Parameters used are T1 mu = 0.1904, T1 sigma = 0.2313, T2 mu = 0.2009 and T2 sigma = 0.2448.
In order to obtain the quantitative imaging features, an open-source Pyradiomics 2.2.0 python package for the radiomics features extraction was utilized (72). Using the dedicated MRI settings, features from following feature classes were extracted: First Order Statistics, Shape-based (2D and 3D), Gray Level Cooccurence Matrix (GLCM), Gray Level Run Length Matrix (GLRLM), Gray Level Size Zone Matrix (GLSZM), Gray Level Dependence Matrix (GLDM), Neighboring Gray Tone Difference Matrix (NGTDM). Along with the original features Laplacian of Gaussian (LoG) (sigma: (2.0,3.0,4.0,5.0) and Wavelet filters were activated resulting in a total of 1197 features per patient. A detailed mathematical feature description as provided by Aerts et al. 2014 (30).

4.4. Molecular Markers

Archival formalin-fixed paraffin-embedded (FFPE) tissue samples were analyzed for tumor percentage by an experienced neuropathologist (JB). DNA was extracted from FFPE tissue using the Cobas method (Roche, Bazel, Switzerland) and DNA concentration was quantified using Qubit Fluorometer (Life Technologies, Waltham, MA, USA). Next-generation sequencing (NGS) was performed using the Ion AmpliSeq Cancer Hotspot Panel v2 (Life Technologies) as previously described (73). For the purpose of this study, the data were analyzed for the presence of an IDH1 (R132H) mutation (minimum coverage 500×) which was manually checked using Integrative Genomics Viewer (IGV).
EGFR amplification was assessed using SNPitty, an open-source web application for interactive B-allele frequency and copy number visualization of NGS data, by comparing the number of reads in the EGFR locus to the surrounding regions (74). MGMT methylation status was assessed using methylation-specific multiplex ligation-dependent probe amplification (MS-MLPA) as previously described (75). In case NGS data was not available for a sample, MLPA was also used to assess IDH1 mutation status and EGFR amplification.

4.5. Statistical Analysis

Statistical analysis for differences between baseline characteristics was performed using double-sided T-test for “age at diagnosis”. Fisher’s exact test was used for all other binary variables (sex, type of surgery, adjuvant treatment and molecular markers).
Overall survival (OS) was defined as the time between the initial surgical intervention after diagnosis and the date of death (confirmed by the Municipal Personal Records Database). Patients that survived were censored at the moment of the last follow-up measurement. To develop a prognostic model, analysis was focused on the IDH-WT GBM samples.
OS analysis was performed using R (version 4.0.2., R Studio, Boston, MA, USA), employing the packages stats, survival, survminer, rms, pec and survcomp. VASARI features were tested in univariate Cox-regression analysis to determine the hazard ratio (HR) of each feature individually on the training cohort. Each feature with a p-value of ≤ 0.2 was considered for inclusion in the multivariable analysis. Resulting VASARI features were used for multivariable Cox-regression analysis with fast backward elimination (removal alpha < 0.2) on the training set. Radiomics features from T1- and T2-weighted images were combined and normalized with the Z-score transformation, where coefficients evaluated on the train set were transferred to the validation set. Highly correlated features exceeding the Spearman’s rank correlation of rs = 0.85 were eliminated.
Resulting radiomics features, were used for multivariable Cox-regression analysis with fast backward elimination for the training set (76) (Model 1–3) All clinical features were entered into the Cox-regression model to develop the Clinical model on the training set. A prognostic Index (PI) for all models developed on the training set was calculated for training and validation datasets, where the PI was defined as $\sum i \beta i x i$ for each individual model. For the combined models, the PI of the individual models was used as a feature along with the PI for the individual model it was combined with in Cox-regression analysis (77). Similarly, for a combined clinical/VASARI/radiomics model (Model 7), VASARI PI was used as a feature along the radiomics PI and clinical PI. Next, the models were validated using multiple-step approach (78). Calibration slope was assessed using the Log-rank (LR) test. Model’s misspecification was evaluated by performing the Cox regression on the individual features of the signature in the validation dataset with offsetting the validation PI (78).
Overall model performance for discriminating survival groups was evaluated with Harrell’s C-index. To display the potential

discrimination between survival groups Kaplan–Meier (KM) curves were used with the threshold value based on 75th percentile of training PI's in order to identify a high-risk group using our model. Significance of the split was estimated using the LR test. In addition, predicted survival curves for each risk group were plotted. The PI is used to estimate the survival curve, which is then averaged over the entire risk-group. These curves are plotted alongside the observed KM-curves. The correlation between radiomics features and tumor volume was assessed using Spearman's rank correlation. This was investigated since previous studies have shown some radiomics features to be surrogate markers for tumor volume and not independent prognostic features (79). Correlation between VASARI features, radiomics features and clinical features were assessed using the point-biserial correlation coefficient.

Python 3.7 was used to develop and validate the predictive models. Patients with unknown outcomes (molecular markers) were excluded from the analysis. At first, highly correlated features (rs > 0.85) were eliminated, in which the feature with the lower AUC value in univariate ROC analysis was removed and resulting features were normalized using Z score on the MUMC+ cohort. Shift/scale parameters of individual features are available upon request. As the second step, the MUMC+ cohort was split randomly into train and test sets with a 70/30 ratio and label stratification. In the third step, to obtain the feature importance scores, a random forest model with the random-sampled initialization of hyper parameters (each iteration parameter was randomly sampled from the hyper parameter ranges: number of estimators (20,300), max depth (2,6)) was fitted 1000 times resulting in the cumulative feature importance histogram. Based on the feature importance rank, the 20 most important features were selected for the further evaluation. In order to find the best performing model in the fourth step, Xgboost, Random Forest and Logistic regression algorithms were initialized with the random-sampling of hyper parameters (Table S6), trained and tested 1000 times. In order to overcome a "lucky split bias", step 2 (the random splitting of the cohorts) followed by model testing was repeated 10 times for the top 5 performing models from step 4, representing the cross validation technique.

Combined model was achieved by ensembling VASARI and radiomics models using averaging of VASARI and radiomics predicted probabilities. To evaluate performance of the predictive models, the area under the receiver operating characteristic (ROC) curve, or AUC, was calculated. Bootstrapping technique with 100 iterations was utilized to estimate ROC AUC 95% confidence intervals on test and validation datasets.

Additionally, to visualize the ability of radiomics features of capturing the outcome heterogeneity in a univariate manner and contribute to concept of explainable radiomics, we visualized the outcome heterogeneity through selected radiomics features by plotting the distribution of feature values for each particular feature of each binary outcome. The significance of the difference in the mean values was evaluated by performing the Mann–Whitney test with Bonferroni correction.

4.6. TRIPOD Statement and Radiomics Quality Score (RQS)

To assess the quality of the conducted study, a radiomics quality score (RQS) was calculated. The RQS is a checklist consisting of 16 components to assess the validity of the radiomics workflow and (external) validation of the models (19,80). Furthermore, the checklist recommended in transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) was assessed(81).

5. Conclusions

In the present study, the potential of non-invasive quantitative and qualitative imaging features to predict prognosis and clinically relevant molecular markers was investigated in a real-life heterogeneous GBM patient cohort. The integrated prognostic model, including clinical and imaging features, showed the most promising performance which was reproducible and most robust between both datasets. However, further improvements and larger prospective studies are needed before this model can be used in daily clinical practice. Using imaging features to predict molecular markers showed promising results in the testing set but could not be validated on the external validation set and warrants additional validation in larger GBM cohorts.

Supplementary Materials
The supplementary materials are available online at https://www.mdpi.com/2072-6694/13/4/722/s1 , or via QR code.

Author Contributions

Conceptualization, M.V., S.P., I.C., H.C.W., M.M.A., M.P.G.B., L.A., O.E.M.G.S., O.T., K.H., D.B.P.E., A.A.P., P.L., A.H.; methodology, M.V., S.P., H.C.W., S.M.J.v.K., B.L.T.R., M.A.V., V.C.G.T., P.L., A.H.; software, S.P., H.C.W., P.L.; validation, M.V., S.P., H.C.W., P.L., A.H.; formal analysis, M.V., S.P., H.C.W., S.M.J.v.K., B.L.T.R., P.L., A.H.; investigation, M.V., S.P., M.t.D., E.G.M.R., S.A.H.P., F.J.A.M., W.W.J.d.L., D.B.P.E., A.A.P., A.H.; resources, S.P., H.C.W., M.t.L., J.B., E.J.S., B.K., W.W.J.d.L., P.L., A.H.; data curation, M.V., I.C., M.t.D., E.G.M.R.; writing—orginal draft preparation, M.V., S.P., A.H.; writing—review and editing, M.V., S.P., I.C., H.C.W., S.M.J.v.K., B.L.T.R., M.t.D., E.G.M.R., M.t.L., S.A.H.P., F.J.A.M., J.B., E.J.S., B.K., W.W.J.d.L., M.M.A., M.P.G.B., L.A., O.E.M.G.S., O.T., K.H., M.A.V., V.C.G.T.-H., D.B.P.E., A.A.P., P.L., A.H.; visualization, M.V., S.P.; supervision, H.C.W., P.L., A.H.; project administration, M.V., A.H.; funding acquisition, P.L., A.H. All authors have read and agreed to the published version of the manuscript.

Funding

M.V./M.A.V.: KWF Kankerbestrijding (Dutch Cancer Society) Unique High Risk (16698/2018-1). A.H.: StopHersentumoren.nl. P.L: ERC advanced grant (ERC-ADG-2015 n° 694812 — Hypoximmuno), ERC-2018-PoC: 813200-CL-IO, ERC-2020-PoC: 957565-AUTO.DISTINCT. SME Phase 2 (RAIL n°673780), EUROSTARS (DART, DECIDE, COMPACT-12053), the European Union's Horizon 2020 research and innovation programme under grant agreement: BD2Decide-PHC30-689715, ImmunoSABR n° 733008,

MSCA-ITN-PREDICT n° 766276, FETOPEN-SCANnTREAT n° 899549, CHAIMELEON n° 952172, EuCanImage n° 952103, TRANSCAN Joint Transnational Call 2016 (JTC2016 CLEARLY n° UM 2017-8295), Interreg V-A Euregio Meuse-Rhine (EURADIOMICS n° EMR4), and Genmab (n° 1044); Dutch Cancer Society (KWF Kankerbestrijding), Project number 12085/2018–2

Institutional Review Board Statement

Ethical review and approval were waived for this study, due to the fact that, besides a few exceptions, most patients were deceased at the moment of this study and that all data collected are not traceable to individual patients. Informed Consent Statement
Patient consent was waived due to the fact that, besides a few exceptions, most patients were deceased at the moment of this study and that all data collected are not traceable to individual patients.

Data Availability Statement

The data presented in this study are available on request from the corresponding author. The data are not publicly available due to privacy restrictions.

Conflicts of Interest

Philippe Lambin reports, within and outside the submitted work, grants/sponsored research agreements from Varian medical, Oncoradiomics, ptTheragnostic, Health Innovation Ventures and DualTpharma. He received an advisor/presenter fee and/or reimbursement of travel costs/external grant writing fee and/or in-kind manpower contribution from Oncoradiomics, BHV, Merck and Convert pharmaceuticals. Lambin has shares in the company Oncoradiomics, Convert pharmaceuticals, MedC2 and LivingMed Biotech and is co-inventor of two issued patents with royalties on radiomics (PCT/NL2014/050248, PCT/NL2014/050728) licensed to Oncoradiomics and one issue patent on mtDNA (PCT/EP2014/059089) licensed to ptTheragnostic/DNAmito, three non-patentable invention (software) licensed to ptTheragnostic/ DNAmito, Oncoradiomics and Health Innovation Ventures. Henry C. Woodruff has (minority) shares in the company Oncoradiomics. Alinda Jacobi-Postma received institutional grants received from

Siemens Healthcare and Bayer Healthcare. The funders had no role in the design of the study; in the collection, analyses or interpretation of data, in the writing of the manuscript or in the decision to publish the results. All other authors declare no potential conflict of interests.

References

1. Urbanska, K.; Sokolowska, J.; Szmidt, M.; Sysa, P. Glioblastoma multiforme—An overview. Contemp. Oncol. (Pozn) 2014,18, 307–312.
2. Stupp, R.; Mason,W.P.; van den Bent, M.J.;Weller, M.; Fisher, B.; Taphoorn, M.J.; Belanger, K.; Brandes, A.A.; Marosi, C.; Bogdahn, U.; et al. Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma. N. Engl. J. Med. 2005, 352, 987–996.
3. Wen, P.Y.; Weller, M.; Lee, E.Q.; Alexander, B.M.; Barnholtz-Sloan, J.S.; Barthel, F.P.; Batchelor, T.T.; Bindra, R.S.; Chang, S.M.; Chiocca, E.A.; et al. Glioblastoma in adults: A Society for Neuro-Oncology (SNO) and European Society of Neuro-Oncology (EANO) consensus review on current management and future directions. Neuro Oncol. 2020, 22, 1073–1113.
4. Louis, D.N.; Perry, A.; Reifenberger, G.; von Deimling, A.; Figarella-Branger, D.; Cavenee, W.K.; Ohgaki, H.; Wiestler, O.D.; Kleihues, P.; Ellison, D.W. The 2016 World Health Organization Classification of Tumors of the Central Nervous System: A summary. Acta Neuropathol. 2016, 131, 803–820.
5. Weller, M.; van den Bent, M.; Preusser, M.; Le Rhun, E.; Tonn, J.C.; Minniti, G.; Bendszus, M.; Balana, C.; Chinot, O.; Dirven, L.; et al. EANO guidelines on the diagnosis and treatment of diffuse gliomas of adulthood. Nat. Rev. Clin. Oncol. 2020.
6. Hegi, M.E.; Diserens, A.C.; Gorlia, T.; Hamou, M.F.; de Tribolet, N.;Weller, M.; Kros, J.M.; Hainfellner, J.A.; Mason,W.; Mariani, L.; et al. MGMT gene silencing and benefit from temozolomide in glioblastoma. N. Engl. J. Med. 2005, 352, 997–1003.
7. Hegi, M.E.; Genbrugge, E.; Gorlia, T.; Stupp, R.; Gilbert, M.R.; Chinot, O.L.; Nabors, L.B.; Jones, G.; Van Criekinge, W.; Straub, J.; et al. MGMT Promoter Methylation Cutoff with Safety Margin for Selecting Glioblastoma Patients into Trials Omitting Temozolomide: A Pooled Analysis of Four Clinical Trials. Clin. Cancer Res. 2019, 25, 1809–1816.
8. Dahlrot, R.H.; Kristensen, B.W.; Hjelmborg, J.; Herrstedt, J.; Hansen, S. A population-based study of high-grade gliomas and mutated isocitrate dehydrogenase 1. Int. J. Clin. Exp. Pathol. 2013, 6, 31–40.

9. Lassman, A.B.; Aldape, K.D.; Ansell, P.J.; Bain, E.; Curran,W.J.; Eoli, M.; French, P.J.; Kinoshita, M.; Looman, J.; Mehta, M.; et al. Epidermal growth factor receptor (EGFR) amplification rates observed in screening patients for randomized trials in glioblastoma.
J. Neuro Oncol. 2019, 144, 205–210.
10. Eskilsson, E.; Rosland, G.V.; Solecki, G.; Wang, Q.; Harter, P.N.; Graziani, G.; Verhaak, R.G.W.; Winkler, F.; Bjerkvig, R.; Miletic, H. EGFR heterogeneity and implications for therapeutic intervention in glioblastoma. Neuro Oncol. 2018, 20, 743–752.
11. Saadeh, F.S.; Mahfouz, R.; Assi, H.I. EGFR as a clinical marker in glioblastomas and other gliomas. Int. J. Biol. Mark. 2018, 33, 22–32. 12. Armocida, D.; Pesce, A.; Frati, A.; Santoro, A.;
Salvati, M. EGFR amplification is a real independent prognostic impact factor between young adults and adults over 45yo with wild-type glioblastoma? J. Neuro Oncol. 2020, 146, 275–284.
13. Hoffman, D.I.; Abdullah, K.G.; McCoskey, M.; Binder, Z.A.; O'Rourke, D.M.; Desai, A.S.; Nasrallah, M.P.; Bigdeli, A.; Morrissette, J.J.D.; Brem, S.; et al. Negative prognostic impact of epidermal growth factor receptor copy number gain in young adults with isocitrate dehydrogenase wild-type glioblastoma. J. Neuro Oncol. 2019, 145, 321–328.
14. Patel, A.P.; Tirosh, I.; Trombetta, J.J.; Shalek, A.K.; Gillespie, S.M.;Wakimoto, H.; Cahill, D.P.; Nahed, B.V.; Curry, W.T.; Martuza, R.L.; et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. Science 2014, 344, 1396–1401.
15. Qazi, M.A.; Vora, P.; Venugopal, C.; Sidhu, S.S.; Moffat, J.; Swanton, C.; Singh, S.K. Intratumoral heterogeneity: Pathways to treatment resistance and relapse in human glioblastoma. Ann.
16. Draaisma, K.; Chatzipli, A.; Taphoorn, M.; Kerkhof, M.;Weyerbrock, A.; Sanson, M.; Hoeben, A.; Lukacova, S.; Lombardi, G.; Leenstra, S.; et al. Molecular Evolution of IDH Wild-Type Glioblastomas Treated With Standard of Care Affects Survival and Design of Precision Medicine Trials: A Report From the EORTC 1542 Study. J. Clin. Oncol. 2020, 38, 81–99.
17. Wangaryattawanich, P.; Hatami, M.;Wang, J.; Thomas, G.; Flanders, A.; Kirby, J.;Wintermark, M.; Huang, E.S.; Bakhtiari, A.S.; Luedi, M.M.; et al. Multicenter imaging outcomes study of The Cancer Genome Atlas glioblastoma patient cohort: Imaging predictors of overall and progression-free survival. Neuro Oncol. 2015, 17, 1525–1537.
Oncol. 2017, 28, 1448–1456.
18. Aerts, H.J. The Potential of Radiomic-Based Phenotyping in

Precision Medicine: A Review. JAMA Oncol. 2016, 2, 1636–1642. Cancers 2021, 13, 722 17 of 19
19. Lambin, P.; Leijenaar, R.T.H.; Deist, T.M.; Peerlings, J.; de Jong, E.E.C.; van Timmeren, J.; Sanduleanu, S.; Larue, R.; Even, A.J.G.; Jochems, A.; et al. Radiomics: The bridge between medical imaging and personalized medicine. Nat. Rev. Clin. Oncol. 2017, 14, 749–762.
20. Rogers, W.; Thulasi Seetha, S.; Refaee, T.A.G.; Lieverse, R.I.Y.; Granzier, R.W.Y.; Ibrahim, A.; Keek, S.A.; Sanduleanu, S.; Primakov, S.P.; Beuque, M.P.L.; et al. Radiomics: From qualitative to quantitative imaging. Br. J. Radiol. 2020, 93, 20190948.
21. Chaddad, A.; Kucharczyk, M.J.; Daniel, P.; Sabri, S.; Jean-Claude, B.J.; Niazi, T.; Abdulkarim, B. Radiomics in Glioblastoma: Current Status and Challenges Facing Clinical Implementation. Front. Oncol. 2019, 9, 374.
22. Peeken, J.C.; Molina-Romero, M.; Diehl, C.; Menze, B.H.; Straube, C.; Meyer, B.; Zimmer, C.; Wiestler, B.; Combs, S.E. Deep learning derived tumor infiltration maps for personalized target definition in Glioblastoma radiotherapy. Radiother. Oncol. 2019, 138, 166–172.
23. Grossmann, P.; Stringfield, O.; El-Hachem, N.; Bui, M.M.; Rios Velazquez, E.; Parmar, C.; Leijenaar, R.T.; Haibe-Kains, B.; Lambin, P.; Gillies, R.J.; et al. Defining the biological basis of radiomic phenotypes in lung cancer. Elife 2017, 6.
24. Su, C.; Jiang, J.; Zhang, S.; Shi, J.; Xu, K.; Shen, N.; Zhang, J.; Li, L.; Zhao, L.; Zhang, J.; et al. Radiomics based on multicontrast MRI can precisely differentiate among glioma subtypes and predict tumour-proliferative behaviour. Eur. Radiol. 2019, 29, 1986–1996.
25. Park, J.E.; Kim, H.S.; Jo, Y.; Yoo, R.E.; Choi, S.H.; Nam, S.J.; Kim, J.H. Radiomics prognostication model in glioblastoma using diffusion- and perfusion-weighted MRI. Sci. Rep. 2020, 10, 4250.
26. Macyszyn, L.; Akbari, H.; Pisapia, J.M.; Da, X.; Attiah, M.; Pigrish, V.; Bi, Y.; Pal, S.; Davuluri, R.V.; Roccograndi, L.; et al. Imaging patterns predict patient survival and molecular subtype in glioblastoma via machine learning techniques. Neuro Oncol. 2016, 18, 417–425.
27. Aquilanti, E.; Miller, J.; Santagata, S.; Cahill, D.P.; Brastianos, P.K. Updates in prognostic markers for gliomas. Neuro Oncol. 2018, 20, vii17–vii26.
28. European Society of, R. ESR Statement on the Validation of Imaging Biomarkers. Insights Imaging 2020, 11, 76.
29. Radiomics Features. Available online: https://pyradiomics.readthedocs.io/en/latest/features.html (accessed on 28 January 2021).
30. Aerts, H.J.; Velazquez, E.R.; Leijenaar, R.T.; Parmar, C.;

Grossmann, P.; Carvalho, S.; Bussink, J.; Monshouwer, R.; Haibe-Kains, B.; Rietveld, D.; et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat. Commun. 2014, 5, 4006.
31. Gittleman, H.; Lim, D.; Kattan, M.W.; Chakravarti, A.; Gilbert, M.R.; Lassman, A.B.; Lo, S.S.; Machtay, M.; Sloan, A.E.; Sulman, E.P.; et al. An independently validated nomogram for individualized estimation of survival among patients with newly diagnosed glioblastoma: NRG Oncology RTOG 0525 and 0825. Neuro Oncol. 2017, 19, 669–677.
32. Peeken, J.C.; Goldberg, T.; Pyka, T.; Bernhofer, M.; Wiestler, B.; Kessel, K.A.; Tafti, P.D.; Nusslin, F.; Braun, A.E.; Zimmer, C.; et al. Combining multimodal imaging and treatment features improves machine learning-based prognostic assessment in patients with glioblastoma multiforme. Cancer Med. 2019, 8, 128–136.
33. Mazurowski, M.A.; Desjardins, A.; Malof, J.M. Imaging descriptors improve the predictive power of survival models for glioblastoma patients. Neuro Oncol. 2013, 15, 1389–1394.
34. Nicolasjilwan, M.; Hu, Y.; Yan, C.; Meerzaman, D.; Holder, C.A.; Gutman, D.; Jain, R.; Colen, R.; Rubin, D.L.; Zinn, P.O.; et al. Addition of MR imaging features and genetic biomarkers strengthens glioblastoma survival prediction in TCGA patients. J. Neuro Radiol. 2015, 42, 212–221.
35. Pope, W.B.; Sayre, J.; Perlina, A.; Villablanca, J.P.; Mischel, P.S.; Cloughesy, T.F. MR imaging correlates of survival in patients with high-grade gliomas. Ajnr. Am. J. Neuro Radiol. 2005, 26, 2466–2474.
36. Schoenegger, K.; Oberndorfer, S.; Wuschitz, B.; Struhal, W.; Hainfellner, J.; Prayer, D.; Heinzl, H.; Lahrmann, H.; Marosi, C.; Grisold,W. Peritumoral edema on MRI at initial diagnosis: An independent prognostic factor for glioblastoma? Eur. J. Neurol. 2009, 16, 874–878.
37. Wu, C.X.; Lin, G.S.; Lin, Z.X.; Zhang, J.D.; Liu, S.Y.; Zhou, C.F. Peritumoral edema shown by MRI predicts poor clinical outcome in glioblastoma. World J. Surg. Oncol. 2015, 13, 97.
38. Henker, C.; Kriesen, T.; Glass, A.; Schneider, B.; Piek, J. Volumetric quantification of glioblastoma: Experiences with different measurement techniques and impact on survival. J. Neuro Oncol. 2017, 135, 391–402.
39. Lemee, J.M.; Clavreul, A.; Menei, P. Intratumoral heterogeneity in glioblastoma: Don't forget the peritumoral brain zone. Neuro Oncol. 2015, 17, 1322–1332.
40. Petrecca, K.; Guiot, M.C.; Panet-Raymond, V.; Souhami, L. Failure pattern following complete resection plus radiotherapy

and temozolomide is at the resection margin in patients with glioblastoma. J. Neuro Oncol. 2013, 111, 19–23.
41. Prasanna, P.; Patel, J.; Partovi, S.; Madabhushi, A.; Tiwari, P. Radiomic features from the peritumoral brain parenchyma on treatment-naive multi-parametric MR imaging predict long versus short-term survival in glioblastoma multiforme: Preliminary findings. Eur. Radiol. 2017, 27, 4188–4197.
42. Shi, J.; Yang, S.;Wang, J.; Huang, S.; Yao, Y.; Zhang, S.; Zhu,W.; Shao, J. Analysis of heterogeneity of peritumoral T2 hyperintensity in patients with pretreatment glioblastoma: Prognostic value of MRI-based radiomics. Eur. J. Radiol. 2019, 120, 108642.
43. Leijenaar, R.T.; Carvalho, S.; Hoebers, F.J.; Aerts, H.J.; van Elmpt, W.J.; Huang, S.H.; Chan, B.; Waldron, J.N.; O0Sullivan, B.; Lambin, P. External validation of a prognostic CT-based radiomic signature in oropharyngeal squamous cell carcinoma. Acta Oncol. 2015, 54, 1423–1429. Cancers 2021, 13, 722 18 of 19
44. Chen, X.; Fang, M.; Dong, D.; Liu, L.; Xu, X.; Wei, X.; Jiang, X.; Qin, L.; Liu, Z. Development and Validation of a MRI-Based Radiomics Prognostic Classifier in Patients with Primary Glioblastoma Multiforme. Acad. Radiol. 2019, 26, 1292–1300.
45. Kickingereder, P.; Neuberger, U.; Bonekamp, D.; Piechotta, P.L.; Gotz, M.; Wick, A.; Sill, M.; Kratz, A.; Shinohara, R.T.; Jones, D.T.W.; et al. Radiomic subtyping improves disease stratification beyond key molecular, clinical, and standard imaging characteristics in patients with glioblastoma. Neuro Oncol. 2018, 20, 848–857.
46. Kickingereder, P.; Burth, S.; Wick, A.; Gotz, M.; Eidel, O.; Schlemmer, H.P.; Maier-Hein, K.H.; Wick, W.; Bendszus, M.; Radbruch, A.; et al. Radiomic Profiling of Glioblastoma: Identifying an Imaging Predictor of Patient Survival with Improved Performance over Established Clinical and Radiologic Risk Models. Radiology 2016, 280, 880–889.
47. Choi, Y.; Nam, Y.; Jang, J.; Shin, N.Y.; Lee, Y.S.; Ahn, K.J.; Kim, B.S.; Park, J.S.; Jeon, S.S.; Hong, Y.G. Radiomics may increase the prognostic value for survival in glioblastoma patients when combined with conventional clinical and genetic prognostic models. Eur. Radiol. 2020.
48. Kickingereder, P.; Bonekamp, D.; Nowosielski, M.; Kratz, A.; Sill, M.; Burth, S.;Wick, A.; Eidel, O.; Schlemmer, H.P.; Radbruch, A.; et al. Radiogenomics of Glioblastoma: Machine Learning-based Classification of Molecular Characteristics by Using Multiparametric and Multiregional MR Imaging Features. Radiology 2016, 281, 907–918.
49. Gupta, A.; Young, R.J.; Shah, A.D.; Schweitzer, A.D.; Graber, J.J.; Shi, W.; Zhang, Z.; Huse, J.; Omuro, A.M. Pretreatment Dynamic Susceptibility Contrast

MRI Perfusion in Glioblastoma: Prediction of EGFR Gene Amplification. Clin. Neuro Radiol. 2015, 25, 143–150.
50. Hu, L.S.; Ning, S.; Eschbacher, J.M.; Baxter, L.C.; Gaw, N.; Ranjbar, S.; Plasencia, J.; Dueck, A.C.; Peng, S.; Smith, K.A.; et al. Radiogenomics to characterize regional genetic heterogeneity in glioblastoma. Neuro Oncol. 2017, 19, 128–137.
51. Han, Y.; Yan, L.F.; Wang, X.B.; Sun, Y.Z.; Zhang, X.; Liu, Z.C.; Nan, H.Y.; Hu, Y.C.; Yang, Y.; Zhang, J.; et al. Structural and advanced imaging in predicting MGMT promoter methylation of primary glioblastoma: A region of interest based analysis. BMC Cancer 2018, 18, 215.
52. Leather, T.; Jenkinson, M.D.; Das, K.; Poptani, H. Magnetic Resonance Spectroscopy for Detection of 2-Hydroxyglutarate as a Biomarker for IDH Mutation in Gliomas. Metabolites 2017, 7, 29.
53. Lohmann, P.; Lerche, C.; Bauer, E.K.; Steger, J.; Stoffels, G.; Blau, T.; Dunkl, V.; Kocher, M.; Viswanathan, S.; Filss, C.P.; et al. Predicting IDH genotype in gliomas using FET PET radiomics. Sci. Rep. 2018, 8, 13328.
54. Park, J.E.; Kim, H.S.; Kim, D.; Park, S.Y.; Kim, J.Y.; Cho, S.J.; Kim, J.H. A systematic review reporting quality of radiomics research in neuro-oncology: Toward clinical utility and quality improvement using high-dimensional imaging features. BMC Cancer 2020, 20, 29.
55. Peeken, J.C.; Hesse, J.; Haller, B.; Kessel, K.A.; Nusslin, F.; Combs, S.E. Semantic imaging features predict disease progression and survival in glioblastoma multiforme patients. Strahlenther Onkol. 2018, 194, 580–590.
56. Bae, S.; Choi, Y.S.; Ahn, S.S.; Chang, J.H.; Kang, S.G.; Kim, E.H.; Kim, S.H.; Lee, S.K. Radiomic MRI Phenotyping of Glioblastoma:
Improving Survival Prediction. Radiology 2018, 289, 797–806.
57. Moradmand, H.; Aghamiri, S.M.R.; Ghaderi, R. Impact of image preprocessing methods on reproducibility of radiomic features in multimodal magnetic resonance imaging in glioblastoma. J. Appl. Clin. Med. Phys. 2020, 21, 179–190.
58. Shiri, I.; Hajianfar, G.; Sohrabi, A.; Abdollahi, H.; Shayesteh, S.P.; Geramifar, P.; Zaidi, H.; Oveisi, M.; Rahmim, A. Repeatability of Radiomic Features in Magnetic Resonance Imaging of Glioblastoma: Test-Retest and Image Registration Analyses. Med. Phys. 2020.
59. Um, H.; Tixier, F.; Bermudez, D.; Deasy, J.O.; Young, R.J.; Veeraraghavan, H. Impact of image preprocessing on the scanner dependence of multi-parametric MRI radiomic features and covariate shift in multi-institutional glioblastoma datasets. Phys. Med. Biol. 2019, 64, 165011.

60. Rai, R.; Holloway, L.C.; Brink, C.; Field, M.; Christiansen, R.L.; Sun, Y.; Barton, M.B.; Liney, G.P. Multicenter evaluation of MRI-based radiomic features: A phantom study. Med. Phys. 2020.
61. Niyazi, M.; Brada, M.; Chalmers, A.J.; Combs, S.E.; Erridge, S.C.; Fiorentino, A.; Grosu, A.L.; Lagerwaard, F.J.; Minniti, G.; Mirimanoff, R.O.; et al. ESTRO-ACROP guideline "target delineation of glioblastomas". Radiother. Oncol. 2016, 118, 35–42.
62. Pavic, M.; Bogowicz, M.; Wurms, X.; Glatz, S.; Finazzi, T.; Riesterer, O.; Roesch, J.; Rudofsky, L.; Friess, M.; Veit-Haibach, P.; et al.
Influence of inter-observer delineation variability on radiomics stability in different tumor sites. Acta Oncol. 2018, 57, 1070–1074.
63. Chang, K.; Beers, A.L.; Bai, H.X.; Brown, J.M.; Ly, K.I.; Li, X.; Senders, J.T.; Kavouridis, V.K.; Boaro, A.; Su, C.; et al. Automatic assessment of glioma burden: A deep learning algorithm for fully automated volumetric and bidimensional measurement. NeuroOncol. 2019, 21, 1412–1422.
64. Rios Velazquez, E.; Meier, R.; Dunn, W.D., Jr.; Alexander, B.; Wiest, R.; Bauer, S.; Gutman, D.A.; Reyes, M.; Aerts, H.J. Fully automatic GBM segmentation in the TCGA-GBM dataset: Prognosis and correlation with VASARI features. Sci. Rep. 2015, 5, 16822.
65. Chaddad, A.; Sabri, S.; Niazi, T.; Abdulkarim, B. Prediction of survival with multi-scale radiomic analysis in glioblastoma patients. Med. Biol. Eng. Comput. 2018, 56, 2287–2300. Cancers 2021, 13, 722 19 of 19
66. Li, Z.C.; Bai, H.; Sun, Q.; Zhao, Y.; Lv, Y.; Zhou, J.; Liang, C.; Chen, Y.; Liang, D.; Zheng, H. Multiregional radiomics profiling from multiparametric MRI: Identifying an imaging predictor of IDH1 mutation status in glioblastoma. Cancer Med. 2018, 7, 5999–6009.
67. Park, J.E.; Kim, H.S.; Kim, N.; Park, S.Y.; Kim, Y.H.; Kim, J.H. Spatiotemporal Heterogeneity in Multiparametric Physiologic MRI Are Associated with Patient Outcomes in IDH-wildtype Glioblastoma. Clin. Cancer Res. 2020.
68. Tustison, N.J.; Avants, B.B.; Cook, P.A.; Zheng, Y.; Egan, A.; Yushkevich, P.A.; Gee, J.C. N4ITK: Improved N3 bias correction. IEEE Trans. Med. Imaging 2010, 29, 1310–1320.
69. van der Walt, S.; Schonberger, J.L.; Nunez-Iglesias, J.; Boulogne, F.; Warner, J.D.; Yager, N.; Gouillart, E.; Yu, T. scikit-image: Image processing in Python. PeerJ 2014, 2, e453.
70. Kaur, H.; Rani, J. MRI brain image enhancement using Histogram Equalization techniques. In Proceedings of the 2016 International Conference onWireless Communications, Signal Processing and Networking (WiSPNET), Chennai, India, 23–25 March 2016; pp. 770–773.

71. Reinhold, J.C.; Dewey, B.E.; Carass, A.; Prince, J.L. Evaluating the Impact of Intensity Normalization on MR Image Synthesis. Proc. SPIE Int. Soc. Opt. Eng. 2019, 10949.
72. van Griethuysen, J.J.M.; Fedorov, A.; Parmar, C.; Hosny, A.; Aucoin, N.; Narayan, V.; Beets-Tan, R.G.H.; Fillion-Robin, J.C.; Pieper, S.; Aerts, H. Computational Radiomics System to Decode the Radiographic Phenotype. Cancer Res. 2017, 77, e104–e107.

CHAPTER 4:

PREDICTING ADVERSE RADIATION EFFECTS IN BRAIN TUMOURS AFTER STEREOTACTIC RADIOTHERAPY WITH DEEP LEARNING AND HANDCRAFTED RADIOMICS

Authors: Simon A. Keek[1], Manon Beuque[1], Sergey Primakov, Henry C. Woodruff, Avishek Chatterjee, Janita E. van Timmeren, Martin Vallières, Lizza E. L. Hendriks, Johannes Kraft, Nicolaus Andratschke, Steve E. Braunstein, Olivier Morin[2] and Philippe Lambin[2]

1 These authors have contributed equally.
2 Share senior authorship.

Adapted from:
Keek SA, Beuque M, Primakov S, Woodruff HC, Chatterjee A, van Timmeren JE, Vallières M, Hendriks LEL, Kraft J, Andratschke N, Braunstein SE, Morin O, Lambin P. Predicting Adverse Radiation Effects in Brain Tumors After Stereotactic Radiotherapy With Deep Learning and Handcrafted Radiomics. Frontiers in Oncology 2022; 12. doi: 10.3389/fonc.2022.920393

Acsess link:
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9326101/

Abstract

Introduction: There is a cumulative risk of 20–40% of developing brain metastases (BM) in solid cancers. Stereotactic radiotherapy (SRT) enables the application of high focal doses of radiation to a volume and is often used for BM treatment. However, SRT can cause adverse radiation effects (ARE), such as radiation necrosis, which sometimes cause irreversible damage to the brain. It is therefore of clinical interest to identify patients at a high risk of developing ARE. We hypothesized that models trained with radiomics features, deep learning (DL) features, and patient characteristics or their combination can predict ARE risk in patients with BM before SRT.

Methods: Gadolinium-enhanced T1-weighted MRIs and characteristics from patients treated with SRT for BM were collected for a training and testing cohort (N = 1,404) and a validation cohort (N = 237) from a separate institute. From each lesion in the training set, radiomics features were extracted and used to train an extreme gradient boosting (XGBoost) model. A DL model was trained on the same cohort to make a separate prediction and to extract the last layer of features. Different models using XGBoost were built using only radiomics features, DL features, and patient characteristics or a combination of them. Evaluation was performed using the area under the curve (AUC) of the receiver operating characteristic curve on the external dataset. Predictions for individual lesions and per patient developing ARE were investigated.

Results: The best-performing XGBoost model on a lesion level was trained on a combination of radiomics features and DL features (AUC of 0.71 and recall of 0.80). On a patient level, a combination of radiomics features, DL features, and patient characteristics obtained the best performance (AUC of 0.72 and recall of 0.84). The DL model achieved an AUC of 0.64 and recall of 0.85 per lesion and an AUC of 0.70 and recall of 0.60 per patient.

Conclusion: Machine learning models built on radiomics features and DL features extracted from BM combined with patient characteristics show potential to predict ARE at the patient and lesion levels. These models could be used in clinical decision making, informing patients on their risk of ARE and allowing physicians to opt for different therapies.

1. Introduction

Brain metastases (BM) are the most common intracranial malignancies, accounting for more than 50% of all brain tumours and occurring in 10 to over 40% of patients with solid malignancies (1–3). BM occur most often in patients with lung cancer, breast cancer, and melanoma, which have a cumulative risk ranging from 20 to 40% of developing BM (4–7). BM can be treated locally by surgery or radiotherapy or with systemic anticancer therapy. Treatment depends on several factors, such as patient performance status, number and volume of metastases, presence of extracranial metastases, symptoms, and presumed efficacy of available systemic therapy "Systemic therapy for brain metastases" (8, 9). The radiotherapy of BM can be either stereotactic radiotherapy (SRT) or whole brain radiotherapy (WBRT), with SRT being the guideline-recommended treatment for a limited number of BM. As WBRT is associated with neurocognitive deterioration, SRT is increasingly used in multiple BM as well (10–12). SRT is delivered either in a single fraction, with stereotactic radiosurgery (SRS), or as fractionated stereotactic radiotherapy (FSRT) and results in a high dose within the target volume with a steep dose gradient to the surrounding healthy tissue (13).

Even though most of the healthy brain is spared from high doses of radiation, a major shortcoming of SRT is a chance of high toxicity in the immediate surrounding tissues, which may lead to adverse radiation effects (ARE) such as radiation necrosis (RN), subacute edema, structural changes in the white matter, and vascular lesions (14). ARE are a relatively late reaction to irradiation of healthy tissues where either reversible or irreversible injury has occurred (15). The risk of ARE after SRT and SRS is found to be similar and ranges from 5 to 10% at patient level (16–19) or approximately 3% at lesion level (15). Known predictors of ARE are tumour volume, isodose volume, and previous SRT to the same lesion (15). ARE of the tumour area and tumour progression (TP) as two different post-therapeutic events require different treatment strategies: while steroids are often indicated for the initial treatment of ARE, true progression or relapse requires repeated radiotherapy, surgery, or effective intracranial systemic therapy for tumour control. Being able to differentiate between ARE and TP is therefore of utmost clinical interest.

Unfortunately, the (neurological) symptoms of ARE and TP are usually indistinguishable. Furthermore, the appearances of ARE and TP are very difficult to discern through qualitative radiological imaging, requiring multiple successive magnetic resonance images

(MRI), specialized MRI sequences such as perfusion-weighted or MR spectroscopy, and trained experts to evaluate the findings (19, 20). The clinical workflow is time- and labor-intensive, and while it is unfeasible to perform for every lesion, a definitive confirmation of the presence of ARE requires tissue acquisition (19).

SRT requires routine pretreatment MRI for accurate target volume delineation. This imaging provides a source of non-invasively acquired information about BM and brain phenotypes that could be investigated for their potential to determine before treatment which patient has a high risk of developing ARE. The early identification of these patients is an unmet clinical need which may help in clinical decision making by informing the patients of the risk of ARE, the early risk stratification of patients that may develop ARE, and the consideration of ARE risk mitigating strategies such as deferring radiotherapy for central nervous system-penetrant systemic therapy.

Advanced quantitative medical image analysis methods such as radiomics and deep learning (DL) extract large amounts of imaging features and associate these with biological and/or clinical outcomes using machine learning (ML) techniques (21–26). Thus, radiological images from routine imaging procedures could potentially be used to non-invasively quantify the lesion phenotype, providing clinically necessary information for patient management decisions. Several studies have indicated that MRI radiomics analysis is able to differentiate BM from glioblastoma (27, 28) to predict local recurrence (29, 30), to predict the origin of metastases (31, 32), and to predict overall survival (33, 34). DL has also shown potential in predicting treatment response on brain MRI (35).

Moreover, DL and radiomics can have a complementary value, potentially establishing a more robust classifier (36).

We hypothesize that models trained with radiomics features, DL features, and patient characteristics or a combination thereof can predict the occurrence of ARE in patients with BM, both lesion specific and patient specific.

2. Materials and methods

2.1. Patient Characteristics

All data from patients with BM treated with SRT between 1997 and 2017 for which imaging, outcome data, and patient data were available were collected retrospectively from the University of California—San Francisco (UCSF) medical center's picture archiving and communication system. Available imaging data, outcome data, and patient data of all patients with BM treated with SRS/SRT between 2014 and 2019 at the University Hospital Zürich (USZ)

were collected retrospectively. The data included clinical and biological information for both the patient and the lesion. The eligibility criteria included radical treatment for metastatic brain cancer using Gamma Knife SRS for the UCSF patients and SRS/FSRT for the USZ patients. The inclusion of patients was regardless of the number of BM, but pathohistological or imaging-based confirmation of ARE during the follow-up was required in addition to pathohistological confirmation of the primary tumour. For the USZ cohort, in case of imaging-based suspicion of RN, positron emission tomography imaging was additionally used to exclude TP. The effort obtained ethical approval for observational research using anonymized linked care data for supporting medical purposes that are in the interests of individuals and the wider public. UCSF Institutional Review Board (https://irb. ucsf.edu) and Cantonal Ethics Committee Zurich approval with waiver of informed consent was obtained.

The UCSF dataset was divided randomly into sub-cohorts for training (70%) and testing (30%) while maintaining the ratios of events to non-events equal in both groups. The USZ dataset was used as an independent external validation dataset, i.e., it was entirely unseen by the models during the training and testing phases. The binary outcome used in training and validation was ARE per lesion, defined as either pathologically or imaging-based confirmation of RN occurring at any time after treatment. For both the USCF and USZ patients, ARE was confirmed by histopathology when treated with open surgery. In all other cases, ARE was confirmed either at routine re-staging 3 months after radiotherapy for asymptomatic patients or at the onset of new symptoms. When patients presented new symptoms, imaging was performed usually after awaiting the effects of cortisone administration. As the time of BM formation is unknown, the outcome was not defined as right-censored. As every lesion is able to independently develop ARE after treatment, every lesion was considered to be an independent sample. The probability of ARE occurring for any lesion within a patient as an outcome was also investigated, whereby each patient was treated as an independent sample instead.

2.2. MR Acquisition Parameters and Lesion Segmentation

All images were axial gadolinium-enhanced T1-weighted MRI acquired prior to the treatment of BM. All included lesions were three-dimensionally delineated for curative Gamma Knife SRS treatment purposes for the UCSF cohort and for curative SRS/ FSRT purposes for the USZ cohort according to local protocols by an experienced radiation oncologist. Figure 1 shows two T1-weighted gadolinium-enhanced MRI with lesions delineated for SRT purposes. To perform segmentations of the brain and the ventricles on the entire dataset, an atlas-based segmentation strategy was chosen. To create the atlas in the MIM software package (MIM v. 6.9.4, MIM Software Inc., Cleveland, OH, USA), 50 randomly chosen MRI were manually segmented by an expert radiologist.

2.3. Pre-Processing of Brain MRI Data

Bias-field correction was performed in the MIM software package using the N4 algorithm, which required brain segmentations (37). A bias field is a low-frequency signal distributed over an MR image, which is caused by inhomogeneities in the magnetic field of the MRI scanner. This causes shifts of intensity value ranges across the image (38). The ventricle mask was subtracted from the brain mask to obtain a white- and gray-matter segmentation. This segmentation was used to determine and correct the bias field present in the image using the N4 algorithm (37) using the MIM software package.
Following the bias correction, all remaining pre-processing, feature extraction, model training, and evaluation were performed in Python (version 3.7). The different Python packages used during this study can be found in Supplementary Table S1. Pre-processing of MRI is essential for ML purposes, for reducing scanner dependence, and for ensuring reproducibility (39–41). As there is, to date, no consensus regarding the best way to pre-process MRI for our purposes, three different pre-processing workflows were applied and compared: “minimalist”, standardization, and “harmonization”. The descriptions of these pre-processing workflows can be found in the Supplementary Materials (Section 1 and in Figure 2).

Pre-processing for radiomics and feature extraction

Feature extraction was performed according to the Image Biomarker Standardization Initiative (IBSI) guidelines (42–44) on the three different sets of processed MRI scans using the BM segmentations. All images were resampled to uniform 1 × 1 × 1- mm3 voxels using the “sitkBSpline” interpolator to correct for differences in pixel size and slice spacing. The choice for voxel dimensions was made based on majority ruling, as it was found that most patients had a pixel

spacing of ~1 mm. To achieve isotropic voxels, the choice for resampling in the z-direction was also chosen as 1 mm. Pixel intensity values were resampled to a fixed number of 64 bins, as the number of gray levels was found to affect the interchangeability of MRI radiomics features, and a fixed bin number of 64 has been found recommended in previous studies (42–44).

A total of 106 IBSI features were extracted from each segmentation. The features were extracted from the BM segmentations of the pre-processed images and can be divided into first-order intensity, histogram statistics, shape, and texture features. A full list and a description of the features can be found in the PyRadiomics documentation ([Radiomics features— PyRadiomics Documentation, (45)], and a description of the feature groups can be found in the Supplementary Materials (Section 2).

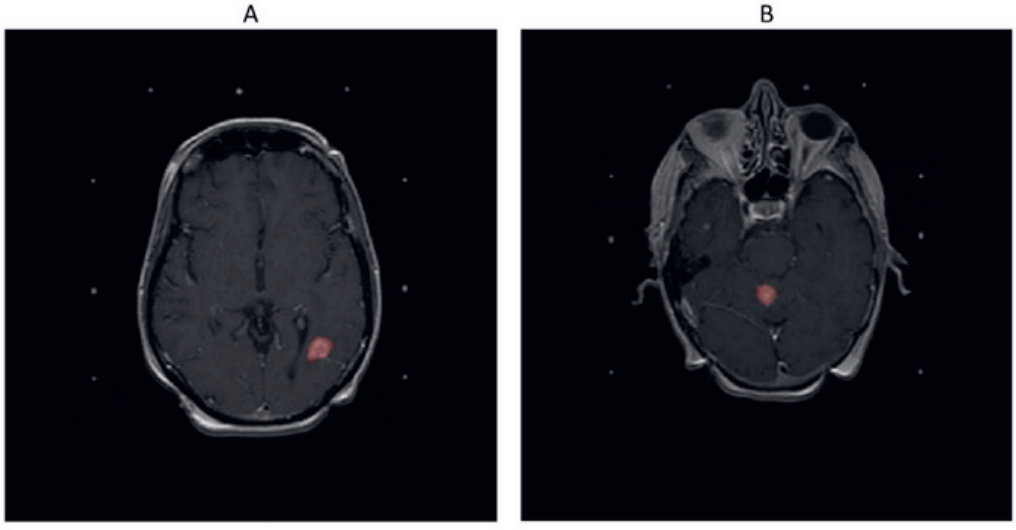


Figure 1: T1-weighted gadolinium-enhanced MRIs of the brain. Delineated in red (A) is a lesion that developed adverse radiation effects after stereotactic radiotherapy and (B) a lesion that did not develop adverse radiation effects after stereotactic radiotherapy.

Pre-processing for deep learning

To inform the DL model on the location and extension of the lesions, lesion masks were used to highlight the ROI. A Gaussian smoothing filter was applied to the image, gradually decreasing the intensity values around the lesion from a factor of 1.0 to 0.2 to still include information of the voxels immediately around the lesion masks.

Otsu thresholding was performed to create a mask containing the brain and the skull. This mask was used to determine the largest three-dimensional bounding box containing the brain and the skull to crop the images. Anything outside this mask was defined as the image background, for which all pixel values were set at 0. For the "minimalist" and the "standardization" datasets, the intensities were resampled in a range between 0 and 255. Finally, the scans were rescaled at 256 × 256 × 64 with spline interpolation order 3. As an example, the steps of the pre- processing workflow for the "minimalist" normalization are illustrated in Figure 3.

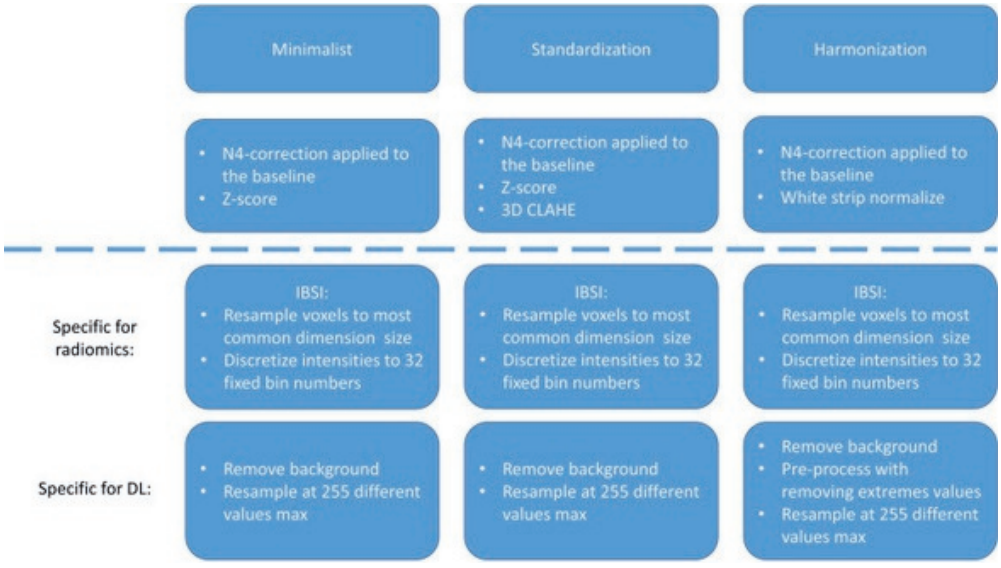


Figure 2: Pre-processing strategies for the "minimalist", "standardization", and "harmonization" approaches.

