Development and application of AI tools to improve diagnostic and prognostic capabilities of medical imaging data

Dissertation

zur Erlangung des Grades

Doktor der Naturwissenschaften

am Fachbereich Biologie

der Johannes Gutenberg-Universität Mainz

Óscar Llorián-Salvador

Geb. Am 24.03.1992 in Oviedo

Mainz, 2024

Dekan: Prof. Dr. Eckhard Thines

1.Berichterstatter: Prof. Dr. Miguel Andrade

2.Berichterstatter: PD Dr. med. Jan C. Peeken

Tag der mündlichen Prüfung: 27.01.2025

Johannes Gutenberg University Mainz

Faculty of Biology

Institute of Organismic and Molecular Evolution (iOME)

Computational Biology and Data Mining Group

AG Andrade

Dissertation

Development and application of AI tools to improve diagnostic and prognostic capabilities of medical imaging data

Óscar Lorián-Salvador

1. Reviewer	Prof. Dr. Miguel A. Andrade-Navarro Faculty of Biology Johannes Gutenberg University Mainz
2. Reviewer	PD Dr. med. Jan C. Peeken Department of Radiation Oncology Klinikum rechts der Isar, Technical University of Munich
Supervisor	Prof. Dr. Miguel A. Andrade-Navarro

October 31, 2024

Óscar Llorián-Salvador

Development and application of AI tools in improving diagnostic and prognostic capabilities of medical imaging data

Dissertation, October 14, 2024

Reviewers: Prof. Dr. Miguel A. Andrade-Navarro and PD Dr. med. Jan C. Peeken

Supervisor: Prof. Dr. Miguel A. Andrade-Navarro

Johannes Gutenberg University Mainz

AG Andrade

Institute of Organismic and Molecular Evolution (iOME)

Faculty of Biology

Hanns-Dieter-Hüsch-Weg 15

55128 Mainz

Abstract

Advances in personalized cancer treatment have been significantly motivated by the integration of advanced computational techniques with medical imaging modalities. These techniques allow for the extraction of quantitative radiomics information, which is then analyzed and correlated with genomic biomarkers, clinical features, and other biological indicators through machine learning (ML) models. This computational approach allows for a better tracking of cancer behavior and treatment response, offering a powerful tool for diagnostic, therapeutic planning and prognosis purposes.

In this thesis, ML models were developed and validated to assess the predictive power of quantitative radiomics and radiogenomic features, extracted from magnetic resonance imaging (MRI) and computed tomography (CT) scans, as well as clinical and semantics information. These models were analyzed in different oncological contexts, using computational methods to process all features, including batch harmonization, outlier detection, normalization, feature selection via redundancy reduction and relevance optimization and class imbalance correction. Classifiers employed included support vector machine (SVM), random forest (RF), least absolute shrinkage and selection operator (LASSO), logistic regression (LR) and multilayer perceptron (MLP) classifiers.

The first study of this thesis focused on the differentiation between atypical lipomatous tumors (ALTs) and lipomas via the detection of the mouse double minute 2 (MDM2) gene biomarker. A LASSO classifier performed best with an area under the receiver-operator characteristic (AUROC) of 0.88. The second study investigated the complete pain response in painful spinal bone metastasis patients after palliative radiotherapy (RT) treatment. A clinical LASSO classifier achieved an AUROC of 0.80. The third study explored common acute side effects on breast cancer patients treated with RT. A LASSO classifier outperformed all other models when predicting the appearance of moist cells epitheliolysis as a surrogate for skin inflammation with an AUROC of 0.74. The fourth study monitored neoadjuvant chemotherapy treatment response to Ewing sarcoma patients via the histological response assessment after surgery. A LR trained on the relative delta of radiomics features achieved an AUROC of 0.62, outperforming the best model trained on information from radiology readings (LR; AUROC of 0.58).

In conclusion, this thesis provides insight into the value of artificial intelligence (AI) and radiomics to address key challenges in oncology by supporting clinical decisions regarding distinguishing tumor types, predicting treatment responses, toxicity and disease evolution. Radiomics features were most effective when differentiating visually similar tumors and as a relative delta of change before and after neoadjuvant chemotherapy treatment. Clinical features have also shown predictive power in other cases of treatment response prediction, while providing useful support and baseline information overall.