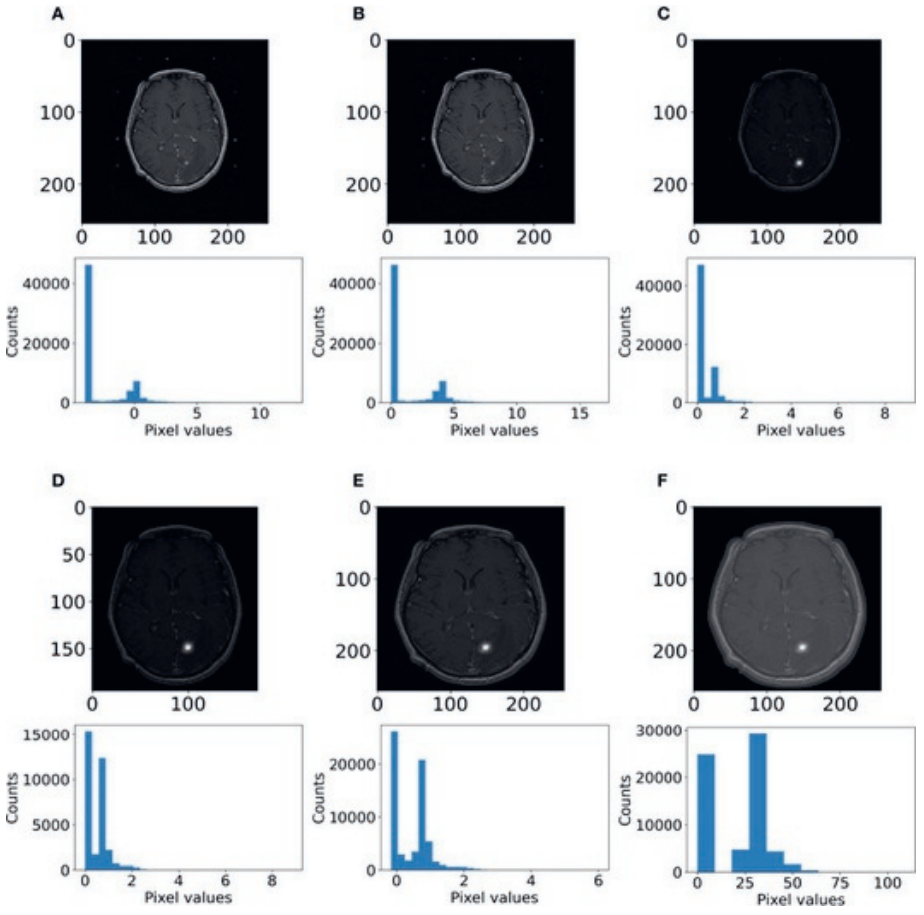


Figure 3: Example of pre-processing strategy: deep learning on the "minimalist" approach. The different steps of preprocessing were (A) z-score normalization, (B) shift to positive values only, (C) pixel attenuations with Gaussian smoothing filtering, (D) cropping around the largest bounding box and background set to 0, (E) resizing at 256 × 256, and (F) rescaling the pixel value range to 0–255.

2.4. Machine Learning Models

The mean and SD of each feature over the entire training population were determined. These values were used to apply z-score normalization to the features of the training, testing, and external validation datasets (46). Next, features with low variance (<0.01) were determined and excluded from the dataset. Lastly, the correlation between features was determined using absolute pairwise Spearman rank correlation. As highly correlated features (>0.85) were assumed to contain overlapping information about the outcome, the feature with the highest mean absolute correlation with the rest of the features was excluded. Lastly, supervised

feature selection was performed through recursive feature elimination (RFE). RFE uses a ML algorithm to build a multivariate model and determine predictive performance using the currently selected features. It recursively drops and adds features, determining the optimal number of features and the selection of most predictive features.
An extreme gradient boosting (XGBoost) model was used for RFE and ARE prediction. A description of the XGBoost architecture and the methodology to determine the optimal hyperparameters for the trained models can be found in the supplementary materials (Section 3).

2.5. Deep Learning Model

An Xception three-dimensional model was trained and tested on the same datasets as the handcrafted radiomics-based model. Xception is the extreme version of an Inception model (47), which uses depth-wise separable convolutions. The architecture can be found in Supplementary Figure S1. Adam optimization was used (48) with an initial learning rate of 10-5, which updated the learning rate during training, and used for loss function binary cross-entropy. This model produced a score ranging from 0 to 1, indicating the estimated probability that a lesion develops ARE. The area under the curve (AUC) of the receiver operating characteristic (ROC) was monitored on the test dataset. The ROC displays the discriminative performance of a model expressed through the sensitivity and specificity as the threshold for binary classification is shifted. The AUC of the ROC is a metric from 0 to 1, where 1 means that the model has perfect predictive performance and 0.5 is equivalent to guessing. To limit the imbalance of the outcomes to affect the model training, the model was only trained on lesions for those patients who had at least a single ARE and tested on the scans of the patients who had ARE in the test dataset. To combine DL and radiomics, the last fully connected layer consisting of 256 features obtained after training the model was extracted. These features were then used to train a ML model similarly to using radiomics features and used in models combining radiomics features and patient characteristics.

2.6. Clinical and Treatment-Related Feature Model

As the training and testing datasets contained patient characteristics not available in the external validation dataset, any feature not overlapping between these datasets was dropped. The list of the remaining features was as follows: primary tumour location, primary tumour histology, primary tumour controlled,

extra-cranial metastases presence, patient age, patient sex, SRS to the same location, prior external beam radiotherapy (EBRT), prior radiosurgery (RS), neurological symptoms, headaches, seizures, hypertension, diabetes, connective tissue disorder, Karnofsky performance score (KPS) status, prescription dose, and isodose lines. For XGBoost to be able to handle categorical variables, one-hot encoding was performed on two categorical clinical features (primary tumour location and primary tumour histology).

Missing values were imputed using MissForest. MissForest is an imputation algorithm that uses RandomForest to train a model on the non-missing data for each feature with missing values to predict the missing values. In the first iteration, all values are set to the mean value present for each variable (i.e., each column). Then, over multiple iterations, each data column with missing values will be predicted using all the data except for the rows containing the missing values in question. This process is repeated over several iterations.

2.7. Metrics Used for Data Analysis

The patient and tumour characteristics in the UCSF and USZ cohorts were assessed through a two-proportion z-test to test for significant differences in categorical variables between the cohorts or the unpaired two-sample t-test to test for significant differences in numerical variables. For the latter, the assumptions of the data having a normal distribution and possessing the same variance in both cohorts were tested through Shapiro–Wilk's test and f-test, respectively. The significance level was set at 5%.

To determine which method ensured best performance for the radiomics-based and DL models, models were trained on the three different pre-processed datasets, and the best AUC of the ROC on the testing set was used to determine the best pre- processing methods for ML and DL separately. The 95% confidence intervals (CI) displayed on the ROC curves were obtained using bootstrapping (n = 2,000). For the radiomics- based model, the results were reported on the full train dataset and the entire test dataset. For the DL model, the results were reported on the balanced train dataset (which served to train the different DL models) and the full test dataset.

Once the best models were selected, the models were validated on the external dataset. The predictive performance of each model was expressed through the ROC curve and its AUC on the training, testing, and external data. By determining an optimal threshold value using Youden's J statistic (49) based on the training dataset, a binary classification was performed on the external dataset. From

this binary classification, the balanced accuracy, precision, recall, and F1-score were determined. The confusion matrices were also derived from the binary classification. To determine model performance and to compare between models, the recall was investigated specifically, which is the proportion of true positives of the total number of true cases. As the number of events was relatively low and not missing any patients at risk of ARE is crucial, a high recall of the models was desirable. The CI obtained for all metrics were obtained using bootstrapping, resampling the results 2,000 times. Moreover, an analysis of the agreement prediction between the DL model and the radiomics-based model was performed. To give a prediction per patient, the maximum prediction of ARE among the different lesion predictions of the patient was selected. The ground truth to which the prediction was compared with was the ARE status of the patient, meaning that the patient had at least one ARE lesion. An overview of the models tested can be found in Figure 4.
We evaluated on the external dataset for which cases the DL model and the best radiomics classifier obtained the same predictions and reported the number of cases for which those models agreed on the label. The metrics based on the data for which the models agreed was also reported.

3. Results

3.1 Patient Characteristics

A total of 1,404 patients with 7,974 lesions from UCSF and 237 patients with 646 lesions from USZ were included. Table 1 shows an overview of the patient characteristics of the UCSF and USZ data. Significant differences between the proportion of male and female patients between the datasets ($P < 0.01$), median age ($P = 0.03$), KPS status ($P < 0.01$), and the number of lesions per patient at treatment ($P < 0.01$) were found. Furthermore, the proportions of primary tumour (lung, melanoma, and breast) were different between the datasets, and the data from USZ did not have kidney, GI, sarcoma, or other types of primary locations that were present in the UCSF dataset. For the histology of the primary tumour, only the melanoma histology subtype was found to be present in a significantly different proportion.

3.2. Radiomics-Based Model and DL Model Results Based on the Three Different Preprocessing Methods of the Dataset

The best AUC on the test dataset for the radiomics-based models was found using the "harmonization" normalization, with an AUC of 0.76 (CI of 0.70–0.81), compared with 0.75 (CI of 0.70– 0.80) and 0.73 (CI of 0.67–0.79) for the "minimalist" and "standardization" methods, respectively.

The best AUC on the test dataset for the DL models was found using the "standardization" normalization, with an AUC of 0.72 (CI of 0.66–0.78), compared with 0.63 (CI of 0.57–0.70) and 0.65 (CI of 0.58–0.71) for the "minimalist" and "harmonization" methods, respectively. Figure 5 shows the ROC curves of the training and testing datasets for the three different pre-processing methods for radiomics-based ML and for DL.

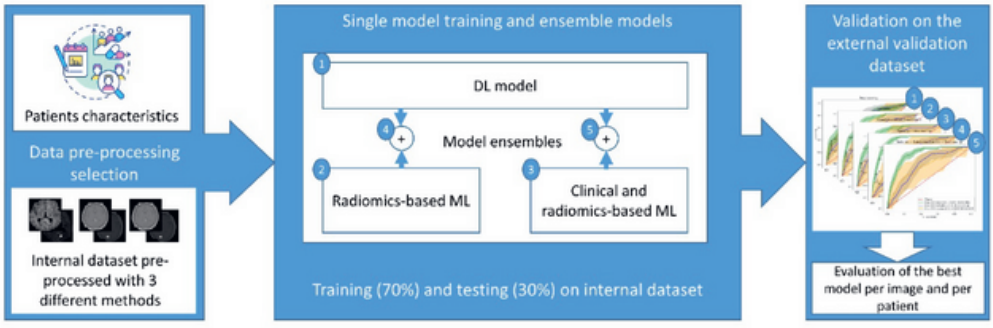


Figure 4: General workflow of the model training process: first, the MRI data was pre-processed using 3 pre-processing methods, the most suitable pre- processed set of images was selected according to the radiomics-based model or the DL model performance on the internal test dataset, then the models were ensembled or trained separately, and finally the performance of each model was computed on the external dataset.

Table 1: Patient characteristics of University of California—San Francisco (UCSF) and University Hospital Zurich (USZ) datasets.

Median age ± SD		59 (13)	62 (12)	0.03
KPS (%)	80–100	1,053 (75)	198 (83)	<0.01
	40–80	351 (25)	37 (16)	<0.01
	10–40	0 (0)	2 (1)	–
Primary tumor location (%)	Lung	530 (38)	136 (58)	<0.01
	Breast	357 (25)	27 (11)	<0.01
	Melanoma	272 (19)	74 (31)	<0.01
	Kidney	91 (7)	0 (0)	–
	Gastrointestinal	57 (4)	0 (0)	–
	Gynecologic	27 (2)	0 (0)	–
	Sarcoma	20 (1)	0 (0)	–
	Other	50 (4)	0 (0)	–
Histology primary tumor (%)	Adenocarcinoma	802 (57)	124 (52)	0.17
	Melanoma	272 (19)	74 (31)	<0.01
	Renal cell carcinoma	88 (6)	0 (0)	–
	Small cell carcinoma	44 (3)	0 (0)	–
	Squamous cell carcinoma	40 (3)	10 (4)	0.26
	Sarcoma	18 (1)	0 (0)	–
	Large cell carcinoma	9 (0.6)	2 (1)	0.72
	Bone carcinoma	8 (0.6)	0 (0)	–
	Adeno squamous carcinoma	6 (0.4)	0 (0)	–
	Broncho alveolar cell carcinoma	5 (0.4)	0 (0)	–
	Germ cell carcinoma	2 (0.1)	0 (0)	–
	Lymphoma	1 (0.1)	0 (0)	–
	Other/NOS	109 (8)	27 (11)	0.06
Primary controlled		974 (70)	149 (63)	0.05
ECM present		1,097 (78)	190 (80)	0.48
Number of lesions per patient at treatment	Median ± SD	3 (7)	2 (3)	<0.01
Symptoms	Headaches	437 (31)	31 (13)	<0.01
	Hypertension	407 (29)	0 (0)	< 0.01
	Seizures	134 (10)	16 (7)	0.17
	Diabetes	98 (7)	13 (6)	0.4
	CTD	21 (2)	2 (1)	0.43
Number of lesions in total		7,974	646	–
Number of ARE cases (% of total lesions)		217 (2.7)	20 (3.1)	0.61
Number of patients with ARE (% of total patients)		155 (11)	19 (8)	0.16
Prescription dose ± SD (Gy)		18.5 (1.5)	20 (5.0)	–

Footnote: P value of two-proportion z-test or unpaired two-sample t-test for significant differences between datasets was reported for each characteristic if applicable. SD = standard deviation;

KPS = Karnofsky performance score: 80-100 good performance, 50-70 medium performance, 10-40 bad performance; ECM = extracranial metastasis; BM = brain metastasis; CTD = connective tissue disorder; ARE = adverse radiation effect; Gy = gray.

3.3. Results of the Combined Best- Performing Models

We calculated the AUC and CI for each model combination on the external validation dataset. The DL model, built on images pre-processed with the "standardization" method, achieved an AUC of 0.64 (CI of 0.50–0.76). The model built on radiomics features, extracted from the images pre-processed with the "harmonization" method, achieved an AUC of 0.73 (CI of 0.63–0.83). The model was built on 20 features selected through RFE. Supplementary Figure S2A provides an overview of the selected features and the corresponding importance in the XGBoost model. Supplementary Table S2 provides an overview of the hyperparameters determined through grid search cross-validation. The model based on the combination of the DL features extracted from the last layer and radiomics features achieved an AUC of 0.71 (CI of 0.60–0.82). The model was built on 10 features selected through RFE. Supplementary Figure S2B provides an overview of the selected features and the corresponding importance in the XGBoost model. The model built on radiomics features, extracted from images pre-processed with the "harmonization" method, combined with patient characteristic features achieved an AUC of 0.70 (CI of 0.57–0.80). The model was built on 19 features selected through RFE. Supplementary Figure S2C provides an overview of the selected features and the corresponding importance in the XGBoost model. Finally, the model built on radiomics features, extracted from images pre-processed with the "harmonization" method, combined with DL features, extracted from images pre-processed with the "standardization" method, and patient characteristics achieved an AUC of 0.69 (CI of 0.56–0.81). The model was built on 20 features selected through RFE. Supplementary Figure S2D provides an overview of the selected features and the corresponding importance in the XGBoost model. Figure 6 shows the ROC curves with CI of the training datasets, testing datasets, and validation datasets for these models.

The combination of radiomics and DL features achieved the highest combination of balanced accuracy and recall of 0.67 (CI of 0.56–0.76) and 0.80 (CI of 0.62–0.96), respectively, of the externally validated models for predictions per lesion. For a patient-level prediction, the DL model achieved an AUC of 0.70 (CI of 0.56–0.80) and that of the radiomics model an AUC of 0.72 (CI of 0.60–0.83).

A combination of radiomics and DL achieved an AUC 0.71 (CI of 0.57–0.83), that of a combination of radiomics and patient characteristics an AUC of 0.71 (CI of 0.59–0.81), and that of a combination of radiomics features, DL features, and patient characteristics an AUC of 0.72 (CI of 0.58– 0.84). The model combining radiomics features, DL features, and patient characteristics achieved the highest combination of balanced accuracy and recall of 0.65 (CI of 0.55–0.74) and 0.84 (CI of 0.65–1.00), respectively, of the externally validated models for predictions per patient. The DL model predictions and the radiomics-based model predictions per lesion agreed for 32% of the external dataset. For the per-patient classification, the DL model predictions and the radiomics combined with clinical feature-based model predictions agreed for 19% of the external dataset. Because the number of patients for which the models agreed was low (47 patients, 6 with ARE), no CI could be derived. Table 2 provides an overview of the AUC, balanced accuracy, precision, recall, and F1 score metrics for all DL and ML models on both lesion and patient levels and for the agreed labels on the external validation. The corresponding confusion matrices are in Supplementary Figures S3, S4, respectively. Supplementary Tables S3, S4 contain the same metrics as that in Table 2 for the training and testing datasets, respectively.

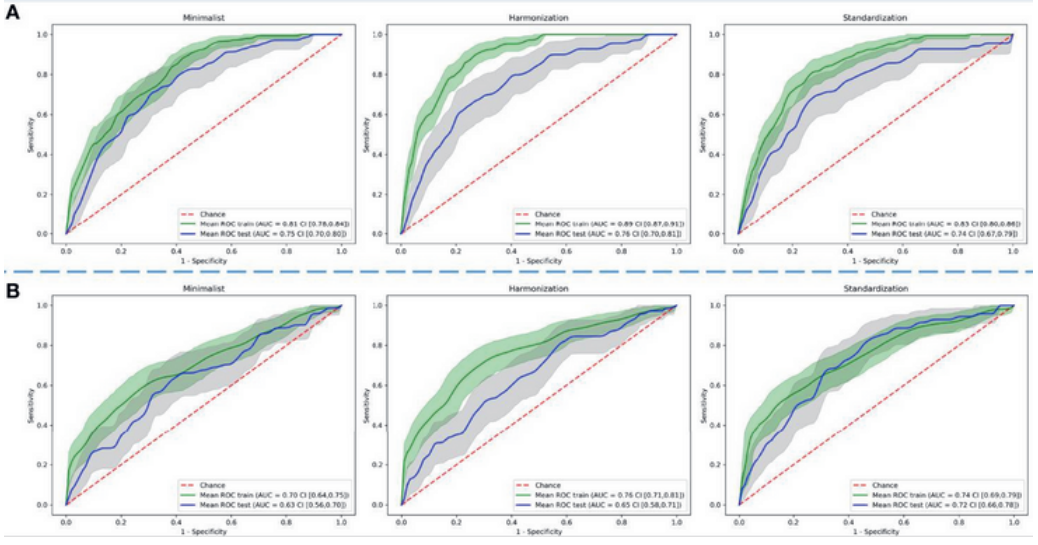


Figure 5: Comparison of predictive performance through receiver operating characteristic curves for (A) radiomics-based machine learning and (B) deep learning models using three different pre-processed image datasets. The shaded areas represent the 95% confidence intervals of the corresponding receiver operating characteristic curves.

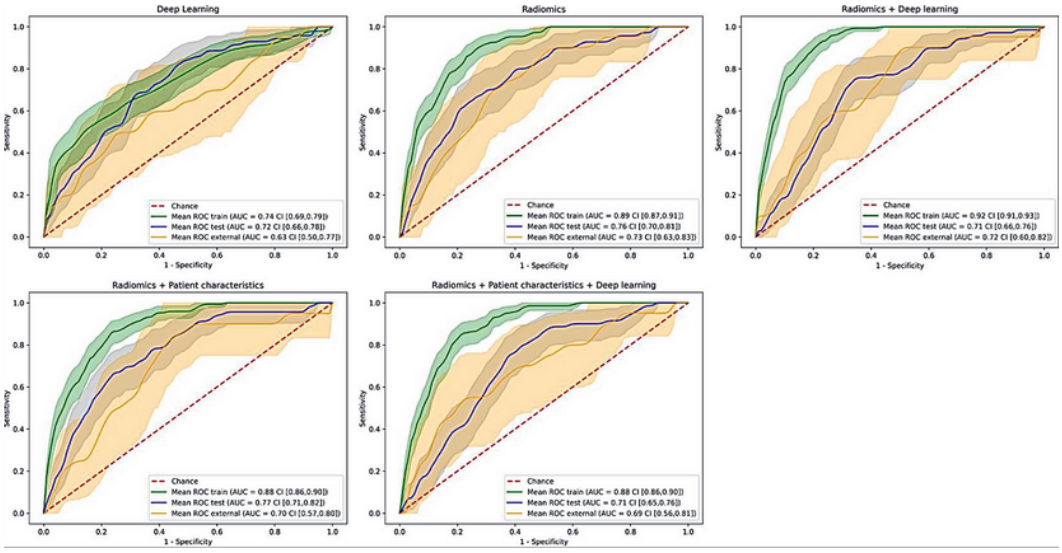


Figure 6: Receiver operating characteristic curves of the training, testing, and external validation datasets for the different model combinations. The shaded areas represent the 95% confidence intervals of the corresponding receiver operating characteristic curves.

4. Discussion

Patients with BM treated with SRT are at risk of developing ARE, such as RN. Early identification of these patients can help in clinical decision making. The MRIs required for SRT planning provide an opportunity to identify these patients through quantitative imaging methods. In this large-scale study, ML models that can successfully predict ARE were trained on T1- weighted MR imaging features from secondary brain tumours treated with SRT. As no consensus to harmonize MR images within and between centers exists, multiple methods were tested for the DL and ML pipeline, resulting in two optimal pre- processing methods (“harmonization” for the ML pipeline and “standardization” for the DL pipeline). A ML model trained with radiomics features combined with DL features yielded the highest predictive performance, with a combination of ROC AUC, balanced accuracy, and recall of 0.71, 0.67, and 0.80, respectively. At the patient level, the best-performing ML model was clearly a combination of radiomics, clinical (age at treatment,

prior RS, and sex), and DL features achieving the highest predictive performance (AUC of 0.72), a balanced accuracy of 0.65, and recall of 0.84.

Performing an aggregate prediction (i.e., using only those predictions that agreed on the outcome) did not improve predictive performance for the lesion-level prediction (AUC of 0.67) nor the binary prediction (balanced accuracy of 0.65). However, using this method, the highest recall of 0.90 was achieved, making this method very robust in detecting true positives. The models pave the way for clinical decision making of patients at risk of ARE before treatment. The information on the risk of an individual patient may be used by clinicians to inform patients of the risk of ARE when SRT is used as treatment. Furthermore, this information may be used to perform an early stratification of those patients at high risk or may allow the patient and clinician to pursue alternative therapy, such as systemic therapy or alternate radiotherapy approaches (e.g., dose de-intensified SRT or WBRT), if the risk of ARE outweighs the possible benefits of SRT (50).

To our knowledge, this is the first study that performs a pre-treatment prediction of ARE using quantitative image analysis. Several studies have investigated the possibility of differentiating between tumour recurrence and RN after treatment, which is nominally similar in purpose to identify those patients who may have ARE. Zhang et al. (51) used radiomics features extracted from four different MR sequences [T1, T1 post-contrast, T2, and fluid-attenuated inversion recovery (FLAIR)] at two different time-points during follow-up to differentiate RN from TP as confirmed pathologically. A model was built on a dataset of 87 patients with 97 lesions using 5 delta-radiomics features from T1 and T2 sequences. The AUC and binary prediction accuracy of the model were both 0.73. However, this result was obtained using leave-one-out cross-validation, as no external validation was used. Similarly, Peng et al. created a model on radiomics features extracted from T1 and T2 FLAIR on 66 patients with 77 lesions in total (52). The model was compared with a neuroradiologist's performance. No external validation was used, and instead a leave- one-out cross-validation was performed, which gave an AUC of 0.81. The sensitivity and specificity of the neuroradiologist were 0.97 and 0.17, compared with 0.65 and 0.87 for the radiomics- based model. In Park etal. (53), the study compared the results obtained after training radiomics-based models using different MRI sequences [T1, T2, and apparent diffusion coefficient (ADC)]. The models were trained using the data from 86 patients and tested on an external dataset of 41 patients. The best AUC was found on the ADC-based data with 0.80, while the other sequences had AUCs of around 0.65. These

results are similar or higher than the results obtained with our model, though within the range of the confidence intervals for the model based on radiomics and DL, and the lack of an external dataset on two of the studies makes the validity of these models difficult to determine (52). Most other studies have a similar lack of external validation and total number of included patients, further making the results difficult to compare with the present study (54). These results show that the model presented in this study is able to perform similarly to or even outperform models that perform classification (post-treatment) instead of prediction (pre-treatment) of ARE.

One of the strengths of the present study is the large number of included patients and subsequent lesions, with 7,974 lesions (2.7% ARE) of 1,404 patients in training and testing and 646 lesions (3.1% ARE) of 237 patients in the external validation. This provides a large volume of data for our models to train on, ensuring that it covers the wide variability found between patients. In addition, the inclusion of an external validation is another strength, especially seeing the general lack of one in most other studies investigating ARE. This ensures that the reported result is not too optimistic and shows that our model can be generalizable to populations from a different hospital in a different country and even with different treatments from the training and testing sets. While the difference in treatment between the training (exclusively SRS) and external validation (a mix of SRS and FSRT) may induce variability due to small differences in treatment planning for these methods, literature has shown that these methods carry the same risk of ARE and were therefore considered interchangeable (16, 17, 19).

The large confidence interval on the external validation is partially due to the low number of positive findings in this dataset (n = 20). This is because of the large imbalance in outcomes for both ARE and tumour failure. One of the major problems that may arise from this imbalance is a skewed view of predictive performance. However, this was addressed in the present study through multiple measures. The DL model was trained on a balanced subset of the data that only included patients that suffered at least 1 ARE. For ML, the XGBoost model was trained while scaling the weights of positive and negative classes and the respective proportion of the labels. Finally, through analysis of the confusion matrix, precision recall curves, and recall metric, we ensured that the performance of the model was not entirely driven by labeling the data as the majority class.

While the models have been successfully validated on a dataset from an external center, further validation on multiple centers is required to ensure that the models are generalizable. Future research could therefore focus on validating the present model on other datasets, potentially with recalibration of the model. At a later stage, a clinical trial to test the efficacy of the model is needed to be able to incorporate the model in a clinical setting. A model combining radiomics features, DL features, and patient characteristics with a high accuracy could help choose other treatment options such as surgery only, systemic therapy, or palliative care (55) if the predicted risk of developing ARE is high. The model could also predict if the patient would be at a low risk of developing ARE, in which case SRT could be preferred over other treatment options.

In the present study, only one sequence of the MRI scan was used. Previous studies showed that a combination of radiomics computed

on T1 and T2 sequences performs best to differentiate ARE and TP (51, 52), and ADC sequence seems to also show a higher performance (53). Investigating more sequences in a future study may therefore improve the performance of the imaging-based models.
Lastly, for ARE (and, to a lesser degree, TP), treatment is one of the primary factors. In this study, multiple-dose-treatment- related variables have been included, such as prior treatments to the same patients as well as dose variables and the volumes encompassing certain dose levels. However, a more thorough "dosiomics" analysis would probably improve the prediction of ARE. Liang et al. (56) described a method to extract the spatial and texture radiomics features from dose maps (56). They found several radiomics features which have a significant predictive value of radiation pneumonitis. Using a similar method for ARE in BM may result in improved prediction results. Our predictions could also be combined with models automatically classifying tumours and RN on brain MRI, such as in Zhang et al. (51), potentially strengthening the results of those studies.

5. Conclusion

Radiomics is able to predict lesions at a high risk of ARE, especially when combined with DL features. When predicting ARE on a patient level, the highest performance was found using a combination of radiomics, DL, clinical, and treatment-related features. These models could potentially be used to aid clinical decision making for patients with BM treated with either gamma knife or EBRT.

Data availability statement

The corresponding author does not own the datasets used (acquired with DTAs). Requests to access the datasets should be directed to olivier.morin@ucsf.edu (for the data from UCSF); Nicolaus.Andratschke@usz.ch (for the data from USZ).

Ethics statement

The studies involving human participants were reviewed and approved by the cantonal ethics committee Zurich and University of California San Francisco (UCSF) Institutional Review Board (IRB). Written informed consent from the participants' legal guardian/next of kin was not required to participate in this study in accordance with the national legislation and the institutional requirements.

Author contributions

MB and SK performed all the ML/DL analysis and wrote the manuscript. SK, MV, SB, and OM collected and curated the imaging and patient data from UCSF. SP helped with the ML/DL analysis and study design. HW supervised the progression of the project and the writing of this article and guaranteed the integrity of the analysis and results presented. AC and MV helped with the ML analysis. JT, JK, and NA collected the imaging and patient data from USZ. LH and SB aided with the clinical aspects of the study. PL and OM devised the project's aim and supervised the progression of the project. All authors contributed to the article and approved the submitted version.

Funding

The research project has been partially funded by the Clinical Research Priority Program "Artificial Intelligence in Oncological Imaging" of the University of Zurich. PL, HW, MB, SK acknowledge financial support from ERC advanced grant (ERC-ADG-2015 n° 694812 - Hypoximmuno), the European Union's Horizon 2020 research and innovation programme under grant agreement: MSCA-ITN-PREDICT n° 766276, CHAIMELEON n° 952172, EuCanImage n° 952103 and IMI- OPTIMA n° 101034347.

Conflict of Interest

LH: none related to the current manuscript, outside of current manuscript: research funding Roche Genentech, Boehringer Ingelheim, AstraZeneca, Takeda (all institution, Beigene under negotiation); advisory board: BMS, Eli Lilly, Roche Genentech,

Pfizer, Takeda, MSD, Merck, Novartis, Boehringer Ingelheim, Amgen, Janssen (all institution, Roche one time self); speaker: MSD, Lilly (institution); travel/conference reimbursement: Roche Genentech (self); mentorship program with key opinion leaders: funded by AstraZeneca; fees for educational webinars: Benecke, Medtalks, VJOncology (self), high5oncology (institution); interview sessions funded by Roche Genentech, Bayer, Lilly (institution); local PI of clinical trials: AstraZeneca, Novartis, BMS, MSD, Merck, GSK, Takeda, Blueprint Medicines, Roche Genentech, Janssen Pharmaceuticals, Mirati; PL: none related to the current manuscript; outside of current manuscript: grants/sponsored research agreements from Radiomics SA, Convert Pharmaceuticals and LivingMed Biotech. He received a presenter fee (in cash or in kind) and/or reimbursement of travel costs/consultancy fee (in cash or in kind) from Radiomics SA, BHV, Varian, Elekta, ptTheragnostic/DNAmito, BMS, and Convert pharma. PL has minority shares in the companies Radiomics SA, Convert pharmaceuticals, Comunicare, and LivingMed Biotech, and he is co-inventor of two issued patents with royalties on radiomics (PCT/NL2014/050248 and PCT/NL2014/ 050728), licensed to Radiomics SA; one issued patent on mtDNA (PCT/EP2014/ 059089), licensed to ptTheragnostic/DNAmito; one non-issued patent on LSRT (PCT/ P126537PC00), licensed to Varian; three non-patented inventions (softwares) licensed to ptTheragnostic/DNAmito, Radiomics SA and Health Innovation Ventures and two non-issued, non-licensed patents on Deep Learning-Radiomics (N2024482, N2024889). He confirms that none of the above entities or funding sources were involved in the preparation of this paper.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Supplementary materials

Supplementary materials are available online and can be accessed via https://www.frontiersin.org/articles/10.3389/fonc.2022.920393/full#supplement ary-material, or the QR code below:

References

1. Walker AE, Robins M, Weinfeld FD. Epidemiology of Brain Tumors: The National Survey of Intracranial Neoplasms. Neurology (1985) 35:219–9. doi: 10.1212/wnl.35.2.219
2. Johnson JD, Young B. Demographics of Brain Metastasis. Neurosurg Clinics North America (1996) 7:337–44. doi: 10.1016/s1042-3680(18)30365-6
3. Wen PY, Loeffler JS. Management of Brain Metastases. Oncology (1999) 13 (7):941–54, 957–61.
4. Schouten LJ, Rutten J, Huveneers HAM, Twijnstra A. Incidence of Brain Metastases in a Cohort of Patients With Carcinoma of the Breast, Colon, Kidney, and Lung and Melanoma. Cancer (2002) 94(10):2698–705. doi: 10.1002/cncr.10541
5. Barnholtz-Sloan JS, Sloan AE, Davis FG, Vigneau FD, Lai P, Sawaya RE. Incidence Proportions of Brain Metastases in Patients Diagnosed, (1993 to 2001) in the Metropolitan Detroit Cancer Surveillance System. J Clin Oncol: Off J Am Soc Clin Oncol (2004) 22(14):2865–72. doi: 10.1200/JCO.2004.12.149
6. Rangachari D, Yamaguchi N, VanderLaan PA, Folch E, Mahadevan A, Floyd SR, et al. Brain Metastases in Patients With EGFR -Mutated or ALK -Rearranged non-Small-Cell Lung Cancers. Lung Cancer (2015) 88:108–11. doi: 10.1016/j.lungcan.2015.01.020

7. Huber RM, Hansen KH, Paz-Ares RL, West HL, Reckamp KL, Leighl NB, et al. Brigatinib in Crizotinib-Refractory ALK+ NSCLC: 2-Year Follow-Up on Systemic and Intracranial Outcomes in thePhase 2 ALTA Trial. J Thorac Oncol: Off Publ Int Assoc Study Lung Cancer (2020) 15(3). doi: 10.1016/ j.jtho.2019.11.004
8. Venur VA, Karivedu V, Ahluwalia MSSystemic Therapy for Brain Metastases. In: Handbook of Clinical Neurology. Elsevier. p. 137–53.
9. Vogelbaum MA, Brown PD, Messersmith H, Brastianos PK, Burri S, Cahill D, et al. Treatment for Brain Metastases: ASCO-SNO-ASTRO Guideline. J Clin Oncol: Off J Am Soc Clin Oncol (2022) 40(5):492–516. doi: 10.1200/JCO.21.02314
10. McTyre E, Scott J, Chinnaiyan P. Whole Brain Radiotherapy for Brain Metastasis. Surg Neurol Int (2013) 4(Suppl 4):S236–44. doi: 10.4103/2152- 7806.111301
11. Kraft J, Zindler J, Minniti G, Guckenberger M, Andratschke N. Stereotactic Radiosurgery for Multiple Brain Metastases. Curr Treat Options Neurol (2019) 21(2):6. doi: 10.1007/s11940-019-0548-3
12. Kraft J, Mayinger M, Willmann J, Brown M, Tanadini-Lang S, Wilke L, et al. Management of Multiple Brain Metastases: A Patterns of Care Survey Within the German Society for Radiation Oncology. J Neuro Oncol (2021) 152 (2):395–404. doi: 10.1007/s11060-021-03714-w
13. Badiyan SN, Regine WF, Mehta M. Stereotactic Radiosurgery for Treatment of Brain Metastases. J Oncol Pract / Am Soc Clin Oncol (2016) 12(8):703–12. doi: 10.1200/JOP.2016.012922
14. Walker AJ, Ruzevick J, Malayeri AA, Rigamonti D, Lim M, Redmond KJ, et al. Postradiation Imaging Changes in the CNS: How can We Differentiate Between Treatment Effect and Disease Progression? Future Oncol (2014) 10 (7):1277–97. doi: 10.2217/fon.13.271
15. Sneed PK, Mendez J, Vemer-van den Hoek JGM, Seymour ZA, Ma L, Molinaro AM, et al. Adverse Radiation Effect After Stereotactic Radiosurgery for Brain Metastases: Incidence, Time Course, and Risk Factors. J Neurosurg (2015) 123(2):373–86. doi: 10.3171/2014.10.JNS141610
16. Gerosa M, Nicolato A, Foroni R, Zanotti B, Tomazzoli L, Miscusi M, et al. Gamma Knife Radiosurgery for Brain Metastases: A Primary Therapeutic Option. J Neurosurg (2002) 97:515–24. doi: 10.3171/jns.2002.97.supplement_5.0515
17. Lawrence YR, Allen Li X, el Naqa I, Hahn CA, Marks LB, Merchant TE, et al. Radiation Dose–Volume Effects in the Brain. Int J Radiat Oncol Biol Phys (2010) 76:S20–7. doi: 10.1016/j.ijrobp.2009.02.091
18. Minniti G, D'Angelillo RM, Scaringi C, Trodella LE, Clarke E,

Matteucci P, et al. Fractionated Stereotactic Radiosurgery for Patients With Brain Metastases. J Neuro Oncol (2014) 117(2):295–301. doi: 10.1007/s11060-014- 1388-3
19. Vellayappan B, Tan CL, Yong C, Khor LK, Koh WY, Yeo TT, et al. Diagnosis and Management of Radiation Necrosis in Patients With Brain Metastases. Front Oncol (2018) 8:395. doi: 10.3389/fonc.2018.00395
20. Petrovich Z, Yu C, Giannotta SL, Oâ€TMDay S, Apuzzo MLJ. Survival and Pattern of Failure in Brain Metastasis Treated With Stereotactic Gamma Knife Radiosurgery. J Neurosurg (2002) 97(5 Suppl):499–506. doi: 10.3171/ jns.2002.97.supplement_5.0499
21. Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, van Stiphout RGPM, Granton P, et al. Radiomics: Extracting More Information From Medical Images Using Advanced Feature Analysis. Eur J Cancer (2012) 48(4):441–6. doi: 10.1016/j.ejca.2011.11.036
22. Aerts HJWL, Velazquez ER, Leijenaar RTH, Parmar C, Grossmann P, Carvalho S, et al. Decoding Tumour Phenotype by Noninvasive Imaging Using a Quantitative Radiomics Approach. Nat Commun 5 (2014) p:4006. doi: 10.1038/ncomms5006
23. Zhou M, Scott J, Chaudhury B, Hall L, Goldgof D, Yeom KW, et al. Radiomics in Brain Tumor: Image Assessment, Quantitative Feature Descriptors, and Machine-Learning Approaches. Am J Neuroradiol (2018) 39:208–16. doi: 10.3174/ajnr.a5391
24. Morin O, Chen WC, Nassiri F, Susko M, Magill ST, Vasudevan HN, et al. Integrated Models Incorporating Radiologic and Radiomic Features Predict Meningioma Grade, Local Failure, and Overall Survival. Neuro Oncol Adv (2019) 1:vdz011. doi: 10.1093/noajnl/vdz011
25. Avanzo M, Wei L, Stancanello J, Vallières M, Rao A, Morin O, et al. Machine and Deep Learning Methods for Radiomics. Med Physics (2020) 47:185–202. doi: 10.1002/mp.13678
26. Rogers W, Thulasi Seetha S, Refaee TAG, Lieverse RIY, Granzier RWY, Ibrahim A, et al. Radiomics: From Qualitative to Quantitative Imaging. Br J Radiol (2020) 93(1108):20190948. doi: 10.1259/bjr.20190948
27. Abidin AZ, Dar I, D'Souza AM, Lin EP, Wismüller A. Investigating a Quantitative Radiomics Approach for Brain Tumor Classification. In. In: Medical Imaging 2019: Biomedical Applications in Molecular, Structural, and Functional Imaging. Medical Imaging 2019: Biomedical Applications in Molecular, Structural, and Functional Imaging, SPIE (2019). p. 36–45.
28. Dong F, Li Q, Jiang B, Zhu X, Zeng Q, Huang P, et al. Differentiation of Supratentorial Single Brain Metastasis and Glioblastoma by Using Peri- Enhancing Oedema Region-Derived Radiomic Features and Multiple Classifiers. Eur Radiol (2020)

30:3015–22. doi: 10.1007/s00330-019- 06460-w
29. Huang C-Y, Lee C-C, Yang H-C, Lin C-J, Wu H-M, Chung W-Y, et al. Radiomics as Prognostic Factor in Brain Metastases Treated With Gamma Knife Radiosurgery. J Neuro Oncol (2020) 146:439–49. doi: 10.1007/s11060- 019-03343-4
30. Mouraviev A, Detsky J, Sahgal A, Ruschin M, Lee YK, Karam I, et al. Use of Radiomics for the Prediction of Local Control of Brain Metastases After Stereotactic Radiosurgery. Neuro-oncology (2020) 22:797–805. doi: 10.1093/ neuonc/noaa007
31. 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. Eur Radiol (2018) 28(11):4514–23. doi: 10.1007/s00330-018-5463-6
32. Kniep HC, Madesta F, Schneider T, Hanning U, Schönfeld MH, Schön G, et al. Radiomics of Brain MRI: Utility in Prediction of Metastatic Tumor Type. Radiology (2019) 28:4514–23. doi: 10.1148/radiol.2018180946
33. Bhatia A, Birger M, Veeraraghavan H, Um H, Tixier F, McKenney AS, et al. MRI Radiomic Features are Associated With Survival in Melanoma Brain Metastases Treated With Immune Checkpoint Inhibitors. Neuro-oncology (2019) 21(12):1578–86. doi: 10.1093/neuonc/noz141
34. Della Seta M, Collettini F, Chapiro J, Angelidis A, Engeling F, Hamm B, et al. A 3D Quantitative Imaging Biomarker in Pre-Treatment MRI Predicts Overall Survival After Stereotactic Radiation Therapy of Patients With a Singular Brain Metastasis. Acta Radiol (2019) 60(11):1496–503. doi: 10.1177/ 0284185119831692
35. Cho J, Kim YJ, Sunwoo L, Lee GP, Nguyen TQ, Cho SJ , et al. Deep Learning- Based Computer-Aided Detection System for Automated Treatment Response Assessment of Brain Metastases on 3D MRI. Front Oncol (2021) 11:739639. doi: 10.3389/fonc.2021.739639
36. Parekh VS, Jacobs MA. Deep Learning and Radiomics in Precision Medicine. Expert Rev Precis Med Drug Dev (2019) 4(2):59–72. doi: 10.1080/ 23808993.2019.1585805
37. Tustison NJ, Avants BB, Cook PA, Zheng Y, Egan A, Yushkevich PA, et al. N4ITK: Improved N3 Bias Correction. IEEE Trans Med Imaging (2010) 29 (6):1310–20. doi: 10.1109/TMI.2010.2046908
38. Juntu J, Sijbers J, Dyck D, Gielen J. Bias Field Correction for MRI Images. Adv Soft Computing (2005), 543–51. doi: 10.1007/3-540-32390-2_64
39. Um H, Tixier F, Bermudez D, Deasy JO, Young RJ, Veeraraghavan H. Impact of Image Preprocessing on the

Scanner Dependence of Multi- Parametric MRI Radiomic Features and Covariate Shift in Multi- Institutional Glioblastoma Datasets. Phys Med Biol (2019) 64(16):165011. doi: 10.1088/1361-6560/ab2f44
40. Moradmand H, Aghamiri SMR, Ghaderi R. Impact of Image Preprocessing Methods on Reproducibility of Radiomic Features in Multimodal Magnetic Resonance Imaging in Glioblastoma. J Appl Clin Med Phys / Am Coll Med Phys (2020) 21(1):179–90. doi: 10.1002/acm2.12795
41. Masoudi S, Harmon SA, Mehralivand S, Walker SM, Raviprakash H, Bagci U, et al. Quick Guide on Radiology Image Pre-Processing for Deep Learning Applications in Prostate Cancer Research. J Med Imaging (Bellingham Wash) (2021) 8(1):010901. doi: 10.1117/1.JMI.8.1.010901
42. Duron L, Balvay D, Vande Perre S, Bouchouicha A, Savatovsky J, Sadik J-C, et al. Gray-Level Discretization Impacts Reproducible MRI Radiomics Texture Features. PLoS One (2019) 14(3):e0213459. doi: 10.1371/journal. pone.0213459
43. CarréA, Klausner G, Edjlali M, Lerousseau M, Briend-Diop J, Sun R, et al. Standardization of Brain MR Images Across Machines and Protocols: Bridging the Gap for MRI-Based Radiomics. Sci Rep (2020) 10(1):12340. doi: 10.1038/s41598-020-69298-z
44. Zwanenburg A, Vallières M, Abdalah MA, Aerts HJWL, Andrearczyk V, Apte A, et al. The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-Based Phenotyping. Radiology (2020) 295:328–38. doi: 10.1148/radiol.2020191145
45. Radiomic Features — Pyradiomics V3.0.1.Post9+Gdfe2c14 Documentation (2019). Available at: https://pyradiomics.readthedocs.io/en/latest/features. html (Accessed 21 October 2021).
46. Chatterjee A, Vallieres M, Dohan A, Levesque IR, Ueno Y, Saif S, et al.) Creating Robust Predictive Radiomic Models for Data From Independent Institutions Using Normalization. IEEE Trans Radiat Plasma Med Sci (2019) 3:210–5. doi: 10.1109/trpms.2019.2893860
47. Chollet F. Xception: Deep Learning With Depthwise Separable Convolutions. Proc IEEE Conf Comput Vision Pattern Recognition (2017), 1251–8. doi: 10.1109/CVPR.2017.195
48. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. arXiv [cs.LG] (2014). Available at: http://arxiv.org/abs/1412.6980.
49. Youden WJ. Index for Rating Diagnostic Tests. Cancer (1950) 3(1):32–5. doi: 10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3
50. Alvarez-Breckenridge C, Remon J, Piña Y, Nieblas-Bedolla E, Forsyth P, Hendriks L, et al. Emerging Systemic Treatment

Perspectives on Brain Metastases: Moving Toward a Better Outlook for Patients. Am Soc Clin Oncol Educ Book Am Soc Clin Oncol Annu Meeting (2022) 42:1–19. doi: 10.1200/EDBK_352320
51. Zhang Z, Yang J, Ho A, Jiang W, Logan J, Wang X, et al. A Predictive Model for Distinguishing Radiation Necrosis From Tumour Progression After Gamma Knife Radiosurgery Based on Radiomic Features From MR Images. Eur Radiol (2018) 28:2255–63. doi: 10.1007/s00330-017-5154-8
52. Peng L, Parekh V, Huang P, Lin DD, Sheikh K, Baker B, et al. Distinguishing True Progression From Radionecrosis After Stereotactic Radiation Therapy for Brain Metastases With Machine Learning and Radiomics. Int J Radiat Oncol Biol Phys (2018) 102:1236–43. doi: 10.1016/j.ijrobp.2018.05.041
53. Park YW, Choi D, Park JE, Ahn SS, Kim H, Chang JH, et al. Differentiation of Recurrent Glioblastoma From Radiation Necrosis Using Diffusion Radiomics With Machine Learning Model Development and External Validation. Sci Rep (2021) 11:2913. doi: 10.1038/s41598-021-82467-y
54. Salvestrini V, Greco C, Guerini AE, Longo S, Nardone V, Boldrini L, et al. The Role of Feature-Based Radiomics for Predicting Response and Radiation Injury After Stereotactic Radiation Therapy for Brain Metastases: A Critical Review by the Young Group of the Italian Association of Radiotherapy and Clinical Oncology (yAIRO). Trans Oncol (2022) 15:101275. doi: 10.1016/ j.tranon.2021.101275
55. Lin X, DeAngelis LM. Treatment of Brain Metastases. J Clin Oncol: Off J Am Soc Clin Oncol (2015) 33(30):3475–84. doi: 10.1200/JCO.2015.60.9503
56. Liang B, Yan H, Tian Y, Chen X, Yan L, Zhang T, et al. Dosiomics: Extracting 3d Spatial Features From Dose Distribution to Predict Incidence of Radiation Pneumonitis. Front Oncol (2019) 9:269. doi: 10.3389/fonc.2019.00269

PART 2:

APPLICATIONS OF DEEP LEARNING IN MEDICAL IMAGING

CHAPTER 5:

BEYOND AUTOMATIC MEDICAL IMAGE SEGMENTATION – THE SPECTRUM BETWEEN FULLY MANUAL AND FULLY AUTOMATIC DELINEATION

Authors: Michael J Trimpl, Sergey Primakov, Philippe Lambin, Eleanor P J Stride, Katherine A Vallis, Mark J Gooding

Adapted from:
Michael J Trimpl, Sergey Primakov, Philippe Lambin, Eleanor P J Stride, Katherine A Vallis, Mark J Gooding. Beyond automatic medical image segmentation—the spectrum between fully manual and fully automatic delineation. Phys. Med. Biol. 67 (2022) 12TR01
doi:
https://doi.org/10.1088/1361-6560/ac6d9c

Acsess link:
https://iopscience.iop.org/article/10.1088/1361-6560/ac6d9c/meta

Abstract

Semi-automatic and fully automatic contouring tools have emerged as an alternative to fully manual segmentation to reduce time spent contouring and to increase contour quality and consistency. Particularly, fully automatic segmentation has seen exceptional improvements through the use of deep learning in recent years. These fully automatic methods may not require user interactions, but the resulting contours are often not suitable to be used in clinical practice without a review by the clinician. Furthermore, they need large amounts of labelled data to be available for training. This review presents alternatives to manual or fully automatic segmentation methods along the spectrum of variable user interactivity and data availability. The challenge lies to determine how much user interaction is necessary and how this user interaction can be used most effectively. While deep learning is already widely used for fully automatic tools, interactive methods are just at the starting point to be transformed by it. Interaction between clinician and machine, via artificial intelligence, can go both ways and this review will present the avenues that are being pursued to improve medical image segmentation.

1. Introduction

Image segmentation is an integral part of many medical tasks. For instance in radiotherapy, image segmentation is essential for treatment planning (Ramkumar et al 2016), enabling the radiation dose delivered to different regions to be calculated and so minimise damage to healthy tissue. More generally, segmentation of anatomical structures allows for detailed volumetric analysis to facilitate various diagnostic and clinical decision-making processes (van Timmeren et al 2020).

Segmentation can be performed using either manual, semi-automatic or fully automatic methods. Manual contouring gives the clinician full control over the process, but it can be tedious and time-consuming work. Additionally, manual contouring is prone to inter- and intra-observer variability (Sharp et al 2014). In contrast, fully automatic segmentation methods do not require any human input and have the potential to be very fast and highly reproducible and ultimately, to reduce the workload for clinicians (Primakov et al 2022). Advances in machine learning (ML) have greatly improved fully automatic approaches (Hamidian et al 2017, Bakator and

Radosav 2018, Hosny et al 2018, Jarrett et al 2019, Renard et al 2020) and ML-based methods are now being implemented in clinical systems (Lustberg et al 2018). Notably, contours generated using fully automatic ML tools have been shown to be indistinguishable from manual contours (Gooding et al 2018, Liu et al 2019, Primakov et al 2022).

Fully automatic methods still face several problems, however. In particular, a large and well-curated labelled training set is required for most ML approaches. Such data sets are not yet available for many regions-of-interest/anatomical structures. Furthermore, the variability of some structures, such as tumours, makes it very difficult to build a sufficiently representative dataset (Tian et al 2021). Technical differences between scanners and reconstruction protocols that are used for image acquisition pose another problem for fully automated ML methods, further increasing the need for large and diverse training data sets (Zhao et al 2014, Mackin et al 2015). The complexity of ML and particularly deep neural networks can also make it difficult to determine the uncertainty associated with segmentation boundaries, so that expert review by clinicians is still required (Wang et al 2020a, Abdar et al 2021). While fully automated ML-based segmentation methods are very successful for some anatomical and pathological regions, but for others remedial human intervention is required, increasing both clinical workload and inter- and intra-observer variability.

Semi-automatic or interactive segmentation methods assist clinicians in cases for which fully automatic methods fail. While interactive methods still require manual input, they can reduce the extent of user interaction by making predictions based on adjustments made by previous users. This gives more control over the contour outcome than with fully automated workflows, while still benefiting from semi-automation in achieving satisfactory segmentation. When compared to fully manual segmentation, interactive techniques have been shown to improve consistency and repeatability and reduce the time spent on contouring (Olabarriaga and Smeulders 2001, Wang et al 2018, Wang et al 2019b, Sakinis et al 2019). The types of user interaction can vary from a few clicks or stylus strokes to produce a contour on a single slice, to drawing selected contours in 2D to produce a 3D contour on a medical image. User interaction can be used to create contours from scratch or can be used to refine the structures of automatic methods. Interest in fully automatic deep learning methods has increased rapidly in recent years, whereas interactive methods have received comparatively little attention, as illustrated in figure 1.

Consequently, there is considerable opportunity for progress in this area.

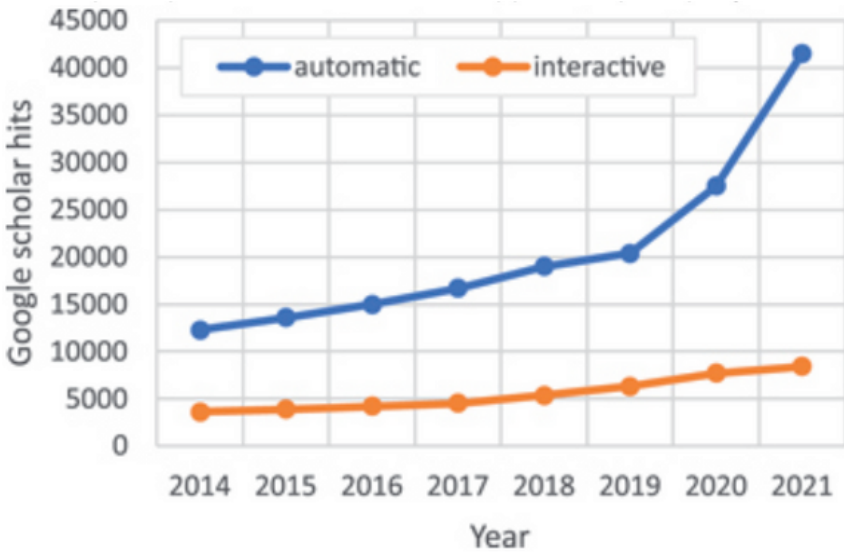


Figure 1. Google scholar search hits for keywords 'segmentation' 'deep learning' 'medical imaging' and additionally 'automatic' (blue) or 'interactive' (orange).

The aim of this review is to describe and evaluate current segmentation methods on a spectrum from manual to fully automatic tools, as shown in figure 2. The review starts with a synopsis of state-of-the-art fully automatic deep learning methods for segmentation, corresponding to the far-right side of the spectrum. Moving to the left, few-shot and transfer learning approaches can be trained with small quantities of annotated data. This distinguishes them from fully supervised methods that require large data sets that are difficult to obtain due to costly annotation time and data privacy regulations. Closer to manual editing, interactive methods provide some of the benefits of automation while retaining user control over the contour output. The review ends with a discussion of emerging techniques such as guiding user interaction through feedback provided by ML to achieve full interactivity.

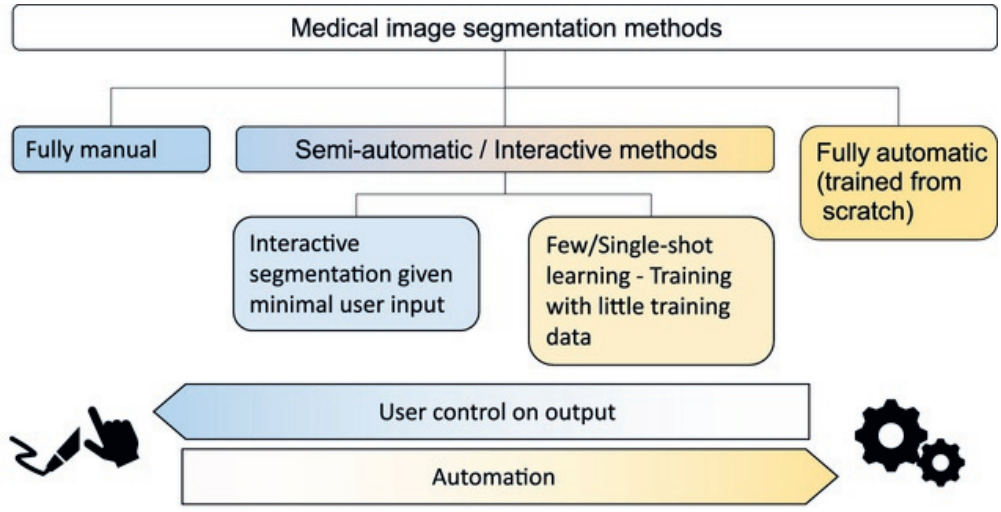


Figure 2. Medical image segmentation methods can be arranged on a spectrum from manual to fully automatic tools. Semi-automatic and interactive methods can be found between the extremes.

2. Deep learning in medical image segmentation

Deep learning has become established as the method of choice for fully automated contouring. The process from creation to application of an automatic deep learning contouring tool is shown in figure 3. This section will give a brief overview of the development of deep learning methods for medical imaging, and the challenges that remain for example in interactive segmentation. A wider introduction to deep learning applied to medical imaging can be found in (Erickson et al 2017, Litjens et al 2017, Shen, Wu and Suk 2017, Suzuki 2017, Anwar et al 2018, Guo et al 2019, Kim et al 2019, Willemink et al 2020, Wang et al 2020b, Seo et al 2020, Tajbakhsh et al 2020).

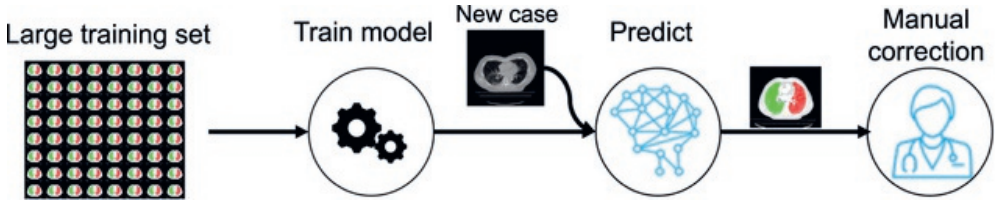


Figure 3. Fully automatic contouring using deep learning. A large training set is required to train the deep learning model. This model can then be used to predict the contours of new cases. The predicted segmentation can then be manually corrected by a clinician.

Deep learning, as a sub-field of ML, has led to many breakthroughs in computer vision tasks. Fundamentally, it uses neural networks to extract features from the provided input data and map these to an output. A neural network consists of information processing units called neurons, that are connected to each other to form the neural network, as shown in figure 4(a). If many layers of neurons are stacked, it is called deep learning. The recent success of deep learning has only been made possible by the availability of increased computing power that can handle the computationally expensive task of training deep neural networks (Schmidhuber 2015).

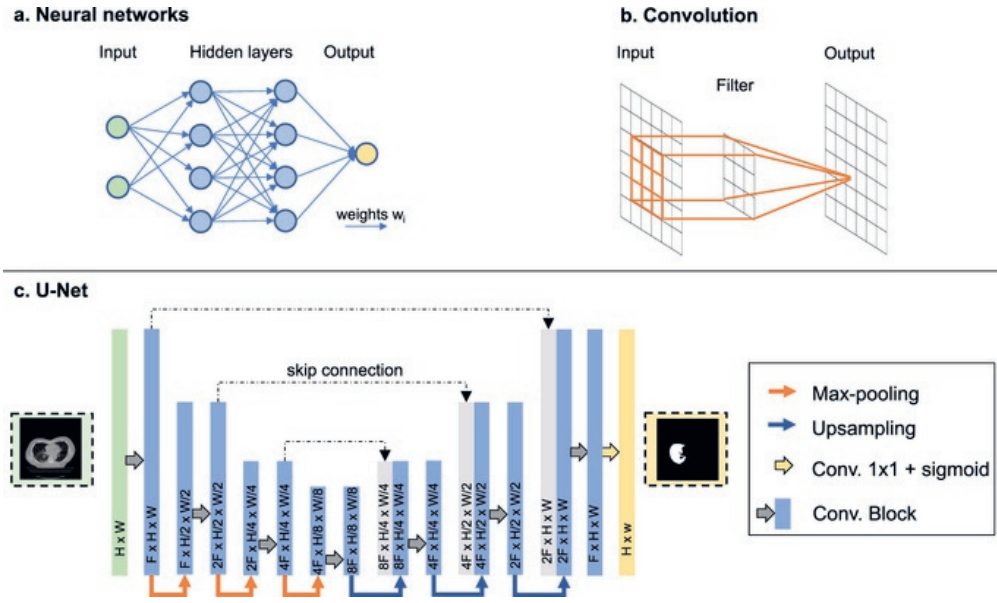


Figure 4. (a). Schematic representation of a neural network. Neurons are connected to each other by weights WI . Multiple layers form a neural net. (b). Convolutional operation. A filter, also called a kernel, is applied to the input layer. The same filter is applied across the entire image to produce an output, called a feature map (c). U-Net. The U-Net is a convolutional neural network architecture that is well suited to medical image segmentation. It consists of a contracting and expanding path with neurons connected by skip connections. The layers in the

neural net vary in their number of feature maps F and spatial resolution (height H and width W).

For computer vision tasks, including medical image segmentation, convolutional neural networks (CNN) are the most successful network architectures. CNNs are inspired by the hierarchical receptive field model of the visual cortex of the human brain (Hubel and Wiesel 1959). The key features of CNNs are convolutional layers that apply a filter to the input to extract features. The appearance of an object is recognized independent of its location in an image. Thus, detection can be performed using convolution across the image. A given neuron gets a weighted input from the units in the previous layer within a small receptive field. In deep learning, multiple layers are stacked to achieve an increasingly wide receptive field. By sharing the weights of the feature mapping in different positions for each layer, the number of parameters can be decreased compared to other types of neural networks. Illustration of a convolutional operation is shown in figure 4(b).

One of the first implementations of CNNs for image segmentation was a fully convolutional network (FCN) (Long et al 2015). This enabled the use of non-fixed input sizes, by using exclusively convolutional layers, as well as the output of a spatial segmentation map. Skip connections were used to merge upsampled feature maps from the final layers with feature maps of earlier layers. The disadvantage of this type of FCN is that the upsampling is crude and not sensitive to the details of the image. The resulting segmentation is therefore of low resolution. Additionally, in FCN, pixels are classified without fully considering spatial consistency between them.

Shortcomings of FCNs were addressed by U-Net architecture (Ronneberger et al 2015), which has since been widely used in medical image segmentation. In principle, the U-Net architecture consists of a contracting path to capture the spatial context and an expanding path for localization (figure 4(c)). Both pathways are connected by skip connections. Skip connections provide an alternative path for the gradient, which helps solve the vanishing gradient problem often faced in deep neural networks (Drozdzal et al 2016). A vanishing gradient during backpropagation can prevent the neural network from updating the weights successfully during training. Furthermore, skip connections allow the U-Net to combine high-level and low-level image information and localize these (Drozdzal et al 2016). The U-net architecture was first applied to 2D microscope images. Many variants of U-Net architecture have

emerged to improve performance of specific tasks. For example, for 3D volumetric images, the 3D U-Net and V-Net architectures (Çiçek et al 2016, Milletari et al 2016) replace 2D convolutional blocks with 3D convolutions. This enables information from neighbouring slices to be used in generation of a contour on a particular slice, which is often critical in medical applications. Other modified neural networks include the Res-UNet (He et al 2016, Wang et al 2017a, Alom et al 2018) which includes so called residual layers to mitigate the vanishing gradient problem, Attention U-Net (Oktay et al 2018, Schlemper et al 2019) which replaces skip connections with self-learned attention gates to determine which image regions matter most, or the hybrid densely connected U-Net (Li et al 2018) which fuses features from a 2D and 3D U-Net to combine intra-slice representations and inter-slice features, respectively.

The above are just a few examples of how deep learning has been applied to fully automatic segmentation in medical imaging. These methods still face challenges that arise from working with deep learning, such as integrating interactivity into the workflow or learning on small datasets. One option to overcome the need for large quantities of data is to use alternative deep learning frameworks, such as few-shot learning or transfer learning. These will be discussed in more detail in the next section.

3. Few-shot learning, transfer learning and fine-tuning

Currently, the majority of the research on automatic segmentation for medical imaging utilizes fully-supervised ML models (Isensee et al 2018, Li et al 2018, Zhou et al 2018, Wei et al 2020). There is, however, an increasing number of articles proposing weakly-supervised methods, including low-, few-, one- and zero-shot learning approaches. An in-depth review of these methods can be found in Kadam and Vaidya (2018), Wang et al (2020d). Fully supervised DL methods require large and representative datasets to maximize their performance. However, in the medical imaging field, acquiring such datasets and corresponding segmentation labels can be problematic, due to restrictive personal data regulations, heterogenous clinical protocols and labelling complexity. In cases when there are only a few training images available, fully supervised methods struggle to generate correct predictions. To tackle this problem few-shot/low-shot learning methods have been proposed. Transfer learning and fine-tuning frameworks can also help when dealing with small annotated datasets of medical images (Pan and

Yang 2010, Karimi et al 2020, Karimi et al 2021).

3.1. Few-shot learning methods in medical imaging segmentation

The name few-shot/low-shot usually refers to methods that use a small number of labelled images, called the 'support set', to assist in solving the segmentation task. Training the model still requires a large training set from which the support set is sampled to simulate the few-shot problem (Shaban et al 2017). The goal of training is not to know what a specific structure is, as is the case in the fully automatic approach discussed in the previous section. Instead, the goal is to learn the similarity and difference between structures. The training set may have a variety of structures in it, expanding the pool of available data for training, even if only a few cases are available for the specific structure to be segmented.

The 'query set' contains the images that are to be segmented. A method that uses a support set with k labelled images and n semantic classes would be called a n-way k-shot learning approach (Shaban et al 2017). The term Zero-shot Learning methods refers to methods where the target class is not present in the support set (Bucher et al 2019). A few annotated cases are all that are needed to achieve the corresponding segmentation on new cases, as illustrated in figure 5.

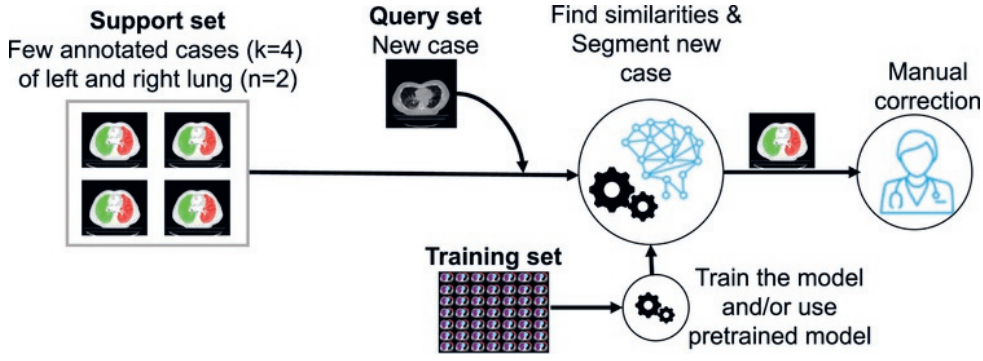


Figure 5. Few-shot learning (n-way k-shot learning). The structures in the training set can be different to the ones provided in the support set, but should be related, e.g. the training set should also be made up of anatomical structures. When presented with a few cases (k) of unseen class—the support set—the corresponding labels (n) are predicted for the new case—the query set. Manual correction may follow to refine segmentation. By applying transfer-learning, a pre-trained model can be used as an initialization of the model.

Early studies on few-shot learning were mainly focused on image classification tasks (Fei-Fei et al 2003, 2006, Santoro et al 2016, Snell et al 2017). However, this approach was soon adopted in natural image computer vision for segmentation tasks (Dong and Xing 2018, Zhang et al 2018) due to its reduced demand for supporting data. To perform one shot semantic segmentation, Shaban et al proposed using a model with two branches: conditioning and segmentation (Shaban et al 2017). The conditioning branch is used to extract the parameters from the masked support set image and the segmentation branch extracts the features from the query set image. The final segmentation mask is then obtained by performing pixel level logistic regression on the query set features using parameters from the conditioning branch (Shaban et al 2017). This approach has since been improved in several ways. Instead of using separate feature extractors for the support and query set, it has been proposed to use shared network weights to extract the features from both sets, reducing the number of parameters in the model (Wang et al 2019c). Following the feature extraction, masked averaged pooling has been shown to better extract foreground and background information from the support set. Additionally, prototype alignment regularization was introduced: When the segmentation mask was produced for the query image, this mask was used to perform the few-shot segmentation in reverse—from query to support—which allowed alignment of the prototype representations between query and support set during training (Wang et al 2019c).

In the medical imaging field, several investigators have tried to adopt few-shot learning, proposing various approaches. Mondal et al (2018) pioneered few-shot medical imaging segmentation and argued that the methods suggested by Shaban et al have limitations, due to the heterogeneity of medical images (Shaban et al 2017, Mondal et al 2018, Wang et al 2019c). Therefore, a modified 3D U-Net was proposed as a discriminator in a generative adversarial network (GAN) setting to perform few-shot infant brain MRI segmentation. This enabled the use of the unlabelled and synthetic

image patches to boost the performance of the 3D U-Net (Mondal et al 2018). As an alternative approach, the one-shot medical imaging segmentation task can be treated as a classical atlas-based segmentation problem. For this, a VoxelMorph framework was used with a GAN for additional supervision (Wang et al 2020c). The proposed method takes the atlas and target images as input and predicts the correspondence map which can be applied to transfer the segmentation label (Wang et al 2020c). More recently, Lu et al presented a one-shot anatomical structure segmentation method (Contour transformer network, CTN) incorporating user intervention (Lu et al 2021). The framework takes only one labelled image and a set of unlabelled images as input, to generate the predicted contour. To bring this solution into clinical application user corrections were incorporated to improve segmentation performance (Lu et al 2021). Training the CTN model requires only one labeled image and leverages additional unlabeled data through loss functions that measure the global shape and appearance consistency of contours. The CTN uses a pre-trained network trained on ImageNet (Deng et al 2010) as the backbone of the encoding block. Optionally, additional labeled data or user annotation on mislabeled outputs of the network can be used to fine-tune the model in order to improve segmentation performance. An example segmentation of this approach is shown in figure 6 and compared to a fully supervised approach. The one-shot learning approach can successfully segment the structures whereas the fully supervised approach fails unless a sufficiently large training set is provided.

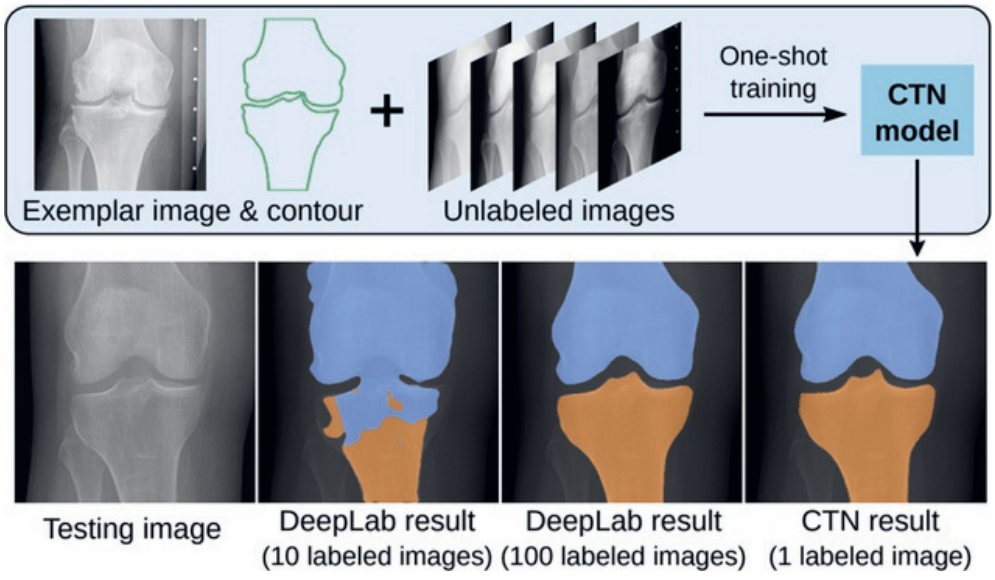


Figure 6. Comparison of few-shot learning and fully supervised methods. The few-shot learning model (Contour transformer network, CTN) learnt to segment the anatomical structure accurately from only one example. In contrast, fully supervised methods such as for example DeepLab (Chen et al 2018) fail when training with insufficient labeled images. © 2021 IEEE. Reprinted, with permission, Lu et al (2021).

3.2. Transfer learning

Deep learning has been used for medical image segmentation with a specific task in mind. Consequently, these methods are built and trained from scratch using a task-specific dataset. However, the human brain - the inspiration for modern neural networks—does not need to be retrained in this same way. By learning how to recognize specific shapes our brain can transfer this knowledge and reuse it for solving more complex tasks (Parisi et al 2019). Transfer learning is the idea that previously learned neuron interaction coefficients, i.e. image features, can be used to solve new tasks instead of starting from scratch (Pan and Yang 2010).

Transfer learning can help in tackling the limited data problem by transferring knowledge from models that were trained previously on the large datasets. This can reduce the amount of data needed to train a model for the new task. Moreover, in some settings, it can drastically reduce the model training time. Compared to few-shot learning, however, a larger training set is required.

Transfer learning can be applied in a variety of settings that depend on the problem domain, model selection, and available data. Pan et al distinguish at least 3 settings based on the availability of the labelled data: Inductive Transfer Learning, Transductive Transfer Learning, and Unsupervised Transfer Learning (Pan and Yang 2010). Inductive Transfer Learning characterizes the condition when both, target and source tasks are different but related, and source and target domains are the same. An example in medical imaging would be segmentation of a different organ in the same image modality; or using a model trained for lung tumour classification to initialize a lung tumour segmentation model. In Transductive Transfer Learning, tasks are the same, but the source and target domains are different, e.g. segmentation of the same organ on different image modalities. In Unsupervised Transfer Learning, both tasks and domains are different but related and there are no labels available in both domains during training. For example, using clustering to identify distinguishing characteristics in lung tumour patient CTs

and utilize these clusters as features in a cancer classification model (Pan and Yang 2010).

In most cases, Deep Learning models utilize an inductive transfer learning strategy, where previously trained weights could be used in two ways, i.e. without retraining as a feature extractor or being fine-tuned for a target dataset. When the pre-trained model is used as a feature extractor, the pre-trained weights are used without being updated when trained on a new task. During training on target data, only the last layer gets trained. If additional layers of the networks are retrained on the new data, it is called fine-tuning.

Fine-tuning is frequently used together with the transfer learning approach; it can include several techniques such as selective layers retraining or pruning. After transferring the weights from a pre-trained network, users can decide to re-train some of the deep layers together with a model's final fully-connected layer (Peng and Wang 2020). These deep layers may contain very specific features that are irrelevant to a new domain/problem. By retraining them on the target data these neurons will learn features specific to the target domain, therefore contributing to the performance of the model (Wang et al 2017b). Pruning is another technique to deal with irrelevant features. Network pruning methods allow redundant neurons to be omitted during inference, resulting in reduced computing costs (Luo et al 2017, Liu et al 2018).

Transfer learning with fine-tuning has been widely adopted in the medical imaging field (Raghu et al 2019, Karimi et al 2020, Wang et al 2020b, Karimi et al 2021). In many of these works, a deep learning model that was pre-trained on a publicly available imaging dataset such as ImageNet (Deng et al 2010) is used. This model is then fine-tuned on the target medical imaging dataset (Rajpurkar et al 2017, Wang et al 2017c, Gulshan et al 2019, Liu and Chan 2021). Several studies have reported that transfer learning-based pipelines were able to achieve performance comparable to a human reader (Esteva et al 2017, Ding et al 2019). For example, the pretrained architecture (InceptionV3) fine-tuned on a PET brain dataset to predict Alzheimer's disease (Ding et al 2019) outperformed the user's performance on the independent testing set (Ding et al 2019). Using a similar approach, a classification of skin lesions, equivalent to that defined by dermatologists, was achieved on a test set consisting of 1942 images (Esteva et al 2017). To compare transfer learning with a fully-supervised training model with no knowledge transfer, Van Opbroek showed that when there are few data available, transfer learning can outperform common supervised

methods and reduce classification errors by up to 60% (Van Opbroek et al 2015). Despite the success of transfer learning in medical imaging, some studies have drawn attention to the possibility of overparameterization and suggested the use of more flexible hybrid approaches to transfer learning for medical imaging tasks (Raghu et al 2019).

4. Interactive methods in medical image segmentation

In contrast to few-shot learning approaches, which move along the spectrum shown in figure 2 by seeking to reduce the need for training data for automated segmentation, interactive methods start with fully manual contouring and seek to reduce the need for manual intervention. Such interactive tools can be applied both when contouring manually and when editing a pre-existing segmented image. User interaction can be introduced in different ways for semi-automatic processing: 2D contours can be generated by using clicks (Sakinis et al 2019, Alemi Koohbanani et al 2020), scribbles (Lin et al 2016, Wang et al 2016, Boers et al 2020) or bounding boxes (Rajchl et al 2017, Wang et al 2018, Redekop and Chernyavskiy 2021); a contour may also be adjusted automatically in real-time while the clinician is drawing (Barrett and Mortensen 1997); or some 2D contours can be used to generate 3D contours of a structure (Léger et al 2018, Michael Trimpl et al 2021). The key to the success of these interactive methods is to try and find a balance between the human interaction and automation. Often this results in an iterative workflow between user interaction and processing, as illustrated in figure 7. Interactive tools have the potential to save the clinician time and effort when contouring, compared to manual annotation alone or reviewing and editing the results of fully automated segmentation.

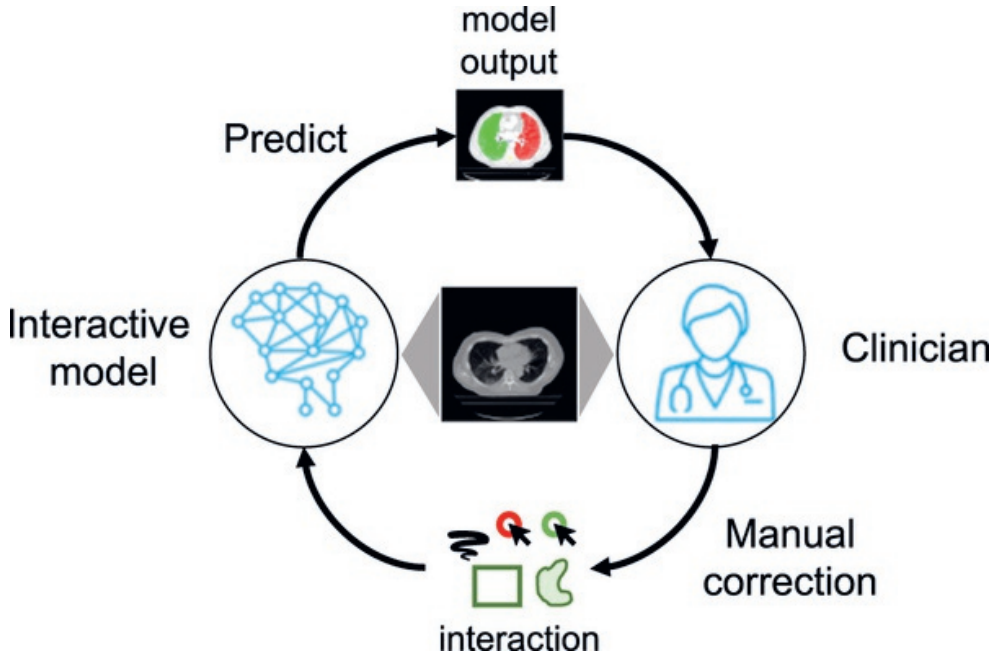


Figure 7. Interactive contouring workflow. The interactive model uses the first user interaction to make an initial prediction. Further user interactions are used to refine the prediction until the clinician is satisfied with the contour.

This section discusses interactive methods for 2D or 3D segmentation. For this, approaches that do not rely on deep learning will be introduced and contrasted with deep learning methods.

4.1. Interactive methods in 2D

In this section, different methods for generating contours in 2D are discussed. An example of how user interactions can result in a contour is illustrated in figure 8. The figure shows a series of clicks that result in the segmentation of the user indicated structure. To achieve such interactive segmentation, a wide range of interactive image segmentation approaches exist that do not rely on deep learning. Here, a selection of these methods and how they affect the development of deep learning methods will be discussed. A fuller description of non-deep learning methods may be found in Camilus and Govindan (2012), Zhao and Xie (2013).

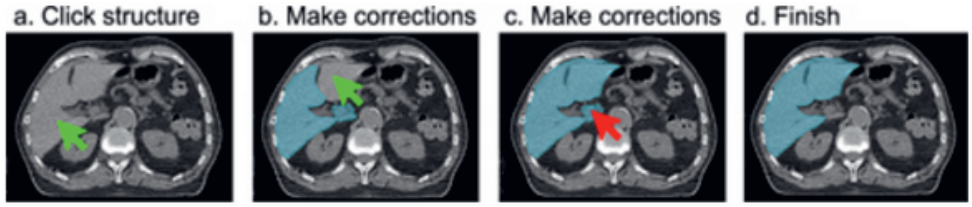


Figure 8. Example of interactive 2D segmentation via clicks. (a) With a first click the structure to be segmented is selected. This resulting initial segmentation may need further corrections which can be edited by clicking on areas that need to be (b) included (c) or excluded. Further clicks can be used until arriving at the (d) final segmentation.

4.2. Traditional methods

Graphical energy minimisation techniques have been a popular method for image segmentation (Xu and Prince 1998, Boykov and Jolly 2001, Komodakis et al 2007, Zhou et al 2016) particularly for interactive segmentation (Boykov and Jolly 2001, Freedman and Zhang 2005, Price, Morse and Cohen 2010, Isensee et al 2018). For example, GrabCut sought to maximize the separation of the foreground and background classes, as initially indicated by the user, by modelling these classes using Gaussian mixture models. (Rother et al 2004). The segmentation was then produced by applying the graph cut method from the user-provided annotation. The model and segmentation are then iteratively refined based on additional user feedback. The strength of this method is the speed of the graph cut segmentation allowing rapid feedback to the user such that the segmentation could be refined easily by providing additional annotation. On the downside, GrabCut often resulted in shrunken structures. This was prevented by adding a topological prior (Lempitsky et al 2009). Furthermore, the performance was improved further by including a Conditional Random Field (CRF) (Lempitsky et al 2009, Cheng et al 2015) to encourage segmentation homogeneity. Instead of bounding boxes, geodesic distance transforms have been introduced to ensure spatial regularization and contrast-sensitivity. This method is called GeoS (Criminisi et al 2008).

Another example of a framework for segmentation is Livewire, a contouring tool that quickly updates the contour based on user interaction. After beginning contouring, the optimal path between the starting point and the current cursor location is

calculated using a lowest cost path algorithm (Dijkstra 1959). The contour changes as the user moves the mouse. Various algorithms may be used. For example, edge detection may be implemented using a Sobel filter. In that case the lowest cost path will be that along the edges. Livewire has been extended to the Intelligent Scissor tool (Barrett and Mortensen 1997). On-the-fly training enables the contour to be applied to the specific type of edge being traced rather than just to the strongest edge in the image area. Furthermore, the Intelligent Scissors tool automatically freezes unchanging segments and inserts additional seed points to increase contouring efficiency and the accuracy of the generated contours. The cost calculation can be adjusted to fit the specific needs of a given image modality. For example, ultrasound images are still difficult to process using fully automatic segmentation due to the presence of speckle and other artefacts—which are inherent to this imaging modality (Rackham et al 2013). Livewire was adapted for ultrasound images by introducing two sets of costs: first, feature asymmetry to improve edge localization, and second, a weak shape constraint cost to improve boundary selection in the presence of missing information or artefacts (Rackham et al 2013). As a result, the fuzzy boundaries in ultrasound images can be detected by identifying structural relevance rather than intensity gradients.

Other examples for non-deep learning based methods include, level-set segmentation (Qiu et al 2013), random walks and random forest (Grady et al 2005). For example, SlicSeg uses online random forests to segment fetal MRI images (Wang et al 2016). Additionally, SlicSeg allows for contour estimation in the remaining image volume and interactivity for refinement following an adapted GraphCut method.

Open-source tool kits have been created based on these methods, for example ITK-SNAP (Yushkevich et al 2016) and trainable WEKA (Waikato Environment for Knowledge Analysis) (Arganda-Carreras et al 2017). Both provide an intuitive graphical user interface and are able to process various image modalities and formats. ITK-SNAP uses active contour methods to produce segmentation given an initial user input in 2D, 3D and multi-modality medical images. The trainable WEKA allows the user to train a model on a given contoured image set. The framework uses and compares any available classifier to perform image segmentation based on pixel classification. Such semi-automatic and interactive tool kits can help minimize the time and effort required for manual contouring task.

4.3. Deep learning methods

More recently, user interactions have been incorporated into CNNs. Unlike standard networks as discussed in section 2, extra input is provided to the model through user interaction. Rajchl et al introduced DeepCut, which uses a CRF to update the parameters of a CNN model to favour segmentation homogeneity—as has been done with GrabCut before. DeepIGeoS combines CNNs, CRFs and geodesic distance transforms (G Wang et al 2019b), bringing deep learning to the GeoS method. This approach uses two models. The first model is used to obtain an initial segmentation. The user can then interact with the result to identify misclassification. A second network is used to refine the result given the user interaction. When applied to placenta and brain tumour segmentation of fetal MRI images, DeepIGeoS reduced user time by 66% compared with segmentation using GeoS. DeepCut performs segmentation using a bounding box input by users (Rajchl et al 2017). To improve performance, Deep Extreme Cut uses extreme points (edges) of the structure as an input to the CNN (Maninis et al 2017). The CNN learns to match the segmentation to the extreme points of the object resulting in improved performance compared to bounding boxes. To further improve performance of this method, other deep neural network frameworks, such as recurrent neural networks, have been proposed (Zheng et al 2015).

Most deep learning based methods that allow for user interaction build on the popular U-Net architecture. For example, Amrhen et al proposed UI-Net for interactive segmentation (Amrehn et al 2017). In addition to allowing each image slice to be contoured, as in the standard U-Net, it also allows for user 'scribbles' to be included in the model as an input. These scribbles indicate which areas should and should not be included in the segmentation. During the segmentation process, the user can continue to provide input to achieve a precise segmentation result. This system has shown superior results compared to networks without the user input channel component when applied to liver lesion segmentation. This is attributed to consistent improvement in segmentation with each user interaction. To make better use of the provided scribbles, Lin et al proposed ScribbleSup (Lin et al 2016). Instead of using the scribbles directly or applying a geodesic distance transform, a graphical model propagates the information from the scribbles to the unmarked pixels, based on spatial constraints and appearance.

Open-source tool kits, similar to ITK-SNAP and trainable WEKA, also exist for deep learning approaches. For example, RIL-Contour

(Philbrick et al 2019) (Radiology Informatics Laboratory) supports medical image annotation using fully automated deep-learning, semi-automated methods, and manual methods. It also enables workflows for continual learning on newly annotated data provided by multiple annotators.

4.4. Interactive methods in 3D

Clinicians frequently need to contour multiple slices, for example in 3D image sets. Manual segmentation of 3D images on a slice-by-slice basis can be very time consuming. A workflow for segmentation of 3D images is illustrated in figure 9. Some of the methods discussed in the section above have been extended into 3D. For example, Livewire in 2D uses two points to define a curve. In 3D, a user needs to indicate one or more closed curves (contours) and the algorithm finds the corresponding minimum cost surface (Grady et al 2005). Bounding box based approaches have been extended by either using individual bounding boxes on various 2D slices or by using a bounding cuboid (Redekop and Chernyavskiy 2021).

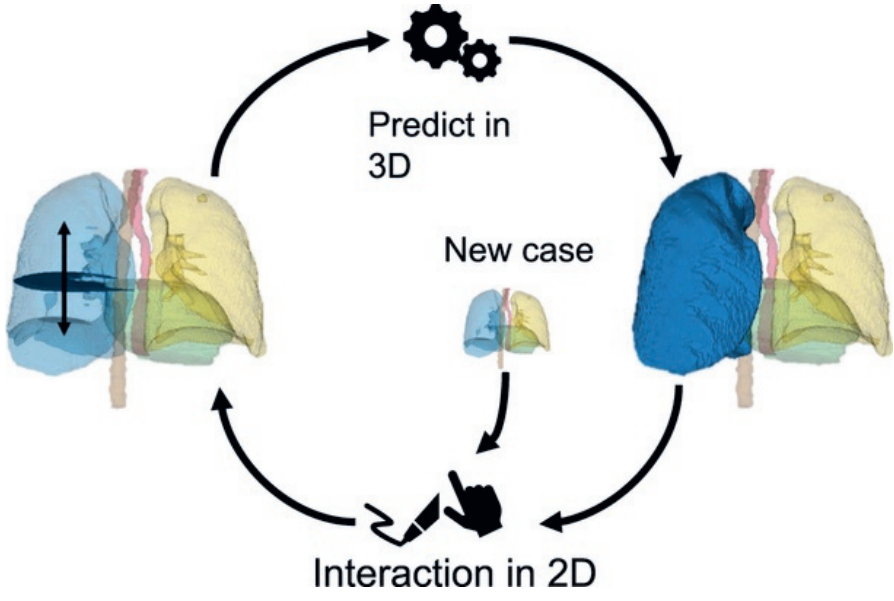


Figure 9. 3D interactive segmentation. 2D segmentation methods can be easily extended to 3D by, for example, propagating the contour through the image. The predicted contours can be reviewed and individual slices edited until a satisfactory segmentation is achieved.

4.5. Contour propagation

If direct extension of a 2D to 3D contouring method is not possible, then there are a number of alternative approaches. These include automatically initializing the segmentation on the adjacent slice via a seed input and performing a segmentation on this slice. Alternatively, the contour can be iteratively propagated throughout consecutive images of a scan. China et al introduced a volumetric segmentation approach that relies on subsequent initialization of the segmentation on adjacent image slices in ultrasound volumes. A gradient vector flow based propagation technique was used to provide an initial segmentation for the next image slice, whereas iterative random walks were used to correct the contours in subsequent steps of the algorithm (China et al 2019). Contour propagation and interpolation has also been achieved by using slice-to-slice registration of contours (Penney et al 2004). For this the deformation between two image slices is calculated and that transformation is then applied to the contour. These propagation approaches can be applied in addition to any of the 2D methods mentioned before to produce 3D segmentation based on 2D input. Alternatively, if a few slices in a 3D image have been contoured, interpolation (for example linear interpolation) between these contours, can provide an estimate for the full 3D volume of a structure.

Using CNNs, contour propagation has been applied to bladder segmentation using a single contoured image slice as the input (Léger et al 2018). A similar propagation approach has been used for multi-class image segmentation of the cardiac system (Zheng et al 2018). These methods have been shown to offer better segmentation performance than fully automatic methods. However, they are only able to segment structures included in the training set and require retraining if used for other structures. This problem is often faced in supervised deep learning, as it is highly dependent on the training set provided. One solution to this is to train CNNs on a large variety of structures simultaneously. In this way the model does not learn structure-specific features, but rather learns to predict the adjacent slice based on the context between input image and contoured slice (Michael Trimpl et al 2021).

4.6. Other deep learning methods in 3D

Sparse user annotations, by providing selected 2D contours, can be used to predict the remaining slices of a 3D structure. A 3D U-Net (Çiçek et al 2016) has been proposed, that uses multiple sparse annotations in a 3D volume. This method can learn from just a few contoured slices of a 3D scan and complete the segmentation by using on-the-fly elastic deformations for efficient data augmentation. The image slices with a user-defined contour are used to fine-tune a deep learning model to segment the remaining non-contoured image slices of the specific scan. This approach in effect applies transfer learning on a case-by-case basis (as discussed in section 3) to enable an interactive contouring approach.

Again, as for 2D methods, user control on the segmentation output through image specific fine-tuning can be increased by providing scribbles to the network. These scribbles are used to create a weighted loss function. User-provided scribbles are associated with higher confidence than the other pixels and are therefore assigned heavier weighting. Additionally, during fine-tuning pixels with low confidence in the test image are given lower weighting. Model based uncertainty has been associated with the softmax output, where low confidence corresponds to a value close to 0.5 (Wang et al 2018). This is discussed further in the next section. Using scribble-based and model uncertainty-based loss function for fine-tuning has been shown to improve interactive segmentation performance in medical imaging for several structures, including placenta, brain, fetal lungs and maternal kidneys (Wang et al 2018). As an alternative to scribbles, seed points have been used to indicate the structure to be segmented into a 2D or 3D U-Net structure (Sakinis et al 2019, Pepe et al 2020).

The interactive 2D and 3D segmentation methods discussed in this section can synergize well with fully automatic segmentation methods. Fully automatic methods can produce a quick first estimate for a contour. Subsequently, interactive methods can be used to get the final increment in accuracy that is needed for medical image segmentation. Thus, interactive methods are an excellent tool during the contour review stage.

4.7. Guiding user interaction using ML feedback

Up to this point the review has covered how user interaction can be interpreted by a computer to assist segmentation. In this section, the possibility of a computer giving feedback to the clinician is discussed. As the performance of computer aided segmentation methods improves, the human role moves from one of active agent to a more supervisory role. In this capacity, the user's role is to check and edit the segmentation where required. Even when using interactive segmentation methods, this checking process can still be time-consuming since the image and segmentation must still be loaded and displayed, and reviewing requires the user's attention and interaction (Michael J Trimpl et al 2021). If a model could estimate on its own where it is certain and uncertain about a prediction, or better yet where it is accurate or likely to be inaccurate, a clinician might not have to review a full image scan but could focus on the few critical regions of the scan (Wang et al 2020a).

To effectively design a feedback system, the measure of uncertainty obtained from the model must correlate with the accuracy of the model. More generally, transparency in how the model came to make a prediction could help clinicians in their decision-making process. It has been suggested that model uncertainty could be used to estimate areas to edit for interactive segmentation (Zheng et al 2021). For example, an Uncertainty-Guided Refinement Framework has been applied to segmentation in motion-corrupted fetal MRI (Wang et al 2020a). In this framework, the clinician is asked to edit slices with the highest segmentation uncertainty following the automatic processing step. This uncertainty estimation was shown to correlate well with mis-segmentations. By guiding the clinician's attention, contouring time was reduced by 30% when compared to using the DeepIGeoS method, while achieving similar final accuracy (Wang et al 2020a). An example of this is shown in figure 10, where model uncertainty is determined for different image slices (figure 10(a)). Based on the model uncertainty selected image slice or image regions may be prompted to the clinician for review (figure 10(b)). Here, a brief overview of different approaches for determining uncertainty in deep learning prediction is given. It should be noted that determining the uncertainty of deep learning, is still a subject of ongoing research and a more extensive discussion may be found in (Abdar et al 2021).

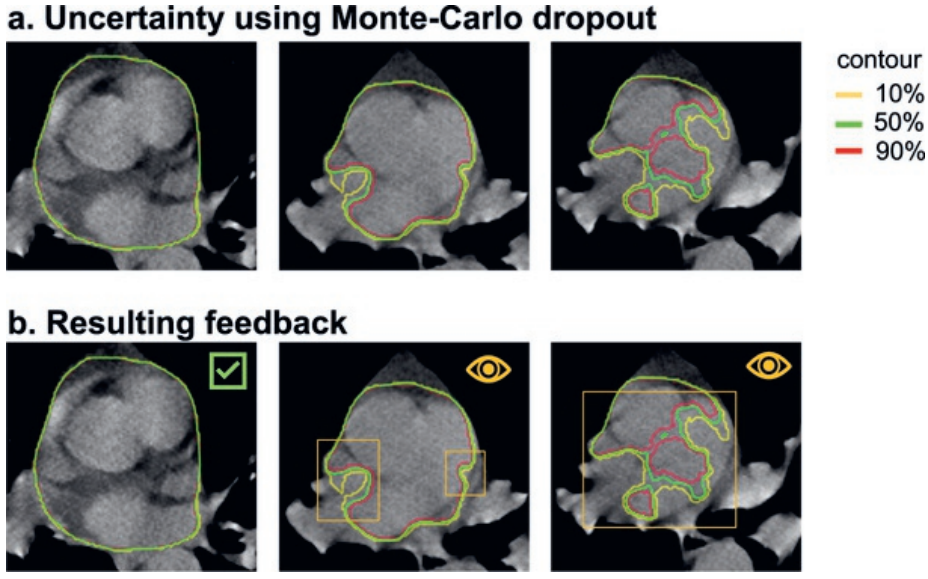


Figure 10. Guiding user interaction based on model uncertainty. (a) Obtaining uncertainty from multiple outcomes via MC dropout or ensembling. The predictions are summarized by contours that surround the area which a certain fraction of the model includes. For example, the 10% contour (yellow) indicates that 10% of outcomes included this area. 50% corresponds to the average contour. (b) Illustration of how model uncertainty could be visualized to the clinician. Image slices with little uncertainty (left image) may not need to be prompted to the clinician for review.

The most straightforward way to estimate uncertainty is to use the final activation layer of a deep learning model as a proxy for confidence in a prediction. The activation layer (e.g. softmax) may output a value between 0 and 1 to predict if a pixel belongs to the background or foreground respectively. If the output is close to 0.5, the model is not sure to which class to assign the pixel. An uncertainty can then be attributed to the prediction using least confidence, marginal confidence or entropy (Shannon 1948, Settles 2009). However, these estimates on model uncertainty suffer from a tendency, common across deep learning methods, to be overconfident in the accuracy of the model predictions.

To estimate the uncertainty of a model, Gal et al introduced dropout in their neural networks as a Bayesian approximation to uncertainty (Gal and Ghahramani 2015). Using dropout means that different connections within the neural network are randomly dropped. This results in a slightly different model each time. Originally, dropout was used during model training to

reduce overfitting and increase the robustness of the model (Srivastava et al 2014). Applying dropout at test time can be used to calculate the outcome of the resulting slightly different models. The resulting distribution of outcomes can then be used as a measure of uncertainty of the model predictions (Gal and Ghahramani 2015, Gal et al 2017, Wang et al 2019a).

Similarly, the uncertainty of a prediction can be estimated if an ensemble of deep learning models is used. Ensembling describes the method of using various models to make a prediction by consensus. Additionally, disagreement between the individual predictions can indicate a region of uncertainty (Settles 2009), illustrated in figure 10. A way to create an ensemble of models is by using convolutions in multiple groups (Wang et al 2020a). Thereby, multiple models following the same architecture can be trained in parallel. The disagreement in the prediction can then be used to obtain an uncertainty measure, which has been shown to be superior to Monte-Carlo dropout when estimating uncertainty (Wang et al 2020a). As above, determining uncertainty in a deep learning model remains a matter of ongoing research. However, focusing a clinician's attention on the most crucial areas and so guiding their user interaction can help make contouring more efficient and accurate (G Wang et al 2020a).

5. Discussion

Deep learning has become the state-of the-art approach to medical image processing and the resulting tools are being adopted into the clinical workflow. However, the clinically implemented approaches have largely been restricted to fully automatic segmentation models. Yet, as outlined in this review, there are many approaches beyond fully automatic tools that could aid clinicians in performing their work.

Fully automatic methods are excellent when presented with a clearly defined task and where a large, labelled training data set is available. In such a scenario, these methods can potentially eliminate the need for editing by clinicians. However, while the task of segmentation might appear to be clearly defined, in reality it is difficult to find a gold standard contour for many structures. Clinicians may disagree on what should and shouldn't be included in a contour; indeed if an individual clinician is asked to outline the same case twice, they will likely produce two different sets of

contours (Sharp et al 2014). A deep learning model trained by a single user or set of users will be biased towards the style of the clinician or centre providing the cases for the training set. Conversely, if the cases in the training set are representative of a wide range of institutions and clinicians, the model may average the different opinions in a way that satisfies no one's requirements. Furthermore, common deep learning methods require large amounts of data to be able to create robust models. The confidential nature of medical imaging data makes sharing and creating large datasets difficult. While there are more and more open access image datasets available (Yang et al 2017, Aerts et al 2019, Simpson et al 2019, Wee and Dekker 2019), they are restricted often to datasets that were acquired under certain conditions for a specific study. Consequently, fully automatic methods are very rigid and not always suitable for application to a general problem.

This review set out to explore possible alternatives to give users greater control over deep learning model output, whilst retaining efficiency. The first step to changing to a more user-controlled model is to provide a minimal training set for a deep learning model. The question: 'How little data can we get away with?' is essential here and this in itself is a large area of ongoing research in computer vision, covered by k-shot learning, transfer-learning and model fine tuning. These methods can produce segmentations requiring only a small number of annotated images of the target class. The performance of these methods strongly depends on the problem domain and annotations in the training/support set and the performance is typically lower than the performance of the fully supervised state of the art methods. Yet, this could give the clinician the option to—for example—define a completely new structure that a model should learn to segment. All that is needed for this is a few prior annotated cases—the support set. From these few cases an automatic contouring pipeline can be established that can then be applied on other cases. Overall, these methods take conventional deep learning and lower the barrier for model training by allowing for small data set sizes. By doing so, a support set can easily be defined by a single person that can than determine what (and how) they want a certain structure to be contoured, given them a greater amount of control on the output.

Despite the promising results of the k-shot learning application in the medical imaging domain reported in several studies (Fei-Fei et al 2003, 2006, Santoro et al 2016, Snell, Swersky and Zemel 2017, Shaban et al 2017, Dong and Xing 2018, Zhang et al 2018, Mondal,

Dolz and Desrosiers 2018, Bucher et al 2019, Wang et al 2019c, Wang et al 2020c, Lu et al 2021), challenges remain. As these methods use one or few annotated images to learn how to produce the segmentation, they are more sensitive to the variation in the medical images than fully supervised methods. To overcome this problem, some methods use unlabelled images or/and GAN's to produce more training samples artificially. Another challenge for using the k-shot learning methods for medical imaging segmentation arises from the segmentation task itself. For example, segmentation of tumours is an arduous task even for fully supervised methods as tumours are very heterogeneous. Thus, tumour segmentation is very challenging for k-shot learning methods as they cannot capture tumour heterogeneity from a small number of samples.

Interactive methods for many applications can provide a good middle ground in terms of the trade-off between automation and user control. They are designed to interpret what the user wants to segment and suggest a segmentation based on that. The interactive methods discussed in this paper highlight the diverse ways that this can be done, ranging from bounding boxes and scribbles to clicks and individual contours. Regardless of the specific interaction type, these methods try to interpret the user interaction and complete the segmentation that the clinician requires. In the case of energy minimization techniques, user interaction is transformed into something the computer understands by including the user interaction as a constraint to the minimization problem. This problem is generally well understood and therefore it is possible to introduce the user interaction directly into the equation. For deep learning this becomes more complicated. The loss function, by which the model is optimized is not directly constrained by the user interaction. The loss function compares how well the prediction matches the ground truth of the training set. The parameters in the neural network are updated based on the discrepancy. With deep learning, the loss function is effectively optimized at training time, when the user interaction on a specific patient cannot be known. It remains a challenge to express the user interaction at interaction time as a clearly defined constraint, similarly as it is done in energy optimization. To make a more effective constraint when training a model, it is necessary to better understand deep learning models themselves.

Model interpretability is an ongoing research topic. A lack of understanding of how user interaction affects a model presents a challenge to its usability. If the clinician does not know what to expect as an output from any interaction, they cannot easily control

the output. This review has briefly addressed how a model can communicate with clinician to guide its attention to the regions, user input is likely to be most needed. Such regions are where confidence in the prediction is low and consequently, the prediction is most likely to be incorrect.

While, to date, it remains challenging to determine uncertainty, there are other ways to improve understanding of a model and guide the clinician when contouring. Given the large number of parameters in a deep learning model it is hard to intuitively understand what is going on throughout a neural network and therefore, it is intrinsically difficult to interpret deep learning models (Reyes et al 2020). This understanding can be beneficial to be able to effectively use deep learning in a feedback loop, through which certain regions of a scan are brought to the attention of the clinician for review. What goes on in the model can be visualized by highlighting image-specific saliency maps - image regions that drive a model to its prediction (Selvaraju et al 2017).

Another way of making deep learning more accessible to human interpretation is by using 'attention' in neural networks (Oktay et al 2018, Schlemper et al 2019). Attention in a network with image input, refers to the regions in the image that receive a larger weighting due to their importance in making a prediction. They also give a more intuitive insight into what is going on inside the neural network. While convolutional filters and the resulting feature maps are very abstract, the deeper the network attention maps are easily interpreted as they simply highlight which regions of the body are weighted highly to make a prediction. While model uncertainty and interpretability are just beginning to be understood, they can be an avenue to a more fully interactive contouring system, where the model not only interprets user inputs but can guide future user interaction where it is most needed.

6. Concluding remarks

Research to date has focused on fully automatic segmentation methods. In practice, clinician review and editing of these outputs is needed. Interactive methods can support clinicians in these tasks. This review has outlined various methods that incorporate user interaction to make contouring for clinicians easier. In particular, the recent advances in integrating user interaction into deep learning have been highlighted. While this is often challenging, due to the often poor interpretability of deep learning models, these

models can already leverage a clinician's expert input and provide support during the contouring workflow. Moving forward, it would be desirable to create interactive deep learning tools that can learn from previous user interactions. Future research should also focus on new methods to combine artificial intelligence and clinical expertise, instead of focusing on one or the other. A clinician can point out segmentation errors to the model, but the model may also communicate areas of high and low uncertainty to the clinician. By combining the strength of artificial intelligence and clinical expertise, patient care can be improved - by elevating contour consistency and quality, and by reducing the time taken on segmentation thus freeing up clinicians to focus on direct patient care.

Acknowledgments

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No.766276. KAV acknowledges support by CRUK (Grant Number A28736).