Zusammenfassung

Fortschritte in der personalisierten Krebstherapie wurden maßgeblich durch die Integration fortschrittlicher rechnergestützter Techniken mit medizinischen Bildgebungsverfahren vorangetrieben. Diese Techniken ermöglichen die Extraktion quantitativer radiomischer Informationen, die dann mithilfe von maschinellen Lernmodellen (ML) analysiert und mit genomischen Biomarkern, klinischen Merkmalen und anderen biologischen Indikatoren korreliert werden. Dieser Ansatz verbessert das Monitoring von Krebsverhalten und Therapieantwort und stellt ein wertvolles Werkzeug für Diagnostik, Therapieplanung und Prognose dar.

In dieser Arbeit wurden ML-Modelle entwickelt und validiert, um die prädiktive Leistungsfähigkeit radiomischer und radiogenomischer Merkmale zu bewerten, die aus Magnetresonanztomographie- (MRT) und Computertomographie-Aufnahmen (CT) sowie klinischen und semantischen Informationen extrahiert wurden. Die Modelle wurden in verschiedenen onkologischen Kontexten analysiert, wobei Methoden zur Harmonisierung, Ausreißererkennung, Normalisierung, Merkmalsauswahl durch Redundanzreduktion sowie Klassenungleichgewichts-Korrektur angewendet wurden. Verwendete Klassifikatoren waren u. a. Support Vector Machine (SVM), Random Forest (RF), Least Absolute Shrinkage and Selection Operator (LASSO) und logistische Regression (LR).

Die erste Studie fokussierte sich auf die Unterscheidung atypischer lipomatöser Tumoren (ALT) und Lipome mittels des Maus-Doppelminuten-2-Gen-Biomarkers (MDM2), wobei LASSO einen Fläche unter der Operationscharakteristik (AUROC) von 0.88 erzielte. Die zweite Studie untersuchte die Schmerzantwort bei spinalen Knochenmetastasen nach palliativer Strahlentherapie. Ein klinischer LASSO-Klassifikator erreichte einen AUROC von 0.80. Die dritte Studie erforschte akute Nebenwirkungen bei Brustkrebspatientinnen, die mit Strahlentherapie behandelt wurden; LASSO übertraf andere Modelle in der Vorhersage von feuchter Epitheliolyse, einem Indikator für Hautentzündung (AUROC 0.74). Die vierte Studie bewertete die Therapieantwort auf neoadjuvante Chemotherapie bei Ewing-Sarkom mittels histologischer Reaktion nach OP. Ein LR-Modell auf Grundlage der relativen Delta-Änderung radiomischer Merkmale erreichte AUROC 0.62 und übertraf das radiologische Modell (AUROC 0.58).

Zusammenfassend gibt diese Arbeit Einblicke in den Wert von KI und Radiomik zur Unterstützung klinischer Entscheidungen in der Onkologie, einschließlich Tumorunterscheidung, Therapieantwort, Toxizität und Krankheitsverlauf. Radiomische Merkmale zeigten besondere Effizienz bei der Differenzierung ähnlicher Tumoren und in der relativen Delta-Änderung vor/nach Chemotherapie.

List of keywords

Radiomics, radiogenomics, oncology, radiology, bioinformatics, machine learning, artificial intelligence, magnetic resonance, computed tomography, radiotherapy, semantics, spinal instability neoplastic score, neoadjuvant chemotherapy, painful spinal bone metastasis, lipoma, atypical lipomatous tumor, breast cancer, side effects, Ewing sarcoma, logistic regression, support vector machine, random forest, multilayer perceptron.

List of abbreviations

ALT	Atypical lipomatous tumor
AI	Artificial intelligence
BA	Balanced accuracy
AUROC	Area under the receiver-operator characteristic
CDK4	Cyclin dependent kinase 4
CNN	Convolutional neural network
CT	Computed tomography
CTV	Clinical target volume
CV	Cross-validation
DICOM	Digital imaging and communication in medicine
DL	Deep learning
EMA	European medicines agency
FDA	Food and drug administration
GT	Glandular tissue
GTV	Gross tumor volume
HPV	Human papillomavirus
IBSI	Imaging biomarker standardization initiative
ICC	Intraclass correlation
LASSO	Least shrinkage and selection operator
LOOCV	Leave-one-out cross validation
LR	Logistic regression
MCC	Matthews correlation coefficient
MRMR	Minimum redundancy-maximum relevance
MDM2	Mouse double minute 2
ML	Machine learning
MLP	Multilayer perceptron
MRI	Magnetic resonance imaging
n-CV	Nested cross-validation
PCA	Principal component analysis
PET	Positron emission tomography
PSBM	Painful spinal bone metastasis
RF	Random forest
RNN	Recurrent neural network
RT	Radiotherapy
SINS	Spinal instability neoplastic score
SMOTE	Synthetic minority oversampling technique
SVM	Support vector machine
T1w	T1 weighted
T2w	T2 weighted
T1fsgd	T1 fat-saturated with contrast agent such as gadolinium
TBV	Total breast volume
VOI	Volumes of interest