References

1. Abdar M et al. 2021 A review of uncertainty quantification in deep learning: techniques, applications and challenges Inform. Fusion 76 243–97 Elsevier
2. Aerts H J W L et al. 2019 Data from NSCLC-radiomics The Cancer Imaging Archive
https://doi.org/10.7937/K9/TCIA.2015.PF0M9REI
3. Alemi Koohbanani N et al. 2020 NuClick: a deep learning framework for interactive segmentation of microscopic images Med. Image Anal. 65
4.
AlomMZetal.2018Recurrentresidualconvolutionalneuralnetworkbase don U-Net (R2U-Net) for medical image segmentation arXiv:1802.06955
5. AmrehnM,GaubeS,UnberathM,SchebeschF,HorzT,StrumiaM,
Steid S, Kowarschik M and Maier A 2017 UI-Net: interactive artificial neural networks for iterative image segmentation based on a user model Eurographics Workshop on Visual Computing for Biology and Medicine (2017) pp 143–47
6. AnwarSMetal.
2018Medicalimageanalysisusingconvolutionalneuralnetworks:

a review J. Med. Syst. 42 226
7. Arganda-CarrerasIetal.2017Trainablewekasegmentation: amachinelearningtool for microscopy pixel classification Bioinform. Oxford Acad. 33 2424–6
8. BakatorMandRadosavD2018Deeplearningandmedicaldiagnosis: areviewof literature Multimodal Technol. Interact. 2 47
9. Barrett W A and Mortensen E N 1997 Interactive live-wire boundary extraction' Med. Image Anal. 1 331–41 Elsevier
10. Boers T G W, Hu Y, Gibson E, Barrat D C, Bonmatti E, Krdzalic J, van der Heijden F, Hermans J J and Huisman H J 2020 Interactive 3D U-net for the segmentation of the pancreas in computed tomography scans Phys. Med. Biol. 65 065002
11. Boykov Y Y and Jolly M-P 2001 Interactive graph cuts for optimal boundary and region segmentation of objects in N-D images Int. Conf. Computer Vision (Vancouver, BC, Canada, 7–14 July 2001) (Picastaway, NJ: IEEE) pp 105–12
12. Bucher M et al. 2019 Zero-shot semantic segmentation arXiv:1906.00817 NeurIPS 2019 (accepted)
13. Camilus K S and Govindan V K 2012 A review on graph based segmentation Graph. Signal Process. 4 1
14. Chen L-C, Zhu Y, Papandreou G, Schroff F and Adam H 2018 Encoder-decoder with atrous separable convolution for semantic image segmentation Proc. of the European Conf. on Computer Vision (Cham: Springer) vol 11211 pp 833–51
15. Cheng M M, Prisacariu V A, Zheng S, Torr P H S and Rother C 2015 Densecut: densely connected crfs for realtime grabcut Comput. Graph. Forum 34 193–201
16. China D et al. 2019 Anatomical structure segmentation in ultrasound volumes using cross frame belief propagating iterative random walks IEEE J. Biomed. Health Inform. 23 1110–8 Institute of Electrical and Electronics Engineers Inc
17. Çiçek Ö et al. 2016 3D U-net: learning dense volumetric segmentation from sparse annotation MICCAI 424–32 (http://lmb.informatik.uni-freiburg.de/resources/opensource/unet.en.html) (Accessed: 12 July 2020)
18. Criminisi A, Sharp T and Blake A 2008 GeoS: geodesic image segmentation 10th European Conf. on Computer Vision Computer Vision—ECCV 2008 vol 5302 pp 99–112
19. Deng J et al. 2010 Imagenet: a large-scale hierarchical image database IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). Institute of Electrical and Electronics Engineers (IEEE) pp 248–55
20. Dijkstra E W 1959 A note on two problems in connexion with graphs Numer. Math. 1 269–71 1959 1:1. Springer

21. Ding Y et al. 2019 A deep learning model to predict a diagnosis of alzheimer disease by using 18 F-FDG PET of the brain Radiol. Radiol. 290 456–64
22. Dong N and Xing E P 2018 Few-shot semantic segmentation with prototype learning British Machine Vision Conf. http://bmvc2018.org/contents/papers/0255.pdf
23. Drozdzal M et al. 2016 The importance of skip connections in biomedical image segmentation Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). (Cham: Springer) pp 179–87 10008 LNCS
24. Erickson B J et al. 2017 Machine learning for medical imaging Radiographics 37 505–15
25. Esteva A et al. 2017 Dermatologist-level classification of skin cancer with deep neural networks Nature 542 115–8
26. Fei-Fei L, Fergus R and Perona P 2003 A Bayesian approach to unsupervised one-shot learning of object categories Proc. of the IEEE Int. Conf. on Computer Vision 2 pp 1134–41
27. Fei-Fei L, Fergus R and Perona P 2006 One-shot learning of object categories IEEE Trans. Pattern Anal. Mach. Intell. 28 594–611
28. Freedman D and Zhang T 2005 Interactive graph cut based segmentation with shape priors Proc.—2005 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition, CVPR 2005. IEEE Computer Society (San Diego, CA, 20–25 June 2005) (Picastaway, NJ: IEEE) I pp 755–62
29. Gal Y and Ghahramani Z 2015 Dropout as a bayesian approximation: representing model uncertainty in deep learning ICML 16 pp 1050–9
30. Gal Y, Islam R and Ghahramani Z 2017 Deep bayesian active learning with image data arXiv:1703.02910
31. Gooding M et al. 2018 PV-0531: multi-centre evaluation of atlas-based and deep learning contouring using a modified turing test Radiother. Oncol. 127 S282–3
32. Grady L et al. 2005 Random walks for interactive organsegmentation in two and three dimensions: implementation and validation Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). (Berlin, Heidelberg: Springer) pp 773–80 in
33. Gulshan V et al. 2019 Performance of a deep-learning algorithm versus manual grading for detecting diabetic retinopathy in india JAMA Ophthalmol. Am. Med. Assoc. 137 987–93
34. Guo Z et al. 2019 Deep learning-based image segmentation on multimodal medical imaging IEEE Trans. Radiat. Plasma Med. Sci. 3 162–9

35. Hamidian S et al. 2017 3D convolutional neural network for automatic detection of lung nodules in chest CT Proc. SPIE 10134, Medical Imaging 2017: Computer-Aided Diagnosis. SPIE vol 10134 pp 54–9
36. He K et al. 2016 Deep residual learning for image recognition Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition pp 770–8 (Accessed: 30 June 2020) (http://image-net.org/challenges/LSVRC/2015)
37. Hosny A et al. 2018 Artificial intelligence in radiology Nat. Rev. Cancer 18 500 NIH Public Access
38. Hubel D H and Wiesel T N 1959 Receptive fields of single neurones in the cat's striate cortex J. Physiol. 148 574–91 Wiley-Blackwell
39. Isensee F et al. 2018 nnU-Net: self-adapting framework for U-Net-based medical image segmentation Med. Segmentation Decathlon Challenge 2018 in
40. Jarrett D et al. 2019 Applications and limitations of machine learning in radiation oncology Br. J. Radiol. Br. Inst. Radiol. 92 20190001
41. Kadam S and Vaidya V 2018 Review and analysis of zero, one and few shot learning approaches Advances in Intellgent Systems and Computing (Cham: Springer) vol 940 pp 100–12
42. Karimi D et al. 2020 Critical assessment of transfer learning for medical image segmentation with fully convolutional neural networks Available at: (https://arxiv.org/abs/2006.00356v1) (Accessed: 2 January 2022)
43. Karimi D, Warfield S K and Gholipour A 2021 Transfer learning in medical image segmentation: New insights from analysis of the dynamics of model parameters and learned representations Artif. Intell. Med. 116 102078 Elsevier
44. Kim M et al. 2019 Deep learning in medical imaging Neurospine. 16 657–68
45. Komodakis N, Tziritas G and Paragios N 2007 Fast, approximately optimal solutions for single and dynamic MRFs Proc. of the IEEE Computer Society Conf. on Computer Vision and Pattern Recognition
46. Léger J et al. 2018 Contour propagation in CT scans with convolutional neural networks Int. Conf. on Advanced Concepts for Intelligent Vision Systems ACIVS 2018: Advanced Concepts for Intelligent Vision Systems pp 380–91 Available at: (https://openreggui.org/) (Accessed: 13 November 2019)in
47. Lempitsky V et al. 2009 Image segmentation with a bounding box prior Proc. of the IEEE Int. Conf. on Computer Vision (Kyoto, Japan, 29 September–2 October, 2009) (Picastaway, NJ: IEEE) pp 277–84

48. Li X et al. 2018 H-denseUNet: hybrid densely connected UNet for liver and tumor segmentation from CT volumes IEEE Trans. Med. Imaging 37 2663–74
49. Lin D et al. 2016 Scribblesup: scribble-supervised convolutional networks for semantic segmentation CVPR 3159–67 (http://research.microsoft.com/en-us/um/people/jifdai/downloads/scribble sup) (Accessed: 12 July 2020)
50. Litjens G et al. 2017 A survey on deep learning in medical image analysis Med. Image Anal. 42 60–88
51. Liu A et al. 2019 Applying the turing test to contouring: are machine-generated contours indistinguishable from human generated ones? Int. J. Radiat. Oncol. Biol. Phys. 105 E136 Elsevier
52. Liu A W and Chan J H 2021 An evaluation of transfer learning with CheXNet on lung opacity detection in COVID-19 and pneumonia chest radiographs 2021 13th Int. Conf. on Information Technology and Electrical Engineering (ICITEE). IEEE (Chiang Mai, Thailand, 14–15 October, 2021) (Picastway, NJ: IEEE) pp 137–42
53. Liu Z et al. 2018 Rethinking the value of network pruning arXiv:1810.05270v2 (Accessed: 8 December 2021)
54. Long J, Shelhamer E and Darrell T 2015 Fully convolutional networks for semantic segmentation IEEE Conf. on Computer Vision and Pattern Recognition. IEEE Computer Society pp 3431–40 in
55. Lu Y et al. 2021 Contour transformer network for one-shot segmentation of anatomical structures IEEE Trans. Med. Imaging 40 2672–84
56. Luo J H, Wu J and Lin W 2017 ThiNet: a filter level pruning method for deep neural network compression Proc. of the IEEE Int. Conf. on Computer Vision (Institute of Electrical and Electronics Engineers Inc.) pp 5068–76
57. Lustberg T et al. 2018 Clinical evaluation of atlas and deep learning based automatic contouring for lung cancer Radiotherapy and Oncology (Ireland: Elsevier) vol 126 pp 312–7
58. Mackin D et al. 2015 Measuring computed tomography scanner variability of radiomics features Invest. Radiol. 50 757–65
59. Maninis K K et al. 2017 Deep extreme cut: from extreme points to object segmentation Proc. of the IEEE Computer Society Conf. on Computer Vision and Pattern Recognition. IEEE Computer Society pp 616–25
60. Milletari F, Navab N and Ahmadi S A 2016 V-Net: fully convolutional neural networks for volumetric medical image segmentation Proc. 2016 4th Int. Conf. on 3D Vision, 3DV 2016 (Institute of Electrical and Electronics Engineers Inc) pp 565–71
61. Mondal A K, Dolz J and Desrosiers C 2018 Few-shot 3D multi-modal medical image segmentation using generative adversarial

learning arXiv:1810.12241v1 (Accessed: 8 December 2021)
62. Oktay O et al. 2018 Attention U-Net: learning where to look for the pancreas Medical Imaging with Deep Learning ()
63. Olabarriaga S D and Smeulders A W M 2001 Interaction in the segmentation of medical images: a survey Med. Image Anal. 5 127–42
64. Pan S J and Yang Q 2010 A survey on transfer learning IEEE Trans. Knowl. Data Eng. 22 1345–59
65. Parisi G I et al. 2019 Continual lifelong learning with neural networks: a review Neural Netw. 113 54–71
66. Peng P and Wang J 2020 How to fine-tune deep neural networks in few-shot learning? arXiv: 2012.00204v1 (Accessed: 7 January 2022)
67. Penney G P et al. 2004 Registration-based interpolation IEEE Trans. Med. Imaging 23 922–6
68. Pepe A et al. 2020 IRIS: interactive real-time feedback image segmentation with deep learning SPIE Med. Imaging
69. Philbrick K A et al. 2019 RIL-contour: a medical imaging dataset annotation tool for and with deep learning Journal of Digital Imaging (New York LLC: Springer) vol 32 pp 571–81
70. Price B L, Morse B and Cohen S 2010 Geodesic graph cut for interactive image segmentation IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. 3161–8
71. Primakov S et al. 2022 Automated detection and segmentation of non-small cell lung cancer computed tomography images
72. Qiu W et al. 2013 Three-dimensional prostate segmentation using level set with shape constraint based on rotational slices for 3D end-firing TRUS guided biopsy Med. Phys. 40
73. Rackham T M et al. 2013 Ultrasound image segmentation using feature asymmetry and shape guided live wire Med. Imaging 2013: Image Process. 8669 86690P
74. Raghu M et al. 2019 Transfusion: understanding transfer learning for medical imaging Adv. Neural Inform. Process. Syst. Neural Inform. Process. Syst. found. 32 Available at: (https://arxiv.org/abs/1902.07208v3) (Accessed: 8 December 2021)
75. Rajchl M et al. 2017 DeepCut: object segmentation from bounding box annotations using convolutional neural networks IEEE Trans. Med. Imaging 36 674–83
76. Rajpurkar P et al. 2017 CheXNet: radiologist-level pneumonia detection on chest x-rays with deep learning Available at: (https://arxiv.org/abs/1711.05225v3) (Accessed: 8 December 2021)
77. Ramkumar A et al. 2016 User interaction in semi-automatic segmentation of organs at risk: a case study in radiotherapy J. Digit. Imaging 29 264–77
78. Redekop E and Chernyavskiy A 2021 Medical image

segmentation with imperfect 3D bounding boxes Lecture Notes Comput. Sci. 13003 193–200
79. Renard F et al. 2020 Variability and reproducibility in deep learning for medical image segmentation Sci. Rep. 10 1–16
80. Reyes M et al. 2020 On the interpretability of artificial intelligence in radiology: challenges and opportunities Radiol.: Artif. Intell. 2 e190043
81. Ronneberger O, Fischer P and Brox T 2015 U-net: convolutional networks for biomedical image segmentation Med. Image Comput. Comput.-Assist. Intervention 9351 234–41
82. Rother C, Kolmogorov V and Blake A 2004 GrabCut'- Interactive foreground extraction using iterated graph cuts ACM Trans. on Graphics -- Proc. of ACM SIGGRAPH vol 2004 pp 309–14 in
83. Sakinis T et al. 2019 Interactive segmentation of medical images through fully convolutional neural networks arXiv:1903.08205
84. Santoro A et al. 2016 Meta-learning with memory-augmented neural networks Proc. of The 33rd Int. Conf. on Machine Learning. PMLR pp 1842–50 Available at: (https://proceedings.mlr.press/v48/santoro16.html) (Accessed: 8 December 2021)
85. Schlemper J et al. 2019 Attention gated networks: learning to leverage salient regions in medical images Med. Image Anal. 53 197–207
86. Schmidhuber J 2015 Deep learning in neural networks: an overview Neural Netw. 61 85–117
87. Selvaraju R R, Cogswell M, Das A, Vedantam R, Parikh D and Batra D 2017 Grad-CAM: visual explanations from deep networks via gradient-based localization 2017 IEEE International Conference on Computer Vision (ICCV) (Venice, Italy, 22–29 October, 2017) (Picastaway, NJ: IEEE) pp 618–26
88. Seo H et al. 2020 Machine learning techniques for biomedical image segmentation: an overview of technical aspects and introduction to state-of-art applications Med. Phys. 47 e148–67
89. Settles B 2009 Active learning literature survey computer sciences technical report Active learning literature survey computer sciences technical report Technical Report #1648 University of Wisconsin, Madison (https://research.cs.wisc.edu/techreports/2009/TR1648.pdf)
90. Shaban A et al. 2017 One-shot learning for semantic segmentation British Machine Vision Conf. 2017, BMVC 2017. BMVA Press
91. Shannon C E 1948 A mathematical theory of communication Bell Syst. Tech. J.
92. Sharp G et al. 2014 Vision 20/20: perspectives on automated image segmentation for radiotherapy Med. Phys. 41

93. Shen D, Wu G and Suk H I 2017 Deep learning in medical image analysis Annu. Rev. 19 221–48
94. Simpson A L et al. 2019 A large annotated medical image dataset for the development and evaluation of segmentation algorithms
95. Google Scholar
96. Snell J, Swersky K and Zemel T R 2017 Prototypical networks for few-shot learning arXiv:1703.05175
97. Srivastava N et al. 2014 Dropout: a simple way to prevent neural networks from overfitting J. Mach. Learn. Res. 15 1929–58
98. Suzuki K 2017 Overview of deep learning in medical imaging Radiol. Phys. Technol. 10 257–73
99. Tajbakhsh N et al. 2020 Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation Med. Image Anal. 63 101693
100. Tian J et al. 2021 Tumour segmentation Radiomics and Its Clinical Application. (Amsterdam: Elsevier) pp 1–18
101. Trimpl M J et al. 2021 Interactive contouring through contextual deep learning Med. Phys. 48 2951–9
102. Trimpl M J et al. 2021 PO-1164 - Clinical evaluation of an interactive deep-learning assisted contouring method for target contouring | (ESTRO) 2021, ESTRO 2021. Available at: (https://estro2021.estro.org/poster/media/po-1164-clinical-evaluation-interactive-deep- learning-assisted-contouring-method) (Accessed: 7 January 2022)
103. Van Opbroek A et al. 2015 Transfer learning improves supervised image segmentation across imaging protocols IEEE Trans. Med. Imaging 34 1018–30
104. Van Timmeren J E et al. 2020 Radiomics in medical imaging—'how-to' guide and critical reflection Insights Into Imaging (Berlin: Springer) vol 11 pp 1–16
105. Wang F et al. 2017a Residual attention network for image classification 6450–8 Available at: (http://arxiv.org/abs/1704.06904) (Accessed: 15 July 2020)
106. Wang G et al. 2016 Slic-seg: a minimally interactive segmentation of the placenta from sparse and motion-corrupted fetal MRI in multiple views Med. Image Anal. 34 137–47
107. Wang G et al. 2018 Interactive medical image segmentation using deep learning with image-specific fine tuning IEEE Trans. Med. Imaging 37 1562–73
108. Wang G et al. 2019a Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks Neurocomputing (Amsterdam: Elsevier) vol 338 pp 34–45
109. Wang G et al. 2020a Uncertainty-guided efficient interactive refinement of fetal brain segmentation from stacks of MRI

slices Lecture Notes Comput. Sci.
110. Wang G et al. 2019b DeepIGeoS: a deep interactive geodesic framework for medical image segmentation IEEE Trans. Pattern Anal. Mach. Intell. 41 1559–72
111. Wang K et al. 2017b Pay attention to features, transfer learn faster CNNs | OpenReview, British Machine Vision Conf. (BMVC) Available at: (https://openreview.net/forum?id=ryxyCeHtPB) (Accessed: 8 December 2021)
112. Wang K et al. 2019c PANet: few-shot image semantic segmentation with prototype alignment Proc. of the IEEE Int. Conf. on Computer Vision. Institute of Electrical and Electronics Engineers Inc 2019 9196–205
113. Wang R et al. 2020b Medical image segmentation using deep learning: a survey Available at: (https://arxiv.org/abs/2009.13120v3) (Accessed: 2 January 2022)
114. Wang S et al. 2020c LT-net: label transfer by learning reversible voxel-wise
correspondence for one-shot medical image segmentation Proc. of the IEEE Computer Society Conf. on Computer Vision and Pattern Recognition. IEEE Computer Society 9159–68
115. Wang X et al. 2017c Chest x-ray8: hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases Proc. 30th IEEE Conf. on Computer Vision and Pattern Recognition, CVPR 2017. Institute of Electrical and Electronics Engineers Inc., 2017-January pp 3462–71
116. Wang Y et al. 2020d Generalizing from a few examples ACM Computing Surveys (CSUR). (New York, NY, USA: ACM PUB27) vol 53
117. Wee L and Dekker A 2019 Data from head-neck-radiomics-HN1. The Cancer Imaging Archive (https://doi.org/10.7937/tcia.2019.8kap372n)
118. Wei J et al. 2020 Genetic U-Net: automatically designed deep networks for retinal vessel segmentation using a genetic algorithm IEEE Trans. Med. Imaging
119. Willemink M J et al. 2020 Preparing medical imaging data for machine learning Radiology 295 4–15
120. Xu C and Prince J L 1998 Snakes, shapes, and gradient vector flow IEEE Trans. Image Process. 7 359–69
121. Yang J et al. 2017 Data from lung CT segmentation challenge The Cancer Imaging Archive. (https://doi.org/10.7937/K9/TCIA.2017.3r3fvz08)
122. Yushkevich P A, Gao Y and Gerig G 2016 ITK-SNAP: an interactive tool for semi-automatic segmentation of multi-modality biomedical images Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and

Biology Society. Annual Int. Conf.. Annu Int Conf IEEE Eng Med Biol Soc vol 2016 pp 3342–5
123. Zhang X et al. 2018 SG-one: similarity guidance network for one-shot semantic segmentation IEEE Trans. Cybernetics 50 3855–65
124. Zhao B et al. 2014 Exploring variability in CT characterization of tumors: a preliminary phantom study Transl. Oncol. 7 88–93
125. Zhao F and Xie X 2013 An overview of interactive medical image segmentation Annals of the BMVA 2013 1–22 (http://bmva.org/annals/2013/2013-0007.pdf)
126. Zheng E et al. 2021 A continual learning framework for uncertainty-aware interactive image segmentation
127. Zheng Q et al. 2018 3D consistent robust segmentation of cardiac images by deep learning with spatial propagation IEEE Trans. Med. Imaging 37 2137–48
128. Zheng S et al. 2015 Conditional random fields as recurrent neural networks 1529–37
129. Zhou S et al. 2016 Active contour model based on local and global intensity information for medical image segmentation Neurocomputing (The Netherlands: Elsevier Science Publishers B. V. PUB568 Amsterdam) vol 186 pp 107–18
130. Zhou Z et al. 2018 UNet++: a Nested U-Net Architecture for Medical Image Segmentation vol 11045 p 3 Available at: (https://arxiv.org/pdf/1807.10165).pdf (Accessed: 16 May 2019)

CHAPTER 6:

AUTOMATED DETECTION AND SEGMENTATION OF NON-SMALL CELL LUNG CANCER COMPUTED TOMOGRAPHY IMAGES

Authors: Sergey P. Primakov, Abdalla Ibrahim, Janita E. van Timmeren, Guangyao Wu, Simon A. Keek, Manon Beuque, Renée W. Y. Granzier, Elizaveta Lavrova, Madeleine Scrivener, Sebastian Sanduleanu, Esma Kayan, Iva Halilaj, Anouk Lenaers, Jianlin Wu, René Monshouwer, Xavier Geets, Hester A. Gietema, Lizza E. L. Hendriks, Olivier Morin, Arthur Jochems, Henry C. Woodruff, Philippe Lambin.

Adapted from:
Primakov, S.P., Ibrahim, A., van Timmeren, J.E. et al. Automated detection and segmentation of non-small cell lung cancer computed tomography images. Nat Commun 13, 3423 (2022), doi: https://doi.org/10.1038/s41467-022-30841-3,

Acsess link:
https://www.nature.com/articles/s41467-022-30841-3

Abstract

Detection and segmentation of abnormalities on medical images is highly important for patient management including diagnosis, radiotherapy, response evaluation, as well as for quantitative image research. We present a fully automated pipeline for the detection and volumetric segmentation of non-small cell lung cancer (NSCLC) developed and validated on 1328 thoracic CT scans from 8 institutions. Along with quantitative performance detailed by image slice thickness, tumor size, image interpretation difficulty, and tumor location, we report an in-silico prospective clinical trial, where we show that the proposed method is faster and more reproducible compared to the experts. Moreover, we demonstrate that on average, radiologists & radiation oncologists preferred automatic segmentations in 56% of the cases. Additionally, we evaluate the prognostic power of the automatic contours by applying RECIST criteria and measuring the tumor volumes. Segmentations by our method stratified patients into low and high survival groups with higher significance compared to those methods based on manual contours.

Introduction

Lung cancer is the deadliest of all cancers afflicting both sexes, accounting for 18.4% of the total cancer deaths worldwide in 2018, almost equal to breast and colon cancers combined (1). Recent advances in treatment (immune checkpoint inhibitors, tyrosine kinase inhibitors) has significantly improved survival times for subgroups of patients. However, much work is still to be done in the field of lung cancer, especially in screening and early detection. Automated detection and segmentation would immediately impact the clinical workflow in radiotherapy, one of the most common treatment modalities for lung cancer (2). Radiotherapy uses medical imaging, especially computed tomography (CT), to obtain accurate tumor localization and electron densities for the purpose of treatment planning dose calculations(3). Accurate segmentation of the tumor and organs at risk are also essential as errors might lead to over- or under-irradiation of both the tumor and/or healthy tissue. It has been estimated that a 1 mm shift of the tumor segmentation could affect the radiotherapeutic dose calculations by up to 15% (4,5). Therefore, automated accurate segmentation can significantly reduce the time needed by clinicians to carryout treatment planning, and adaptive re-planning of treatment depending on the changes in the tumor.

Equally important are the lesion and organ at risk segmentation process for radiation oncologists for radiotherapy planning, and the measurement of lesions within the Response Evaluation Criteria in Solid Tumors (RECIST) 1.1 framework for radiologists, both laborious manual routines which impose an avoidable workload (6). Currently, such segmentations and appropriate RECIST measurements are performed manually or semi-automatically, consuming valuable time and resources, as well as being prone to inter- and intra-observer variability (7).

Another field to profit directly from automated detection and delineation of lesions is radiomics, the high-throughput mining of quantitative features from medical images and their subsequent correlation with clinical and/or biological endpoints (8,9). Radiomics has the potential to facilitate personalized medicine via diagnostic and predictive models based on phenotypic properties of the region of interest (ROI) being analyzed (10). ROI segmentation is currently considered to be one of the most time-intensive and laborious steps within the entire radiomics workflow (11).

The recent advancement of machine learning techniques, combined with improvements in the quality and archiving of medical images, have fueled intensive research in the field of artificial intelligence (AI) for medical imaging analysis (12,13). Deep learning, a branch of AI-based artificial neural networks, has been successfully applied on images to solve problems such as classification or segmentation (14,15). Several attempts have been made to adapt these methods for medical imaging problems, including tumor detection and segmentation on CT images (16,17,18,19). A major hurdle in developing fully automated software that can be applied to any CT is the heterogeneity of the datasets, especially when acquired from multiple centers (20). CT scans with different acquisition- or reconstruction parameters present lung structures differently. The methods described in the current literature usually lack a CT preprocessing module in the pipeline, and the problem of data harmonization is left to be solved by a data-driven approach, requiring large datasets representing all aspects of this inhomogeneity.

Taking into consideration these clinical and research needs for lung tumor segmentation, the implementation of automated detection software that is capable of fast and accurate delineation of NSCLC on thoracic CT scans is desirable, bordering on necessity. The applications and benefits include, but are not limited to: (1) CT-based automated screening of lung cancer; (2) Retrospective

analysis of entire databases of patients who underwent thoracic CT in daily care for research purposes; (3) Consistent and reproducible segmentations, which are important in planning and monitoring (radio)therapy, and in research; (4) Follow-up of treated primary lung cancer; (5) Automation and acceleration of certain aspects of the clinical radiotherapy workflow, making adaptive re-planning more feasible.

Automated segmentation of NSCLC tumors requires prior identification of the lesion as NSCLC. Invasive tissue biopsy is currently the clinical gold standard in identifying NSCLC. However, an accurate automated segmentation tool requires high detection accuracy. Therefore, software that can automatically segment NSCLC tumors could also be used as a detection method, decreasing the need for invasive biopsies.

In this work, we present a fully automated lung tumor detection and 3D volumetric segmentation pipeline that is capable of handling a large variety of CT acquisition and reconstruction parameters. Furthermore, we externally validate our method on three datasets, compare the volumetric prognostic factor to an existing clinical standard, compare the quantitative performance to a similar published method, and compare the preference score, speed, and reproducibility of our method to those of experts in a prospective clinical trial setting.

Results

Overall, 1328 thoracic volumetric CT scans with corresponding 3-dimensional tumor segmentations were used in order to train, test, and externally validate a fully automated method for detection and segmentation of NSCLC in standard-of-care images. Datasets 1–7 were combined and randomly divided into training and testing datasets with 999 patients and 93 patients, respectively (see Table 1). Datasets 8–10, comprising 236 patients were used for external validation of the method. A summary of the data is provided in Table 1, description of patient characteristics is provided in Supplementary Table 2.

Table 1. Description of the datasets used in this study

Ref.	#	Name	Use	Medical Center	#CT scans	#CT scans used	#CT slices with tumor (%)	#CT slices without tumor (%)*	Mean tumor volume (ml)
[40]	1	Maastro-CT-Lung-1	Training/ Testing	Open source (TCIA)	422	419	4262 (16)	22490 (84)	71.0
N/A	2	UCL-CT-Lung	Training/ Testing	Université catholique de Louvain	39	39	400 (16)	2096 (84)	53.44
N/A	3	UCSF-CT-Lung	Training/ Testing/ Clinical trial	University of California - San Francisco	101	101	689 (11)	5775 (89)	19.35
N/A	4	MUMC+ Inoperable Lung	Training/ Testing	Maastricht University Medical Center+	92	91	1247 (21)	4577 (79)	94.99
N/A	5	AZHDU Lung	Training/ Testing	Affiliated Zhongshan Hospital of Dalian University	222	222	464 (4)	9456 (96)	2.08
[41]	6	Stanford Lung	Training/ Testing	Open source (TCIA)	211	137	796 (10)	7396 (90)	22.37
[42]	7	TCIA-CT-Lung-3	Training/ Testing	Open source (TCIA)	89	83	630 (12)	4618 (88)	51.39
[43]	8	The Maastro interobserver reproducibility test	External validation	Open source (BMIA XNAT)	22	20	210 (16)	1070 (84)	88.03
N/A	9	Radboud Lung 2	External validation	Radboud University Medical Center	132	132	3493 (22)	12460 (78)	92.04
N/A	10	MUMC/Heerlen lung	External validation	MUMC/Heerlen	89	84	1120 (13)	7317 (87)	77.79
-	-	Overall training/test	-	-	1176	1092	8488 (13)	56408 (87)	49.07
-	-	Overall validation	-	-	238	236	4823 (19)	20843 (81)	88.93

*CT slices without a segmentation were considered as not containing tumor

Tumor detection and segmentation

A three-step workflow was developed and successfully implemented (Fig. 1): (i) image preprocessing, a crucial step as datasets collected for this work were obtained from different scanners with various image acquisition and reconstruction protocols (Fig. 1 suppl.). The

data inhomogeneity necessitated the harmonization of CT data in order to achieve comparable representations of the tumor region, reduce computational power requirements and image noise, and to optimize contrast; (ii) lung isolation, which allows the model to focus on the ROI and the input of the entire CT scans; (iii) automated tumor detection and segmentation, employing the convolutional neural network.

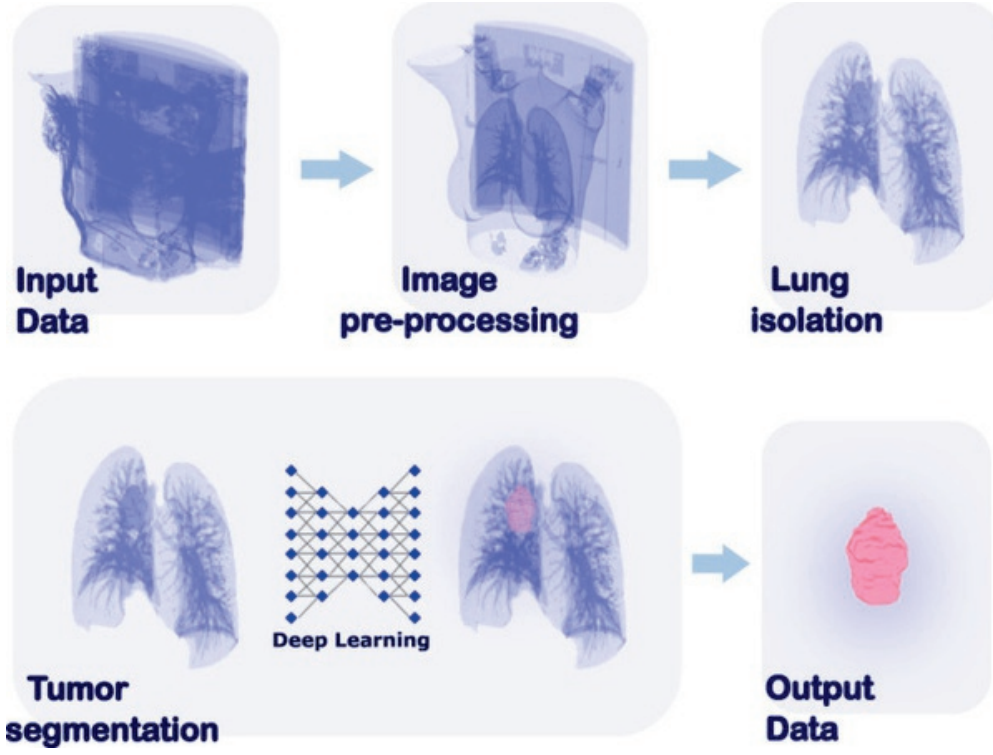


Fig. 1: Graphic representation of the major steps in the proposed workflow.

The ability of the system to detect tumors was assessed lung-wise and yielded a sensitivity of 0.97 and specificity of 0.99 in the external validation dataset and an area under the receiver operating characteristic curve (AUC) of 0.98. Confusion matrices for the detection performance can be found in supplementary materials (Fig. 2 Suppl.). The median contouring performance in the external validation dataset as assessed by the volumetric Dice similarity coefficient (DSC) was 0.82, while the 95th percentile of the Hausdorff distance (H95th) was 9.43 mm. Further metrics, associated uncertainties, as well as test

dataset results are reported in Table 2. Using dataset 8 we have established the tolerance level τ for NSCLC manual segmentation variability (τ = 1.18 mm), allowing the application of the Surface DSC for the NSCLC segmentation task.

Table 2. Overview of quantitative model performance. IQR = Interquartile range, DSC = Dice similarity coefficient, Ji = Jaccard index, H95th = 95th percentile, Hausdorff distance.

Data, # of patients	Detection performance			Segmentation performance				
	Lung-wise AUC (CI)	Specificity	Sensitivity	DSC (IQR)	Ji (IQR)	H95th, mm	Surf DSC [τ=1.18] (IQR)	APL, cm
Test, 93	0.96 (0.94-0.98)	0.97	0.96	0.85 (0.15)	0.74 (0.22)	5	0.75 (0.29)	106 (274)
External Validation, 236	0.98 (0.97-0.99)	0.99	0.97	0.82 (0.17)	0.70 (0.24)	9.43	0.63 (0.28)	306 (984)

Model performance was also separately assessed in regard to groupings of image slice-thickness, tumor size, expert-reported tumor complexity, and tumor location. The sub-cohorts were analyzed for significant differences in model performance, with the results reported in Table 3. As some of the tumors had two or more unconnected components (satellite lesions, or edges of the tumor), the Hausdorff metric can yield unreliable distances when the distance between different volume fragments are calculated. Therefore, the interquartile range (IQR) for H95th was not provided. Histograms showing the distributions of detection and segmentation results are provided in the supplementary materials (Fig. 2 suppl. and Fig. 3 supply.).

Table 3. Overview of quantitative model performance with regard to various factors. Statistical significance were calculated within the factor groups using a two-sided Mann-Whitney-Wilcoxon test with Bonferroni correction wand referred as follows: “ns” refers to the p-value in the range: 5.00e-02 < p <= 1.00e+00; * refers to the p-value in the range: 1.00e-02 < p <= 5.00e-02; ** refers to the p-value in the range: 1.00e-03 < p <= 1.00e-02; *** refers to the p-value in the range: 1.00e-04 < p <= 1.00e-03; **** refers to the p-value in the range: p <= 1.00e-04.

Factors	Test				External Validation			
	DSC (IQR)	Significance			DSC (IQR)	Significance		
Slice thickness, 0-2.5 (mm)	0.86 (0.1)	-	ns	ns	0.90 (0.08)	-	ns	ns
Slice thickness, 2.5-5 (mm)	0.88 (0.17)	ns	ns	-	0.81 (0.18)	**	ns	-
Slice thickness, >5 (mm)	0.83 (0.1)	ns	-	ns	0.86 (0.13)	**	-	ns
Complexity label, 0 (No PET needed)	0.88 (0.16)	****	-	-	0.87 (0.12)	****	-	-
Complexity label, 1 (PET needed)	0.84 (0.15)	****	-	-	0.79 (0.19)	****	-	-
Tumor size, <20 (ml)	0.84 (0.11)	-	ns	ns	0.79 (0.26)	-	ns	ns
Tumor size, 20-150 (ml)	0.86 (0.15)	ns	ns	-	0.82 (0.16)	*	ns	-
Tumor size, >150 (ml)	0.89 (0.12)	ns	-	ns	0.86 (0.15)	*	-	ns
Tumor location, parenchyma	0.82 (0.15)	-	ns	ns	0.83 (0.14)	-	ns	ns
Tumor location, mediastinum	0.87 (0.15)	ns	ns	-	0.80 (0.19)	****	ns	-
Tumor location, chest wall involvement	0.88 (0.09)	ns	-	ns	0.89 (0.08)	****	-	ns

Box plots showing DSC distributions in the sub cohort's tumor size and tumor complexity for both test and validation datasets are shown in Fig. 2. There is a clear trend toward better performance and less variability for larger and less complex tumors. More comparisons for differing slice-thickness groups, complexity classes, tumor location, and tumor sizes performed on the test and external validation dataset are provided in the supplementary materials (Figs. 4–7 suppl.).

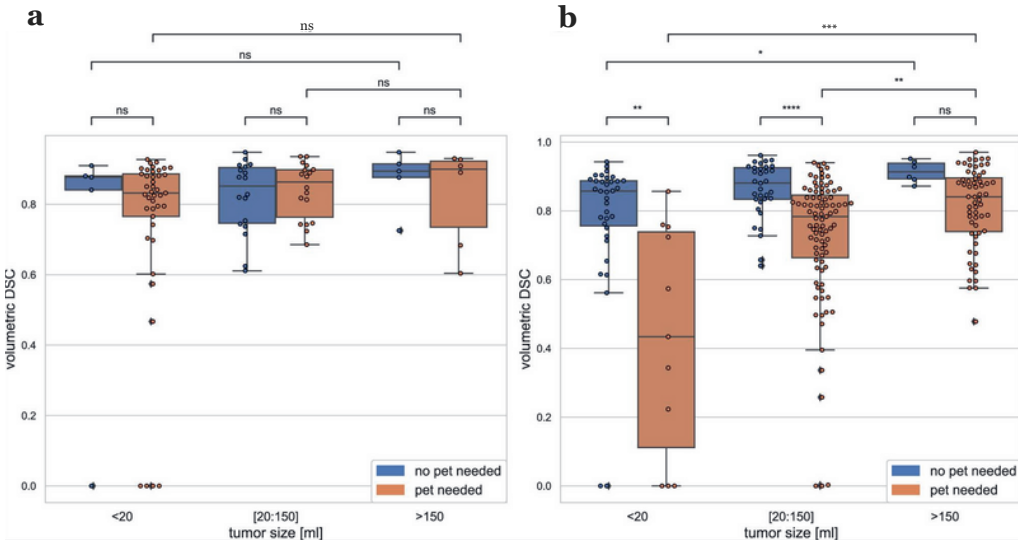


Fig. 2: Quantitative performance with regards to tumor size and complexity. Quantitative performance is measured in volumetric dice similarity coefficient (DSC). Tumor complexity is defined through the necessity of using PET to produce segmentation. Data were presented as box plots with overlaid swarm plots, where boxes are representing the interquartile range (IQR), extending from Q1 to Q3 and centered on the median value. Upper whiskers represent the highest data point that is less than Q3 + 1.5 × IQR. Lower whiskers represent the smallest data point that is greater than Q1 − 1.5 × IQR. Data points outside whiskers are considered as outliers. P values were calculated using a two-sided Mann–Whitney–Wilcoxon test with Bonferroni correction and referred as follows: “ns” on the plot refers to the p value in the range: 5.00e-02 < p ≤ 1.00e+00; *refers to the p value in the range: 1.00e-02 < p ≤ 5.00e-02; **refers to the p value in the range: 1.00e-03 < p ≤ 1.00e-02; ***refers to the p value in the range: 1.00e-04 < p ≤ 1.00e-03; ****refers to the p value in the range: p ≤ 1.00e-04. The exact p values are reported in the order from left to right and from the top to the bottom as they are displayed on the figures. Calculations provided for: a the test dataset of 93 independent NSCLC CT scans, corresponding p values are: 1.000e+00, 1.000e+00, 1.000e+00, 1.000e+00, 1.000e+00, and 1.000e+00; b the external validation dataset of 236 independent NSCLC CT scans, corresponding p values are: 4.120e-04, 4.022e-02, 8.471e-03, 2.259e-03, 1.662e-05, and 1.117e-01.

Examples of the automatically generated segmentations (from the validation set) in comparison to contours segmented by

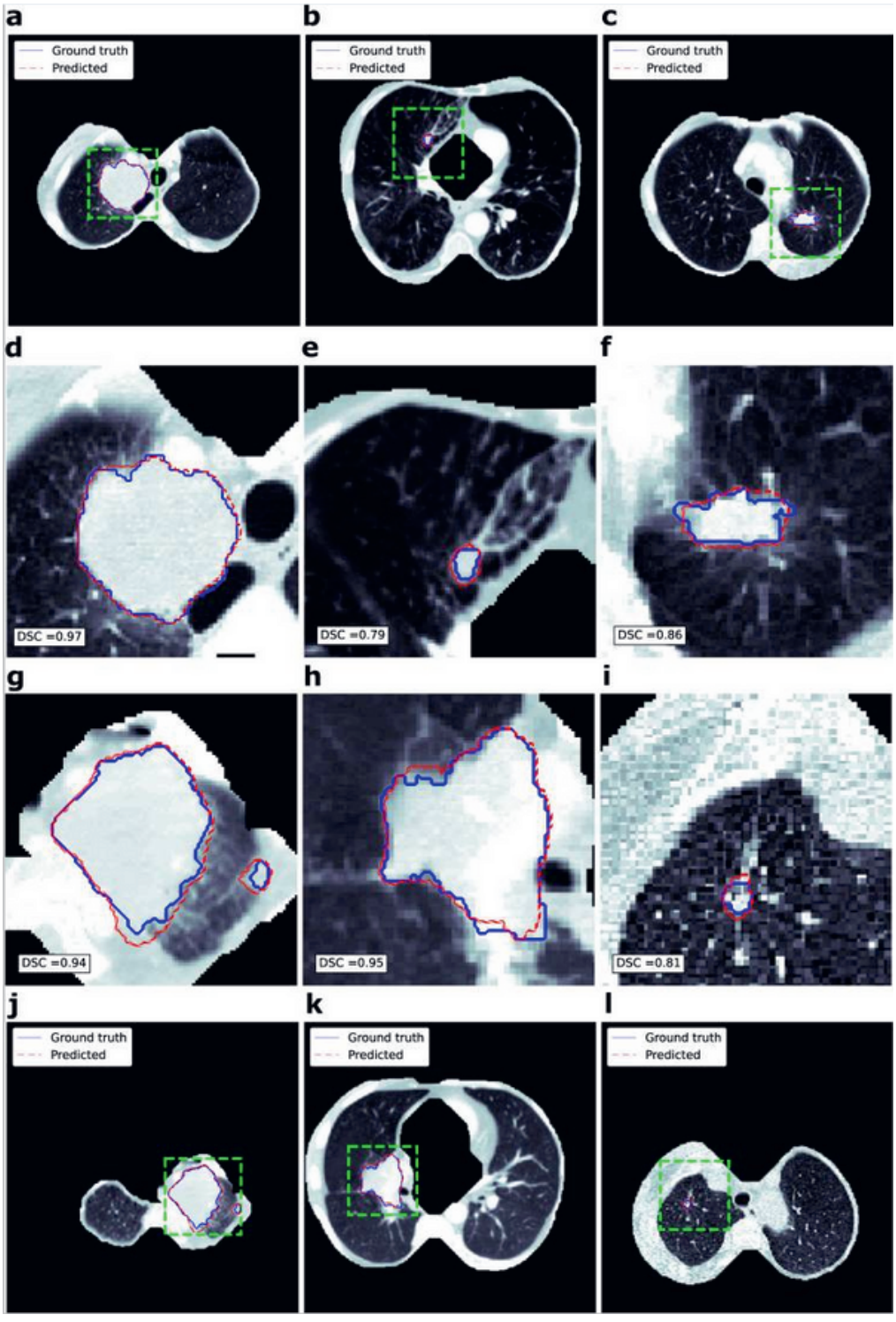

a
b
c
Ground truth
Predicted
d
e
f
DSC =0.97
DSC =0.79
DSC =0.86
g
h
i
DSC =0.94
DSC =0.95
DSC =0.81
j
k
l

Fig. 3: Visualization of segmentations. Automatically generated tumor segmentations are shown as red lines while manual segmentations are shown in blue, the green dashed box shows the area to be magnified for better visuals. d–i display magnified area for the (a–c, j–l) respectively. Corresponding 2D dice similarity coefficient is provided in the bottom left corner on the (d–i).

Comparison to a published method

A previously published external segmentation model (19) was evaluated on dataset 8 and compared to our model. The performance of the published model was evaluated using two different inputs: (i) as described in the original article (using patches of 256 × 256 pixels centered on the tumor); (ii) using the whole slice. For that dataset, our method achieved a DSC of 0.87 (IQR = 0.12), whereas the published method achieved a DSC of 0.83 (IQR = 0.16) when the cropped tumor regions were used and a DSC of 0.09 (IQR = 0.19) in the fully automated configuration (no pre-cropping). Figures for DSC, Ji, and H95th are provided in the supplementary materials (Fig. 8 suppl.).

Prognostic power of automatic segmentation

Datasets 1 and 6 were used to compare the prognostic power of measurements extracted from automatically generated and manual contours, as they had available survival data. We calculated the RECIST largest diameter and the tumor volume for both the expert and the automatic segmentation and found that for both metrics the automatically generated segmentations have more prognostic power. Statistical differences in the probability of survival for two groups separated by the median values of these measurements for automated and manual segmentations are reported in Table 4. Kaplan–Meier curves for survival split based on the tumor volume are shown in Fig. 4. KM curves for survival split based on RECIST score can be found in the supplementary materials (Fig. 9 suppl.). Additionally, we have also evaluated the difference using univariate cox analysis to report the cut-off independent results and looked at the scatter plot for tumor volumes. C-index, hazard ratio, and p values for a univariate Cox regression are reported in Table 3 in the supplementary materials. Scatter plots for tumor volume based on manual vs automated segmentations can be found in the supplementary materials (Fig. 10 supply.).

Table 4. Statistical difference between survival groups separated by the median values of RECIST and tumor volume. Statistical comparison were performed using log rank test.

Data, (# of patients)	RECIST manual segmentation (*p* value)	RECIST automatic segmentation (*p* value)	Tumor volume manual segmentation (*p* value)	Tumor volume automatic segmentation (*p* value)
1,419	0.0003	<0.0001	0.0017	<0.0001
6,137	0.0038	0.0031	0.031	0.013

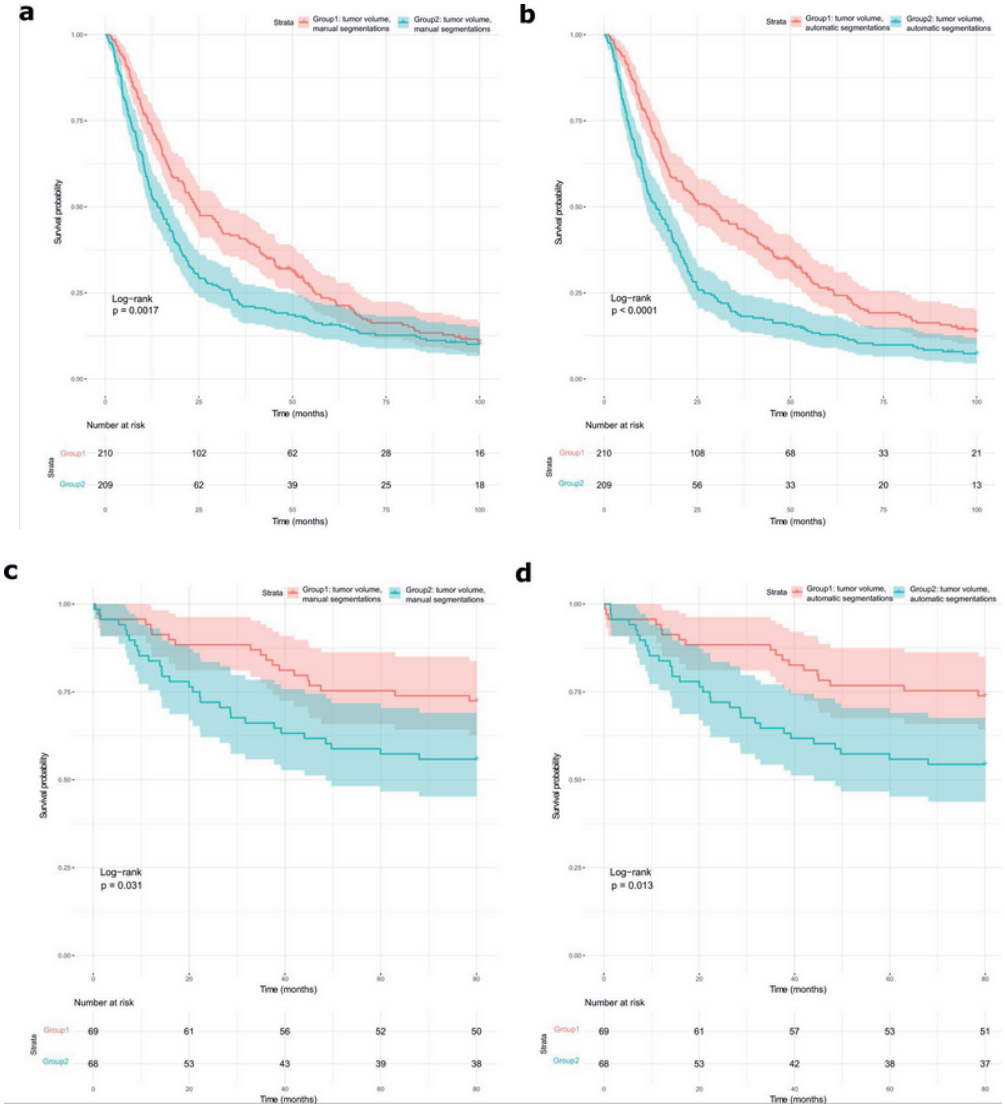


Fig. 4: Prognostic power of NSCLC segmentations measured with tumor volume.

Comparison of prognostic power of non-small cell lung cancer (NSCLC) segmentation is measured through tumor volume. Tumor volume is calculated based on the manual (a, c) and automatically generated contours (b, d). Kaplan–Meyer curves for survival groups based on tumor volume are displayed with 95% pointwise confidence intervals. P values are calculated using the log-rank test. Vertical hash marks indicate censored data. a, b KM curves for Maastro-CT-Lung-1 cohort of 419 NSCLC patients. c, d KM curves for Stanford Lung cohort of 137 NSCLC patients.

In silico clinical trial

A registered in silico clinical trial was performed to assess the following endpoints: (1) the time needed for the processes of manual and automated segmentation; (2) inter and intra-observer variability; (3) the preference of experts for manual or automatically generated segmentations.

For the first and second endpoints, seven medical imaging specialists experienced in NSCLC contouring were asked to contour the tumors of 25 patients from dataset 3 while being timed. Our automated method was significantly faster than the fastest participant ($p < 0.0001$). The mean time for the automated method was 2.78 s/patient (SD = 0.44), whereas the mean time for manual segmentation was 172.19 s/patient (SD = 158.99) (Fig. 5a).

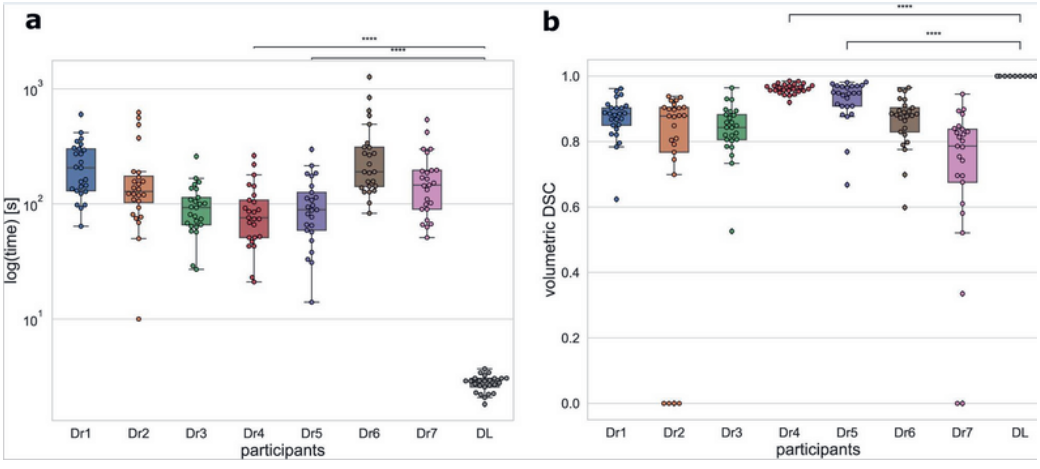


Fig. 5: Contouring time and intra-observer variability. Data were presented as box plots with overlaid swarm plots, where boxes are representing the interquartile range (IQR), extending from Q1 to Q3 and centered on the median value. Upper whiskers represent the highest data point that is less than Q3 + 1.5 × IQR. Lower whiskers represent the smallest data point that is greater than Q1 − 1.5 × IQR. Data points outside whiskers are considered outliers. P values were calculated using a two-sided Mann–Whitney–Wilcoxon test with Bonferroni correction and referred as follows: ****refers to the p value in the range: $p \leq 1.00e\text{-}04$. The exact p values are reported in the order from the top to the bottom as they are displayed on the figures. Dr1, Dr2, Dr3, Dr4, Dr5, Dr6, and Dr7—represent contours made by the medical doctors, DL—represents automatically generated contours. a Distribution of contouring time was obtained on the 25 NSCLC patients by seven participants and the automated method, corresponding p values are 2.816e-09 and 2.824e-09. b Volumetric dice similarity coefficient (DSC) representing intra-observer variability, across participants and the automated method, obtained on the 25 NSCLC patients, corresponding p values are: 1.946e-10 and 1.946e-10.

The median DSC for intra-observer variability among all experts was 0.88 (IQR = 0.12) whereas automated segmentations were 100% reproducible. Individual intra-observer variability scores are reported in Fig. 5b and the JI and H95th are reported in the supplementary materials (Fig. 11a, b suppl.). The median DSC for interobserver variability was 0.81 (IQR = 0.24) (see Fig. 12 suppl.).

The results for assessment of the variability between expert clinicians and the proposed automatic segmentation method achieved on the validation dataset 8 are presented in Fig. 6. Our method achieved an average DSC of 0.82 (IQR = 0.14), whereas the average DSC of experts inter-variability was 0.84 (IQR = 0.12).

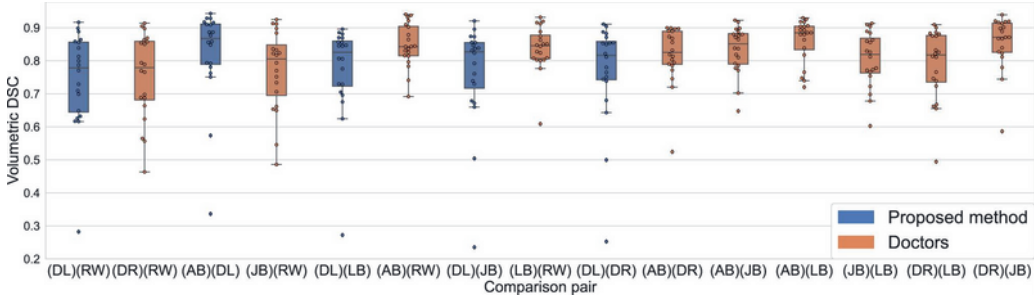


Fig. 6: Method performance vs interobserver variability. Quantitative segmentation performance and interobserver variability is measured using volumetric DSC across comparison pairs obtained on 20 NSCLC patients. DR1, DR2, DR3, DR4, and DR5—represent contours made by the doctors (expert clinicians), DL—represents automatically generated contours. Orange box plots correspond to manual segmentation vs manual segmentation comparison and display interobserver variability. Blue box plots correspond to the proposed method vs manual segmentations comparisons and display the proposed method performance. Data were presented as box plots with overlaid swarm plots, where boxes are representing the interquartile range (IQR), extending from Q1 to Q3 and centered on the median value. Upper whiskers represent the highest data point that is less than Q3 + 1.5 × IQR. Lower whiskers represent the smallest data point that is greater than Q1 − 1.5 × IQR. Data points outside whiskers are considered outliers.

For the third endpoint, we had 40 participants from four different backgrounds: four health/medicine master students, 17 computer scientists, 12 medical doctors working in the field of medical imaging, and seven medical specialists (radiologists or radiation oncologists). In order to quantitatively evaluate the qualitative preferences of experts regarding automated vs manual contours, we developed a software tool which allowed experts to visually compare the segmentation and choose their preferences.

On average, the participants preferred the automatic segmentation above the expert's contour in 55% (IQR = 12%) of the cases (Fig. 13a suppl.). Among the groups the qualitative preference scores were as follows: students = 51% (IQR = 4%) computer scientists = 52% (IQR = 14%), medical doctors = 56% (IQR = 12%) and radiologists and radiation oncologists = 59% (IQR = 13%) (Fig. 13b suppl.).

Discussion

We presented a deep learning-based approach that is able to achieve state-of-the-art detection and 3D volumetric segmentation of NSCLC on CT scans. Although several attempts to develop lung cancer CT detection and segmentation methods have been previously made, we believe our work is standing out, especially in its external validation and ability to work on full thoracic CT scans without further input needed by a human operator. To improve detection and segmentation performance, we introduced several complementary steps to the automatic segmentation pipeline: (1) a harmonization routine for the preprocessing of CT scans in order to more comprehensively unify patterns on the images for the models to learn from; (2) a robust computer vision-based method to isolate the lung area, allowing the subsequent deep learning step to focus on the region of interest; (3) a dynamically changing loss function for the training procedure, allowing us to control and modify the quality of produced segmentation; (4) CTs of lung abnormalities other than NSCLC were included in the training dataset as negative examples, allowing our method to exclude them from the detection and segmentation process; (5) lung CT slices without contours were also used in the training process as negative samples, thereby increasing the number of unique training samples and decreasing the false-positive rate of the model; (6) although a 2D DL architecture was employed, a 3D post-processing routine produced volumetric segmentation. A prospective, registered in silico clinical trial showed that the performance of the automatic segmentation model is acceptable by modern clinical standards and that participants preferred automatic segmentations more often than the manual contours. Furthermore, RECIST and tumor volume based on the automatic contours were able to generate a more significant split of survival groups than manual contours.

To set our model in the context of similar published work, Kamal et al. (2018) (17) used a Recurrent 3D-DenseUNet architecture to segment lung cancers which allowed them to obtain a DSC of 0.74 on a validation dataset of 40 patients. Jue et al. (2019) (19) evaluated several 2D convolutional neural network (CNN) architectures such as U-net, Segnet, full-resolution residual neural network (FRRN), and incremental multiple resolution residual network (MRRN) to segment patches of 160 × 160 pixels centered around the tumor, achieving DSC of 0.68 on the external validation dataset. Zhang et al. (2020) (25) used a modified version of ResNet to automatically segment GTV and achieved an averaged dice similarity coefficient (DSC) of 0.73 on the test set, lacking however external validation

of the model. Ardila et al. (2019) (16) developed a deep learning-based software, which can detect lung cancer on low dose CTs with an AUC of 94.4%. In our study we were not able to evaluate a patient based AUC for lung cancer detection since all patients had cancer, instead, we have demonstrated that our model was able to detect lungs containing cancer on low dose CTs with a robust AUC of 0.96 in the test and 0.98 in the external validation datasets. Additionally, we evaluated the performance of a published 3D U-net-based approach on our validation dataset, where our model outperformed the published method.

The state-of-the-art detection accuracy and the fact that it accepts any CT containing the lungs as input means the software can be used as a method for screening and detection of lung cancer. This is further corroborated by the fact that CT scans acquired using different parameters can be directly put in, making our method multi-vendor and multi-reconstruction compliant to a certain degree. The inclusion of cases that were hard to segment without a co-registered PET scan allows the deep learning networks to learn how to differentiate tumors from other lung abnormalities such as atelectasis and tumors with mediastinal involvement, which in conjunction with the accurate segmentation of the 3D tumor volume means it can be used clinically in radiotherapy settings or for big data radiomics (and potentially other) research. The robust automatic volumetric and RECIST measurements will subsequently have a positive impact on sample size calculations for clinical trials (26).

Although we attempted to address the flaws and limitations of previous research while developing our software, there were limitations to our work. The ground truth segmentations were originally made on primary NSCLC. Therefore, although the software has a high detection accuracy, it is hypothetically limited to the detection and segmentation of primary NSCLC tumors. Moreover, by considering medical expert contours as the ground truth and taking into account the high interobserver variability of the contouring process (27), the deep learning network was also learning inaccuracies, such as contoured air (that certainly is not cancerous). However, this effect can be alleviated by increasing the training dataset size.

In future work we will utilize the evaluated image factors (slice-thickness, complexity class, predicted tumor size, and tumor

location) in order to give a confidence score to each segmentation produced, providing added information to the user about which segmentations might need more attention. Additionally, we think it would be interesting to evaluate our method in a prospective clinical trial setting for tumor response to treatment evaluation utilizing the automatic volumetric RECIST measurement. Since our method was trained only on the planning/pretreatment CT scans, post-treatment changes in the tumor and lung structures may impose extra challenges on our automated segmentation approach.

Further tuning of the model on NSCLC CT scans, and other independent NSCLC datasets can improve the performance of the software, and advance it towards clinical implementation.

The ability of the software developed in this study to handle full thoracic CT scans with different acquisition and reconstruction parameters and without further human intervention represents the pillar for its clinical transition. Clinical application of this software following prospective validation can have a positive impact on the management of lung cancer patients, as it will improve the detection accuracy, and provide a fast, consistent, and reliable volumetric segmentation for treatment (evaluation) purposes. Furthermore, the use of the software in large radiomics studies will allow automation and will reduce the time needed to complete the studies in a robust manner, as it will significantly decrease the time needed for the rate-limiting part of the workflow—tumor segmentation.

Methods

Description of data

The pretreatment CT scans of 1414 NSCLC patients were retrospectively collected and anonymized by each center and approved by the respective institutional review boards. A description of the data were provided in Table 1, and a description of patient characteristics is provided in Supplementary Table 2.

In this study, which followed the Standards for Reporting of Diagnostic Accuracy Studies statement (28), the requirement for written informed consent was waived. The institutional review board of Maastricht University Medical Center has waived the need for informed consent since the data were anonymized and retrospectively collected with no intervention planned for

participants based on the study, and no compensations were provided. The images in dataset 8 were segmented by five radiation oncologists, which allowed us to compare the performance of the deep learning segmentation model to multiple manual delineations. All other segmentations were performed by a radiologist or radiation-oncologist at the center where the diagnosis was made and checked by at least one segmentation expert at our site. The expert segmentations were considered the ground truth for training and further evaluations. Eighty-six patients from various datasets were excluded due to missing tumor contours and the lack of a PET scan to perform the segmentations according to a clinical protocol. Survival data and CT scans for datasets 1 and 6 were collected from the open sources.

Image preprocessing

Data inhomogeneity necessitated the harmonization of CT data in order to achieve comparable representations of the tumor region. Furthermore, several steps were introduced to reduce computational power requirements and image noise and to optimize the contrast. The first step is the extraction of a 3D array with voxel intensity values represented as Hounsfield Units (HU) from Digital Imaging and Communications in Medicine (DICOM) data. Next, the image contrast is enhanced using a lung window setting (window width (WW) of 1500 HU and window level (WL) of −600 HU) to highlight lung structures. All voxel intensities outside of the upper and lower limits are assigned the value of the closest limit. Following this, nearest-neighbor interpolation is applied to obtain isotropic spatial resolution in the axial plane so that each pixel has a size of 1 × 1 mm2. After spatial normalization, an image with standard bone window settings (WW: 1800, WL: 400) is saved, as it is used as an input in the lung isolation step of the workflow. In order to smooth the effect of different reconstruction methods on the image and to reduce the computational burden, intensity values are aggregated into bins of equal width. This also allows optimization of storage and image processing by packing the images into a much shorter 8-bit integer range and by filtering high-frequency noise. Hereafter, the image is cropped or padded with air intensity values to arrive at a resolution of 512 × 512 pixels, which is chosen as input for the selected deep learning architecture. All image processing and deep learning modeling steps were performed in Python 3.7 with the libraries and respective versions detailed in supplementary materials Table suppl. 1.

Lung region isolation

A robust algorithm for the isolation of the lung region was developed in order to focus on the ROI and allow for the use of whole body CT scans as input. First, the CT couch is detected and removed from the image volume. Air-filled connected volumes are detected and region growing and morphological operations are applied in order to remove small vessels and to connect adjacent regions, resulting in a 3D binary lung mask. The spine axis is identified and the lung mask is halved and symmetrically flipped about the sagittal plane, keeping the union of the flipped and the original lung masks. By doing so, the algorithm is optimized for handling lung abnormalities such as atelectasis, pulmonary infiltration, consolidation, and fibrosis. To accurately identify the spine axis, a further algorithm was developed which identifies the center of the spine using the stored preprocessed image with bone window settings as described in the previous section (Fig. 14 a suppl.). A “bone image” slice containing the lung is projected onto the coronal plane and filtered with a seventh order moving average filter (Fig. 14 b-c suppl.). This is repeated for the first five slices in which the lung mask is present in order to find a starting point for the center spine position S_0. The axis of the spine is positioned normally to this point (Fig. 14 d suppl.).

$$S_0 = \frac{1}{n}\sum_{z=0}^{n} P_z \qquad (1)$$

Where P is a central spine point for the current axial slice, n is the number of slices (= 5).
Due to irregularities of patient positioning and anatomy, the central spine position St is recalculated slice-wise by using exponential smoothing:

$$S_t = \propto \cdot x_t + (1 - \propto) \cdot S_{t-1} \qquad (2)$$

Where x is a central spine point based on the filtered signal for the current axial slice, and α is the weighting coefficient (= 0.3).
This method of flipping the lung mask allows for the inclusion of regions that contain large-sized abnormalities, such as lung collapse, which obscure parts of the lung, whereas commonly used methods exclude those regions (Fig. 14 f-g suppl.).
A morphological dilation with the circle kernel (r=5) is applied to the resulting lung mask in order to have a margin around the lung area. The final binary lung mask is used to isolate the lung region within the original image by setting all the voxel values outside the mask to the normalized air value.

Tumor detection and segmentation

The widely used 2D U-net convolutional neural network (CNN) was employed for slice-wise tumor segmentation (30–33). The axial projection was used to train the network due to the higher resolution of image representation in this plane. To improve segmentation performance, several changes were made to the original CNN architecture. First, rectified linear unit (ReLU) activations were replaced with Exponential Linear Unit (ELU) in order to alleviate the gradient vanishing problem and kick-start the training process (34). Second, dropout layers with the dropout rate ($p = 0.5$) were introduced prior to the 2 last layers of U-net encoder to prevent overfitting (35).
A 2D CNN architecture was chosen for several reasons: 1) by using a 2D input the training dataset can be increased by more than a factor of 60, as overall more than 60000 unique slices were available in the training set; 2) due to calculation costs, most present deep 3D architectures could analyze only a sub volume of the medical image (36,37), or they require a dimensionality reduction using interpolation or other image processing methods. 2D architectures do not have this problem and can process CT scans in the original resolution; 3) our main goal was to develop a pipeline that can be used in a clinical setting, and a 2D architecture allows for significantly lower requirements for executing PC. Our software does not require GPUs and can run on a regular laptop (Intel Core i5, 2.5GHz, 8GB RAM).
In order to increase robustness of the system to a wide range of imaging parameters, the training dataset was expanded using augmentation techniques with the following parameters: random rotation around the image center pixel in a range of 0-25 degrees with a probability of 60%, random horizontal and vertical shifts of

the image in the range of 12% of image shape with a probability of 25%, random zooming of the image with a maximum of 3% of the image shape with a probability of 10%.
The loss function was calculated by combining the Dice similarity coefficient (DSC) loss and the binary cross-entropy, and privilege was given to the DSC loss during the first 50 epochs. The privilege was defined by the coefficients before the DSC and cross-entropy terms in the loss function. By adding the binary cross-entropy component to the loss function, negative samples (slices without contour) could also contribute to the training.
The model was trained for 300 epochs using eight NVIDIA GTX 1080 Ti GPUs. The Adam algorithm was used for the stochastic optimization of the loss function (38). The cosine annealing scheduler was used to adjust the learning rate during the training process. A checkpoint function tracking the DSC on the test dataset was used to keep the best weights.
Predicted 2D binary masks are stacked into a 3D volume and connected component extraction is applied as a post-processing step, whereby only spatially connected mask regions are extracted (39). The connected region containing the most voxels is defined as the primary gross tumor target volume (GTV-1) for quantitative assessment. The final mask is resampled to the original image shape using cv2.INTER_BITS interpolation.

Evaluation metrics

In order to evaluate tumor detection performance we generated lung-based labels, where lungs containing a tumor segmentation were assigned a positive label and lungs without were labeled negative. For cases where a tumor was present in both lungs of a patient, both were labeled positive. The ability of the system to detect tumors was assessed by calculating the area under receiver operating characteristic curve (ROC AUC) and generating a confusion matrix. Automatically generated binary masks were resampled to the original image resolution using cv2.INTER_BITS interpolation before comparing with manual segmentations. The contouring performance of the proposed pipeline, as well as the doctors variability, were assessed by using the volumetric Dice similarity coefficient (DSC), Jaccard index (Ji) and 95th percentile Hausdorff distance (H95th). Additionally, we have evaluated quantitative contouring performance using Surface DSC and Added Path Length (APL).

The DSC is a measure of overlap between two volumes and was computed as:

$$DSC(A,B) = \frac{2\cdot|A\cap B|}{|A|+|B|} = \frac{2\cdot TP}{2\cdot TP+FP+FN} \qquad (3)$$

Jaccard index, used for gauging the similarity between two volumes, was computed as:

$$Ji(A,B) = \frac{|A\cap B|}{|A\cup B|} = \frac{TP}{TP+FP+FN} \qquad (4)$$

where A and B are the sets of voxels corresponding to the ground truth and the automatic segmentation, respectively. TP is the number of true positive voxels, FP is the number of false positive voxels and FN is the number of false negative voxels.
To evaluate the maximum deviation between the automatic segmented surface boundary and the ground truth surface boundary, the 95th percentile of Hausdorff distance (H95th) was used. Hausdorff distance (H) is defined as:

$$H(A,B) = max\{\underset{a\in Sa\, b\in Sb}{sup\ inf}\ d(a,b), \underset{b\in Sb\, a\in Sa}{sup\ inf}\ d(b,a)\} \qquad (5)$$

where a and b are the points on the voxel sets A and B, which represent the ground truth and the automatic segmentation, respectively. Sa and Sb are the surfaces of A and B.
Surface DSC at tolerance τ was computed as:

$$SurfDSC(A,B,\tau) = \frac{\left|S_a\cap \beta_b^{(\tau)}\right|+\left|S_b\cap \beta_a^{(\tau)}\right|}{|S_a|+|S_b|} \qquad (6)$$

Where Sa and Sb are the surfaces of A and B, β^{τ}_{a} and β^{τ}_{b} are the border regions of A and B at a given tolerance τ, where τ is a maximum deviation from the ground truth contour which would not be penalized (40). Tolerance τ for the NSCLC segmentation task have been evaluated on the dataset 8 using segmentations of 5 experts.
APL was defined as follows:

$$APL\left(A,B,PS_{xy}\right) = 10 * PS_{xy}\sum B - A\cap B \qquad (7)$$

Where A and B are the voxel sets of automatic and manual segmentation respectively and PS_{xy} is the pixel spacing in the axial plane in mm (41).

In addition to the model performance evaluation on the test and validation datasets, the variability between expert clinicians was assessed and displayed against the performance of our method by comparing the volumetric DSC among all possible comparison pairs, i.e. experts were compared with each other as well as with the proposed method.
To better gauge the performance of our model under varying circumstances, it was evaluated with regard to slice-thickness, tumor complexity, tumor size, and tumor location. Tumor size sub groups were chosen based on the overall tumor size distribution in the training set. Furthermore, expert subjective tumor complexity labels were defined. To describe the complexity of the tumor, two medical doctors were asked to label the test and validation dataset as follows: for tumors where segmentation cannot be performed without a corresponding PET scan the labels were set to “1”, and “0” otherwise. In case of disagreement, the label “1” was chosen. Additionaly, one medical doctor have also labeled the tumor locations on the test and validation datasets, where tumor locations were defined as follows: lung parenchyma, mediastinum and chest-wall involvement. Tumor locations were selected based on the discussion with clinical experts and existing published research (42).

Statistical analysis

For all non-normally distributed scores the median and interquartile range (IQR) were reported, as well as the frequency histograms (29). Statistical significance was assessed using a two-sided Mann-Whitney-Wilcoxon test with Bonferroni correction. Survival evaluation was done in R (version 4.0.2) using survival (version 3.1-12) and survminer (version 0.4.7) packages. To estimate the difference between survival groups a log-rank test was applied. High and low survival groups were separated by the median tumor volume or median RECIST measurement respectively. A random sampling with replacement bootstrapping strategy was used to compute confidence intervals for AUC values.

An in-silico clinical trial

This trial was registered at clinicaltrials.gov (NCT04164186). For the first and second endpoints (the time needed for the processes of manual and automated segmentation, and inter and intra-observer variability), participants used a state of the art commercial software (MIM version 7.0.4) to produce the segmentations. In order to make the conditions of the trial close to the real clinical practice, experts had CT and PET scans available for each patient and they were able to use a semi-automated segmentation solution provided by MIM, while the proposed method generated the segmentation using only CT scans.

For the third endpoint (preference of experts for manual or automatically generated segmentations), a software tool was developed in-house. The tool has two interactive screens with the first screen showing the description of the experiment and a small questionnaire. In order to analyze preferences at different levels of expertise, the participants were asked to specify their training (e.g. radiologist, radiation-oncologist, medical doctor). The second screen displays comparisons between pairs of segmented axial CT slices (automatic vs. expert) with randomized screen positions, blinded to the participant. For each comparison pair, the participants were asked to select the more accurate contour. Finally, a table was generated containing the choices made. Screenshots of this tool are provided in supplementary materials (Fig. 15-16 suppl.).
The software tool presents scans and contours from the external validation datasets 8. It randomly selects 100 pairs of contoured CT slices, where the DSC between the contours was higher than 0.7. During the assessment, participants were able to adjust the image contrast by changing window settings (WW and WL), and to leave comments.

The preference of the experts was evaluated using the qualitative preference score, defined as:

$$PS = \frac{n_m}{n_o} \times 100\% \quad (8)$$

where n_m is a number of times where preference was given to the proposed method, n_o is a number of cases in total.

Data availability

The datasets 1,6,7,8 used in this study are available open source and can be acessed through the corresponding sources: dataset 1 - https://wiki.cancerimagingarchive.net/display/Public/NSCLC-Radiomics; dataset 6 - https://wiki.cancerimagingarchive.net/display/Public/NSCLC+Radiogenomics#28672347a99a795ff4454409862a398ffc076b98; dataset 7 - https://wiki.cancerimagingarchive.net/display/Public/NSCLC-Radiomics-Genomic s#16056856db10d39adf704eefa53e41edcf5ef41c; dataset 8 - https://wiki.cancerimagingarchive.net/display/Public/NSCLC-Radiomics-Interobserver1#52756590171ba531fc374829b21d3647e95f532c.
The processed datasets 2,3,4,5,9,10 are available under restricted access as they were provided under Data Transfer Agreements from corresponding centers, and are not yet public due to data privacy laws, access can be obtained through the corresponding author upon request subject to ethical review. Approximate time for processing the data request is one month. The raw datasets 2,3,4,5,9,10 are protected and are not available due to data privacy laws. The minimum dataset is available on the GitHub repository of this project: https://github.com/primakov/DuneAI-Automated-detection-and-segmentation-of -non-small-cell-lung-cancer-computed-tomography-images/tree/main/Software %20for%20qualitative%20assesment/test_data. Philippe Lambin should be addressed for correspondence and material requests (email: philippe.lambin@maastrichtuniversity.nl)

Code availability

Code, model files, extra software used in this manuscript and derived data to reproduce the results are available on the GitHub page: https://github.com/primakov/DuneAI-Automated-detection-and-segmentation-of -non-small-cell-lung-cancer-computed-tomography-images. Code for the conversion of DICOM to NRRD format is available through Precision medicine toolbox43 GitHub page: https://github.com/primakov/precision-medicine-toolbox

Acknowledgements

SP, MB and IH acknowledge the financial support of Marie Skłodowska-Curie grant (PREDICT - ITN - No. 766276). AI acknowledges the financial support from Liege-Maastricht imaging valley grant. PL,HW acknowledge financial support from ERC advanced grant (ERC-ADG-2015 n° 694812 - Hypoximmuno), ERC-2018-PoC: 813200-CL-IO, ERC-2020-PoC: 957565-AUTO.DISTINCT, SME Phase 2 (RAIL n°673780), EUROSTARS (DART, DECIDE, COMPACT-12053), the European Union's Horizon 2020 research and innovation programme under grant agreement: ImmunoSABR n° 733008, FETOPEN- SCANnTREAT n° 899549, CHAIMELEON n° 952172, EuCanImage n° 952103, TRANSCAN Joint Transnational Call 2016 (JTC2016 CLEARLY n° UM 2017-8295) and Interreg V-A Euregio Meuse-Rhine (EURADIOMICS n° EMR4).

Author contributions

S.P., A. J., A. I., P.L. conceived the idea of the article. M. B, S. K., E. K., A. I., S. S., I. H., J. W., R. M., H. G., L. H., O. M., M. S., R. G., G. W., A. L., E.L., X.G. participated in the data acquisition and clinical trial. S.P. implemented the analysis. E.L.,S.K. reproduced the results. J. V. T., A. J., A. I., H. C. W., P. L., S. K., R. G. contributed to the writing of the manuscript. H.C.W, A.J., P.L. supervised the work. P.L. approved the submitted version and has agreed both to be personally accountable for the author's own contributions and to ensure that questions related to the accuracy or integrity of any part of the work, even ones in which the author was not personally involved, are appropriately investigated, resolved, and the resolution documented in the literature.

Competing interests

Sergey Primakov reports, within the submitted work a non-issues, non licensed patent in the field of medical imaging segmentation: IMAGE DATA PROCESSING METHOD, METHOD OF TRAINING A MACHINE LEARNING DATA PROCESSING MODEL AND IMAGE PROCESSING SYSTEM; year 2020; application number (PCT/NL2020/050794); inventors: SP, HW, PL. Dr. Philippe Lambin reports, within and outside the submitted work, grants/sponsored research agreements from Varian medical, Oncoradiomics,

ptTheragnostic, Health Innovation Ventures andExomnis. He received an advisor/presenter fee and/or reimbursement of travel costs/external grant writing fee and/or in-kind manpower contribution from Oncoradiomics, BHV, Merck and Convert pharmaceuticals. Dr Lambin has minority shares in the company Oncoradiomics SA, Comunicare Solutions SA, LivingMed Biotech and Convert pharmaceuticals SA and is co-inventor of two issued patents with royalties on radiomics (METHOD AND SYSTEM FOR DETERMINING A PHENOTYPE OF A NEOPLASM IN A HUMAN OR ANIMAL BODY; year 2014; publication number WO/2014/171830; IMAGE ANALYSIS METHOD SUPPORTING ILLNESS DEVELOPMENT PREDICTION FOR A NEOPLASM IN A HUMAN OR ANIMAL BODY; year 2016; publication number WO/2016/060557) licensed to Oncoradiomics, one issue patent on mtDNA (METHOD FOR DETERMINING THE RISK OF DEVELOPING RADIATION-INDUCED TOXICITY AFTER EXPOSURE TO RADIATION; year 2014; publication number WO/2014/184028) licensed to ptTheragnostic/DNAmito, three non-patentable invention (software) licensed to ptTheragnostic/ DNAmito, Oncoradiomics and Health Innovation Ventures. He confirms that none of the above entities or funding was involved in the preparation of this paper.
Dr. Woodruff reports, outside of current manuscript, (minority) shares in the company Oncoradiomics and non-issues, non licensed patent in the field of medical imaging segmentation.
Dr. Lizza Hendriks reports, none related to current manuscript, outside of current manuscript: research funding Roche Genentech, Boehringer Ingelheim, AstraZeneca (all institution); advisory board: Boehringer, BMS, Eli Lilly, Roche Genentech, Pfizer, Takeda, MSD, Boehringer Ingelheim, Amgen (all institution); speaker: MSD (institution); travel/conference reimbursement: Roche Genentech (self); mentorship program with key opinion leaders: funded by AstraZeneca; fees for educational webinars: Quadia (self); interview sessions funded by Roche Genentech (institution); local PI of clinical trials: AstraZeneca, Novartis, BMS, MSD /Merck, GSK, Takeda, Blueprint Medicines, Roche Genentech, Janssen Pharmaceuticals, Mirati. The remaining authors declare no competing interests.

Supplementary materials

The supplemetary materials are provided online and can be accessed via the following link:
https://static-content.springer.com/esm/art%3A10.1038%2Fs41467-022-30841-3/MediaObje cts/41467_2022_30841_MOESM1_ESM.pdf, or via following QR code:

References

1. Bray, F. et al. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: A Cancer Journal for Clinicians vol. 68 394–424 (2018).
2. Postmus, P. E. et al. Early and locally advanced non-small-cell lung cancer (NSCLC): ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Ann. Oncol. 28, iv1–iv21 (2017).
3. Jaffray, D. A. Image-guided radiotherapy: from current concept to future perspectives. Nat. Rev. Clin. Oncol. 9, 688–699 (2012).
4. Barrett, A., Dobbs, J. & Roques, T. Practical Radiotherapy Planning Fourth Edition. (CRC Press, 2009).
5. Stroom, J. C. & Heijmen, B. J. M. Geometrical uncertainties, radiotherapy planning margins, and the ICRU-62 report. Radiother. Oncol. 64, 75–83 (2002).
6. Wolchok, J. D. et al. Guidelines for the Evaluation of Immune Therapy Activity in Solid Tumors: Immune-Related Response Criteria. Clinical Cancer Research vol. 15 7412–7420 (2009).
7. Erasmus, J. J. et al. Interobserver and Intraobserver Variability in Measurement of Non–Small-Cell Carcinoma Lung Lesions: Implications for Assessment of Tumor

Response. J. Clin. Oncol. 21, 2574–2582 (2003).
8. Lambin, P. et al. Radiomics: extracting more information from medical images using advanced feature analysis. Eur. J. Cancer 48, 441–446 (2012).
9. Lambin, P. et al. Radiomics: the bridge between medical imaging and personalized medicine. Nat. Rev. Clin. Oncol. 14, 749–762 (2017).
10. Aerts, H. J. W. L. et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat. Commun. 5, 4006 (2014).
11. Kumar, V. et al. Radiomics: the process and the challenges. Magn. Reson. Imaging 30, 1234–1248 (2012).
12. Ibrahim, A. et al. Radiomics Analysis for Clinical Decision Support in Nuclear Medicine. Seminars in Nuclear Medicine (2019) doi:10.1053/j.semnuclmed.2019.06.005.
13. Kalmet, P. H. S. et al. Deep learning in fracture detection: a narrative review. Acta Orthop. 91, 362 (2020).
14. Ronneberger, O., Fischer, P. & Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. Lecture Notes in Computer Science 234–241 (2015) doi:10.1007/978-3-319-24574-4_28.
15. Szegedy, C. et al. Going Deeper with Convolutions. arXiv [cs.CV] (2014).
16. Ardila, D. et al. End-to-end lung cancer screening with three-dimensional
deep learning on low-dose chest computed tomography. Nat. Med. 25, 954–961 (2019).
17. Kamal, U., Rafi, A. M., Hoque, R. & Hasan, M. K. Lung Cancer Tumor Region Segmentation Using Recurrent 3D-DenseUNet. arXiv [eess.IV] (2018).
18. Ray, A. Lung Tumor Segmentation via Fully Convolutional Neural Networks. (2016).
19. Jiang, J. et al. Multiple Resolution Residually Connected Feature Streams for Automatic Lung Tumor Segmentation From CT Images. IEEE Trans. Med. Imaging 38, 134–144 (2019).
20. Mackin, D. et al. Measuring Computed Tomography Scanner Variability of Radiomics Features. Invest. Radiol. 50, 757–765 (2015).
21. Aerts, H. et al. Data from NSCLC-radiomics. Cancer Imaging Archive (2015).
22. Bakr, S. et al. A radiogenomic dataset of non-small cell lung cancer. Scientific Data vol. 5 180202 (2018).
23. Clark, K. et al. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. J. Digit. Imaging 26, 1045–1057 (2013).

24. van Baardwijk, A. et al. PET-CT-based auto-contouring in non-small-cell lung cancer correlates with pathology and reduces interobserver variability in the delineation of the primary tumor and involved nodal volumes. Int. J. Radiat. Oncol. Biol. Phys. 68, 771–778 (2007).
25. Zhang, F., Wang, Q. & Li, H. Automatic Segmentation of the Gross Target Volume in Non-Small Cell Lung Cancer Using a Modified Version of ResNet. Technol. Cancer Res. Treat. 19, 1533033820947484 (2020).
26. Revel, M.-P. et al. Are two-dimensional CT measurements of small noncalcified pulmonary nodules reliable? Radiology 231, 453–458 (2004).
27. Velazquez, E. R. et al. A semiautomatic CT-based ensemble segmentation of lung tumors: Comparison with oncologists' delineations and with the surgical specimen. Radiotherapy and Oncology vol. 105 167–173 (2012).
28. Cohen, J. F. et al. STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. BMJ Open 6, e012799 (2016).
29. Habibzadeh, F. How to report the results of public health research. J Public Health Emerg 1, 90–90 (2017).
30. Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T. & Ronneberger, O. 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016 424–432 (2016) doi:10.1007/978-3-319-46723-8_49.
31. Norman, B., Pedoia, V. & Majumdar, S. Use of 2D U-Net Convolutional Neural Networks for Automated Cartilage and Meniscus Segmentation of Knee MR Imaging Data to Determine Relaxometry and Morphometry. Radiology 288, 177–185 (2018).
32. Livne, M. et al. A U-Net Deep Learning Framework for High Performance Vessel Segmentation in Patients With Cerebrovascular Disease. Front. Neurosci. 13, 97 (2019).
33. Hashimoto, F., Kakimoto, A., Ota, N., Ito, S. & Nishizawa, S. Automated segmentation of 2D low-dose CT images of the psoas-major muscle using deep convolutional neural networks. Radiol. Phys. Technol. (2019) doi:10.1007/s12194-019-00512-y.
34. Clevert, D.-A., Unterthiner, T. & Hochreiter, S. Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs). arXiv [cs.LG] (2015).
35. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. 15, 1929–1958 (2014).
36. Yu, L., Yang, X., Chen, H., Qin, J. & Heng, P.-A.

Volumetric ConvNets with mixed residual connections for automated prostate segmentation from 3D MR images. in Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence 66–72 (AAAI Press, 2017).
37. Wu, W. et al. Segmentation of pulmonary nodules in CT images based on 3D-UNET combined with three-dimensional conditional random field optimization. Medical Physics vol. 47 4054–4063 (2020).
38. Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization. arXiv [cs.LG] (2014).
39. Dillencourt, M. B., Samet, H. & Tamminen, M. A general approach to connected-component labeling for arbitrary image representations. Journal of the ACM vol. 39 253–280 (1992).
40. Nikolov, S. et al. Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy. arXiv [cs.CV] (2018).
41. Vaassen, F. et al. Evaluation of measures for assessing time-saving of automatic organ-at-risk segmentation in radiotherapy. Phys Imaging Radiat Oncol 13, 1–6 (2020).
42. Wu, J. et al. Radiological tumour classification across imaging modality and histology. Nature Machine Intelligence vol. 3 787–798 (2021).
43. Primakov, S., Lavrova, E., Salahuddin, Z., Woodruff, H. C. & Lambin, P. Precision-medicine-toolbox: An open-source python package for facilitation of quantitative medical imaging and radiomics analysis. arXiv [eess.IV] (2022).

CHAPTER 7:

DEEP LEARNING BASED IDENTIFICATION OF BONE SCINTIGRAPHIES CONTAINING METASTATIC BONE DISEASE FOCI

Authors: Abdalla Ibrahim, Akshayaa Vaidyanathan, Sergey Primakov, Flore Belmans, Fabio Bottari, Turkey Refaee, Pierre Lovinfosse, Alexandre Jadoul, Celine Derwael, Fabian Hertel, Henry C. Woodruff, Helle D. Zacho, Sean Walsh, Wim Vos, Mariaelena Occhipinti, François-Xavier Hanin, Philippe Lambin, Felix M. Mottaghy & Roland Hustinx

Adapted from:
Ibrahim, A., Vaidyanathan, A., Primakov, S. et al. Deep learning based identification of bone scintigraphies containing metastatic bone disease foci. Cancer Imaging 23, 12 (2023). https://doi.org/10.1186/s40644-023-00524-3

Acsess link:
https://link.springer.com/article/10.1186/s40644-023-00524-3

Abstract

Purpose

Metastatic bone disease (MBD) is the most common form of metastases, most frequently deriving from prostate cancer. MBD is screened with bone scintigraphy (BS), which have high sensitivity but low specificity for the diagnosis of MBD, often requiring further investigations. Deep learning (DL) - a machine learning technique designed to mimic human neuronal interactions- has shown promise in the field of medical imaging analysis for different purposes, including segmentation and classification of lesions. In this study, we aim to develop a DL algorithm that can classify areas of increased uptake on bone scintigraphy scans.

Methods

We collected 2365 BS from three European medical centres. The model was trained and validated on 1203 and 164 BS scans respectively. Furthermore we evaluated its performance on an external testing set composed of 998 BS scans. We further aimed to enhance the explainability of our developed algorithm, using activation maps. We compared the performance of our algorithm to that of 6 nuclear medicine physicians.

Results

The developed DL based algorithm is able to detect MBD on BSs, with high specificity and sensitivity (0.80 and 0.82 respectively on the external test set), in a shorter time compared to the nuclear medicine physicians (2.5 min for AI and 30 min for nuclear medicine physicians to classify 134 BSs). Further prospective validation is required before the algorithm can be used in the clinic.

Background

Metastatic bone disease (MBD) is the most common form of metastatic lesions (1,2). The incidence of bone metastasis varies depending on the cancer type (3), yet around 80% of MBD arise from breast and prostate cancers (4). MBD, as the name implies, is due to the propensity of these tumours to metastasize to bones, and it results in eventually difficulty treating painful lesions. Henceforth, early diagnosis is necessary for individualized management that could significantly improve a patient's quality of life (5).

MBD is usually detected using radionuclide bone scintigraphy (or bone scans, BS). BS are nuclear medicine images, which are used frequently to evaluate the distribution of active bone formation, related to benign or malignant processes, in addition to physiological processes. BS scans are indicated in a spectrum of clinical scenarios including exploring unexplained symptoms, diagnosing a specific bone disease or trauma, and the metabolic assessment of patients prior to and during the treatment (6,7). BS combining whole-body planar images and tomographic acquisition (SPECT – single photon emission computed tomography) on selected body parts are highly sensitive, as they detect metabolic changes earlier than conventional radiologic images, with lower sensitivity to lytic lesions. However, depending on the pattern it may lack the specificity to identify the underlying causes. Therefore, a SPECT/CT that correlates the findings of bone scintigraphy anatomically is often useful and leads to a more specific diagnosis of the changes noted (8), although MRI scans may also be additionally requested to clarify the diagnosis. Hence, a tool to improve the specificity of decisions based on BS, and reduce the need for further imaging is a relevant unmet clinical need.

Deep learning (DL) is a branch of machine learning (ML), and refers to data driven modelling techniques, which applies the principles of simplified neuron interactions (9). The application of imaging analysis techniques using artificial neurons on medical imaging started to draw attention decades ago [10], but it only became a major research focus recently due to the advancement in computational capacities and imaging techniques (11, 12). The artificial neuron model is used as a foundation unit to create complex chains of interactions - DL layers. These layers are used to generate even more complex structures - DL architectures. The neural network (NN) training procedure is typically a cost-function minimization process. The cost function measures the error of predictions based on the ground truth labels (13), and the DL network learns how to solve a problem directly from existing data, and apply it to data it has never seen. These complex models contain the parameters (weights) for millions of neurons, which can be trained for the recognition of problem-related patterns in the data being analysed. Several studies investigated the potential of DL-based algorithms for analysing bone scintigraphy scans (14,15,16). The majority of these studies applied DL-algorithms on BS scans of diagnosed (specific) cancer patients, which could limit the learning ability of the DL-algorithm to differentiate MBD from other bone diseases.
In this study, we hypothesize that DL-based algorithms can learn

the pattern of metastatic bone disease on bone scintigraphy scans, and differentiate it from other non-metastatic bone diseases. We investigate the potential of a DL-based algorithm to detect MBD on BS, not limited to those of cancer patients, based on activation maps obtained using the gradient weighted class activation mapping (Grad-CAM) method (17, 18). By doing so, we aim to develop a generalizable tool that can classify scans containing metastases and detect MBD on BS. Moreover, extracting activation maps with the Grad-CAM method (19) and superimposing these maps to the original BD scans, we explored the explainability of the deep learning model's predictions. This is very important to promote the application of these methods in the clinic and avoid the common misconception that sees DL models as "black boxes" without any real connection to clinical and imaging characteristics.

Methods

Imaging data

The imaging data were retrospectively collected from different European centres: Aachen RWTH University Clinic (Aachen, Germany), Aalborg University Hospital (Aalborg, Denmark), and Namur University Hospital (Namur, Belgium). The scans were acquired at each center, following local protocols and with different scanner and acquisition parameters. The electronic medical records of these hospitals were searched for patients who underwent BS between 2010 and 2018. Patients for whom a definitive classification of the foci was available, mostly through further investigations, were further included. All images were acquired with anteroposterior (AP) and posteroanterior (PA) whole-body views. The imaging analysis was approved by the Aachen RWTH institutional review board (No. EK 260/19). According to Danish National Legislation, the Danish Patient Safety Authority can waive informed consent for retrospective studies (approval 31-1521-110). All methods were carried out in accordance with the relevant guidelines and regulations (20). The study protocol for the in silico trial was published on clinicaltrials.gov (NCT: NCT05110430). Manual segmentation of the metastatic spots was performed on 25 BS scans coming from Namur University Hospital by the treating radiation oncologists.

Image pre-processing

Every datapoint containing acquisition at two views (AP and PA) was resized to size (length = 256, height = 512) and the intensities were normalized to range [0–1] using the minimum and maximum intensity of each image. For all the data points, image acquisitions at both views are appended besides each other as shown in Fig. 1.

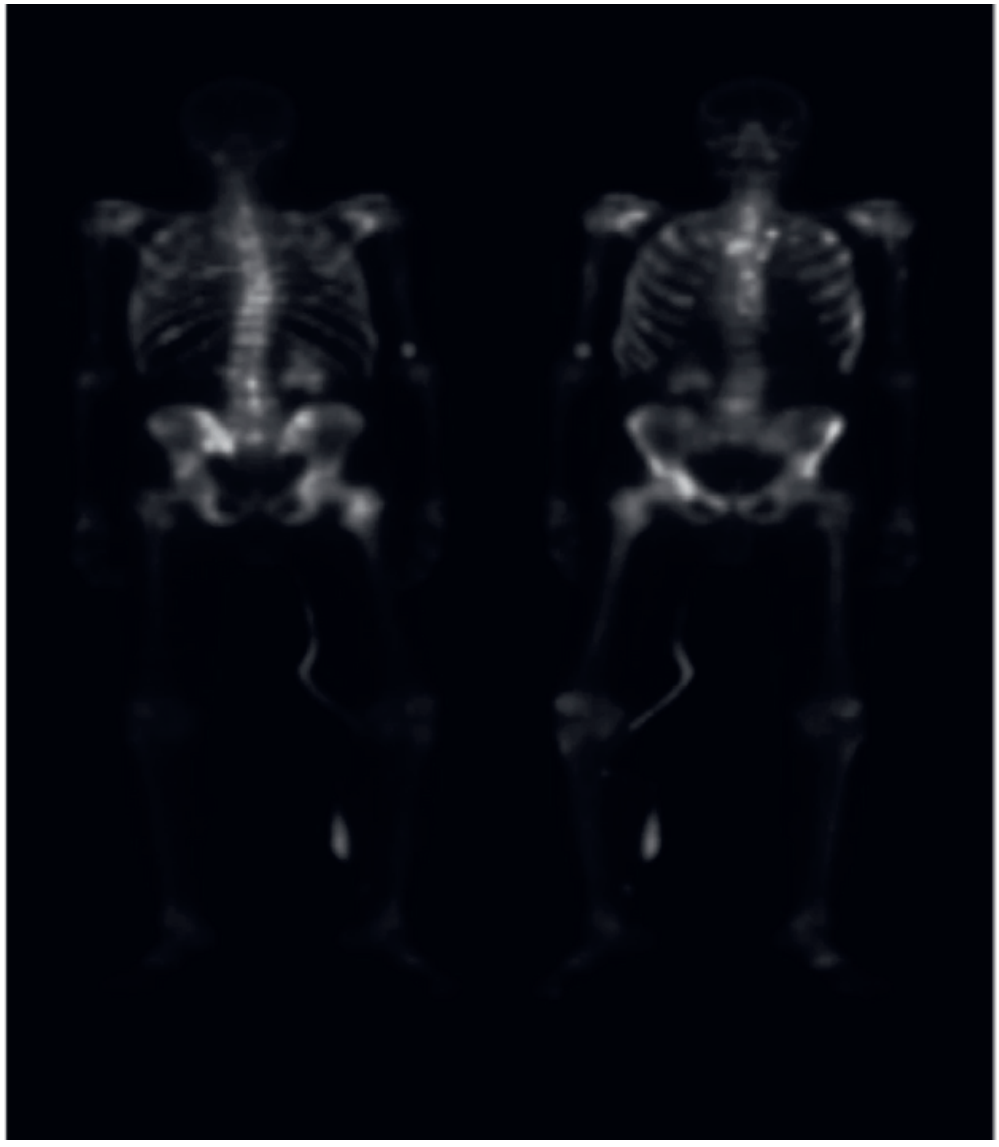


Fig.1 Example of pre-processed BS scans used as input for model training

Model architecture, training and testing

The training and validation datasets are composed of 1203 and 164 images respectively, coming from Centre A (Aachen) and B (Aalborg). The external test cohort is composed of 998 images collected at centre C (Namur). A full overview of the patients cohort division between the different datasets is reported in Table 1.

Table 1 Division of the patients cohort between training, validation and external test

	Training (n = 1203)	Validation (n = 164)	External test (n = 998)
Centre A (Aachen)	235 with metastasis 668 normal	58 with metastasis 58 normal	-
Centre B (Alborg)	94 with metastasis 206 normal	24 with metastasis 24 normal	-
Centre C (Namur)	-	-	411 with metastasis 587 normal

The model was trained on 329 images containing metastasis from Centre B (94) and A (235). At each epoch, the 874 images without any metastasis were shuffled and 329 images were randomly selected to train the model with balanced labels. VGG16 architecture with ImageNet pretrained weights (21) was trained with categorical cross entropy loss for 6 epochs with 200 steps per epoch. The model was trained with 3 channel input. The pre-processed input was duplicated in all the channels, concatenating the inputs along the whole channels dimension to match the size of the pretrained ImageNet. During the training, the images were augmented (22) by flipping along the vertical axis so that the views at AP and PA were randomly represented in the left or right in the images.

The last Max Pooling layer in the VGG16 model was followed by a Global Average pooling layer, followed by a fully connected layer with 512 units and ReLu activation, which is followed by a classification layer containing 2 units with Softmax activation (23) as shown in Fig. 2. The network weights are updated by using the Adam optimizer at learning rate of 1e− 4 (24). The model's performance was evaluated on an external test dataset (n = 998).

Fig. 2 The architecture used in the study. Pre-processed BS scans resized to 512 * 512 dimensions were provided as input to the network. The network outputs a probability score for presence and absence of metastasis on BS images. X = block repetitions, Conv = Convolution kernel, ReLU = rectified linear unit, 3 × 3 = the size of the 2D CNN kernels

The following software packages were used: Python v3.6, Keras v2.0.6 for modelling, training and validation and Sklearn v1.1.1 for metrics calculation and results visualization. The model was trained and validated on a 11GB NVidia GeForce GPU.

Quantitative metrics

The quantitative model performance in this study was assessed using ROC AUC, sensitivity and specificity of the classifier and confusion matrix (true positive rate (TPR), true negative rate (TNR), false negative rate (FNR) and false positive rate (FPR)). The model was evaluated according to the Checklist for AI in Medical Imaging (CLAIM) (25) and Standards for Reporting Diagnostic accuracy studies (STARD) (26).

In silico clinical trial

To better gauge the proposed DL model performance, we developed an application allowing the creation of a reference performance point by collecting nuclear medicine physician's feedback based on the visual assessment of BS scans. We have enrolled 6 nuclear medicine physicians (from one to ten years' experience) to measure their performance on the evaluation dataset of 134 BS images. This dataset was sampled from the Centre C images with an equal number of negative and positive cases. In order to collect participant's feedback, the application was displaying BS image, comment window and window filtering settings (Fig. 3). In the end of the feedback assessment an excel file was generated. For better visual comparison we have evaluated DL based AUC on the same dataset that has been used for visual assessment (134 BS images). Bootstrapping technique, involving 100 resamples obtained via random sampling with replacement from the same dataset, was utilized to estimate ROC AUC 95% confidence interval. Also F1 scores have been calculated and reported for the performance of both the model and the reader study.

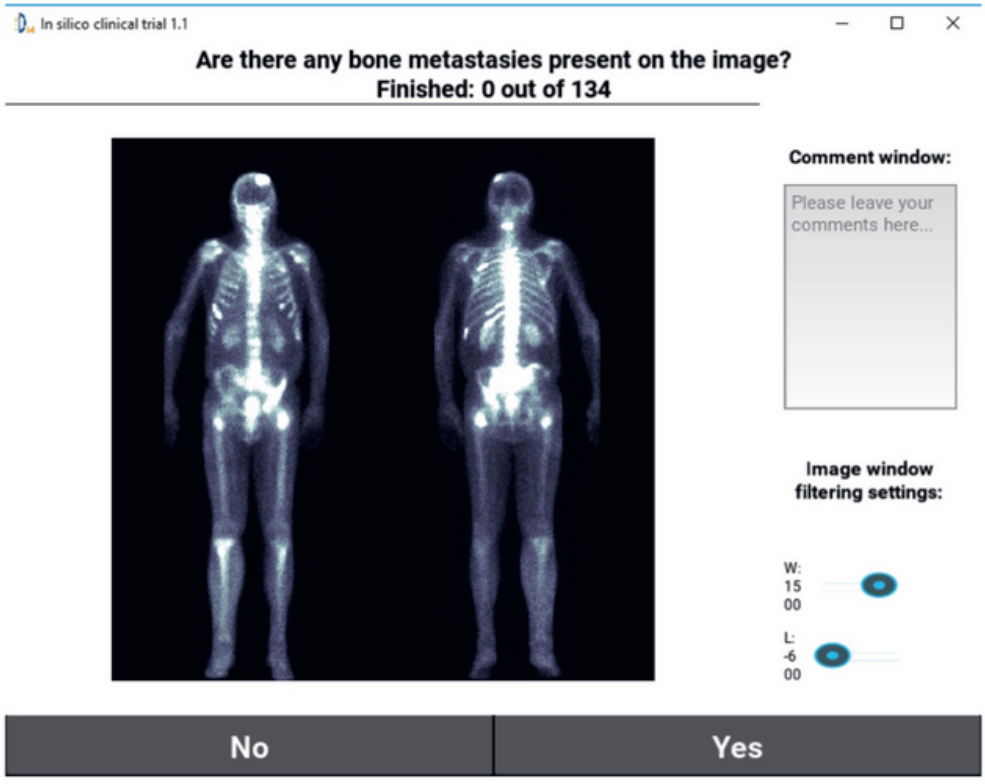


Fig. 3 Screenshot of the application feedback window used in the in silico trial

Results

Model performance

The classification performances of the DL model were evaluated on the external test set coming from Centre C, in terms of Area under the Curve (AUC). The AUC gives the diagnostic ability of a binary classifier to discriminate between true and false values, in this case metastatic and non-metastatic bone disease. Figure 4 (left) represents the ROC curve of the DL classification model, while Fig. 4 (right) is the confusion matrix, which reports the percentages of correct and incorrect classification for each class (metastatic and non-metastatic).