List of figures

List of tables

Contents

1 General introduction

1.1 Cancer

Cancer, representing a diverse group of related diseases, is characterized by its uncontrolled growth and the expansion of the atypical cells that compose it. Though every type of cancer is different in behavior and impact, they all share their fundamentally disruptive nature. The understanding of cancer as a disease has significantly evolved over the years, from ancient times where it was perceived as a death sentence, until the present era, addressed as a disease with multiple ways of treatment and, in some cases, curable [1]. However, it is still a widespread and life-threatening disease, with 19.3 million new cases and almost 10 million deaths due to cancer, as approximated in 2020 by the Global Cancer Observatory [2]. It is expected that cancer cases will keep increasing in the next decades. By 2040, it is expected that new cases will increase to 28.4 million, 47% higher than in 2020.

This disease starts at the cellular level. When the control of normal cell proliferation and division is lost, cells begin expanding and replicating in a chaotic manner, forming the cancerous tissue [3]. Avoidance of the natural cell death mechanisms, activation of oncogenes, deactivation of tumor suppressor genes and altering of the DNA reparation processes, are among the main causes of the abnormal growth that leads to tumors [4,5]. These tumors can be benign, non-invasive; or malignant, with an aggressive growth rate and the possibility of spreading to neighboring tissues, a process known as metastasis. This complication can make the treatment process significantly more difficult, thus reducing the survival probability of patients [6,7].

There are multiple factors that contribute to the development of this disease, or to a tumoral growth, many of them related to specific lifestyles. Smoking is the primary cause of lung cancer, and is also associated with other various types, such as mouth, neck, esophagus and bladder cancer [2,8,9]. Diets high in processed meat and low on fruit have also been associated with a higher risk of developing colorectal cancer [6,10]. Sedentarism and obesity, on the other hand, have been linked to some types of malignancies, such as breast, endometrium and, again, colorectal cancer [6,11]. Some infections are known factors for specific types of cancer, such as human papillomavirus (HPV), Helicobacter pylori, and hepatitis B and C [12–14]. Further, exposure to carcinogens, such as asbestos, some industrial chemical compounds, or prolonged UV radiation, also increase the risk of developing several types of cancer [15–17].

Not all types of the disease affect the global population equally, but patterns have emerged in the different continents. In Asia there is a higher incidence registered of

cancers related to the digestive system, such as esophagus, stomach or liver cancer, likely due to dietary factors. In Africa, malignancies related to infections have been notably common, such as cervical cancer (from HPV infection), or liver cancer (from hepatitis B and C). Central and South America have registered mixed patterns of malignancies, with an increasing incidence of breast and prostate cancer, and the prevalence of infection-related cancers. In North America and Europe, breast, prostate and colorectal types of cancer have the highest incidences. Some variables contribute to a higher incidence on this global perspective: notably, with progressively higher life expectancies overall, a higher number of individuals reach later ages, when cancer is more common. However, their mortality has decreased in many of these malignancies mentioned thanks to advances in early detection and treatment [2,8,18]. Incidence rate is currently also being addressed in multiple ways, for instance by programs to prevent smoking, or vaccinations against cancer-inducing infections [19–21].

Cancer treatment involves a variety of approaches, such as surgical intervention, RT, chemotherapy, targeted therapy and immunotherapy [3,22,23]. Each treatment can have a different efficacy based on the type and stage of cancer, as well as individual features from the patient. A combination of treatments is also common to increase the overall effectiveness [24–26].

Presently, cancer research is experiencing a quick expansion thanks to advances in molecular biology, genetics and technology, and multidisciplinary fields such as biotechnology and bioinformatics. These advances, more effective and tailored to each type of case, are already increasing survival rates and the quality of life of patients [27,28].