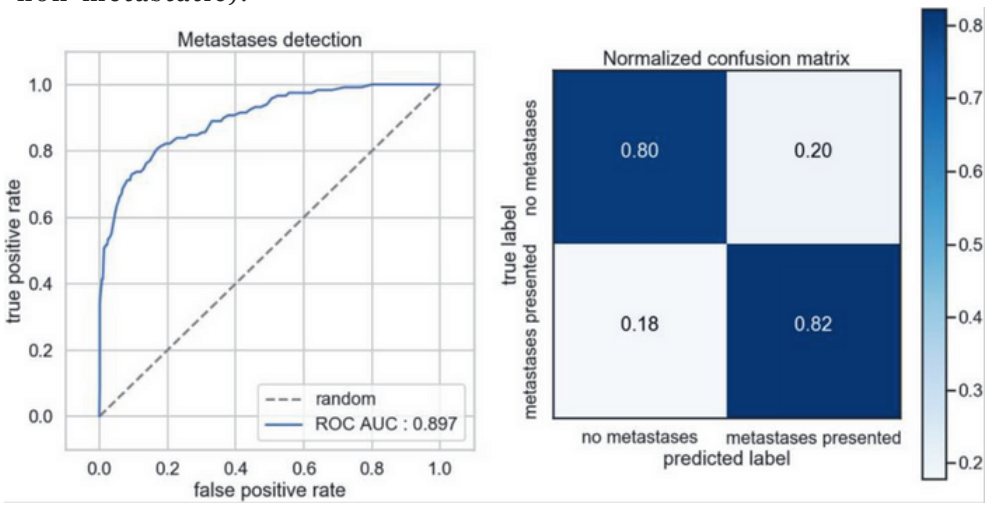


Fig. 4 ROC curve for the classification DL model (left) and Confusion matrix (right)

The model achieved an AUC of 0.897, TPR of 82.2%, TNR of 80.45%, FPR of 19.55% and FNR of 17.79% on the external test set (n = 998). The model achieved a CLAIM score of 64% (27 out of 42 items) and STARD of 50% (15 out of 30 items).

Explainability of trained model based on activation maps

During the testing phase of the trained model, for the scans that were predicted positive (i.e. metastatic disease), activation maps were extracted using the Grad-CAM method. The method uses the gradients extracted corresponding to the class with highest predicted probability, flowing through the last convolutional layer, to produce the activation map. The map was then resized to the size of the input image and

superimposed on the original BS scan, allowing visual inspection of activated zones on the image as shown in Figs. 5 and 6.

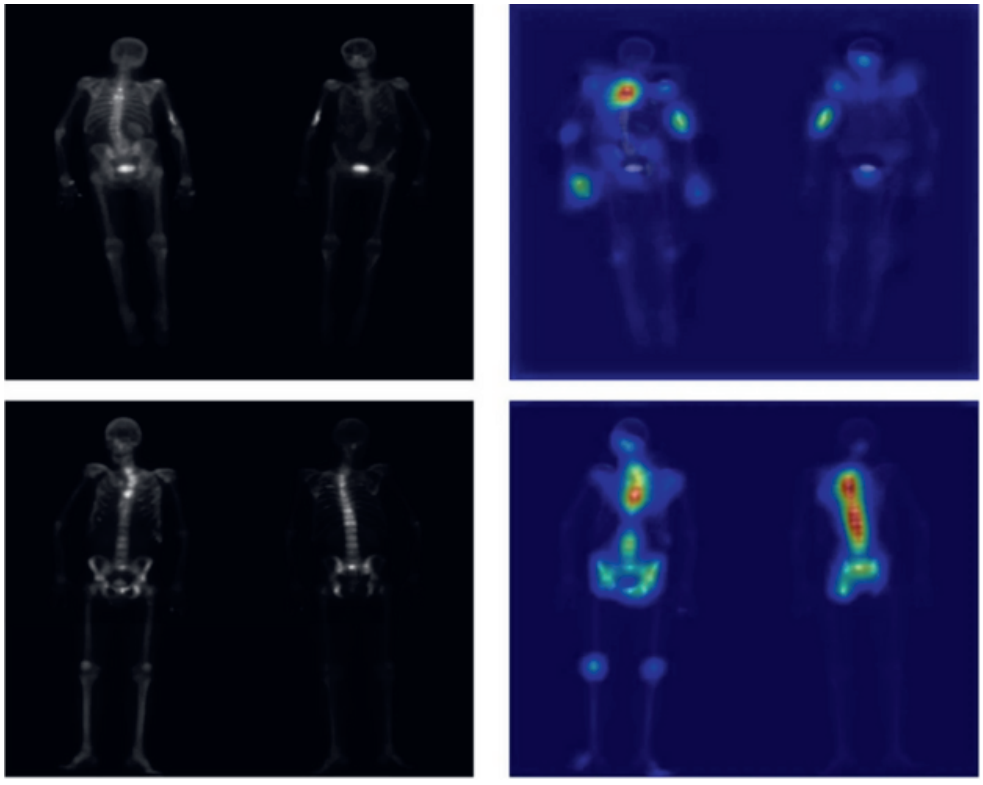


Fig.5 BS images which are correctly classified along with their corresponding activation maps extracted using the GRAD-CAM method. Left) original BD scan, Right) Grad-CAM activation maps obtained from the DL model. Scan correctly classified with a probability of 0.78 (top) and 0.99 (bottom)

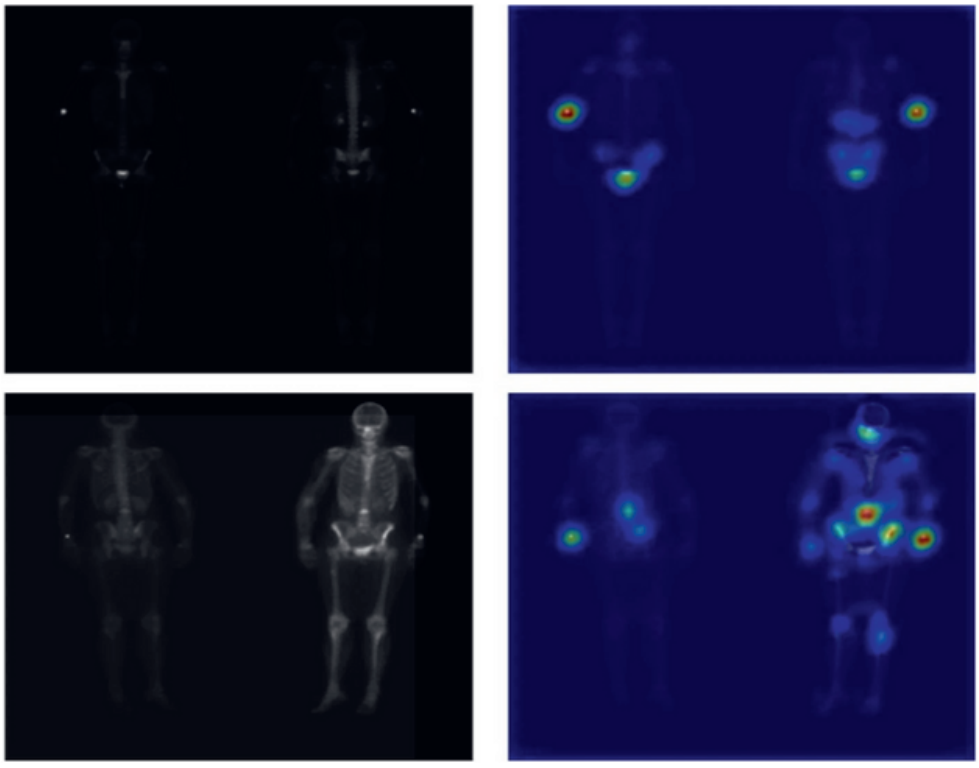


Fig. 6 BS images which are wrongly classified along with their corresponding activation maps extracted using the GRAD-CAM method. Left) original BD scan, Right) Grad-CAM activation maps obtained from the DL model. Scan incorrectly classified with a probability of 0.79 (top) and 0.63 (bottom)

In silico clinical trial

The performance of nuclear medicine physicians based on the BS images was evaluated using AUC (Fig. 7, left), where median performance of the nuclear medicine physician was 0.895 (IQR = 0.087) with F1 score of 0.865 and median performance of DL based method was 0.95 (IQR = 0.024) with F1 score of 0.866.

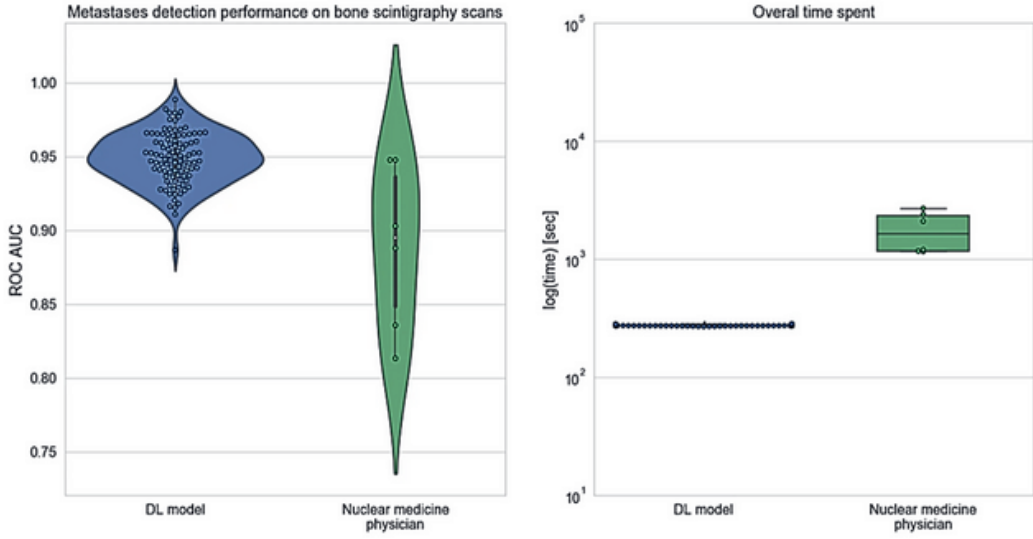


Fig. 7 Violin plots showing the distributions of AUC scores for DL based and manual (across physicians) metastases detection on BS (left); boxplots of the log of the time needed by DL algorithm and nuclear medicine physicians (right)

On average, nuclear medicine physicians spent 30 min to classify all the 134 scans (Fig. 7, right). Given that the physicians had no access to clinical information about the patients, it takes on average 15 s to review one scan. In comparison, the developed algorithm takes 2 and half minutes to classify all the 134 scans, which is around 2 s per patient/ scan.

Discussion

In this study, we investigated the potential of DL-based algorithms to detect MBD on BSs collected from different centres without limiting the study population to cancer patients. All BS scans were acquired at each center, following the standard of care, with different scanners brands and acquisition protocols, assuring the robustness and generalizability of the resulting DL model. Our results show that DL-based algorithms have a great potential to be applied as clinical decision aid tools, which could minimize the time needed by a nuclear physician to assess BSs, and increase the diagnostic specificity of BSs. The application of the state-of-the-art classification techniques has yielded a performance similar to nuclear physicians with no background about the patients' history,

which was further endorsed by the results of the in silico clinical trial.

Some studies previously investigated the potential of DL algorithms to classify lesions on BSs (27). A study investigated the potential of a DL algorithm trained on 139 patients to detect MBD on BSs of prostate cancer patients (16). The authors reported that the nuclear medicine physicians participating in the study achieved a higher sensitivity and specificity compared to the DL algorithm, though the differences were not statistically significant, and highlighted the possibility of involving DL in this clinical aspect. Another study also investigated the ability of DL algorithms to detect MBD in BS of prostate cancer patients (15).

However, the authors did not report on the comparison with the performance of nuclear medicine physicians. Another study investigated the performance of two DL architectures for classifying BS of prostate cancer patients (28). The study included a large number of scans, and the authors reported that the best model achieved an overall accuracy of 0.9. Anand et al. reported on the performance of EXINI bone software, a classification tool for classifying BS of prostate cancer patients based on bone scan index, on simulated and patient scans (29). The authors reported that the software was more consistent in classifying BS compared to visual assessment. Uniquely, we trained our model on patients with and without a history of cancer. The use of our developed algorithm resulted in better classification results on the external test set compared to the median nuclear medicine physician performance, in a significantly shorter time. These results highlight the potential of such algorithms to become reliable clinical decision support tools that minimize the time a clinician needs to review bone scintigraphy scans. Furthermore, Grad-CAM maps allow the nuclear physicians to rapidly check the spots based on which the classification was made. The activated regions are compared with radiologists' segmentation of metastatic spots for qualitative assessment of the explainability of the model's predictions on 25 BS scans (centre C) manually segmented by clinicians (Figs. 5 and 6). The activated regions superimposed on the image can be used in a clinical setting for qualitative assessment by radiologist which further impacts precise diagnosis. In the case of misclassification, Grad-CAM activation maps can help to quicky identify the area of the scan on which the model based its decision. In the reported case in Fig. 6, the image clearly evidence the injection spot located in the hand of the patients and other hyper intense regions in the pelvic bone as reasons for misclassification. This suggests the model

which shows model's overfitting (30) on features that are not relevant to the metastatic spot to classify presence or absence of metastasis in images.

While our study included a relatively large number of scans for training and externally testing the algorithm, several limitations of this study should be noted. Although explainability of model's predictions were explored with qualitative assessment, this study lacks quantitative assessment of the activations due to the limited number of manual segmentations of metastasis (25) on the external test dataset. This could represent a strong point in the future development of the tool, with the availability of larger annotated datasets. Secondly, a prospective validation is required to properly assess the possible impact of the algorithm on the current standard of care, and considering other clinical characteristics of the patients (for example age, sex or primary tumour) that could influence classification performances. This is especially important given the current retrospective nature of the study, to prove beyond reasonable doubts that the classification performances are due to imaging features and not based on clinical/demographic data instead. Lastly, the physicians performances in the in silico trial are only indicative, as they were provided only with planar images, without corresponding SPECT and CT images, and without any clinical covariates available. Obviously, this approximates the actual routine in clinical settings, but it provides a fair indication of the potential added value of the proposed DL model.

Conclusion

We developed a DL based algorithm that is able to detect MBD on BSs, with high specificity and sensitivity. This tool can be used also as a didactic support for radiologists in training. Further prospective validation is required before the algorithm can be used in the clinic

Availability of data and materials

The data that support the findings of this study are not publicly available.

Abbreviations

AP:
Anteroposterior

BS:
Bone scintigraphy

DL:
Deep learning

FNR:
False negative rate

FPR:
False positive rate

IQR:
Interquartile range

MBD:
Metastatic bone disease

PA:
Posteroanterior

ROC:
Receiver operating curve

TNR:
True negative rate

TPR:
True positive rate

Funding

Authors acknowledge financial support from ERC advanced grant (ERC-ADG-2015 n° 694812 - Hypoximmuno), ERC-2020-PoC: 957565-AUTO.DISTINCT, Authors also acknowledge financial support from the European Union's Horizon 2020 research and innovation programme under grant agreement: ImmunoSABR n° 733008, MSCA-ITN-PREDICT n° 766276, CHAIMELEON n° 952172, EuCanImage n° 952103,Interreg V-A Euregio Meuse-Rhine (EURADIOMICS n° EMR4) and Maastricht-Liege Imaging Valley grant, project no. "DEEP-NUCLE".

Contributions

Abdalla Ibrahim, Akshayaa Vaidyanathan: Conceptualization, Methodology, Validation, Formal analysis, Investigation, Data Curation, Writing - Original Draft, Visualization. Sergey Primakov, Flore Belmans, Fabio Bottari, Turkey Refaee: Methodology, Validation, Formal analysis, Data Curation, Writing - Review & Editing, Visualization. Pierre Lovinfosse, Alexandre Jadoul, Celine Derwael, Fabian Hertel, Helle D. Zacho: Validation, Investigation, Resources, Data Curation, Writing - Review & Editing. Henry C. Woodruff, Sean Walsh, Wim Vos, Mariaelena Occhipinti, Francois-Xavier Hanin: Investigation, Resources, Writing - Review & Editing, Funding acquisition. Philippe Lambin, Felix M. Mottaghy, Roland Hustinx: Conceptualization, Resources, Writing - Review & Editing, Supervision, Project administration, Funding acquisition. The author(s) read and approved the final manuscript.

Competing interests

Akshayaa Vaidyanathan, Flore Belmans, Fabio Bottari are salaried employees of Radiomics.

Philippe Lambin reports, within and outside the submitted work, grants/sponsored research agreements from Radiomics SA, ptTheragnostic/DNAmito, and Health Innovation Ventures. He received an advisor/presenter fee and/or reimbursement of travel costs/consultancy fee and/or in kind manpower contribution from Radiomics SA, BHV, Merck, Varian, Elekta, ptTheragnostic, BMS and Convert pharmaceuticals. Lambin has minority shares in the companies Radiomics SA, Convert pharmaceuticals, Comunicare,

and LivingMed Biotech, and he is co-inventor of two issued patents with royalties on radiomics (PCT/NL2014/050248, PCT/NL2014/050728) licensed to Radiomics SA and one issue patent on mtDNA (PCT/EP2014/059089) licensed to ptTheragnostic/DNAmito, three non-patented invention (softwares) licensed to ptTheragnostic/DNAmito, Radiomics SA and Health Innovation Ventures and three non-issues, non-licensed patents on Deep LearningRadiomics and LSRT (N2024482, N2024889, N2024889). He confirms that none of the above entities or funding was involved in the preparation of this paper.

Henry Woodruff has minority shares in the company Radiomics.

Felix M. Mottaghy received an advisor fee and reimbursement of travel costs from Radiomics. He reports institutional grants from GE and Nanomab outside the submitted work.

Mariaelena Occhipinti reports personal fees from Radiomics, outside the submitted work.

Wim Vos and Sean Walsh have shares in the company Radiomics.

The rest of co-authors declare no competing interest.

Supplementary Information

Supplementary information is available through:
https://static-content.springer.com/esm/art%3A10.1186%2Fs40644-023-00524-3/MediaObjects/40644_2023_524_MOESM1_ESM.docx

References

1. Coleman RE. Clinical features of metastatic bone disease and risk of skeletal morbidity. Clin cancer Res an Off J Am Assoc Cancer Res United States. 2006;12:6243s–9.
2. Migliorini F, Maffulli N, Trivellas A, Eschweiler J, Tingart M, Driessen A. Bone metastases: a comprehensive review of the literature. Mol Biol Rep [Internet]. Department of Orthopaedics, University Clinic Aachen, RWTH Aachen University Clinic, Pauwelsstraße 30, 52074, Aachen, Germany. migliorini.md@gmail.com.; 2020;47:6337–45. Available from: http://europepmc.org/abstract/MED/32749632
3. Huang J-F, Shen J, Li X, Rengan R, Silvestris N, Wang M et al. Incidence of patients with bone metastases at diagnosis of solid tumors in adults: a large population-based study. Ann Transl Med [Internet]. AME Publishing Company; 2020;8:482. Available from: https://pubmed.ncbi.nlm.nih.gov/32395526
4. Coleman RE. Metastatic bone disease: clinical features, pathophysiology and treatment strategies. Cancer Treat Rev Netherlands. 2001;27:165–76.
5. Macedo F, Ladeira K, Pinho F, Saraiva N, Bonito N, Pinto L, et al. Bone metastases: an overview. Oncol Rev. 2017;11:321.
6. Ryan PJ, Fogelman I. Bone scintigraphy in metabolic bone disease. Semin Nucl Med United States. 1997;27:291–305.
7. Ziessman HA, O'Malley JP, Thrall JHBT-NM, Fourth E, editors., editors. Chapter 7 - Skeletal Scintigraphy. Philadelphia: W.B. Saunders; 2014. p. 98–130. Available from: https://www.sciencedirect.com/science/article/pii/B9780323082990000079
8.Van den Wyngaert T, Strobel K, Kampen WU, Kuwert T, van der Bruggen W, Mohan HK, et al. The EANM practice guidelines for bone scintigraphy. Eur J Nucl Med Mol Imaging. 2016;43:1723–38.
9. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature [Internet]. 2015;521:436–44. Available from: https://doi.org/10.1038/nature14539
10. McCulloch WS, Pitts W. A logical calculus of the ideas immanent in nervous activity. Bull Math Biophys [Internet]. 1943;5:115–33. Available from: https://doi.org/10.1007/BF02478259
11. Deng L. A tutorial survey of architectures, algorithms, and applications for deep learning. APSIPA Trans Signal Inf Process [Internet]. 2014/01/22. Cambridge University Press; 2014;3:e2. Available from: https://www.cambridge.org/core/article/tutorial-survey-of-architectures-algorithms-and-applications-for-deep-learning/023B6ADF962FA37F8EC684B 209E3DFAE
12. Aslam YNS. A Review of Deep Learning Approaches for Image

Analysis. Int Conf Smart Syst Inven Technol. 2019;2019:709–14.
13. Janocha K, Czarnecki WM. On loss functions for deep neural networks in classification. Schedae Informaticae. 2016;25:49–59.
14. Cheng D.-C, Hsieh T.-C, Yen K.-Y, Kao C.-H. Lesion-Based Bone Metastasis Detection in Chest Bone Scintigraphy Images of Prostate Cancer Patients Using Pre-Train, Negative Mining, and Deep Learning. Diagnostics. 2021;11:518. https://doi.org/10.3390/diagnostics11030518.
15. Papandrianos, N.; Papageorgiou, E.; Anagnostis, A.; Papageorgiou, K. Efficient Bone Metastasis Diagnosis in Bone Scintigraphy Using a Fast Convolutional Neural Network Architecture. Diagnostics. 2020;10:532. https://doi.org/10.3390/diagnostics10080532.
16. Aoki Y, Nakayama M, Nomura K, Tomita Y, Nakajima K, Yamashina M, et al. The utility of a deep learning-based algorithm for bone scintigraphy in patient with prostate cancer. Ann Nucl Med Japan. 2020;34:926–31.
17. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. Int J Comput Vis [Internet]. 2020;128:336–59. Available from: https://doi.org/10.1007/s11263-019-01228-7
18. Dubost F, Adams H, Yilmaz P, Bortsova G, van Tulder G, Ikram MA et al. Weakly supervised object detection with 2D and 3D regression neural networks. Med Image Anal [Internet]. 2020;65:101767. Available from: https://www.sciencedirect.com/science/article/pii/S1361841520301316
19. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. Int J Comput Vis Springer. 2016;128:336–59.
20. World Medical Association. Declaration of Helsinki: ethical principles for medical research involving human subjects. JAMA United States. 2013;310:2191–4.
21. Simonyan, K. and Zisserman, A. (2015) Very Deep Convolutional Networks for Large-Scale Image Recognition. The 3rd International Conference on Learning Representations (ICLR2015). https://arxiv.org/abs/1409.1556.
22. Shorten C, Khoshgoftaar TM. A survey on Image Data Augmentation for Deep Learning. J Big Data [Internet]. 2019;6:60. Available from: https://doi.org/10.1186/s40537-019-0197-0
23. Calin O, Activation Functions BT. - Deep Learning Architectures: A Mathematical Approach. In: Calin O, editor. Cham: Springer International Publishing; 2020. p. 21–39. Available from: https://doi.org/10.1007/978-3-030-36721-3_2

24. Kingma, D. and Ba, J. (2015) Adam: A Method for Stochastic Optimization. Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015). https://arxiv.org/abs/1412.6980.
25. Mongan J, Moy L, Kahn CE. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): A Guide for Authors and Reviewers. Radiol Artif Intell [Internet]. Radiological Society of North America; 2020;2:e200029. Available from: https://doi.org/10.1148/ryai.2020200029
26. Cohen JF, Korevaar DA, Altman DG, Bruns DE, Gatsonis CA, Hooft L et al. STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. BMJ Open [Internet]. 2016;6:e012799. Available from: http://bmjopen.bmj.com/content/6/11/e012799.abstract
27. Liu S, Feng M, Qiao T, Cai H, Xu K, Yu X, et al. Deep learning for the Automatic diagnosis and analysis of bone metastasis on bone scintigrams. Cancer Manag Res. 2022;14:51–65.
28. Han S, Oh J.S, Lee J.J. Diagnostic performance of deep learning models for detecting bone metastasis on whole-body bone scan in prostate cancer. Eur J Nucl Med Mol Imaging. 2022;49:585–595. https://doi.org/10.1007/s00259-021-05481-2.
29. Anand A, Morris MJ, Kaboteh R, Båth L, Sadik M, Gjertsson P, et al. Analytic Validation of the automated bone scan index as an imaging biomarker to standardize quantitative changes in bone scans of patients with metastatic prostate Cancer. J Nucl Med. 2016;57:41–5.
30. Narasinga Rao MR, Venkatesh Prasad D, Sai Teja V, Zindavali P, Phanindra Reddy M. A Survey on Prevention of Overfitting in Convolution neural networks using machine learning techniques. Int J Eng Technol. 2018;7:177.

PART 3:

OPEN SOURCE AND PATENTED CONTRIBUTIONS TO THE FIELD

CHAPTER 8:

PRECISION-MEDICINE-TOOLBOX: AN OPEN-SOURCE PYTHON PACKAGE FOR FACILITATION OF QUANTITATIVE MEDICAL IMAGING AND RADIOMICS ANALYSIS

Authors: Sergey Primakov[1], Elizaveta Lavrova[1], Zohaib Salahuddin, Henry C Woodruff, Philippe Lambin

1 These authors have contributed equally.

Adapted from:
Sergey Primakov, Elizaveta Lavrova, Zohaib Salahuddin, Henry C Woodruff, Philippe Lambin., Precision-medicine-toolbox: An open-source python package for facilitation of quantitative medical imaging and radiomics analysis arXiv:2202.13965 [eess.IV] (or arXiv:2202.13965v1 [eess.IV] for this version) doi: https://doi.org/10.48550/arXiv.2202.13965

Acsess link:
Pre-print:https://arxiv.org/abs/2202.13965
Publication:https://www.sciencedirect.com/science/article/pii/S2665963823000453
GitHub repo:https://github.com/primakov/precision-medicine-toolbox
PyPi:https://pypi.org/project/precision-medicine-toolbox/
Readthedocs:https://precision-medicine-toolbox.readthedocs.io/en/latest/

Highlights

•Medical imaging demands automation but is lacking methodology standardization.

•Medical imaging data curation and exploration are performed in an in-house manner.

•Our toolbox is aimed to fill these gaps and enable automated pipelines in radiomics.

•The toolbox will increase reproducibility in quantitative medical imaging research.

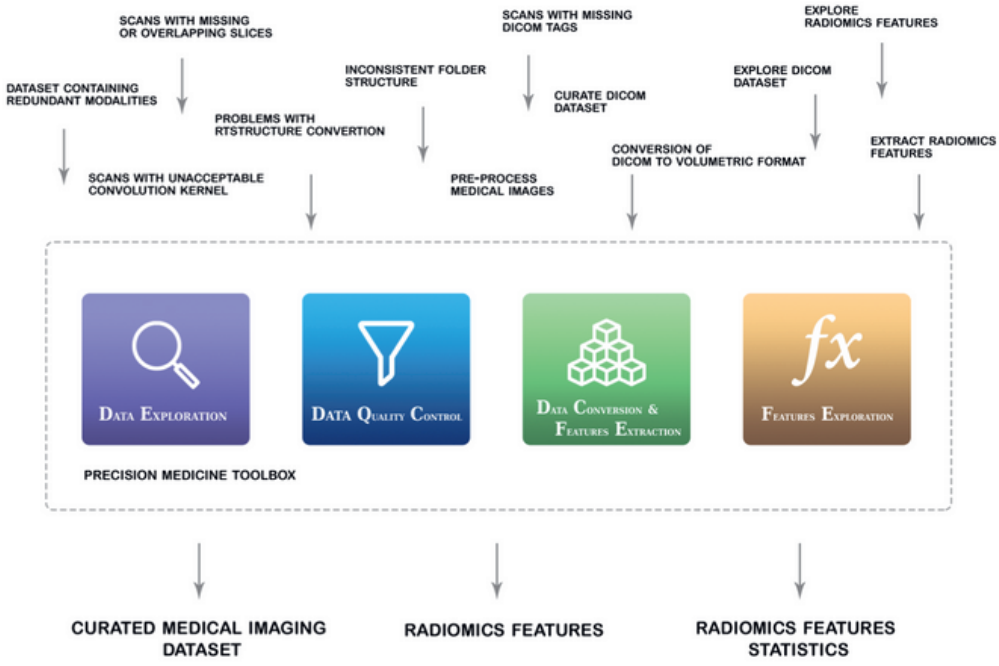
•The community is encouraged to contribute to develop a powerful tool for radiomics.

Abstract

Medical image analysis plays a key role in precision medicine. Data curation and pre-processing are critical steps in quantitative medical image analysis that can have a significant impact on the resulting performance of machine learning models. In this work, we introduce the Precision-medicine-toolbox, allowing clinical and junior researchers to perform data curation, image pre-processing, radiomics extraction, and feature exploration tasks with a customizable Python package. With this open-source tool, we aim to facilitate the crucial data preparation and exploration steps, bridge the gap between the currently existing packages, and improve the reproducibility of quantitative medical imaging research.

Keywords

Medical imaging research, DICOM, Radiomics, Statistical analysis, Features, Image pre-processing



1. Introduction

Medical imaging allows the visualisation of anatomical structures and metabolic processes of the human body and plays an integral part in clinical decision-making for diagnostic, prognostic, and treatment purposes (Beheshti and Mottaghy [2021], Wei et al. [2019]). Medical imaging is becoming increasingly popular in clinical practice due to increasing accessibility of hardware, rising population and growing confidence in the utility of multiple imaging modalities (Smith-Bindman et al. [2008]). Precision medicine aims to enhance individual patient care by identifying subgroups of patients within a disease group using genotypic and phenotypic data for better understanding of the disease characteristics and consequently targeting the disease with more precise treatment (Niu et al. [2019], Carrier-Vallières [2018]). Medical image analysis plays a key role in precision medicine as it allows the clinicians to identify anatomical abnormalities and it is routinely used in clinical assessment (Acharya et al. [2018]). The amount of healthcare imaging data from disparate imaging sources is exploding and it is not possible for radiologists to cope up with the increasing demand. Multiple studies have shown that there is a significant inter-observer variability for various clinical tasks (Kinkel et al. [2000],

Luijnenburg et al. [2010]). Hence, there is a need for quantitative image analysis tools to aid the clinicians in meeting the challenges of rising demand and better clinical performance. Radiomics is the extraction of quantitative image features and correlating them with biological and clinical outcomes (Lambin et al. [2017]). The field of radiomics is gaining traction each year due to increase in computational power and increasing amount of multimodal data (Oren et al. [2020], Aggarwal et al. [2021], Zhou et al. [2021]) as illustrated by Figure 1. The field of radiomics has demonstrated promising results in various clinical applications including diagnostics, prognosis and decision support systems (Tagliafico et al. [2020], Zhang et al. [2017], Wang et al. [2021], Mu et al. [2020]). Radiomics can broadly be classified into two different categories: handcrafted radiomics and deep learning. Handcrafted radiomics utilises machine learning techniques and image biomarker standardisation initiative (IBSI)-compliant handcrafted features (such as shape, intensity, and texture features) extracted from a specific region of interest (Rogers et al. [2020]). Pyradiomics is one of the available open source tools that allows the extraction of IBSI-compliant handcrafted radiomics features (van Griethuysen et al. [2017]).

Deep learning automatically learns representative image features from the high dimensional image data without the need of feature engineering by using non-linear modules that constitute a neural network (Schmidhuber [2015]). Convolutional neural networks are deep neural networks that became popular in 2012 after the AlexNet architecture demonstrated state-of-the-art performance for image recognition (Krizhevsky et al. [2017]). Since then, convolutional neural networks have demonstrated state-of-the-art performance for many clinical tasks (Murtaza et al. [2020], Bhatt et al. [2021], Mazurowski et al. [2019]). Tensorflow (Abadi et al. [2016]) with Keras interface (Gulli and Pal [2017]) and Pytorch (Paszke et al. [2019]) are popular deep learning frameworks for the implementation of deep neural networks.

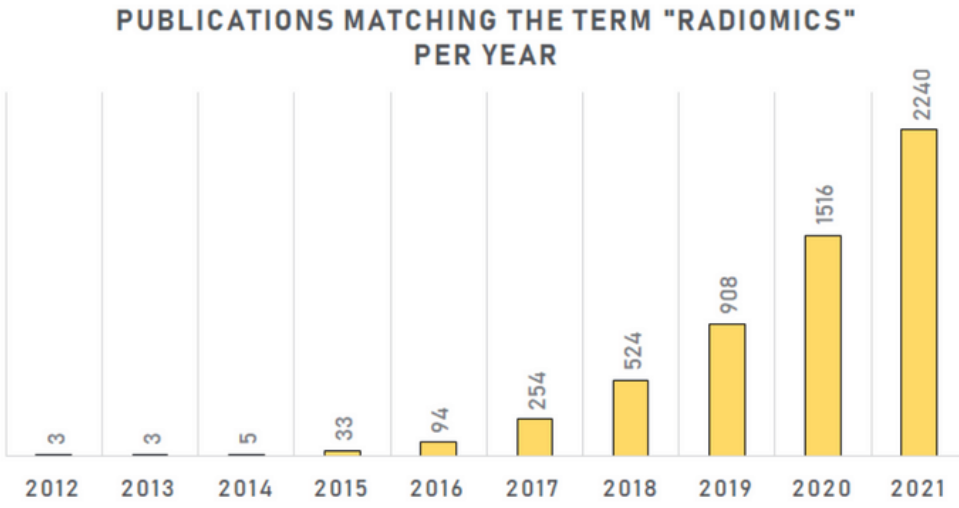


Figure 1 - Number of publications, by year, containing the keyword 'radiomics' in PubMed database.

Data Curation and pre-processing of medical images are time-taking and critical steps in the radiomics workflow that can have a significant impact on the resulting model performance (Fave et al. [2016], Hosseini et al. [2021], Zhang et al. [2019]). The data curation step usually comprises several steps such as image format conversion, out-of-distribution detection and checks for redundant modalities, unacceptable convolution kernel, and missing or overlapping slices. These steps may be performed manually or using low level python libraries such as Pydicom (Mason [2011]), Nibabel (Brett et al. [2020]), SimpleITK (Yaniv et al. [2018]), Numpy (van der Walt et al. [2011]), Pandas (McKinney and Others [2011]), Scipy (Virtanen et al. [2020]), Scikit-image (van der Walt et al. [2014]), and Scikit-learn (Kramer [2016]). The re-implementation of the above-mentioned data curation steps by the researchers makes it error-prone and results in increased difficulty for reproducibility. It is also important to investigate the potential of image processing during the development of a radiomics workflow. Image biomarker standardisation initiative (IBSI) also emphasises on the need of image processing before the extraction of radiomics features (Zwanenburg et al. [2020]).
Moreover, it is also important to perform an exploratory analysis on the handcrafted radiomics features and visualise discriminatory statistics. While there are available tools for the implementation of entire radiomics pipeline such as Nipype (Gorgolewski et al. [2016]), Pymia (Jungo et al. [2021]), and MONAI (MONAI

Consortium [2020]), there is still a need of a toolbox that allows researchers to perform critical tasks such as data curation, image pre-processing and handcrafted radiomics feature exploration during the development of the radiomics study.
In this paper, we introduce the precision-medicine-toolbox that allows researchers to perform data curation, image preprocessing and handcrafted radiomics feature exploration tasks. This toolbox will also benefit the researchers without a strong programming background to implement these critical steps and increase the reproducibility of quantitative medical imaging research. In this paper, we discuss the functionality of the first release of this open source project. In future, more functionality will be added to the toolbox.

2. Methods

2.1 Example data

The functionality of the toolbox is demonstrated on the Lung1 open-source dataset. The Lung1 dataset (Jungo et al. [2021], MONAI Consortium [2020], Aerts et al. [2014]) contains pretreatment CT scans of 422 non-small cell lung cancer (NSCLC) patients, as well as manually delineated gross tumor volume (GTV) for each patient, and clinical outcomes. The imaging data is presented in Digital Imaging and Communications in Medicine (DICOM) format. The delineations are available in Radiotherapy Structure (RT Structure) format. The clinical data is present in Comma-Separated Values (CSV) format.

2.2 Design and implementation

2.2.1 Organisation of the toolbox

The toolbox allows for the preparation of the imaging datasets and exploration of the feature datasets. As illustrated in Figure 2, dedicated base classes have been implemented for each dataset type (imaging or features) to extract the corresponding data, as well as the associated metadata. The functionality classes inherit from the base classes. This approach allows for the separation of reading and processing tasks and makes it readily available for new data formats or functions.

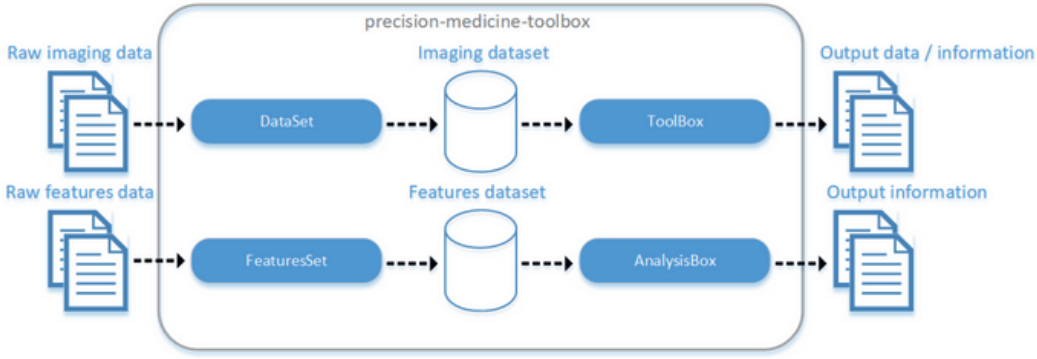


Figure 2 - Organisation of the precision-medicine-toolbox: The DataSet class takes an imaging dataset as in input and is inherited by the ToolBox class; the FeaturesSet class takes a features dataset as an input and is inherited by
the AnalysisBox class.

2.2.2 Imaging module

This module allows for pre-processing and exploration of the imaging datasets. It consists of the base DataSet class and the inheriting ToolBox class. The DataSet class reads the imaging data and the corresponding metadata and initializes a dataset object. The ToolBox class allows for high-level functionality while working with the raw computed tomography (CT) or magnetic resonance (MR) imaging data. Currently, the following functions are implemented:

- dataset parameters exploration by parsing of the imaging metadata,
- dataset basic quality examination, including check of imaging modality, slice thickness, number of slices, in-plane resolution and pixel spacing, and reconstruction kernel,
- conversion of DICOM dataset into volumetric Nearly Raw Rusted Data (NRRD) dataset,
- image basic pre-processing, including bias field correction, intensity rescaling and normalization, histogram matching, intensities resampling, histogram equalization, image reshaping,
- unrolling NRRD images and ROI masks into Joint Photographic experts Group (JPEG) slices for a quick check of the converted images or any existing NRRD or MetaImage Medical Format (MHA) dataset,
- radiomics features extraction from NRRD/MHA data using PyRadiomics package (van Griethuysen et al.[2017]).

2.2.3 Features module

This module allows for the exploration of the feature datasets. It consists of the base FeaturesSet class and the inheriting AnalysisBox class. The FeaturesSet class reads the features data and the corresponding metadata, and initializes a FeaturesSet object. The AnalysisBox class allows for the basic analysis of the features. Currently, the following functions are implemented:

- visualization of feature values distributions in classes,
- visualization of features mutual Spearman correlation matrix,
- calculation of corrected p-values for Mann-Whitney test for features mean values in groups,
- visualization of univariate receiver operating characteristic (ROC) curves for each feature and calculation of the area under the curve (AUC),
- volumetric analysis, including visualization of volume-based precision-recall curve and calculation of Spearman correlation coefficient between every feature and volume,
- calculation of basic statistics (number of missing values, mean, std, min, max, Mann-Whitney test p-values for binary classes, univariate ROC AUC for binary classes, Spearman's correlation with volume if volumetric feature name is sent to function) for every feature.

2.3 Online documentation and tutorials

The online documentation for the precision-medicine-toolbox contains information about the source code, third-party packages, package installation, quick start, instructions for contribution, information about the authors, code licence, and acknowledgments. The examples of the toolbox functionality implementation are presented in tutorials. The full description of the classes and methods is presented in Application Programming Interface (API) specifications.

3 Results

3.1 Design and implementation

3.1.1 Organisation of the toolbox

The precision-medicine-toolbox is implemented in Python (Python Software Foundation, Wilmington, DA, U.S.) and requires version 3.6 or higher. The source code is hosted on GitHub (https://github.com/primakov/precision-medicinetoolbox) and Zenodo platform (DOI 10.5281/zenodo.6126913). It depends on the following packages: NumPy (Harris et al. [2020]), SimpleITK (Lowekamp et al. [2013]), Tqdm (Lowekamp et al. [2013], da Costa-Luis [2019]), Pydicom (Mason [2011]), Pandas (Mason [2011], McKinney [2010]), PyRadiomics (van Griethuysen et al. [2017]), Scikit-image (van der Walt et al. [2014]), Ipywidgets (jupyter-widgets), Matplotlib (Hunter [2007]), Pillow (Clark [2015]), Scikitlearn (Buitinck et al. [2013]), Scipy (Buitinck et al. [2013], Virtanen et al. [2020]), Plotly (noa), Statmodels (Seabold and Perktold [2010]). The precision-medicine-toolbox package has been released under the BSD-3-Clause License and is available from the Python Package Index (PyPI) repository (https://pypi.org/project/precision-medicine-toolbox/). An easy installation of the latest version is possible with “pip install precision-medicine-toolbox” command. At the time of submission of this work, precision-medicine-toolbox is at 0.0 release. The project has the following structure:

- README.MD: file with the project overview,
- Requirements.txt: file with the list of the packages to be installed,
- LICENSE: statement of the license applicable to the project’s software and manuscripts,
- .gitignore: specification of the files, intentionally untracked by Git,
- .readthedocs.yaml: Read the Docs configuration file,
- Mkdocs.yml: Mkdocs configuration file,
- Setup.cfg and setup.py: configuration files for PyPi package,
- Data: folder with the raw data for the examples as well as generated files,
- Docs: folder with the documentation files,
- Examples: folder with the examples:
  - Imaging_module.ipynb: tutorial illustrating functionality for the imaging datasets,
  - Features_module.ipynb: tutorial illustrating functionality for the features datasets,

- Pmtool: folder with the toolbox scripts:

– __init__.py: initialization file,
– data_set.py: script defining the base class for imaging datasets,
– tool_box.py: script defining the inheriting class for imaging datasets methods,
– features_set.py: script defining the base class for features datasets,
– analysis_box.py: script defining the inheriting class for features datasets methods.

The next sections shortly summarize the examples that cover the current functionalities of the precision-medicinetoolbox.

3.1.2 Imaging module

The example 'Imaging module' illustrates how to explore the imaging parameters retrieved from the DICOM tags, perform data quality check, convert DICOM slices to volume format, perform image basic pre-processing, check ROI segmentation, and extract the radiomic features. At first, the ToolBox class needs to be initialized with the user-defined parameters, such as the path to the dataset, data format, mask availability, mask file names, and image file names. After the ToolBox object is created, the corresponding methods can be called. In the example, to speed up the process, we read only one mask per patient. To get an insight of the data and plan the following pre-processing routines, we perform the exploration of the dataset by collecting its imaging metadata. After the initialization of the ToolBox object, the get_dataset_description method is implemented. The outcome is stored in the dataset_description DataFrame. If we call this method with the default parameters, we get modality, slice thickness, pixel spacing, date, and manufacturer for every DICOM file. After calling this method with the indication of the imaging modality (CT), the collected information contains patient name, CT convolution kernel, slice thickness, pixel spacing, kilovoltage peak, exposure, X-Ray tube current, and series date. For a better understanding of the data we are dealing with, we are using Python Matplotlib functionality to plot the distributions of these parameters. The results are presented in Figure 3.

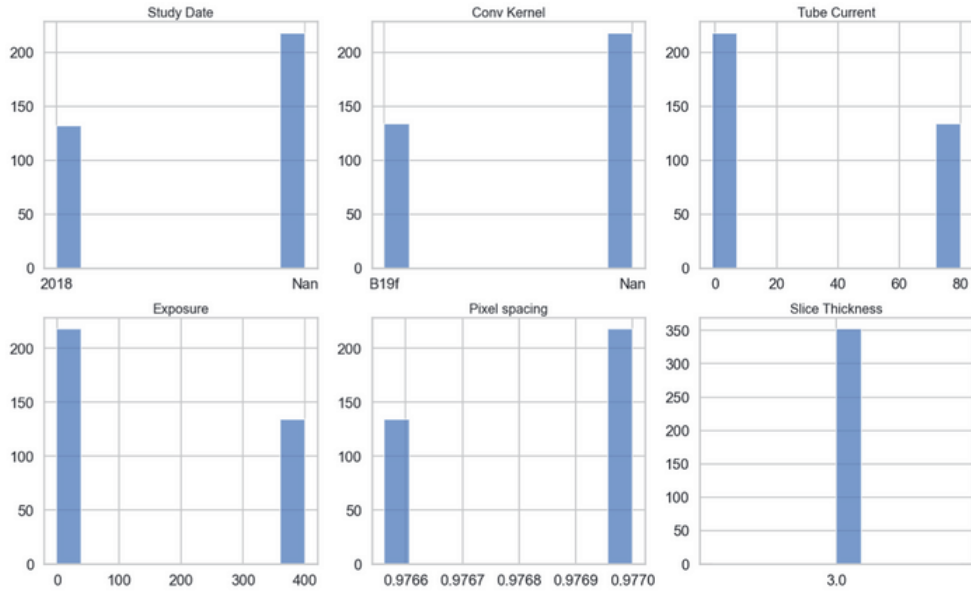


Figure 3 - Distributions of some of the CT imaging parameters in Lung1 data set.

The get_quality_checks method allows to perform a simple quality check of the data and possibly detect irrelevant scans. These might be scans of wrong imaging modality, with wrong imaging projections, with non-consistent (missing/overlapping) slices, with insufficient amount of slices, with slice thickness inconsistent or out of the defined range, with pixel spacing out of range, with unknown or unacceptable convolution kernel, with wrong axial plane resolution, with missing slope/intercept tags. To perform this check, the target scanning parameters are to be passed to the function. While removing some of the input parameters, the corresponding checks are disabled. For each patient, the output DataFrame contains the following flags: '1' - check passed, '0' - check failed. The convert_to_nrrd method is converting the DICOM data into volumetric NRRD format and saves it into the created folder ('.../data/converted_nrrd/'). Currently supported modalities are: CT, MRI, PET, RTSTRUCT. In the example, we performed conversion of the DICOM dataset with CT and corresponding RTSTRUCTs containing GTV contours. Image basic pre-processing is performed by the pre_process method. According to IBSI recommendations, radiomic analysis should be performed for the raw images, except for the modalities, represented in arbitrary units (e.g., MRI, ultrasound). For these modalities, Z-scoring is recommended.

Nevertheless, some image pre-processing can be performed to keep the same data shape within the dataset, decrease diversity of the data, or harmonize the images from different datasets. The following functionality is available in the pre_process method: N4 bias field correction (Tustison et al. [2010]), intensity rescaling and normalization, histogram matching and histogram equalization, intensity resampling, image reshaping. The pre-processing step is not executed, if the corresponding parameter is not passed to the method. It is possible to visualise every pre-processing step for every patient and print out the processing parameters and basic intensity statistics for input and output scans. To perform a sanity check of the converted images and masks and their co-alignment, we initialize a ToolBox class for the newly converted NRRD dataset. Then we call the get_jpegs method, which saves the converted JPEG slices into the .../data/'images_quick_check/' folder. The example of the output is presented in Figure 4.

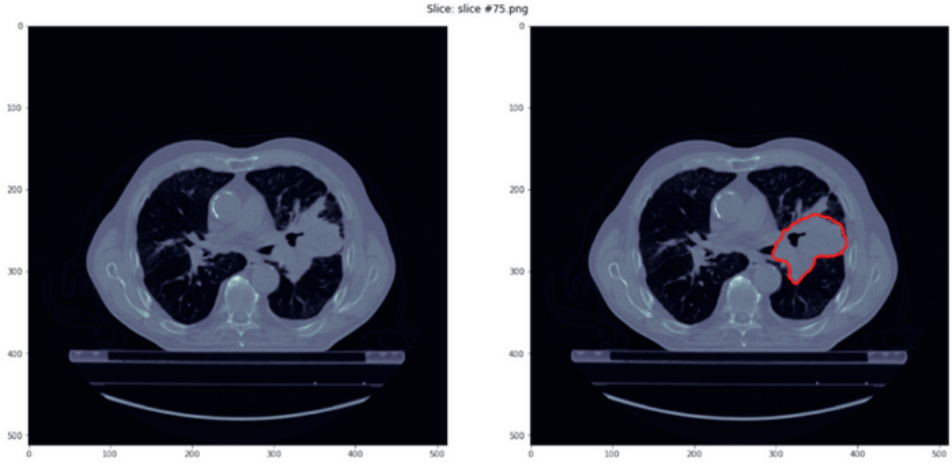


Figure 4 - Example of the quick segmentation check.

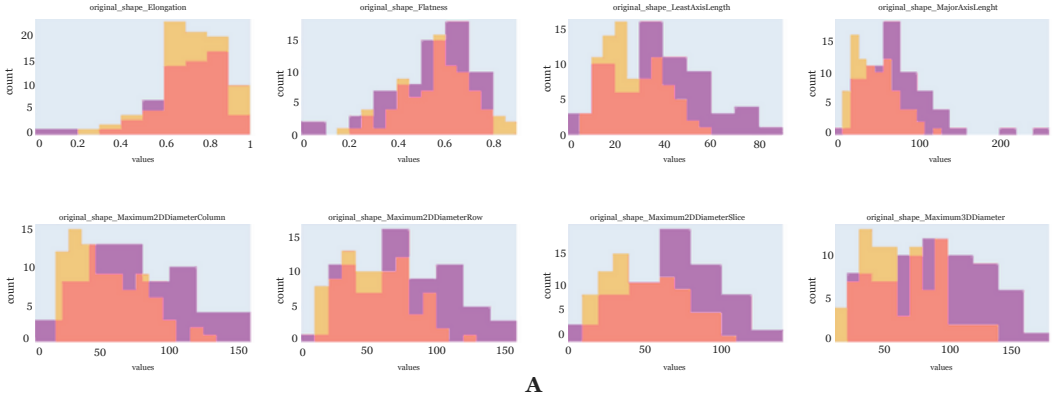
To extract the PyRadiomics features, a ToolBox class for the newly converted NRRD dataset needs to be initialized. Then the features are extracted with the extract_features method. The extraction parameters are imported from the example_ct_parameters.yaml file. The parameter file is required by the PyRadiomics package and contains information about the preferred image types, features to extract, resampling and discretization settings. In the example we use the parameters suggested for the CT features extraction provided by the PyRadiomics repository. The extract_features method returns a

Pandas DataFrame with radiomics features which can be further exported into the Excel file, CSV or any other format supported by Pandas.

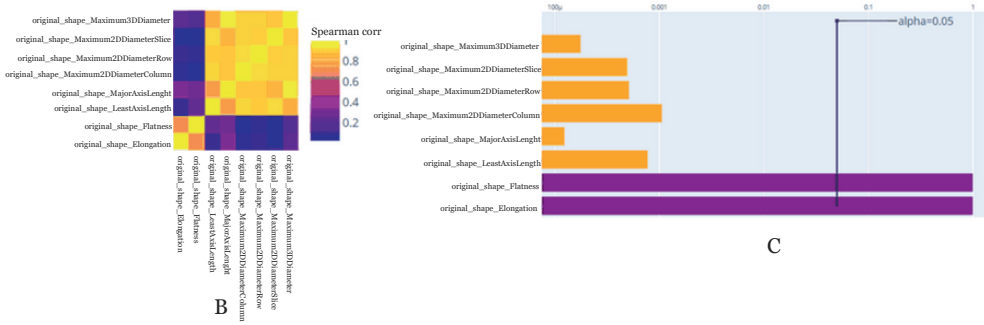
3.1.3 Features module

The example 'Features module' illustrates how to visualize features distribution in classes, plot the feature correlation matrix, check Mann-Whitney U-test p-values, plot univariate ROC and calculate AUC for each feature, perform volumetric analysis, and save all the scores. The tutorial is using the radiomics features, extracted from the Lung1 dataset, and the clinical data file, provided with the dataset. Using the clinical data, we generated three binary outcomes of 1-, 1.5-, and 2-years survival. In the tutorial, we present two cases: binary class dataset and multi-class dataset. The AnalysisBox class is calling a FeaturesSet initialization with the user-defined parameters, such as paths to the tabular data with the features and outcomes, a list of the features to be included or excluded, names of the patient and outcome columns, and a list of the patients to be excluded. The dataset estimated parameters are the available class labels and dataset balance in terms of the outcome values. For the binary class dataset, we declared 1yearsurvival as an outcome column. After AnalysisBox object initialization, we get the class labels ('0' and '1') and class balance (0.42 and 0.58). After using the handle_nan method for the patients, there were no changes in the dataset, which means we did not have any missing values. After calling the plot_distribution method, for each feature, the value distributions were plotted as bin histograms. The result is presented in Figure 5A. The class affiliation is highlighted with a color. The class label is presented on the plot. After calling the plot_correlation_matrix method, the mutual feature correlation coefficient (Spearman's) matrix is visualized. The result is presented in Figure 5B. The values are the absolute values. Colorbar is presented on the right side of the matrix. After calling the plot_MW_p method, Mann-Whitney (with Bonferroni correction) p-values for binary classes test are visualized as a barplot. The result is presented in Figure 5C. The p-value scale is logarithmic. If the p-value for some feature is below the set significance level (alpha=0.05), the corresponding bar is highlighted with a yellow color, whereas the other bars are purple. After calling the plot_univariate_roc method, the univariate ROC curves are visualized for all the features. The ROC AUC scores are reported as well. The result is presented in in Figure 5D. If the ROC AUC score is exceeding the set threshold (auc_threshold=0.70), the curve is highlighted with the purple color.

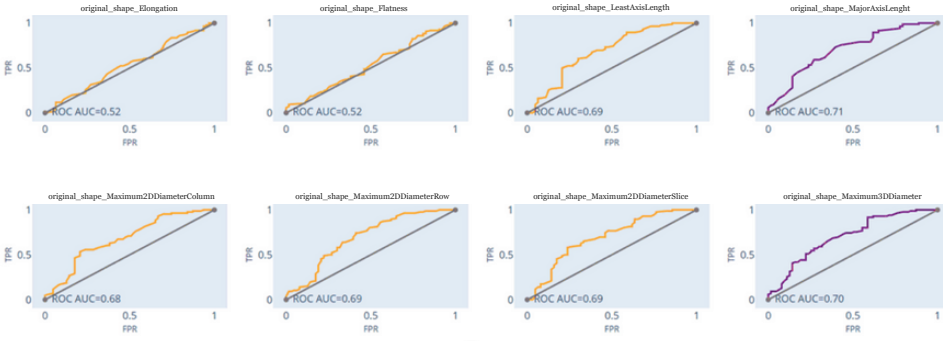
Otherwise, it is yellow. After calling the volume_analysis method with sending there a volumetric feature name ('original_shape_VoxelVolume'), a volume precision-recall curve is visualized (with AUC calculated) as well as a barplot with volume Spearmen's correlation coefficient absolute values with all the features. The resulting plots are presented in Figures 5E and 5F. If the correlation coefficient exceeds a threshold value (corr_threshold=0.75), the bar is highlighted with the purple color. Otherwise, it is yellow. After calling the calculate_basic_stats method, the basic statistics are calculated for all the features. As the dataset has two classes, Mann-Whitney test p-values and univariate ROC AUC scores are calculated. We also define the feature, which is representing the volume ('original_shape_VoxelVolume'), thus Spearman's correlation coefficient with volume is calculated. The results are saved into the 'extracted_features_full_basic_stats.xlsx' file, which belongs to the same directory as the features file.


original_shape_Elongation
original_shape_Flatness
original_shape_LeastAxisLength
original_shape_MajorAxisLenght
original_shape_Maximum2DDiameterColumn
original_shape_Maximum2DDiameterRow
original_shape_Maximum2DDiameterSlice
original_shape_Maximum3DDiameter
count
values

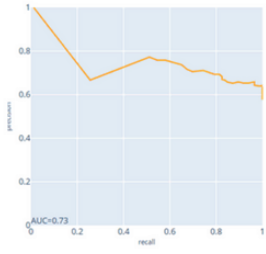
A


original_shape_Maximum3DDiameter
original_shape_Maximum2DDiameterSlice
original_shape_Maximum2DDiameterRow
original_shape_Maximum2DDiameterColumn
original_shape_MajorAxisLenght
original_shape_LeastAxisLength
original_shape_Flatness
original_shape_Elongation
Spearman corr
alpha=0.05

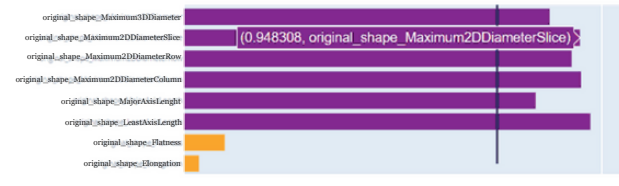
B C


original_shape_Elongation
original_shape_Flatness
original_shape_LeastAxisLength
original_shape_MajorAxisLenght
original_shape_Maximum2DDiameterColumn
original_shape_Maximum2DDiameterRow
original_shape_Maximum2DDiameterSlice
original_shape_Maximum3DDiameter
TPR
FPR
ROC AUC=0.52
ROC AUC=0.52
ROC AUC=0.69
ROC AUC=0.71
ROC AUC=0.68
ROC AUC=0.69
ROC AUC=0.69
ROC AUC=0.70

D


precision
recall
AUC=0.73

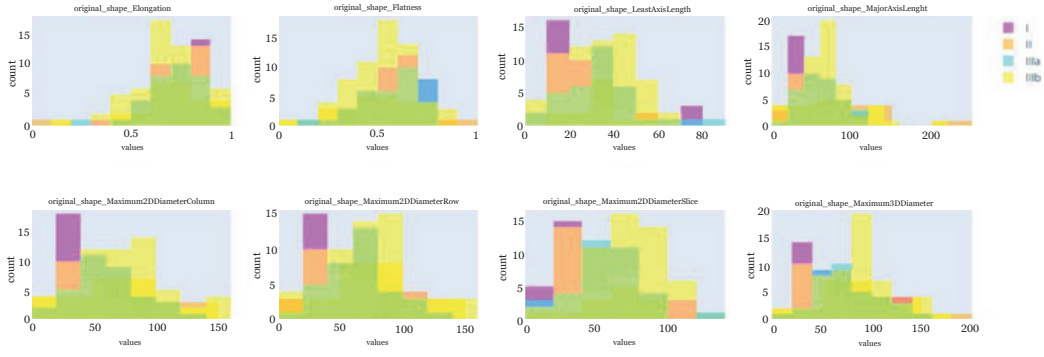
E


original_shape_Maximum3DDiameter
original_shape_Maximum2DDiameterSlice
original_shape_Maximum2DDiameterRow
original_shape_Maximum2DDiameterColumn
original_shape_MajorAxisLenght
original_shape_LeastAxisLength
original_shape_Flatness
original_shape_Elongation
(0.948308, original_shape_Maximum2DDiameterSlice)

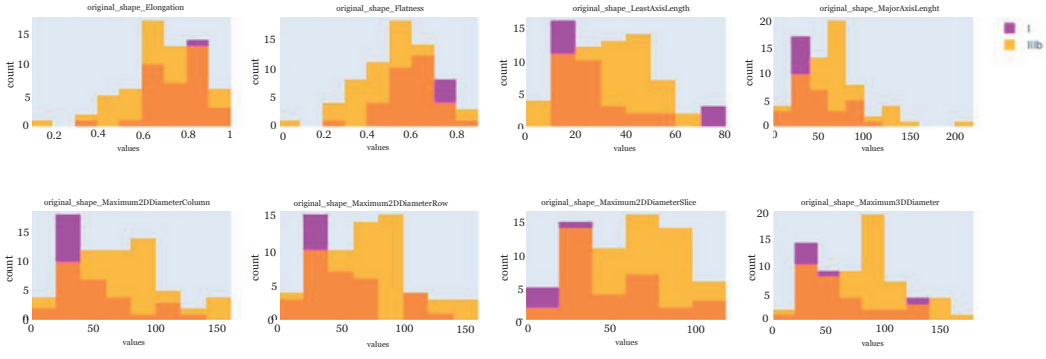
F

Figure 5 - Feature analysis plots for binary outcomes for eight features:
A - feature value distributions in binary classes, B - Spearman's correlation matrix between features, C - Mann-Whitney test (Bonferroni corrected) p-values,
D - univariate ROC curves for binary classification, E - volume based precision-recall curve, F - features Spearman's correlation with volume.

The next part of the tutorial is devoted to multi-class analysis. The AnalysisBox is initialized in the same way, but the outcome column is changed to 'Overall.Stage'. The available class labels are 'I', 'II', 'IIIa', 'IIIb', and the empty value. The class balance is 0.24, 0.09, 0.23, 0.42, and 0.01, respectively. While implementing the handle_nan method at the patient's level, one patient with an unknown outcome is dropped.
Therefore, after the re-initialization of the class object, we have 148 patients with 'I', 'II', 'IIIa', and 'IIIb' labels. The class proportions are 0.24, 0.09, 0.23, and 0.43, respectively. The plot_distribution method works for all the presented classes as well as for the selected classes. The result is presented in Figure 6.



A



B

Figure 6 - Feature value distributions in multiple classes: A - for all the presented classes, B - for the selected classes I and IIIb.

The plot_MW_p and plot_univariate_roc methods are not supported for the multi-class data, but they can still be implemented for the observations from any of the selected two classes. The other methods are working in the same way, as for binary classes. The calculate_basic_stats method does not calculate Mann-Whitney test p-values and univariate ROC AUC scores.

3.2 Online documentation and tutorials

The documentation (http://precision-medicine-toolbox.readthedocs.io/) is built with Mkdocs (https://www.mkdocs.org/) and hosted on the Read the Docs platform (http://readthedocs.io). Code quality is reviewed with CodeFactor (http://codefactor.io). The API specifications for all the classes and methods are generated automatically from the source code annotations with Mkdocstrings (https://mkdocstrings.github.io/). This enables keeping documentation up to date with the latest developments of the package. The documentation also contains links to tutorials with examples generated from Jupyter notebooks. These notebooks are included in the precision-medicine-toolbox package (https://github.com/primakov/precision-medicine-toolbox/tree/master/examples) and are available for any user.

4 Discussion

This paper introduced the open-source precision-medicine-toolbox for imaging data preparation and exploratory analysis. It aims to address the data preparation and exploration problem, bridge the gap between the currently existing packages and improve the reproducibility of quantitative medical imaging research.

The functionality of the toolbox aims to meet some challenges that are specific to the radiomics field. One of these challenges is the lack of data and pipelines standardisation (van Timmeren et al. [2020], Ibrahim et al. [2021], Zwanenburg et al. [2020]). Therefore, reproducibility is one of the key criterias for the radiomics studies. The precision-medicine-toolbox has the functionality for the preliminary data check, including both investigation of the imaging parameters and features properties. This enables a rapid evaluation of the existing data, models, and studies. The other challenge is related to the large amount of the volume surrogate features. This means that many features are highly correlated with volume and do not add any value. In order to identify such features, volumetric analysis functions have been implemented in the precision-medicine-toolbox.

The toolbox is mostly dedicated for the radiomics analysis, as it allows for handling of both raw imaging data and derivative features. Nevertheless, its modules can be used separately for other medical imaging research applications. Imaging module is applicable for the deep learning tasks to prepare the imaging data and get the information about its inhomogeneity. Features module can be used for any tabular data analysis, such as health records variables, or histology-derived features.

The precision-medicine-toolbox was successfully utilised and tested during the development of multiple projects including DUNE.AI, automatic NSCLC segmentation on the CT (Primakov et al. [2021]), repeatability of breast MRI radiomic features (Granzier et al. [2021]), prognostic and predictive analysis of Glioblastoma MRI (Verduin et al. [2021]), quantitative MRI biomarkers discovery in multiple sclerosis (Lavrova et al. [2021]).The development of precision-medicine-toolbox not only aims for democratisation of the machine learning and deep learning pipelines for the researchers without strong programming skills but also drives a programming community effort to improve this package and add its own variables and methods. Therefore, user contributions are very welcome.

5 Conclusions

The development of precision-medicine-toolbox aims to lower the entry barrier for researchers who are starting to work in medical imaging and provide an open source solution for the researchers who already have their inhouse workflowof managing data to increase the reproducibility of the quantitative medical imaging research. We would also like to encourage the community to improve this open-source toolbox by contributing to it.

6 Acknowledgements

The authors would like to thank the Precision Medicine department colleagues for their support and feedback. We also would like to thank PyRadiomics authors for a reliable open-source tool for radiomic features extraction. And we would like to thank TCIA for the open-source Lung1 dataset we used to demonstrate our functionality.

References

1. Mohsen Beheshti and Felix M Mottaghy. Special issue: Emerging technologies for medical imaging diagnostics, monitoring and therapy of cancers. J. Clin. Med. Res., 10(6), March 2021.
2. Hong Wei, Hanyu Jiang, and Bin Song. Role of medical imaging for immune checkpoint blockade therapy: From response assessment to prognosis prediction. Cancer Med., 8(12):5399–5413, September 2019.
3. Rebecca Smith-Bindman, Diana L Miglioretti, and Eric B Larson. Rising use of diagnostic medical imaging in a large integrated health system. Health Aff., 27(6):1491–1502, November 2008.
4. Tianye Niu, Xiaoli Sun, Pengfei Yang, Guohong Cao, Khin K Tha, Hiroki Shirato, Kathleen Horst, and Lei Xing. Pathways to radiomics-aided clinical decision-making for precision medicine, 2019.
5. Martin Carrier-Vallières. Radiomics: Enabling Factors Towards Precision Medicine. McGill University Libraries,2018.
6. U Rajendra Acharya, Yuki Hagiwara, Vidya K Sudarshan, Wai Yee Chan, and Kwan Hoong Ng. Towards precision medicine: from quantitative imaging to radiomics. J. Zhejiang Univ. Sci. B, 19(1):6–24, January 2018.
7. K Kinkel, T H Helbich, L J Esserman, J Barclay, E H Schwerin, E A Sickles, and N M Hylton. Dynamic high-spatialresolution MR imaging of suspicious breast lesions: diagnostic criteria and interobserver variability. AJR Am. J.Roentgenol., 175(1):35–43, July 2000.
8. Saskia E Luijnenburg, Daniëlle Robbers-Visser, Adriaan Moelker, Hubert W Vliegen, Barbara J M Mulder, and Willem A Helbing. Intra-observer and interobserver variability of biventricular function, volumes and mass in patients with congenital heart disease measured by CMR imaging. Int. J. Cardiovasc. Imaging, 26(1):57–64, January
9. Philippe Lambin, Ralph T H Leijenaar, Timo M Deist, Jurgen Peerlings, Evelyn E C de Jong, Janita van Timmeren, Sebastian Sanduleanu, Ruben T H M Larue, Aniek J G Even, Arthur Jochems, Yvonka van Wijk, Henry Woodruff, Johan van Soest, Tim Lustberg, Erik Roelofs, Wouter van Elmpt, Andre Dekker, Felix M Mottaghy, Joachim E Wildberger, and Sean Walsh. Radiomics: the bridge between medical imaging and personalized medicine. Nat. Rev. Clin. Oncol., 14(12):749–762, December 2017.
10. Ohad Oren, Bernard J Gersh, and Deepak L Bhatt. Artificial intelligence in medical imaging: switching from radiographic pathological data to clinically meaningful endpoints. Lancet Digit

Health, 2(9):e486–e488, September 2020.
11. Ravi Aggarwal, Viknesh Sounderajah, Guy Martin, Daniel S W Ting, Alan Karthikesalingam, Dominic King, Hutan Ashrafian, and Ara Darzi. Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. NPJ Digit Med, 4(1):65, April 2021.
12. S Kevin Zhou, Hayit Greenspan, Christos Davatzikos, James S Duncan, Bram Van Ginneken, Anant Madabhushi, Jerry L Prince, Daniel Rueckert, and Ronald M Summers. A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises. Proc. IEEE Inst. Electr. Electron. Eng., 109(5):820–838, May 2021.
13. Alberto Stefano Tagliafico, Michele Piana, Daniela Schenone, Rita Lai, Anna Maria Massone, and Nehmat Houssami. Overview of radiomics in breast cancer diagnosis and prognostication. Breast, 49:74–80, February 2020.
14. Yucheng Zhang, Anastasia Oikonomou, Alexander Wong, Masoom A Haider, and Farzad Khalvati. Radiomics-based prognosis analysis for Non-Small cell lung cancer. Sci. Rep., 7:46349, April 2017.
15. Shouchao Wang, Feng Xiao, Wenbo Sun, Chao Yang, Chao Ma, Yong Huang, Dan Xu, Lanqing Li, Jun Chen, Huan Li, and Haibo Xu. Radiomics analysis based on magnetic resonance imaging for preoperative overall survival prediction in isocitrate dehydrogenase Wild-Type glioblastoma. Front. Neurosci., 15:791776, 2021.
16. Wei Mu, Lei Jiang, Jianyuan Zhang, Yu Shi, Jhanelle E Gray, Ilke Tunali, Chao Gao, Yingying Sun, Jie Tian, Xinming Zhao, Xilin Sun, Robert J Gillies, and Matthew B Schabath. Non-invasive decision support for NSCLC treatment using PET/CT radiomics. Nat. Commun., 11(1):5228, October 2020.
17. William Rogers, Sithin Thulasi Seetha, Turkey A G Refaee, Relinde I Y Lieverse, RenéeWY Granzier, Abdalla Ibrahim, Simon A Keek, Sebastian Sanduleanu, Sergey P Primakov, Manon P L Beuque, Damiënne Marcus, Alexander M A van der Wiel, Fadila Zerka, Cary J G Oberije, Janita E van Timmeren, Henry C Woodruff, and Philippe Lambin. Radiomics: from qualitative to quantitative imaging. Br. J. Radiol., 93(1108):20190948, April 2020.
18. Joost J M van Griethuysen, Andriy Fedorov, Chintan Parmar, Ahmed Hosny, Nicole Aucoin, Vivek Narayan, Regina G H Beets-Tan, Jean-Christophe Fillion-Robin, Steve Pieper, and Hugo J W L Aerts. Computational radiomics system to decode the radiographic phenotype. Cancer Res., 77(21):e104–e107, November 2017.
19. Jürgen Schmidhuber. Deep learning in neural networks: an overview. Neural Netw., 61:85–117, January 2015.

20. Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet classification with deep convolutional neural networks, 2017.
21. Ghulam Murtaza, Liyana Shuib, Ainuddin Wahid Abdul Wahab, Ghulam Mujtaba, Ghulam Mujtaba, Henry Friday Nweke, Mohammed Ali Al-garadi, Fariha Zulfiqar, Ghulam Raza, and Nor Aniza Azmi. Deep learning-based breast cancer classification through medical imaging modalities: state of the art and research challenges. Artificial Intelligence Review, 53(3):1655–1720, March 2020.
22. Chandradeep Bhatt, Indrajeet Kumar, V Vijayakumar, Kamred Udham Singh, and Abhishek Kumar. The state of the art of deep learning models in medical science and their challenges. Multimedia Systems, 27(4):599–613, August 2021.
23. Maciej A Mazurowski, Mateusz Buda, Ashirbani Saha, and Mustafa R Bashir. Deep learning in radiology: An overview of the concepts and a survey of the state of the art with focus on MRI. J. Magn. Reson. Imaging, 49(4):939–954, April 2019.
24. Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanja Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mane, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viegas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-Scale machine learning on heterogeneous distributed systems. March 2016.
25. Antonio Gulli and Sujit Pal. Deep Learning with Keras. April 2017.
26. Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, High-Performance deep learning library. Adv. Neural Inf. Process. Syst., 32, 2019.
27. Xenia Fave, Lifei Zhang, Jinzhong Yang, Dennis Mackin, Peter Balter, Daniel Gomez, David Followill, A Kyle Jones, Francesco Stingo, and Laurence E Court. Impact of image preprocessing on the volume dependence and prognostic potential of radiomics features in non-small cell lung cancer. Transl. Cancer Res., 5(4):349–363, August 2016.

28. Seyyed Ali Hosseini, Isaac Shiri, Ghasem Hajianfar, Pardis Ghafarian, Mehrdad Bakhshayesh Karam, and Mohammad Reza Ay. The impact of preprocessing on the PET-CT radiomics features in non-small cell lung cancer, 2021.
29. Ruiping Zhang, Lei Zhu, Zhengting Cai, Wei Jiang, Jian Li, Chengwen Yang, Chunxu Yu, Bo Jiang, Wei Wang, Wengui Xu, Xiangfei Chai, Xiaodong Zhang, and Yong Tang. Potential feature exploration and model development based on 18F-FDG PET/CT images for differentiating benign and malignant lung lesions. Eur. J. Radiol., 121:108735, December 2019.
30. D Mason. SU-E-T-33: Pydicom: An open source DICOM library. Med. Phys., 38(6Part10):3493–3493, June 2011.
31. Matthew Brett, Christopher J Markiewicz, Michael Hanke, Marc-Alexandre Côté, Ben Cipollini, Paul McCarthy, Dorota Jarecka, Christopher P Cheng, Yaroslav O Halchenko, Michiel Cottaar, Eric Larson, Satrajit Ghosh, Demian Wassermann, Stephan Gerhard, Gregory R Lee, Hao-TingWang, Erik Kastman, Jakub Kaczmarzyk, Roberto Guidotti, Or Duek, Jonathan Daniel, Ariel Rokem, Cindee Madison, Brendan Moloney, Félix C Morency, Mathias Goncalves, Ross Markello, Cameron Riddell, Christopher Burns, Jarrod Millman, Alexandre Gramfort, Jaakko Leppäkangas, Anibal Sólon, Jasper J F van den Bosch, Robert D Vincent, Henry Braun, Krish Subramaniam, Krzysztof J Gorgolewski, Pradeep Reddy Raamana, Julian Klug, B Nolan Nichols, Eric M Baker, Soichi Hayashi, Basile Pinsard, Christian Haselgrove, Mark Hymers, Oscar Esteban, Serge Koudoro, Fernando Pérez-García, Nikolaas N Oosterhof, Bago Amirbekian, Ian Nimmo-Smith, Ly Nguyen, Samir Reddigari, Samuel St-Jean, Egor Panfilov, Eleftherios Garyfallidis, Gael Varoquaux, Jon Haitz Legarreta, Kevin S Hahn, Oliver P Hinds, Bennet Fauber, Jean-Baptiste Poline, Jon Stutters, Kesshi Jordan, Matthew Cieslak, Miguel Estevan Moreno, Valentin Haenel, Yannick Schwartz, Zvi Baratz, Benjamin C Darwin, Bertrand Thirion, Carl Gauthier, Dimitri Papadopoulos Orfanos, Igor Solovey, Ivan Gonzalez, Jath Palasubramaniam, Justin Lecher, Katrin Leinweber, Konstantinos Raktivan, Markéta Calábková, Peter Fischer, Philippe Gervais, Syam Gadde, Thomas Ballinger, Thomas Roos, Venkateswara Reddy Reddam, and freec.nipy/nibabel: 3.2.1, 2020.
32. Ziv Yaniv, Bradley C Lowekamp, Hans J Johnson, and Richard Beare. SimpleITK Image-Analysis notebooks: a collaborative environment for education and reproducible research. J. Digit. Imaging, 31(3):290–303, June 2018.
33. Stefan van der Walt, S Chris Colbert, and Gael Varoquaux. The NumPy array: A structure for efficient numerical computation. Computing in Science Engineering, 13(2):22–30, March 2011.

34. Wes McKinney and Others. pandas: a foundational python library for data analysis and statistics. Python for high performance and scientific computing, 14(9):1–9, 2011.
35. Pauli Virtanen, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J van der Walt, Matthew Brett, Joshua Wilson, K Jarrod Millman, Nikolay Mayorov, Andrew R J Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, ̇Ilhan Polat, Yu Feng, Eric W Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E A Quintero, Charles R Harris, Anne M Archibald, Antônio H Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: fundamental algorithms for scientific computing in python. Nat. Methods, 17(3): 261–272, March 2020.
36. Stéfan van der Walt, Johannes L Schönberger, Juan Nunez-Iglesias, François Boulogne, Joshua D Warner, Neil Yager, Emmanuelle Gouillart, Tony Yu, and scikit-image contributors. scikit-image: image processing in python. PeerJ, 2: e453, June 2014.
37. Oliver Kramer. Scikit-Learn. In Oliver Kramer, editor, Machine Learning for Evolution Strategies, pages 45–53.Springer International Publishing, Cham, 2016.
38. Alex Zwanenburg, Martin Vallières, Mahmoud A Abdalah, Hugo J W L Aerts, Vincent Andrearczyk, Aditya Apte, Saeed Ashrafinia, Spyridon Bakas, Roelof J Beukinga, Ronald Boellaard, Marta Bogowicz, Luca Boldrini, Irène Buvat, Gary J R Cook, Christos Davatzikos, Adrien Depeursinge, Marie-Charlotte Desseroit, Nicola Dinapoli, Cuong Viet Dinh, Sebastian Echegaray, Issam El Naqa, Andriy Y Fedorov, Roberto Gatta, Robert J Gillies, Vicky Goh, Michael Götz, Matthias Guckenberger, Sung Min Ha, Mathieu Hatt, Fabian Isensee, Philippe Lambin, Stefan Leger, Ralph T H Leijenaar, Jacopo Lenkowicz, Fiona Lippert, Are Losnegård, Klaus H Maier-Hein, Olivier Morin, Henning Müller, Sandy Napel, Christophe Nioche, Fanny Orlhac, Sarthak Pati, Elisabeth A G Pfaehler, Arman Rahmim, Arvind U K Rao, Jonas Scherer, Muhammad Musib Siddique, Nanna M Sijtsema, Jairo Socarras Fernandez, Emiliano Spezi, Roel J H M Steenbakkers, Stephanie Tanadini-Lang, Daniela Thorwarth, Esther G C Troost, Taman Upadhaya, Vincenzo Valentini, Lisanne V van Dijk, Joost van Griethuysen, Floris H P van Velden, Philip Whybra, Christian Richter, and Steffen Löck. The image biomarker standardization initiative: Standardized quantitative radiomics for High-Throughput image-based phenotyping. Radiology, 295(2):328–338, May 2020.
39. Krzysztof J Gorgolewski, Oscar Esteban, Christopher Burns, Erik Ziegler, Basile Pinsard, Cindee Madison, Michael Waskom, David

Gage Ellis, Dav Clark, Michael Dayan, Alexandre Manhães-Savio, Michael Philipp Notter, Hans Johnson, Blake E Dewey, Yaroslav O Halchenko, Carlo Hamalainen, Anisha Keshavan, Daniel Clark, Julia M Huntenburg, Michael Hanke, B Nolan Nichols, Demian Wassermann, Arman Eshaghi, Christopher Markiewicz, Gael Varoquaux, Benjamin Acland, Jessica Forbes, Ariel Rokem, Xiang-Zhen Kong, Alexandre Gramfort, Jens Kleesiek, Alexander Schaefer, Sharad Sikka, Martin Felipe Perez-Guevara, Tristan Glatard, Shariq Iqbal, Siqi Liu, David Welch, Paul Sharp, Joshua Warner, Erik Kastman, Leonie Lampe, L Nathan Perkins, R Cameron Craddock, René Küttner, Dmytro Bielievtsov, Daniel Geisler, Stephan Gerhard, Franziskus Liem, Janosch Linkersdörfer, Daniel S Margulies, Sami Kristian Andberg, Jörg Stadler, Christopher John Steele, William Broderick, Gavin Cooper, Andrew Floren, Lijie Huang, Ivan Gonzalez, Daniel McNamee, Dimitri Papadopoulos Orfanos, John Pellman, William Triplett, and Satrajit Ghosh. Nipype: A flexible, lightweight and extensible neuroimaging data processing framework in python. 0.12.0-rc1, 2016.

40. Alain Jungo, Olivier Scheidegger, Mauricio Reyes, and Fabian Balsiger. pymia: A python package for data handling and evaluation in deep learning-based medical image analysis. Comput. Methods Programs Biomed., 198:105796, January 2021.

41. MONAI Consortium. MONAI: Medical open network for AI, 2020.

42. Hugo JWL Aerts, Emmanuel Rios Velazquez, Ralph T H Leijenaar, Chintan Parmar, Patrick Grossmann, Sara Carvalho, Johan Bussink, René Monshouwer, Benjamin Haibe-Kains, Derek Rietveld, Frank Hoebers, Michelle M Rietbergen, C René Leemans, Andre Dekker, John Quackenbush, Robert J Gillies, and Philippe Lambin. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat. Commun., 5:4006, June 2014.

43. Charles R Harris, K Jarrod Millman, Stéfan J van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández Del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E Oliphant. Array programming with NumPy. Nature, 585(7825):357–362, September 2020.

44. Bradley C Lowekamp, David T Chen, Luis Ibáñez, and Daniel Blezek. The design of SimpleITK. Front. Neuroinform., 7:45, December 2013.

45. Casper O da Costa-Luis. tqdm: A fast, extensible progress meter for python and CLI. J. Open Source Softw., 4(37): 1277, May 2019.

46. Wes McKinney. Data structures for statistical computing in python. In Proceedings of the 9th Python in Science Conference. SciPy, 2010.
47. jupyter-widgets. GitHub - jupyter-widgets/ipywidgets: Interactive widgets for the jupyter notebook. https://github.com/jupyter-widgets/ipywidgets. Accessed: 2022-2-16.
48. John D Hunter. Matplotlib: A 2D graphics environment. Comput. Sci. Eng., 9(3):90–95, 2007.
49. Alex Clark. Pillow (PIL fork) documentation. readthedocs, 2015.
50. Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake Vanderplas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. API design for machine learning software: experiences from the scikit-learn project. September 2013.
51. Plotly: The front end for ML and data science models. https://plot.ly. Accessed: 2022-2-15.
52. Skipper Seabold and Josef Perktold. Statsmodels: Econometric and statistical modeling with python, 2010.
53. Nicholas J Tustison, Brian B Avants, Philip A Cook, Yuanjie Zheng, Alexander Egan, Paul A Yushkevich, and James C Gee. N4ITK: improved N3 bias correction. IEEE Trans. Med. Imaging, 29(6):1310–1320, June 2010.
54. Janita E van Timmeren, Davide Cester, Stephanie Tanadini-Lang, Hatem Alkadhi, and Bettina Baessler. Radiomics in medical imaging-“how-to” guide and critical reflection. Insights Imaging, 11(1):91, August 2020.
55. A Ibrahim, S Primakov, M Beuque, H C Woodruff, I Halilaj, G Wu, T Refaee, R Granzier, Y Widaatalla, R Hustinx, F M Mottaghy, and P Lambin. Radiomics for precision medicine: Current challenges, future prospects, and the proposal of a new framework. Methods, 188:20–29, April 2021.
56. S Primakov, A Ibrahim, J van Timmeren, G Wu, S Keek, M Beuque, R Granzier, M Scrivener, S Sanduleanu, E Kayan, and Others. OC-0557 AI-based NSCLC detection and segmentation: faster and more prognostic than manual segmentation. Radiother. Oncol., 161:S441–S443, 2021.
57. R W Y Granzier, A Ibrahim, S Primakov, S A Keek, I Halilaj, A Zwanenburg, S M E Engelen, M B I Lobbes, P Lambin, H C Woodruff, and M L Smidt. Test-Retest data for the assessment of breast MRI radiomic feature repeatability. J. Magn. Reson. Imaging, December 2021.
58. Maikel Verduin, Sergey Primakov, Inge Compter, Henry C Woodruff, Sander M J van Kuijk, Bram L T Ramaekers, Maarten te

Dorsthorst, Elles G M Revenich, Mark ter Laan, Sjoert A H Pegge, Frederick J A Meijer, Jan Beckervordersandforth, Ernst Jan Speel, Benno Kusters, Wendy W J de Leng, Monique M Anten, Martijn P G Broen, Linda Ackermans, Olaf E M G Schijns, Onno Teernstra, Koos Hovinga, Marc A Vooijs, Vivianne C G Tjan-Heijnen, Danielle B P Eekers, Alida A Postma, Philippe Lambin, and Ann Hoeben. Prognostic and predictive value of integrated qualitative and quantitative magnetic resonance imaging analysis in glioblastoma. Cancers, 13(4),February 2021.
59. Elizaveta Lavrova, Emilie Lommers, Henry C Woodruff, Avishek Chatterjee, Pierre Maquet, Eric Salmon, Philippe Lambin, and Christophe Phillips. Exploratory radiomic analysis of conventional vs. quantitative brain MRI: Toward automatic diagnosis of early multiple sclerosis. Front. Neurosci., 15:679941, August 2021.

CHAPTER 9:

PATENT: IMAGE DATA PROCESSING METHOD, METHOD OF TRAINING A MACHINE LEARNING DATA PROCESSING MODEL AND IMAGE PROCESSING SYSTEM

Authors: Sergey Primakov, Henry C Woodruff, Philippe Lambin

Adapted from:
NL2024889/WO2021125950, Title: Image data processing method, method of training a machine learning data processing model and image processing system.

Acsess link:
Access links: NL:https://patentscope.wipo.int/search/en/detail.jsf?docId=NL337460213&_fid= NL337460216
WO:https://patentscope.wipo.int/search/en/detail.jsf?docId=WO2021125950&_ cid=P22-LIW6Q8-48170-1

Abstract

The present document relates to an image data processing method for processing imaging data from different imaging systems, providing a harmonized three-dimensional data set for enabling analysis independent of the image system. The method comprises: obtaining the imaging data which comprises a array of input voxels, and segmenting the data to provide at least one two-dimensional data slice. The method also comprises reconstructing a three-dimensional data set from the data slices. Prior to segmenting, a step of preprocessing is performed, which includes image normalization of the imaging data. This includes at least a transforming or processing of the imaging data for mapping the imaging data onto an image data standard. The document also descries a method of training a machine learning data processing model and an image processing system. The methods and system may be applied to perform harmonized classification and/or radiomics for a plurality of systems.

Field of the invention

The present invention is directed at an image data processing method for processing, by a controller of an analysis system, imaging data obtained from an imaging system. The invention is further directed at a method of training a machine learning data processing model as well as an image processing system for processing medical image data.

Background

With respect to the diagnosing and treating of life-threatening medical conditions, such as cancer or pre-malignant lesions and other illnesses such as non-malignant infectious or inflammatory disease – benign tumours, that may result in the appearance of lesions in a human or animal body, an important role is played by the technological field of medical imaging. In the past decade, the importance of imaging has grown substantially from being primarily a diagnosis and monitoring tool to becoming a tool that supports the treatment of a condition in multiple areas. Furthermore, medical imaging may be applied instead of, or in addition to, autopsy in order to gain more insight in a deceased patient's medical condition at the moment of dying, or for medical studies.

Typically, imaging data needs to be processed as fast and efficient as possible, such that the images are available to medical staff promptly for evaluation. Naturally, above anything else, accuracy is very important during this process. One of the tasks presently being performed by a medical specialist, e.g. the radiologist or radiation oncologist, is the identification of a tumor or a non-malignant pathological lesion and the delineation in two or three dimensions of a region that contains the tumor, neoplasm or non-malignant pathological lesion. Accuracy during such an evaluation is important for several reasons:

- to facilitate the right diagnosis on a radiological image or a pathological image;
- to prevent focusing of the treatment on an area or volume that is too small or too large - an area that is too small may result in the tumor not being treated sufficiently and an area that is too large may result in the treatment of healthy tissue;
- to follow the response after a treatment more specifically to calculate the regression of the target lesion or, in other words, the response; and
- to enable more accurately investigating a cause of death on a dead body in order to prevent the need for an invasive and costly autopsy.

With the trained eye of the medical specialist (typically radiologist, nuclear medicine specialist, radiation oncologist, pathologist and others), this process can normally be performed rather well. However, this process is typically time consuming and not reproducible: there is a large inter and intra doctor heterogeneity. Furthermore, although the medical specialist is typically used to recognize the relevant image features even in the presence of noise and visual artefacts in imaging data, incorrect characterization of the image due to such noise or artefacts as well as human error in general cannot fully be excluded and the implications thereof can be critical. Furthermore there are key image features that are not visible, and/or not quantifiable by the human eye. In particular in case imaging data is obtained using different imaging systems, each system being of a certain type and with it's own system settings, the chance on error increases. Apart from the above, for a three dimensional data set the process has to be conducted for each layer of the image, which is time consuming and expensive.

Apart from identification and delineation, analysis of the imaging data by means of data mining and statistical methods is another technological area that is increasingly applied prior and during treatment, e.g. to make the right diagnosis on a pathological slide or to benefit precision medicine or to estimate potential effectiveness of various treatment plans for certain patient or to evaluate response or to replace an autopsy on a dead body. One of these

areas, for example, is the area of radiomics or histomics (a synonym for quantitative imaging of rrespectively radiological images or histological images) – a statistical data analysis method using handcrafted imaging features or deep learning that has grown popularity over the past years. Radiomics is based on extracting and qualifying descriptive features from image data, and comparing these with data (clinical or biological endpoints) from a database to recognize a certain radiomics signature (a set of distinctive imaging features which is of prognostic relevance). Ideally, such analyses are performed according to certain standards, giving a similar result in similar cases. This may be complicated in case imaging data is obtained from different imaging systems and/or human/semi-automatic segmentation of the tumour or the lesion is not reproducible.

Summary of the invention

It is an object of the present invention to provide an image data processing method that overcomes these disadvantages and enables automated processing of image data from a plurality of different imaging systems with reliable and standardized results.

To this end, there is provided herewith an image data processing method for processing, by a controller of an analysis system, imaging data obtained from an imaging system of a plurality of different imaging systems, for providing a harmonized three-dimensional data set of the image data which data set is harmonized for enabling analysis thereof independent of the respective image system of the plurality of image systems, the method comprising the steps of: obtaining, from the imaging system, the imaging data, wherein the imaging data comprises data for visualizing at least a part of the human or animal body, wherein the imaging data comprises a array of input voxels; segmenting the image data to provide at least one two¬-dimensional data slice, wherein the data slice comprises an array of pixels; and reconstructing a three-dimensional data set from the at least one data slice, wherein the data set is reconstructed by providing a plurality of output voxels, each output voxel being based on an associated pixel of the at least one data slice; wherein the method, prior to the step of segmenting, further comprises a step of preprocessing of the imaging data, wherein the preprocessing at least includes: image normalization of the imaging data, including at least one step of transforming or processing of the imaging data for mapping the imaging data onto an image data standard for enabling said analysis.

The method of the present invention, prior to the step of segmenting

that is typically performed, provides a preprocessing step including a step of image normalization of the imaging data. The image normalization may include one or more data preprocessing steps that enable the imaging data coming from different imaging systems to be processed by a single and same processing method performed by the same processing system. Such steps, for example, may simply include the normalization of parameter values associated with each voxel or pixel of an image, but may also or otherwise include the transformation of data in order to conform it to a certain standard. As a result, the data coming from different imaging systems may easily be processed by single image processing system for example in order to perform identification, delineation or analysis as described hereinabove, and providing a standardized result independent of the imaging system used. In contrast, the conventional methods based on manual or semi-automatic processes do not enable such standardization due to the fact that the process is not reproducible and insufficiently accurate.

For example, in some embodiments, the method further comprises a step of: recognizing, in the harmonized three dimensional data set, using a trained machine learning data processing model, at least one contour of an organ or a neoplasm or another pathological lesion. Due to the image normalization, the standardized evaluation of imaging data by an analysis system in order to perform contour recognition of an organ or a neoplasm or another pathological lesions becomes possible. For example, after normalization, from any differences in color or intensity a difference in tissue density may be calculated or a difference in type of tissue may be established. This enables accurate recognition of such contours in an automated fashion. In addition to increased accuracy, a large amount of data may also be processed much faster.

In some of these embodiments, the at least one contour comprises a contour of the neoplasm, and the method includes: associating a subset of the image data with the neoplasm and defining a gross tumour volume to include the subset of the image data; determining that at least a first part of the image data of the subset has an intensity or textural difference with at least a second part of the image data of the subset, wherein the intensity or textural difference exceeds a predetermined threshold; and identifying a solid tumour volume as including the first part of the image data which is associated with the largest intensity values or homogeneous textures, and identifying a ground glass tumour volume by including the subset of image data and subtracting or excluding therefrom the first part of the image data associated with the solid tumour volume. This allows to analyze one or more of the solid tumour volume (STV), the ground glass tumour volume (GGTV) and the gross

tumour volume (GTV) separately, for example by applying radiomics separately to these parts of the subset of imaging data. It has been found that analyzing these parts separately improves the performance of the method in terms of prognostic value and/or classification of the lesion or neoplasm.

In some of these embodiments, the method further comprises extracting, from the harmonized three dimensional data set, a subset of voxels associated with the organ or the neoplasm. Hence, the system will deliver the subset of data containing the relevant information for the area of interest (e.g. only the data for one or both lungs, a liver, an intestine or a part of a blood vessel, or only the subset of data for a tumor). Other data from parts lying outside the area of interest may then be discarded from further analysis. This not only increases data processing and storage efficiencies, but also enables to perform more sophisticated and time consuming data analysis algorithms to be performed in a smaller amount of time.

In accordance with some embodiments, the step of image normalization includes a step of spatial resolution normalization wherein the image data is transformed for increasing or reducing an input voxel size of the input voxels such as to correspond to a standard input voxel size. In fact, in these embodiments the step of normalization includes the harmonization of voxel sizes across various images obtained from different imaging systems. In order to enable correct interpretation of imaging data from different imaging systems, harmonizing the voxel size enables harmonized interpretation of image information, such as voxel values or parameter values associated with voxels. Although the same may be achieved in absence of the step of increasing or reducing the input voxel size, the present embodiments prevent the need for applying different algorithms to each image to interpret the information from the image.

In some particular of these embodiments, the method further includes a step of receiving from the respective imaging system, image metadata indicating at least one of a spatial voxel size or a spatial voxel spacing, wherein the step of spatial resolution normalization includes adapting the at least one of the spatial voxel size or the spatial voxel spacing to a standard spatial voxel size or a standard spatial voxel spacing. For example, in order to determine the size of a tumor or neoplasm, it is important to know the voxel size and voxel spacing. Such adaptation may include resampling of the image, an interpolation of image parameters, and/or a reduction/increase of voxels by enhancement, discarding, averaging or other algorithms.

In cases wherein voxel sizes need to be adapted only slightly (e.g. from an original size between 0.7 and 1.2 mm to a new size of 1

mm), resampling the pixels to e.g. 1mm (e.g. in two or three dimensions) is a good option in terms of accuracy. However, for large sized voxels or pixels, using interpolation will not be a solution. Interpolation only fills in a gap by assuming a gradual change between two data points, but it does not add information that is not available from the measurements during imaging. In some cases this may be particularly troublesome. For example, slice thickness values may be in a range from 1 to 7 mm, and resampling from 7mm to 1mm is not possible. In these cases, use may possibly be made of a generative adversarial network (GAN) to generate new data with the same statistics as a training set in order to estimate samples in between measured data points for resampling. Naturally, this still suffers from the disadvantage that the estimated data is not obtained by measurement and thus cannot be considered real data. However, it allows to stretch the range over which resampling may be applied.

In accordance with some embodiments, the image normalization includes a step of voxel parameter value normalization, wherein for each input voxel of the plurality of input voxels at least one input voxel parameter value is normalized relative to a reference range for said voxel parameter value, for harmonizing the three-dimensional data set. One exemplaric manner of normalizing an arbitrary voxel parameter value includes determining a standard range of parameter values by determining a minimum and maximum value from the available images (or a subset thereof) in the database, and thereafter normalizing the parameter values from the respective image to be analyzed by mapping these values onto the standard range. This enables a direct comparison between the measured parameter values in the image to be analyzed with other values of images in the database. In one particular example, the image normalization includes normalizing a voxel intensity value of each respective input voxel relative to a standard voxel intensity range. In some further embodiments, the image data comprises a plurality of image frequencies in a frequency domain, the image frequencies spanning a frequency range, and wherein the image normalization includes normalizing each image frequency of the plurality of image frequency relative to the frequency range.

In accordance with some embodiments, the method further includes a contrast enhancement step prior to the image normalization, wherein the contrast enhancement step includes at least one of: windowing, a gray-level mapping, contrast stretching, histogram modification, or de-noising. Performing contrast enhancement during preprocessing reduces the general noise level in each of the images to be analyzed, and enables the comparison of grey levels between different images.

In some embodiments, the method further includes an artefact recognition step performed after the step of preprocessing, wherein the artefact recognition step includes a step of pattern recognition performed on the image data such as to identify one or more image features having a non-biological origin. Non-limiting examples of these are foreign objects (e.g. a wallet) or artefacts stemming from hardware or software restrictions upstream in the image creation pipeline. This additional step enables to correct or compensate for the existence of artefacts in the imaging data, or to enable the system to take corrective action, e.g. modify the imaging data such as to exclude the artefact data therefrom or remove/filter the artefact from the image.

In some embodiments, the method further includes analyzing the output voxels and classifying at least one of an organ or a neoplasm based on said step of analyzing. The harmonized data enables to perform automatic classification using, for example, a trained machine learning data model such as a convolutional neural network, a generative adversarial network or a random forest model.

In some embodiments, the method further includes a radiomics analysis step wherein a set of distinctive imaging features may be determined from the image data such as to form a radiomics signature. The advantages of enabling automatic radiomics analysis on a large number of images coming from a plurality of imaging systems are enormous. This enables to improve the exact definitions of radiomics signatures by being based on a larger number of images from different systems in order to provide greater prognostic value. It also enables to perform such analysis in a reproducible and harmonized manner, independent of the imaging system used. This benefits the public by enabling more uniform diagnosis.
In order to perform a method of automatic contour recognition in accordance with some of the abovementioned embodiments, use may be made by a machine learning data processing model to implement the step of contour recognition. Such a machine learning data processing model, in that case, must be trained to perform this method, and the training method for enabling automatic contour recognition greatly benefits from implementing the method of the present invention as described above, i.e. from performing the preprocessing steps of the invention or any of the particular described embodiments thereof. Therefore, in accordance with an embodiment of the present invention there is provided a method of training a machine learning data processing model for performing a step of automatic contour recognition on image data visualizing at least a part of the human or animal body and obtained from at least

one of a plurality of different imaging systems, for recognizing a contour of an organ or a neoplasm, wherein the method includes: a. receiving at least one three-dimensional data set, wherein the data set is based on processed imaging data, wherein the imaging data is obtained from at least one imaging system of the plurality of different imaging systems; b. receiving contour data for the at least one three-dimensional data set, wherein the contour data is indicative of a contour that delineates a spatial region that contains the organ or the neoplasm or other non-malignant pathological lesions; c. training of the machine learning data processing model based on the contour data received in step b. and the at least one three-dimensional data set received in step a., for enabling, after completion of the training method, the step of automatic contour recognition for producing contour data of the contour delineating the spatial region that contains the organ or the neoplasm; wherein prior to step a. the method further includes a method as described above for providing the at least one three-dimensional data set.

The above training method includes the method steps of the invention described above, and therefore forms an embodiment of the above described method of the present invention. However, at the same time, once the machine learning data processing model has been trained in accordance with this training method, it may be applied to the above mentioned embodiments wherein automatic contour recognition is performed on the imaging data. Therefore, the application of a machine learning data processing model for performing automatic contour recognition, where machine learning data processing model is trained in accordance with the abovementioned training method, forms a further embodiment of the present invention.

As may be appreciated, using a machine learning data processing model for performing the automatic contour recognition of organs or neoplasms, enables to perform this task very accurately for a large number of images coming from different imaging systems, in relatively short time. Moreover, by doing so for plurality of different imaging systems, a large database of imaging data can be build that enables to perform reliable statistical analysis such as radiomics on the basis of a large amount of data. For example, it allows the building of a centralizing database wherein imaging data from a large number of different imaging system is required, and radiomics analysis or other statistical analysis is performed on the images to provide a harmonized an reliable outcome. This greatly benefits treatment of such neoplasms in a large number of medical facilities that make use of such a centralized system. It also benefits research activities to, for example, treatment methods of certain neoplasms.

In view of the above, in accordance with some embodiments, a

method is performed for a plurality of different three dimensional datasets based on imaging data from two or more of the plurality of different imaging systems.

Similarly, a method of the present invention may also benefit from focusing the training from specific times of neoplasm or non-malignant lesions. In accordance therewith, in some embodiments the method is performed for a plurality of different three-dimensional datasets, and wherein each three-dimensional dataset of the three dimensional data sets is based on imaging data of a same specific type of neoplasm, such that the method of training is performed for enabling automatic contour recognition for delineating a neoplasm of said specific type. In accordance with some of these embodiment the specific type of neoplasm or lesion is an element of a group comprising: malignant lesions such as any of: glioblastoma multiforma; glioma grade i-iii; meningioma; head and neck cancer such as squamous cell carcinoma; esophageal cancer; lung cancer such as non-small cell lung carcinoma, small cell lung carcinoma, or lung neuroendocrine tumours; breast cancer; stomach cancer; pancreas cancer; primary liver cancer; colon cancer; rectal cancer; ovarian cancer; endometrium cancer; cervical cancer; soft tissue sarcoma; melanoma; paediatric cancers such as neuroblastoma, Wilms' tumor; brain cancers such as gliomas and medulloblastomas; osteosarcoma; Ewing's sarcoma; squamous cell carcinoma skin cancer; Merkel cell cancer; mesothelioma; pancreatic ductal adenocarcinoma; benign tumours such as any of: polyps in the colon; fibroadenomas; hepatic adenomas; fibroadenomas of the breast; uterine fibroids; angiofibromas; fibromas dermatofibroma; hemangioma; lipomas; benign parotid tumours such as pleomorphic adenomas or Warthin's tumours; premalingnant lesions or intraepithelial neoplasia such as any of: actinic keratosis; cervical dysplasia; metaplasia of the lung; leukoplakia; premalignant lesion of the pancreas such as pancreatic lesions into: intraductal papillary mucinous neoplasms (IPMN), mucinous cystic neoplasms (MCN) with varying prevalence of invasive carcinoma, or pancreatic intraepithelial neoplasia (PanIN); middle ear lesions such as Meniere disease, grey/white matter or hippocampus or brain lesions such as any of multiple sclerosis, Alzheimer disease, Parkinson diseases, cerebrovascular accident (CVA); infectious lesions such as any of: abscess, tuberculosis lesions, expansive lobar consolidation causing fissural bulging or displacement by copious amounts of inflammatory exudate within the affected organ, inhomogeneous enhancement with or without cavitation, halo sign, the air crescent sign, finger-in-glove sign , crazy-paving sign, grape-skin sign, miliary pattern, reverse halo sign, the meniscus, Cumbo sign, water lily sign, Burrow sign of paragonimiasis; lesion of idiopathic lung

fibrosis or chronic obstructive pulmonary disease (COPD), emphysema, sarcoidosis, auto-immune lung disease, pneumonia, pulmonary embolism, pleural effusion; lesions of any non infectious inflammatory diseases such as sarcoidosis including a wide spectrum of pulmonary parenchymal changes: perilymphatic micronodules, airspace opacities/consolidation (e.g. alveolar sarcoidosis), lung masses, pulmonary fibrosis, pleural effusion; rheumatological diseases such as any of: osteoarthritis, rheumatoid arthritis (RA), lupus, spondyloarthropathies, ankylosing spondylitis (AS), psoriatic arthritis (PsA), Sjogren's syndrome, gout, scleroderma, infectious arthritis, juvenile idiopathic arthritis, polymyalgia rheumatic.

The abovementioned list is not exclusive, but includes at least the most relevant and common types of neoplasms and neoplasm or non-malignant lesions pathological lesions that may be analyzed using radiological or pathological or any medical imaging and for which the present invention may therefore be applied.

In accordance with a second aspect of the present invention there is provided an imaging processing system for processing medical image data, the system comprising data communication unit for receiving image data from at least one imaging system, including microscopes, wherein the imaging data comprises data for visualizing at least a part of a human or animal body, wherein the system further comprises a controller and a memory, the memory storing instruction which, when executed by the controller, enable the controller to perform a method according to any one or more of claims 1-8, the method comprising the steps of: obtaining, by the communication unit from the at least one imaging system, the imaging data, wherein the imaging data comprises a array of input voxels; segmenting, by the controller, the image data to provide at least one two¬-dimensional data slice, wherein the data slice comprises an array of pixels; and reconstructing, by the controller, a three-dimensional data set from the at least one data slice, wherein the data set is reconstructed by the controller by providing a plurality of output voxels, each output voxel being based on an associated pixel of the at least one data slice; wherein the method, prior to the step of segmenting, further comprises a step of preprocessing of the imaging data by the controller, wherein the preprocessing at least includes: image normalization of the imaging data, including at least one step of transforming or processing of the imaging data for mapping the imaging data.

Claims

1. Image data processing method for processing, by a controller of an analysis system, imaging data obtained from an imaging system of a plurality of different imaging systems, for providing a harmonized three-dimensional data set of the image data which data set is harmonized for enabling analysis thereof independent of the respective image system of the plurality of image systems, the method comprising the steps of: obtaining, from the imaging system, the imaging data, wherein the imaging data comprises data for visualizing at least a part of the human or animal body, wherein the imaging data comprises a array of input voxels; segmenting the image data to provide at least one two¬-dimensional data slice, wherein the data slice comprises an array of pixels; and reconstructing a three-dimensional data set from the at least one data slice, wherein the data set is reconstructed by providing a plurality of output voxels, each output voxel being based on an associated pixel of the at least one data slice;
wherein the method, prior to the step of segmenting, further comprises a step of preprocessing of the imaging data, wherein the preprocessing at least includes: image normalization of the imaging data, including at least one step of transforming or processing of the imaging data for mapping the imaging data onto an image data standard for enabling said analysis.