However, multiple challenges still remain, beyond the success of the treatment: some types of cancer are difficult to identify and differentiate from others, leading to incorrect diagnoses and, as a consequence, treatment. In other cases, the choice of treatment can lead to adverse repercussions such as side effects or secondary cancers that may potentially be avoided with enough understanding of the circumstances of the disease and patient. On the other hand, the evolution of several treatments can only be observed afterwards (for instance, during surgery after some chemotherapy treatments), leading to delays in its evaluation and, therefore, risking a metastatic spread if the strategy was not adequate [29,30]. In those situations where distant metastasis is already present, the selection of an appropriate palliative treatment can significantly improve the quality of life of patients [27,28,31]. However, there are numerous factors and uncertainties to consider when dealing with heterogeneous cancers and distant metastases, possibly leading to suboptimal therapies.

1.1.1 Distinction of biologically similar tumors

Correct tumor identification is essential in oncology: in some cases, slight biological or clinical behavioral characteristics can be the difference between a benign or a malignant variant of cancer. Diagnostic precision not only affects the choice of treatment and patient handling, but also their prognosis and quality of life. One manifest example are lipomatous tumors: they are the most common neoplasm that can be found as soft-tissue tumors primarily in extremities. Even though they are a heterogeneous group of tumors, almost half of them are either benign adipocytic tumors (lipomas), or atypical lipomatous tumors (ALTs) [32–35]. While the former only needs treatment in cases where quality of life is compromised (in the form of pain or functional disorders), ALTs are prone to an aggressive growth, possibly evolving into high-grade sarcomas [36–39].

Lipomas are usually present as soft tissue masses, usually painless, in the subcutaneous tissue of the extremities. Histologically, they are made of uniform, mature adipocytes, which are then covered by a thin fibrous capsule. In a magnetic resonance image (MRI) or computed tomography (CT) scan, they appear as a homogeneous and fat-enriched mass. The former is more preferable given its capability to enhance contrast of soft-tissues [40–42].

On the other hand, ALTs, even though they are clinically similar to lipomas, display a different biological behavior, entailing pain and possibly infiltrating adjacent tissues. Histologically, they appear as larger and more hyperchromatic fat masses with atypical adipose cells and lipoblasts. They contain more cellularity and, generally, larger fibrosis areas [43,44].

Their distinction has significant clinical implications. Lipomas, given their benign nature, may not need to be addressed if there is no pain, though their removal implies usually just a surgical resection. In contrast, ALTs require a more aggressive treatment, given their malignant potential. This treatment usually involves a wider surgical resection and sometimes radiotherapy (RT). Early detection and accurate characterization are therefore crucial for an optimal therapy planning and minimize recurrence [43,44].

Precise differentiation between both can be challenging when only taking into account clinical and medical imaging data, with radiology residents achieving approximately 65% accuracy. This performance contrasts with experimented attending radiologists, who reach accuracies of 90%, showcasing the importance of experience for discerning each type of tumor [45]. For this reason, and to bridge this gap of expertise,

clinical and medical imaging data are often combined with genetic analyses, with the goal of finding relevant biomarkers such as the mouse double minute 2 (MDM2) or the cyclin dependent kinase 4 (CDK4) genes [46].

1.1.2 Response prediction to palliative treatment for distant metastases

When an existing cancer starts invading neighboring tissues, including the intravasation into both blood and lymph vessels to then extravasate and grow in the neighboring tissue, distant metastases can develop [47,48]. In such cases, patient response to metastasis can drastically vary depending on a number of factors, including the type and source of primary cancer, the metastasis location, and the specific characteristics of the individual. An effective management of pain response in such patients is crucial to improve their quality of life, and one of the most commonly treatments used is palliative RT. An accurate following of the pain response would allow for a better therapy planning and an optimized personalized treatment [49].

Bone metastasis and, particularly, painful spinal bone metastasis (PSBM), are a frequent complication of multiple types of primary cancers, especially breast, prostate and lung cancer [50–54]. Bone metastatic lesions can cause severe pain reactions, immobility and an overall reduction of the quality of life of the patient. Given its innate painful nature due to factors such as the direct bone destruction, nerve compression or inflammatory response, accurate response measures for PSBM patients after palliative RT has become an important area of research focus [55,56].

Pain caused by oncological complications is first addressed by combining pharmacological and non-pharmacological strategies. Among the former, usual treatments include analgesics, opioids, bisphosphonates and denosumab. However, when this strategy provides insufficient pain relief, it becomes supplementary to palliative RT treatment. Palliative RT works by shrinking the tumor size, relieving pressure in the area and slowing the destruction of the bone tissue, ultimately relieving of bone pain [57]. Even though there is a positive outcome in two thirds of patients, the decrease in pain may take from days to weeks from the therapy, and the effective response may vary, increasing the need of a personalized treatment.