2. Image data processing method according to claim 1, wherein the method further comprises a step of: recognizing, in the harmonized three dimensional data set, using a trained machine learning data processing model, at least one contour of an organ or a neoplasm.

3. Image data processing method according to claim 2, wherein the at least one contour comprises a contour of the neoplasm, and wherein the method includes:
associating a subset of the image data with the neoplasm and defining a gross tumour volume to include the subset of the image data; determining that at least a first part of the image data of the subset has an intensity difference with at least a second part of the image data of the subset, wherein the intensity difference exceeds a predetermined threshold;

and identifying a solid tumour volume as including the first part of the image data which is associated with the largest intensity values, and identifying a ground glass tumour volume by including the subset of image data and subtracting or excluding therefrom the first part of the image data associated with the solid tumour volume.

4. Image data processing method according to claim 2 or 3, wherein the method further comprises a step of:
extracting, from the harmonized three dimensional data set, a subset of voxels associated with the organ or the neoplasm.

5. Image data processing method according to any of the preceding claims, wherein the step of image normalization includes a step of: spatial resolution normalization wherein the image data is transformed for increasing or reducing an input voxel size of the input voxels such as to correspond to a standard input voxel size.

7. Image data processing method according to any of the preceding claims, wherein the image normalization includes a step of voxel parameter value normalization, wherein for each input voxel of the plurality of input voxels at least one input voxel parameter value is normalized relative to a reference range for said voxel parameter value, for harmonizing the three-dimensional data set.

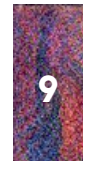
8. Image data processing method according to claim 7, wherein the image normalization includes normalizing a voxel intensity value of each respective input voxel relative to a standard voxel intensity range.

9. Image data processing method according to claim 7 or 8, wherein the image data comprises a plurality of image frequencies in a frequency domain, the image frequencies spanning a frequency range, and wherein the image normalization includes normalizing each image frequency of the plurality of image frequency relative to the frequency range.

10. Image data processing method according to any of the preceding claims, further including a contrast enhancement step prior to the image normalization, wherein the contrast enhancement step includes at least one of: windowing, a gray-level mapping, contrast stretching, histogram modification, or de-noising.

11. Image data processing method according to any of the preceding claims, further including an artefact recognition step performed after the step of preprocessing, wherein the artefact recognition step includes a step of pattern recognition performed on the image data such as to identify one or more image features having a non-biological origin.

12. Image data processing method according to any of the preceding claims, further including analyzing the output voxels and



classifying at least one of an organ or a neoplasm based on said step of analyzing.

13. Image data processing method according to any of the preceding claims, further including a radiomics analysis step wherein a set of distinctive imaging features may be determined from the image data such as to form a radiomics signature.

14. Method of training a machine learning data processing model for performing a step of automatic contour recognition on image data visualizing at least a part of the human or animal body and obtained from at least one of a plurality of different imaging systems, for recognizing a contour of an organ or a neoplasm, wherein the method includes:
a. receiving at least one three-dimensional data set, wherein the data set is based on processed imaging data, wherein the imaging data is obtained from at least one imaging system of the plurality of different imaging systems;
b. receiving contour data for the at least one three-dimensional data set, wherein the contour data is indicative of a contour that delineates a spatial region that contains the organ or the neoplasm;
c. training of the machine learning data processing model based on the contour data received in step b. and the at least one three-dimensional data set received in step a.,
for enabling, after completion of the training method, the step of automatic contour recognition for producing contour data of the contour delineating the spatial region that contains the organ or the neoplasm; wherein prior to step a. the method further includes a method according to any of the claims 1-10 for providing the at least one three-dimensional data set.

15. Method according to claim 14, wherein the method is performed for a plurality of different three-dimensional data sets based on imaging data from two or more of the plurality of different imaging systems.

16. Method according to claim 14 or 15, wherein the method is performed for a plurality of different three-dimensional data sets, and wherein each three dimensional data set of the three dimensional data sets is based on imaging data of a same specific type of neoplasm, such that the method of training is performed for enabling automatic contour recognition for delineating a neoplasm of said specific type.

17. Method according to claim 16, wherein the specific type of

neoplasm is an element of a group comprising: : malignant lesions such as any of: glioblastoma multiforma; glioma grade i-iii; meningioma; head and neck cancer such as squamous cell carcinoma; esophageal cancer; lung cancer such as non-small cell lung carcinoma, small cell lung carcinoma, or lung neuroendocrine tumours; breast cancer; stomach cancer; pancreas cancer; primary liver cancer; colon cancer; rectal cancer; ovarian cancer; endometrium cancer; cervical cancer;

soft tissue sarcoma; melanoma; paediatric cancers such as neuroblastoma, Wilms' tumor; brain cancers such as gliomas and medulloblastomas; osteosarcoma; Ewing's sarcoma; squamous cell carcinoma skin cancer; Merkel cell cancer; mesothelioma; pancreatic ductal adenocarcinoma; benign tumours such as any of: polyps in the colon; fibroadenomas; hepatic adenomas; fibroadenomas of the breast; uterine fibroids; angiofibromas; fibromas dermatofibroma; hemangioma; lipomas; benign parotid tumours such as pleomorphic adenomas or Warthin's tumours; premalingnant lesions or intraepithelial neoplasia such as any of: actinic keratosis; cervical dysplasia; metaplasia of the lung; leukoplakia; premalignant lesion of the pancreas such as pancreatic lesions into: intraductal papillary mucinous neoplasms (IPMN), mucinous cystic neoplasms (MCN) with varying prevalence of invasive carcinoma, or pancreatic intraepithelial neoplasia (PanIN); middle ear lesions such as Meniere disease, grey/white matter or hippocampus or brain lesions such as any of multiple sclerosis, Alzheimer disease, Parkinson diseases, cerebrovascular accident (CVA); infectious lesions such as any of: abscess, tuberculosis lesions, expansive lobar consolidation causing fissural bulging or displacement by copious amounts of inflammatory exudate within the affected organ, inhomogeneous enhancement with or without cavitation, halo sign, the air crescent sign, finger-in-glove sign , crazy-paving sign, grape-skin sign, miliary pattern, reverse halo sign, the meniscus, Cumbo sign, water lily sign, Burrow sign of paragonimiasis; lesion of idiopathic lung fibrosis or chronic obstructive pulmonary disease (COPD), emphysema, sarcoidosis, auto-immune lung disease, pneumonia, pulmonary embolism, pleural effusion; lesions of any non infectious inflammatory diseases such as sarcoidosis including a wide spectrum of pulmonary parenchymal changes: perilymphatic micronodules, airspace opacities/consolidation (e.g. alveolar sarcoidosis), lung masses, pulmonary fibrosis, pleural effusion; rheumatological diseases such as any of: osteoarthritis, rheumatoid arthritis (RA), lupus, spondyloarthropathies, ankylosing spondylitis (AS), psoriatic arthritis (PsA), Sjogren's syndrome, gout, scleroderma, infectious arthritis, juvenile idiopathic arthritis, polymyalgia rheumatic.

18. Image processing system for processing medical image data, the system comprising data communication unit for receiving image data from at least one imaging system, wherein the imaging data comprises data for visualizing at least a part of a human or animal body, wherein the system further comprises a controller and a memory, the memory storing instruction which, when executed by the controller, enable the controller to perform a method according to any one or more of claims 1-13, the method comprising the steps of:
obtaining, by the communication unit from the at least one imaging system, the imaging data, wherein the imaging data comprises a array of input voxels;
segmenting, by the controller, the image data to provide at least one two-dimensional data slice, wherein the data slice comprises an array of pixels; and reconstructing, by the controller, a three-dimensional data set from the at least one data slice, wherein the data set is reconstructed by the controller by providing a plurality of output voxels, each output voxel being based on an associated pixel of the at least one data slice;
wherein the method, prior to the step of segmenting, further comprises a step of preprocessing of the imaging data by the controller, wherein the preprocessing at least includes: image normalization of the imaging data, including at least one step of transforming or processing of the imaging data for mapping the imaging data onto an image data standard for enabling said analysis.

Supplementary information

Drawings and detailed description of the figures are available through the patentscope:https://patentscope.wipo.int/search/en/detail.jsf?docId=WO202112 5950&_cid=P22-LIW6Q8-48170-1, or via the QR code:

PART 4:

GENERAL DISCUSSION AND FUTURE PERSPECTIVES

CHAPTER 10:

GENERAL DISCUSSION AND FUTURE PERSPECTIVES

This final chapter provides a summary and discussion of the individual studies presented in the thesis as well as the obstacles that need to be overcome for the widespread use of Artificial Intelligence (AI) tools in the clinical setting and future prospective.
After a short introduction, this thesis, addresses the use of Machine Learning (ML) models based on Handcrafted Radiomics Features (HRFs). It aims to investigate the complementary value of HRFs to clinical features, deep learning based features and qualitative features for the task of prognosis and prediction. The second part addresses the use of Deep Learning (DL) in medical imaging, with the objective of exploring its potential to enhance clinical routines through detection and automatic segmentation. The third section highlights open-source and patented contributions to the field, resulting from the research work combined in this thesis. Finally, we examine the future prospects and outline the existing challenges that need to be addressed to facilitate the adoption of AI tools in clinical settings.

To briefly summarize, medical imaging has played a pivotal role in cancer management for several decades. It has remained a crucial aspect of cancer diagnosis and treatment, enabling healthcare professionals to detect abnormalities, determine the best available treatment and monitor disease progression (1,2). Recent advancements in imaging hardware have significantly enhanced the capabilities of medical imaging technologies. Improved image sensitivity and resolution have enabled the identification of even subtle differences in tissue densities, aiding in the early detection and accurate characterization of various medical conditions (3).
AI has emerged as a transformative force in the field of medical imaging, offering new avenues for optimizing clinical routines and providing efficient and minimally invasive clinical decision support. The integration of AI techniques into clinical medical imaging workflows has the potential to streamline processes, enhance diagnostic accuracy, and optimize treatment planning (4,5).

Applications of HRFs based ML methods in medical imaging

In Chapter 2 of the thesis we conducted a literature review on the application of quantitative AI methods in medical imaging and identified several obstacles that need to be overcome for their translation into clinical practice. Numerous studies have demonstrated promising results of HRFs based ML models in areas such as lesion classification, disease progression prediction and

prognosis prediction (6-8). Nevertheless, the application of HRF-based approaches comes with its own set of challenges that can have a profound impact on the reproducibility of HRF-based models. These challenges primarily revolve around the variability encountered in MI data acquisition and reconstruction, stability, and reproducibility of HRFs (9,10).
Aiming to address these issues, we proposed a framework to improve the robustness of radiomics analysis. Furthermore, we suggested that development of standardized protocols are imperative to ensure the robustness and reliability of HRF-based methods in clinical practice. Addressing these challenges remains a critical area of ongoing research that is essential for the successful integration and advancement of HRF-based models.

In Chapter 3 we aimed to explore the potential of non-invasive quantitative and qualitative medical imaging features in a heterogeneous Glioblastoma (GBM) patient cohort to predict prognosis and clinically relevant molecular markers. We used a cohort of 188 GBM patients for the analysis. The data included T1 + Gadolinium and T2- weighted Magnetic Resonance Imaging (MRI) from different centers, molecular features (isocitrate dehydrogenase-mutation; 06-methylguanine-DNA-methyltransferase-methylation; epidermal growth factor receptor amplification), clinical features, and qualitative Visually Accessible Rembrandt Images features. A pre-processing routine was suggested and applied to the imaging data to address the variation in MRI data across different centers. To our knowledge, this study is the first to combine both quantitative and qualitative MRI features with clinical features to assess their combined effect on prognosis and prediction. The results of this study showed that the addition of quantitative HRFs features complemented the model based on the clinical and qualitative features for prognosis. It had the most promising performance and was robust across both GBM cohorts. However, no complementary value of the HRFs for predicting molecular features was identified. Which we have also observed in other published research (11,12).

In Chapter 4, we conducted a comprehensive study comparing and integrating a HRFs-based ML model with a DL model to predict adverse radiation effects (ARE) in patients with brain metastasis who underwent radiotherapy using pre-treatment brain magnetic resonance imaging (MRI) data. To address the variability in MRI data we employed multiple pre-processing strategies with various methods including white-stripe, z-score, and CLAHE. We found that the combined approach of utilizing radiomics and DL models

outperformed individual models in predicting ARE. To the best of our knowledge it was the first study to utilize pre-treatment brain MRI images for predicting the risk of ARE, integrating radiomics with DL predictions to achieve more robust and accurate results. However, despite these encouraging results, the prediction score obtained in our study is not yet sufficiently high to be confidently used for treatment planning. Further investigation is necessary, particularly with a dataset containing a higher proportion of the scans representing patients with ARE.

Applications of Deep learning in medical imaging

In Chapter 5 we presented an extensive literature review that explored various clinical segmentation approaches, ranging from manual to fully automatic methods. These approaches encompassed diverse techniques from the spectrum of present clinical segmentation approaches, including fully automatic deep learning models, few-shot learning models capable of learning from limited data, transfer learning and fine-tuning, and interactive methods. In this chapter, we provided an explanation of the underlying principles behind each approach and discussed their respective advantages and limitations. Additionally, we proposed the optimal utilization scenarios for each method, taking into consideration the clinical context and data availability.

Chapter 6 is a centerpiece of this thesis. This chapter incorporates a range of research objectives and serves as foundation for the development of clinical software designed for the automatic segmentation of non-small cell lung cancer (NSCLC) on CT images, which subsequently obtained CE marking.
Within this chapter, we present a fully automated pipeline designed for the detection and volumetric segmentation of NSCLC on CT images. To ensure the robustness and effectiveness of our approach, we collected a large multi-centric dataset, consisting of 1328 pre-treatment thoracic CT scans from patients diagnosed with NSCLC. This dataset served for developing and externally validating our approach.
To overcome the challenges posed by the diversity of imaging acquisition and reconstruction protocols in CT data, we proposed and implemented a multi-step pre-processing routine, which included lung extraction, spatial normalization, and image intensity normalization. By incorporating these measures, we aimed to address the inherent heterogeneity present in CT data, thereby

enhancing the reliability of produced segmentations.
In order to provide a more detailed overview of the method's performance, we conducted a comprehensive quantitative analysis that considered multiple factors associated with the CT scans and the cancers. These factors included variables such as image slice thickness, tumor size, image interpretation difficulty, and the location of the tumor. Additionally, to facilitate comparisons with other published studies, we expanded our set of quantitative metrics to include recently published measures, such as Added Path Length and Surface DICE.
However, given the significant intra-/inter-observer variability observed in the ground truth segmentations, which was also confirmed during the prospective in-silico clinical trial conducted as part of this study, we felt it necessary to go beyond reporting solely quantitative metrics. To gain insights into the qualitative performance of our method, we recruited 40 participants, including radiologists and radiation oncologists, and developed a specialized software tool. This tool enabled the participants to visually assess the segmentations side by side, without indicating which segmentations were manually created and which ones were generated automatically.
Remarkably, results of the qualitative assessment revealed that, on average, even among the group of radiologists and radiation oncologists, the automatic segmentations were preferred in 56% of the cases.
Additionally, we assessed the prognostic capability of the automatic contours by applying the RECIST criteria and measuring tumor volumes. Notably, our method's segmentations stratified patients into low and high survival groups with a higher level of significance compared to methods relying on manual contours. The results of the qualitative evaluation, along with the promising quantitative performance that aligned with the observed inter/intra-contouring variability, emphasized the potential clinical significance of our automated approach. These results served as a catalyst for the software development process for a clinically usable version of this approach.
The work conducted in this chapter received numerous recognitions, including the best research presentation in 2019 at the MUMC+ research day (Maastricht, Netherlaands) and receiving the ESTRO Jack Fowler award in 2021 (Madrid, Spain).
The entire code associated with this study has been openly published and is available for access through GitHub (13).

Chapter 7 presents a deep learning (DL) algorithm designed for the

identification of metastatic bone lesions on bone scintigraphy images. The data used in this study was collected from three medical centers, providing a diverse and representative sample. A total of 1367 images from two different centers were combined and used to train and validate the model. Another 998 images from a different medical center was used as external test to assess the algorithm's generalizability. Our model achieved a promising quantitative performance comparable to that of nuclear physicians, even in the absence of background knowledge about the patients' medical history. Additionally, we have performed an in-silico clinical trial where we developed an application allowing for collecting nuclear medicine physician's feedback based on the visual assessment of bone scintigraphy scans. We have recorder the feedback and the time spent per each scan. We used this data to compare the quantitative performance along with the time spent to the performance of our method. We have shown that our model solely based on the image has outperformed the nuclear medicine physicians in the similar setting (without the access to clinical information about the patient), at the same time being significantly faster.

To enhance the explainability of the DL algorithm we have utilized the Grad-CAM (14) method that allowed us to highlight the regions within the image that contributed to the positive decision. By providing visual cues, we make the algorithm's decision-making process more transparent and interpretable, enabling clinicians to gain insights into the underlying features influencing the classification.

Open source and patented contributions to the field

In Chapter 8 of this thesis, we introduce an open-source initiative aimed at enhancing reproducibility in Quantitative Medical Imaging (QMI) research. This initiative seeks to address two present challenges prevalent in the QMI research: the lack of methodology standardization and the diversity of in-house data curation and exploration methods. Data curation and pre-processing of medical images are crucial and time-consuming steps in the QMI workflow. The quality of these steps significantly impacts the resulting model performance and reproducibility (15,16). The data curation process incorporate multiple tasks, including image format conversion, outlier detection, verification of different image DICOM tags, and handling of missing or overlapping slices. Currently, these steps are often carried out using in-house developed software or individually

implemented by researchers, without undergoing community scrutiny and may introduce errors. Furthermore, the implementation of these steps can vary methodologically. To address these issues we proposed an open-source standardized implementation of these data curation methods, which can be accessed and validated by the medical imaging community. Furthermore, the toolbox offers functionality for conducting exploratory analysis, which is vital in the development of a radiomics workflow. It is important to explore the potential of image processing techniques to enhance the extraction of radiomics features. The Image Biomarker Standardization Initiative (IBSI) underscores the significance of image processing in this context (17). Additionally, performing an exploratory analysis on handcrafted radiomics features and visualizing discriminatory statistics is critical for gaining insights and understanding the data better. By adopting the use of open-source tools, reviewed by community we aim to increase the transparency and reliability of the data curation process, ultimately contributing to improved reproducibility in QMI studies. We also believe that proposed toolbox will benefit researchers without strong programming backgrounds and lower the entry barrier for the students who want to start their journey in QMI research.

Chapter 9 is a summary of the patent issued to Maastricht University for the work on Image data processing method, method of training a machine learning data processing model and image processing system. The patent claims were made as a part of the clinical software development process for the automatic segmentation of NSCLC on CT. The work described in the chapter 6 served as a foundation for the prototype of the clinical software (DUNE.AI/ DUNE.BIO). By following the clinical needs and regulations, the resulted web based prototype was developed for both local and server deployment so that the data would not leave the clinic side. Moreover, user privacy and anonymization functionality was implemented for the processed DICOM images (Figure 1). Extra functionality including quick check of the generated segmentations in 3D and range of editing tools in 2D was implemented alongside, so that the clinicians could quickly adjust the results (Figure 2). The software has automatically calculated multiple features derived from the tumor segmentation, such as RECIST and volumetric RECIST measurements that could be used for the tumor response to treatment evaluation (Figure 2).

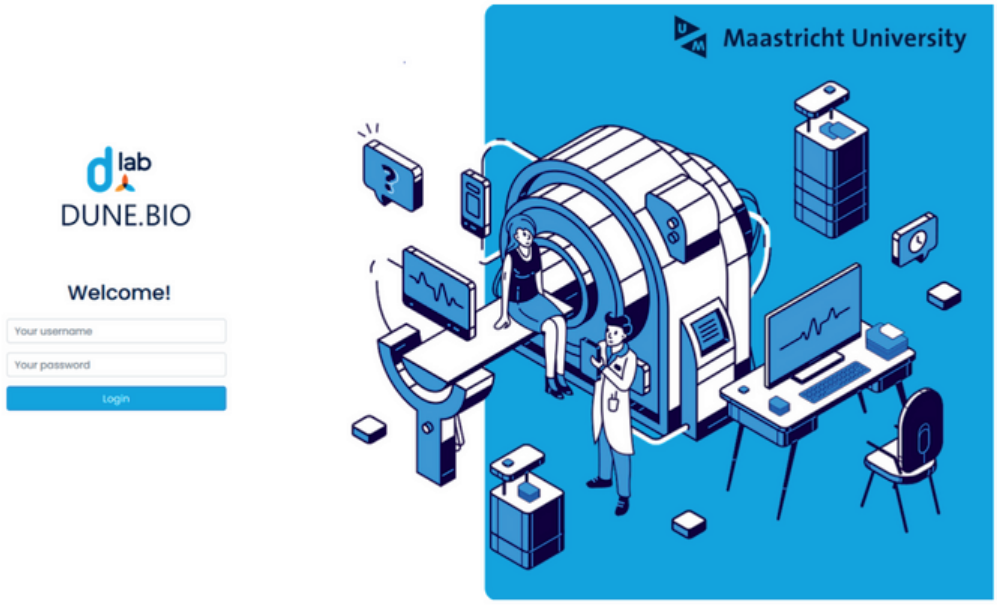
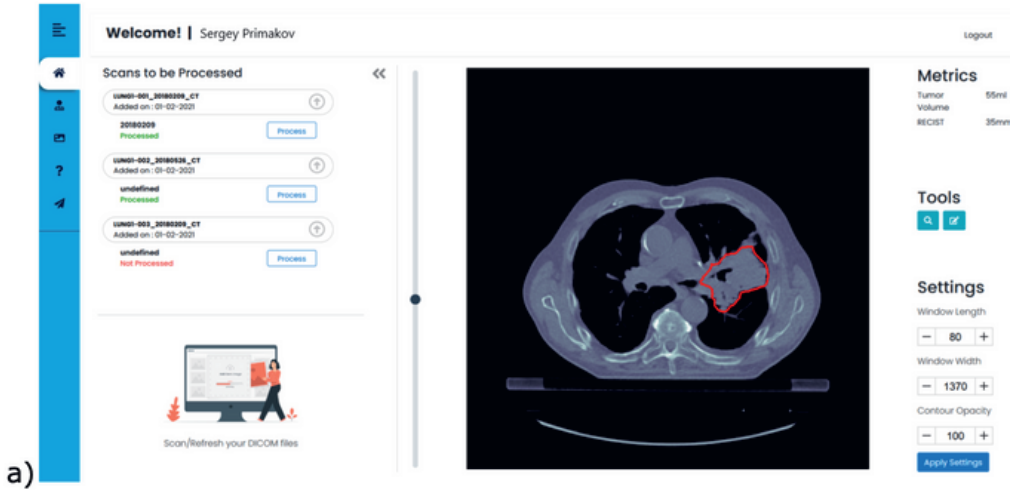


Figure 1 Login screen of the developed automatic NSCLC detection and segmentation software (DUNE.BIO)



a)

b)

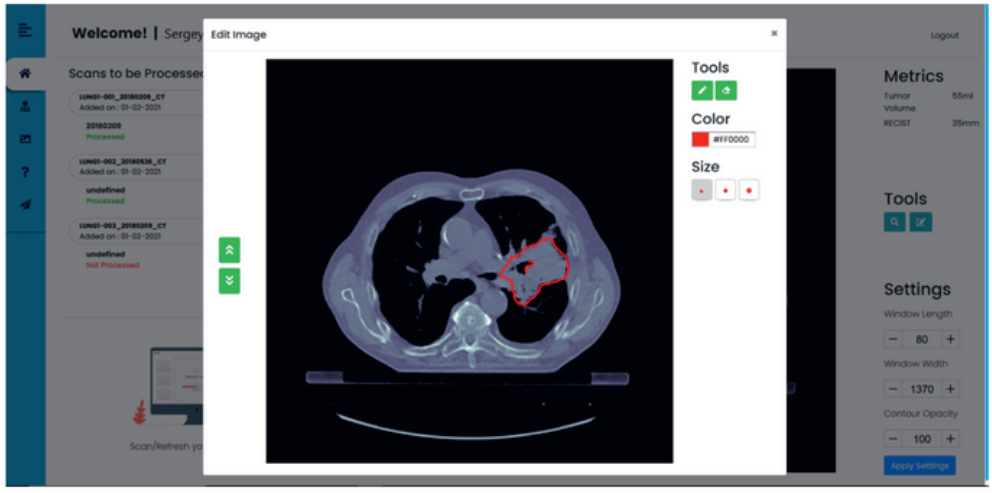


Figure 2 a) Patient view and b) quick editing screens of the automatic NSCLC detection and segmentation software (DUNE.BIO)

Current challenges in translating the research into clinical practice

Bringing medical imaging research to clinical implementation poses several challenges that need to be addressed for successful integration into healthcare practice. These challenges arise from various aspects, including technical, regulatory, and data.

One of the primary challenges, as also outlined in Chapter 2 is the need for robust and reliable validation of the imaging algorithms and methodologies. While research studies often demonstrate promising results in controlled environments, the translation of these findings to real-world clinical settings requires rigorous validation. Clinical implementation necessitates the evaluation of algorithms on diverse patient populations, with variations in imaging equipment, protocols, and disease presentations. This validation process involves addressing issues related to data heterogeneity, generalizability, and the establishment of clinically relevant performance metrics.

Second challenge in bringing medical imaging research to clinical implementation is the variability in data. Medical images

are originally intended for human interpretation and are reconstructed with specific parameters that can vary a lot across different scanner manufacturers. To address this challenge, the development of an open-source reconstruction protocol specifically designed for the AI applications in medical imaging research could be a significant step forward. Such protocol would aim to establish a unified and standardized approach to image reconstruction, independent of the manufacturer. Such an initiative would make a significant stride in fighting the variability issues and promote consistency in image data, facilitating the development and deployment of AI algorithms in clinics.

Another significant challenge is the integration of AI algorithms into existing clinical workflows. Most of the radiotherapy departments typically operate within complex systems such as Picture Archiving and Communications System (PACS) and electronic medical records (EMR), and incorporating new AI tools in a seamless way with existing infrastructure could be a real issue. The unified format adopted by major PACS that would allow the integration of AI research applications that qualify pre-defined security requirements in a clinical trial setting would be highly desired. Otherwise researchers need to take a long route by developing standalone software applications including GUI efforts, and a lot of platform work e.g. security and communication with databases to support the AI algorithm.

Additionally, the regulatory landscape surrounding medical imaging technologies is a challenging and location sensitive topic that could exclude the possibility of using some of the available tools and services e.g. cloud based services. The development and deployment of AI-based algorithms for medical imaging are subject to regulatory approval, which involves demonstrating safety, efficacy, and adherence to relevant standards. Meeting these regulatory requirements can be a time-consuming and resource-intensive process, involving extensive documentation, clinical trials, and collaboration with regulatory bodies.

Privacy and data security also present challenges in the implementation of medical imaging research applications. Ensuring data anonymization, secure storage, and compliance with ethical guidelines is crucial to protect patient privacy while enabling robust research and development.

Future prospects

In the coming years, AI is expected to be a major player in advancing medical imaging technologies, revolutionizing healthcare practices, and improving patient outcomes. Here are just some of the prospectives for the AI in the medical imaging field:

Continuous improvement of the AI technology

One of the emerging directions currently is the vision transformers for object detection, image classification, and image segmentation. Vision transformers, also known as ViTs, are a groundbreaking development in the field of computer vision18. Unlike traditional convolutional neural networks (CNNs), which is still the go-to architecture for most of the MI tasks, vision transformers offer a novel approach by utilizing the power of self-attention. These ViTs, inspired by their success in natural language processing, bring the benefits of sequential modeling to the visual domain. The self-attention mechanism allows the model to capture global dependencies between different regions of the input image, enabling it to extract contextual relationships and long-range dependencies. This is in contrast to CNNs, which rely on local receptive fields and convolutional filters. Numerous studies, including the recent Meta paper on SAM, have demonstrated impressive performance, challenging the current benchmark set by CNNs (18–21,22).

More affordable and improved diagnosis and decision-making

AI algorithms can analyze medical images with remarkable speed and accuracy, aiding radiologists and other healthcare professionals in diagnosing diseases at an early stage (23,24). AI systems can also assist radiologists in a range of other tasks including the detection, classification and segmentation of various conditions, including cancers, diabetes, cardiovascular diseases and more (25). The widespread use of AI powered diagnostics can help to improve early detection and lead to timely interventions and better treatment outcomes. It can also increase the quality and make such diagnostic tools available for the people who cannot afford it due to financial or location-based barriers.

Radiology Workflow Optimization
As we demonstrated in Chapter 6, 7 AI can streamline radiology workflows by automating time-consuming tasks, such as image curation, lesions and OAR segmentation, and automatic measurements. By automating these processes, AI algorithms can reduce the radiologist's workload, allowing them to focus more on interpreting results, communicating with patients, and making crucial clinical decisions. This can enhance efficiency, speed up diagnosis, and reduce the chances of human error.

Personalized medicine
Once there is enough structured data, AI can enable personalized medicine by analyzing medical images along with other clinical and patient data, such as genetic information, electronic health records, and lifestyle factors. By considering the cross-disciplinary data, AI algorithms can predict individual patient responses to specific treatments and help find better-tailored therapies, which has the potential to optimize treatment plans, minimize side effects, and enhance patient care (26).

Conclusion

In this thesis, we conducted multiple studies to explore various AI applications using different medical imaging modalities and clinical problems. We placed particular emphasis on ensuring the robustness of our models and the reproducibility of our results. To achieve this, we utilized harmonization and pre-processing procedures for the imaging data and shared our code and results open source. Regarding the application of ML models based on Handcrafted Radiomics Features (HRFs), we demonstrated their potential as complementary approach to interdisciplinary methods. However, challenges persist in this area. In the context of DL applied to medical imaging, we demonstrated its ability to perform at a level comparable to that of clinicians for multiple applications. This was supported by quantitative and qualitative metrics obtained during in-silico clinical trials. We also highlighted the potential of DL to assist medical professionals and enhance clinical routines. Additionally, we have outlined the challenges encountered during the implementation of the clinical AI based software. Some of these challenges remain, preventing the sustainable implementation of the AI applications in clinical settings. While ongoing research is steadily pushing the barriers on the AI side, the change should also

happen on the clinical infrastructure side. Much effort should be done to support the integration of AI tools and make them available in the current clinical workflows. By collaboratively addressing this challenges, we can bring AI-based personalized medicine closer to becoming a reality.

References

1. Shreve, P. & Townsend, D. W. Clinical PET-CT in Radiology: Integrated Imaging in Oncology. (Springer New York, 2010).
2. Dynamic Contrast-Enhanced Magnetic Resonance Imaging in Oncology. (Springer Berlin Heidelberg).
3. Zaidi, H. & Alavi, A. Recent Advances in Imaging with PET, CT, and MR Techniques, An Issue of PET Clinics EBook. (Elsevier Health Sciences, 2020).
4. Trimpl, M. J., Primakov, S. & Lambin, P. Beyond automatic medical image segmentation—the spectrum between fully manual and fully automatic delineation. Phys. Med. Biol. (2022).
5. Ibrahim, A. et al. Radiomics for precision medicine: Current challenges, future prospects, and the proposal of a new framework. Methods 188, 20–29 (2021).
6. van Timmeren, J. E. et al. Survival prediction of non-small cell lung cancer patients using radiomics analyses of cone-beam CT images. Radiother. Oncol. 123, 363–369 (2017).
7. Keek, S. A. et al. A Prospectively Validated Prognostic Model for Patients with Locally Advanced Squamous Cell Carcinoma of the Head and Neck Based on Radiomics of Computed Tomography Images. Cancers 13, (2021).
8. Sanduleanu, S. et al. Tracking tumor biology with radiomics: A systematic review utilizing a radiomics quality score. Radiother. Oncol. 127, 349–360 (2018).
9. Zhao, B. et al. Reproducibility of radiomics for deciphering tumor phenotype with imaging. Sci. Rep. 6, 23428 (2016).
10. Midya, A., Chakraborty, J., Gönen, M., Do, R. K. G. & Simpson, A. L. Influence of CT acquisition and reconstruction parameters on radiomic feature reproducibility. J Med Imaging (Bellingham) 5, 011020 (2018).
11. Kickingereder, P. et al. Radiogenomics of Glioblastoma: Machine Learning-based Classification of Molecular Characteristics by Using Multiparametric and Multiregional MR Imaging Features. Radiology 281, 907–918 (2016).
12. Gupta, A. et al. Pretreatment Dynamic Susceptibility Contrast MRI Perfusion in Glioblastoma: Prediction of EGFR Gene

Amplification. Clin. Neuroradiol. 25, 143–150 (2015).
13. Primakov, S. DuneAI-Automated-detection-and-segmentation-of-non-small-cell-lung-cancer-computed-tomography-images: Repository supporting the original
research paper in Nature Communications (Primakov et al. 2022). (Github).
14. Selvaraju, R. R. et al. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. arXiv [cs.CV] (2016).
15. Fave, X. et al. Impact of image preprocessing on the volume dependence and prognostic potential of radiomics features in non-small cell lung cancer. Transl. Cancer Res. 5, 349–363 (2016).
16. Hosseini, S. A. et al. The impact of preprocessing on the PET-CT radiomics features in non-small cell lung cancer. Frontiers in Biomedical Technologies 8, 261–272 (2021).
17. Zwanenburg, A. et al. The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping. Radiology 295, 328–338 (2020).
18. Ma, J. et al. Visualizing and Understanding Patch Interactions in Vision Transformer. IEEE Trans Neural Netw Learn Syst PP, (2023).
19. Springenberg, M. et al. From modern CNNs to vision transformers: Assessing the performance, robustness, and classification strategies of deep learning models in histopathology. Med. Image Anal. 87, 102809 (2023).
20. Zhang, Z. et al. TC-Net: A joint learning framework based on CNN and vision transformer for multi-lesion medical images segmentation. Comput. Biol. Med. 161, 106967 (2023).
21. Oh, Y., Bae, G. E., Kim, K.-H., Yeo, M.-K. & Ye, J. C. Multi-Scale Hybrid Vision Transformer for Learning Gastric Histology: AI-Based Decision Support System for Gastric Cancer Treatment. IEEE J Biomed Health Inform PP, (2023).
22. Kirillov, A. et al. Segment Anything. arXiv [cs.CV] (2023).
23. Grewal, M., Srivastava, M. M., Kumar, P. & Varadarajan, S. RADnet: Radiologist level accuracy using deep learning for hemorrhage detection in CT scans. in 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018) 281–284 (2018).
24. Primakov, S. et al. OC-0557 AI-based NSCLC detection and segmentation: faster and more prognostic than manual segmentation. Radiother. Oncol. 161, S441–S443 (2021).
25. Vaidyanathan, A. et al. Deep learning for the fully automated segmentation of the inner ear on MRI. Sci. Rep. 11, 2885 (2021).
26. Johnson, K. B. et al. Precision Medicine, AI, and the Future of Personalized Health Care. Clin. Transl. Sci. 14, 86–93 (2021).

APPENDICES

SUMMARY

Summary

This thesis explored applications of AI in medical imaging for enhancing and streamlining cancer management. It comprises a composition of comprehensive review articles as well as research studies using various medical imaging data. Additionaly, it outlines the current challenges encountered when implementing AI in clinical settings and explores future prospects in the field.

Part 1: Applications of HRFs based ML methods in medical imaging

Part 1 starts with an introduction through **Chapter 2** where application of HRFs based models were explained and discussed, along with the challenges, limitations, and future prospects. **Chapter 3** utilizes the HRF in conjunction with clinical, molecular, and qualitative imaging data to explore the integrated performance of these features for prediction and prognosis in patients with Gliablostoma. **Chapter 4** continues the investigation of complimentary value of HRFs extracted from the MRI, it compares and combines handcrafted feature based models with models based on the automatically extracted deep features for predicting the ARE.

Part 2: Applications of Deep learning in medical imaging

Part 2 also starts with an introduction, **Chapter 5** explores existing methods for the medical imaging segmentation, ranging from fully manual to fully automatic. It provides an in depth explanation for the methods behind each solution and suggests the best suitable option based on the clinical scenario. **Chapter 6** continues the research of automatic medical imaging segmentation methods using AI. It incorporates multiple research objectives for enhancing the NSCLC management. It demonstrates that AI can be used for automatic NSCLC detection and segmentation on CT with the performance comparable to the manual annotators. It also serves as the evidence that AI could be used to streamline and enchance the radiotherapy workflows. **Chapter 7** shifts application of AI in medical imaging from automatic segmentation to classification. It explores the use of DL for detection of bone metastases on the bone scintigraphy images. It demonstrated the potential of AI to be used as clinical decision aid tool that could minimize the time needed by a nuclear physician to assess bone scans.

Part 3: Open source and patented contributions to the field

Part 3 shifts the focus from the research to the development of AI based and auxiliary applications. **Chapter 8** describes an open source initiative to improve the reproducibility of quantitative medical imaging research through standardisation of data curation and pre-processing. The developed python package provides various functionality for handling medical and clinical data including data exploration, curation, outlier detection and verification. **Chapter 9** presents the summary of the patent for the work on image data processing method, method of training a machine learning data processing model and image processing system. The patent claims were made as a part of the clinical software development process for the automatic segmentation of NSCLC on CT.

Part 4: General discussion and future perspectives

Part 4 and **Chapter 10** addresses the present challenges in the integration of the AI based applications in the clinic and concludes the thesis by discussing the future prospects.

IMPACT PARAGRAPH

Impact paragraph

This thesis explored various applications of AI in medical imaging for enhancing and streamlining cancer management through detection, localization, prognosis, outcome prediction, and automatic cancer segmentation. The comprehensive review articles included in the thesis provide insights into the current state of AI applications in medical imaging field, along with the existing challenges and future prospects. The methods, findings and results provided in these thesis have been externally validated, peer reviewed, and openly shared with the community to insure their reproducibility and robustness.

Scientific impacts

In **Chapter 2** we proposed a new framework that guides development of robust HRFs pipelines. In the **Chapter 3** and **4** we have demonstrated that HRFs extracted from the MRI images have complimentary value to the ML models based on the clinical, molecular, qualitative or deep features for prognosis and predicting. In **Chapter 5** we reported extensively on the different segmentation methods currently used in the medical imaging field. In **Chapters 6, 9** we provided multiple research endpoints. Firstly, we demonstrated that AI based NSCLC automatic segmentation could reach the quantitative performance comparable to clinicians. Secondly, we performed an insilico clinical trial where we estimated the variance of manual contouring of NSCLC and showed that segmentations produced by our method were preferred by the group of radiologists/radiation oncologists more often than manual segmentations. We have also estimated the tolerance parameter for the manual segmentation task of NSCLC allowing for computation of variance avare Surface DICE metric in further research. The work in chapter 6, 9 was used as a foundation for the development of the clinical application taht had received a CE marking. Lastly we have shared all the model data open source allowing the possible transfer learning applications. In **Chapter 7** we shown the potential of AI based applications to improve clinical decision aid tools, increase diagnostic specificity and minimize the time needed by a nuclear physician to assess bone scintigraphy scans. In **Chapter 8** we developed a precision medicine toolbox that aims to increase the reproducibility of quantitative medical imaging research through standartisation of data curation and pre-processing.

Societal impacts

Cancer has a major impact on society. Although the overall mortality rate has declined, it remains a leading cause of death worldwide. Advancements in cancer managment are crucial to maintain the decrease in mortality rates. Currently clinical decisions are still subjective and prone to variability (1,2). They depend on multiple factors including the level of expertise and experience of the clinicians, geographical location and clinical infrastructure. AI can help optimize the current cancer management workflows, assist clinicians with the objective decision support and make the advanced cancer management tools available for the regions with poor clinical infrastructure.

The research outcomes, findings, and tools that are presented and implemented in the **Chapter 3,4** have the potential to guide researchers and clinicians in leveraging AI technology for more efficient and effective cancer management. **Chapter 6** presents an open source AI based solution for automatic detection and segmentation of NSCLC, it can be used to assist the clinicians in detecting and segmenting the NSCLC on CT, decreasing the time and effort needed for this laborious process. It could also assist in evaluating the tumor response to treatment through automatic calculation of RECIST and volumetric RECIST. Chapter 7 proposes a method that could benefit nuclear medicine clinicians in detecting the metastatic spots on the bone scintigraphy scans. Once implemented it could help in reducing the time needed for the initial assessment and also be used as a radiologist training support tool. **Chapter 8 and 9** highlighted some of the contribution to the open science and a real clinical application for NSCLC segmentation, that subsequently received a CE marking.

The publications presented in this thesis along with the analyses and code, were peer reviewed and published in reputable open acess journals, including Nature communications, Cancers, Physics in Medicine & Biology, etc. This should increase the transparency and transmitabillity of our research. The research conducted in this thesis has been extensively shared and discussed with medical imaging and radiology community at various national and international conferences, including Big Data For Imaging conference (2018), GROW science day of Maastricht University (2018, 2020), Maastricht University Medical Center MUMC+ science day (2019), Dutch Week wan de longen (2019), the European Congress of Radiology ECR (2020) and the European SocieTy for Radiotherapy and Oncology ESTRO (2021). Additionaly the

work presented in this thesis has received recognition through multiple awards, including the best research presentation in 2019 at the MUMC+ research day (Maastricht, Netherlaands) and receiving the ESTRO Jack Fowler award in 2021 (Madrid, Spain) (3).

References

1. Berry, S. L., Boczkowski, A., Ma, R., Mechalakos, J. & Hunt, M. Interobserver variability in radiation therapy plan output: Results of a single-institution study. Pract. Radiat. Oncol. 6, 442–449 (2016).
2. Bond, M. J. G. et al. Intersurgeon Variability in Local Treatment Planning for Patients with Initially Unresectable Colorectal Cancer Liver Metastases: Analysis of the Liver Expert Panel of the Dutch Colorectal Cancer Group. Ann. Surg. Oncol. (2023) doi:10.1245/s10434-023-13510-7.
3. University award.
https://www.estro.org/About/Hall-of-fame/University-award.

ACKNOWLEDGMENTS

Acknowledgements

This PhD journey has been a substantial chapter in my life that undeniably transformed me, influencing the way I perceive, assess, and respond to various life scenarios. As any journey it had its inevitable highs and lows, and I can't be more grateful for all the invaluable support and trust from those who accompanied me on this path.

First of all, I would like to thank my promoter Prof. Dr. Philippe Lambin for giving me the opportunity to pursue my Ph.D. degree at Maastricht University and for serving as a continuous source of motivation throughout all of my research. Your visionary input, support and inspiration shaped my academic interests and helped me grow as a researcher.

I would like to extend my gratitude to my co-promoter Dr. Henry Woodruff and Dr Arthur Jochems, for all the support and supervision I have received on this journey and for making this time fun and memorable through lots of social activities.

Prof. Dr. Andre Dekker, Prof. Dr. Regina Beets-Tan, Prof. Dr. Wiro Niessen, Dr. Alberto Traverso and Dr. Wouter van Elmpt I would like to thank you for taking the time to review my thesis. Maikel and Ann I would like to thank you for co-authoring my first research paper, and providing me with lots of valuable medical insights. It was a pleasure to work with you both.

To all of my D-lab friends and colleagues in Maastricht: Renee, Yvonka, Janita, Ralph, Robin, Avisheek, Zohaib, Shruti, Yi, Monideepa, Relinde, Turkey, Yousif, William, Lisa, thank you for making my PhD experience brighter, sharing struggles and celebrating all the beautiful moments together. Thank you Avi for hosting countless BBQs and dinners that we shared. Thank you, Will, Lisa, Yvonka, and Robin for all the activities that we did. My journey wouldn't be nearly as enjoyable without you!

To all of my "Predict" friends and colleagues: Anke, Michael, Patrick, Francesco, Akshaya, Sithin, Nora, and Fadila from the very first meeting I got a strong sense of community with you, it was a pleasure to work together and come visit you during the secondments. I hope to see you more often in the coming years!

Prabash, Binosha, Kejsi, Nilofar, Sean, Rick, Slava and Nastya thank

you for being part of my Maastricht life. It's always a pleasure to meet with you guys! Your positive attitude helped me to stay optimistic and face low moments with a smile.

Abdalla, Iva, Ivan, Manon, Simon, your contributions to my journey have been immense, and I want to express my heartfelt gratitude to all of you! You are the ones who made Maastricht feel like home for me. Abdalla you were the first friend that I made in Maastricht. I can't even count the number of events and activities that we shared since then. It was a great pleasure to also know you as a researcher and collaborate with you on multiple projects. Iva thank you for being an amazing, bright, social person that you are. Your positive energy, hospitality and drive for adventures has really filled our lives with hundreds of beautiful moments. Your support and care had a tremendous impact on mine, Kate's and Baas' life in Maastricht. Ivan we started our PhD journeys almost at the same time, sharing the same struggles of immigration, integration and later renovation. Your and Natalya's support and motivation helped me to stay afloat in tough times, grow and succeed. I can't be more grateful to both of you! Manon, thank you for being a great friend and an amazing colleague. I am very impressed by your dedication, professionalism and insane organisation skills. As a colleague, you motivated me to be more productive and organised and had a major impact on my personal growth. As a friend we shared so many great moments with you and Alex and I'm eagerly looking forward to the next adventures ahead. Simon, I was really happy to meet you and be your friend since then. I love your dedication and approach to things, and sometimes even a bit of "the dude" attitude towards life's challenges. It really teached me to take some things easier and do it: 'stap voor stap'.

To all of my university friends: Vsem yashuram bolshoi privet, vseh obnyal pripodnyal!

Мама, Папа, Саша, без вас я бы не дописывал сейчас свою диссертацию. Вы сделали меня тем кто я есть и помогли пройти через огромное количество сложностей и всевозможных жизненных ситуаций. Хочу выразить огромную благодарность за вашу поддержку и заботу!

Finally Kate, I consider myself incredibly fortunate to have you in my life. You are making my life complete, and I wouldn't have reached this point without your infinite support and care.

CURRICULUM VITAE

Sergey Primakov was born in April 1993 in Potsdam, Germany. After finishing high-school at Moscow State school no. 632 he was admitted to Bauman Moscow State Technical University (BMSTU) through the scientific olympiad "Step into the future".

In 2016, Sergey graduated BMSTU with the Biomedical Engineering diploma and started working as an engineer in Moscow Scientific Center of Otorhinolaryngology.

In 2018 Sergey moved to Maastricht, Netherlands to pursue his PhD degree at the Precision Medicine department in Maastricht University. During his research Sergey focused on developing computer vision algorithms for improving cancer management through use of Artificial Intelligence (AI) in Medical Imaging (MI).

While working as a PhD researcher, Sergey was involved in every step of the research process, from data collection and data curation to pre-processing, in-silico clinical trial management, AI model development, validation, testing, and a Proof Of Concept (POC) software development. Sergey has been actively presenting his research on the local and international conferences including Week van de longen, GROW SD, MUMC+ RD, ECR, ESTRO, and published his research projects open source. In the course of his PhD journey, Sergey engaged in several secondments, collaborating as a visiting researcher with both academic institutions (RWTH Uniklinik, DKFZ) and industry partners (Mirada).

Additionally, during his research tenure at Maastricht University, Sergey actively participated in educational activities. This included conducting seminars and workshops, supervising bachelor students, and creation of course materials for Big Data for Imaging 2018 (BD4I) and Artificial Intelligence for Imaging in 2019 (AI4I).

Awards

- 2019 - Best research presentation: Using Deep Learning for nodule detection and automatic segmentation of lung cancer.; Maastricht UMC+ Research Day 2019. Maastricht, Netherlands.

- 2021 - Jack Fowler Award: AI-based NSCLC detection and segmentation: faster and more prognostic than manual segmentations.; ESTRO 2021. Madrid, Spain.

LIST OF PUBLICATIONS

1. Sergey P Primakov, Abdalla Ibrahim, Janita E van Timmeren, Guangyao Wu, Simon A Keek, Manon Beuque, Renée WY Granzier, Elizaveta Lavrova, Madeleine Scrivener, Sebastian Sanduleanu, Esma Kayan, Iva Halilaj, Anouk Lenaers, Jianlin Wu, René Monshouwer, Xavier Geets, Hester A Gietema, Lizza EL Hendriks, Olivier Morin, Arthur Jochems, Henry C Woodruff, Philippe Lambin (2022). Automated detection and segmentation of non-small cell lung cancer computed tomography images;
Nature communications: 3423 (2022).

2. Maikel Verduin* Sergey Primakov*, Inge Compter, Henry C Woodruff, Sander MJ van Kuijk, Bram LT Ramaekers, Maarten Te Dorsthorst, Elles GM Revenich, Mark Ter Laan, Sjoert AH Pegge, Frederick JA Meijer, Jan Beckervordersandforth, Ernst Jan Speel, Benno Kusters, Wendy WJ de Leng, Monique M Anten, Martijn PG Broen, Linda Ackermans, Olaf EMG Schijns, Onno Teernstra, Koos Hovinga, Marc A Vooijs, Vivianne CG Tjan-Heijnen, Danielle BP Eekers, Alida A Postma, Philippe Lambin, Ann Hoeben (2021). Prognostic and predictive value of integrated qualitative and quantitative magnetic resonance imaging analysis in glioblastoma;
Cancers, 13(4), 722 (2021).

3. Sergey Primakov, Elizaveta Lavrova, Zohaib Salahuddin, Henry C Woodruff, Philippe Lambin. (2022). Precision-medicine-toolbox: an open-source python package for facilitation of quantitative medical imaging and radiomics analysis;
arXiv preprint arXiv:2202.13965 / Software Impacts 16, 100508 (2023)

4. A Ibrahim, S Primakov, M Beuque, HC Woodruff, I Halilaj, G Wu, T Refaee, R Granzier, Y Widaatalla, Roland Hustinx, FM Mottaghy, P Lambin. Radiomics for precision medicine: Current challenges, future prospects, and the proposal of a new framework;
Methods 188, 20-29

5. MJ Trimpl, S Primakov, P Lambin, EPJ Stride, KA Vallis, MJ Gooding. Beyond automatic medical image segmentation—the spectrum between fully manual and fully automatic delineation.;
Physics in Medicine & Biology 67 (12), 12TR01

6. Pishtiwan HS Kalmet*, Sebastian Sanduleanu*, Sergey Primakov, Guangyao Wu, Arthur Jochems, Turkey Refaee, Abdalla Ibrahim, Luca v Hulst, Philippe Lambin, Martijn Poeze. Deep learning in fracture detection: a narrative review;
Acta orthopaedica 91 (2), 215-220

7. Abdalla Ibrahim, Turkey Refaee, Sergey Primakov, Bruno Barufaldi, Raymond J Acciavatti, Renee WY Granzier, Roland Hustinx, Felix M Mottaghy, Henry C Woodruff, Joachim E Wildberger, Philippe Lambin, Andrew DA Maidment. The effects of in-plane spatial resolution on CT-based radiomic features' stability with and without ComBat harmonization;
Cancers 13 (8), 1848

8. Abdalla Ibrahim, Akshayaa Vaidyanathan, Sergey Primakov, Flore Belmans, Fabio Bottari, Turkey Refaee, Pierre Lovinfosse, Alexandre Jadoul, Celine Derwael, Fabian Hertel, Henry C Woodruff, Helle D Zacho, Sean Walsh, Wim Vos, Mariaelena Occhipinti, Francois-Xavier Hanin, Philippe Lambin, Felix M Mottaghy, Roland Hustinx. Deep learning based identification of bone scintigraphies containing metastatic bone disease foci;
Cancer Imaging 23 (1), 12

9. Simon A Keek, Manon Beuque, Sergey Primakov, Henry C Woodruff, Avishek Chatterjee, Janita E van Timmeren, Martin Vallières, Lizza EL Hendriks, Johannes Kraft, Nicolaus Andratschke, Steve E Braunstein, Olivier Morin, Philippe Lambin. Predicting adverse radiation effects in brain tumors after stereotactic radiotherapy with deep learning and handcrafted radiomics
Frontiers in Oncology 12, 920393

10. Abdalla Ibrahim, Martin Vallieres, Henry Woodruff, Sergey Primakov, Mohsen Beheshti, Simon Keek, Sebastian Sanduleanu, Sean Walsh, Olivier Morin, Philippe Lambin, Roland Hustinx, Felix M Mottaghy. Radiomics analysis for clinical decision support in nuclear medicine
Seminars in nuclear medicine 49 (5), 438-449
...

Full list of publications:

SCAN ME