Pain response assessment after palliative RT treatment is an area of clinical and research interest. Recent approaches for its estimation involve the prediction based on a number of factors that are known to be influential regarding pain response. Besides oncological details such as the type and location of the primary tumor or the number of metastatic bone lesions, patient age, overall health condition and functional impairment measured by the Karnofsky performance scale are known clinical characteristics linked

to pain response [50–53]. In addition, relevant biomarkers and genetic profiles can also be used to determine the effectiveness of a given palliative RT treatment.

1.1.3 Therapeutic strategies and side effects

The principal cancer treatments include surgery, chemotherapy, targeted therapy and RT, each one having their unique profile and optimal uses. However, each of these indispensable strategies may cause secondary effects that affect the quality of life of patients both in gravity and duration. Understanding these effects take a significant role in tailoring their treatment strategy.

Surgical interventions are the most traditional and one of the most effective treatments against cancer, especially in the early stages of the disease, when the tumor is more easily located and can be completely surgically removed. It is also used in combination with other therapies to remove the primary tumor, perform biopsies to evaluate the extension of the disease, and in some cases to provide relief to several symptoms in more advanced stages. The principal drawbacks to surgical interventions encompass pain, potential infections, scarring-related problems and limitations or complications of the body functions related to the location. For instance, mastectomy for breast cancer may cause lymphedema, causing swelling of the arm due to an accumulation of lymphatic fluid [58].

Chemotherapy treatments are based on the use of chemotherapeutic drugs to target cancerous cells, and it is especially useful in cases where the disease has metastasized, or when there is a high probability of recurrence. It is often used in combination with surgical interventions and RT, providing a systemic response against the disease. However, the impact of chemotherapeutic drugs is not limited to cancerous cells, but also to healthy cells of rapid growth. Its side effects as a result are comprised of nausea, vomits, hair loss, myelosuppression (reduction of blood cell production due to a reduced activity of the bone marrow), peripheric neuropathy, renal and hepatic disorders, and risk of secondary cancer and leukemia [59,60].

Targeted therapy is a strategy focusing on combating the molecules involved in the abnormal growth and propagation of cancer. Targeted therapy is mainly applied on cancers with well-defined and specific molecular characteristics, such as genetic mutations or protein overexpression. Thus, their use is particularly effective in cases where specific biomarker targets have been identified, for example in HER2-positive breast cancer and some types of leukemia and melanoma. Side effects of targeted therapy largely depend on the molecular target and the activation mechanism of the drug employed. Some of these adverse effects are hypertension, skin complications,

gastrointestinal effects and the hand-foot syndrome or palmar-plantar erythrodysesthesia [59,60].

RT uses ionizing radiation to eliminate cancerous cells, and it is especially useful when the disease is more localized. It is commonly used in combination before surgical resection to shrink the tumor and make the surgery less invasive (neoadjuvant therapy), or after surgery to remove residual cancerous cells (adjuvant therapy). RT is also used as palliative treatment, providing pain relief and slowing other symptoms such as the osteoclast activity in bone metastasis cases. Even though RT is a crucial strategy as the main therapy, adjuvant therapy or palliative therapy options, it can also cause acute and late secondary effects. Acute effects include erythema, mucositis and general skin inflammation, edema, fatigue, and desquamation or lysis of epithelial moist and dry cells. Late adverse effects, on the other hand, entail fibrosis or fibrotic scarring of tissues, necrosis of tissues, and the development of secondary cancers due to the exposure to radiation [61,62].

Breast cancer is one of the most common types of cancer in women, and RT is one of the standard treatment options. This is especially the case after lumpectomy or surgical intervention conserving the breast, as a way to eliminate residual cancerous cells. Here, the spectrum of adverse effects is largely similar. Skin inflammation, edema and moist cells epitheliolysis stand out as the most common acute side effects after RT on breast cancer cases [63–65].

There are several strategies that can be employed in order to minimize the adverse effects of these treatments. Intensity-modulated RT, deep-inspiration breath hold gated RT, and image-guided RT allow for a higher precision in administering doses, avoiding significant exposures to healthy tissues. Some other strategies include pain relief therapies, physiotherapy for lymphedema cases, and a regular check and prognosis for an early detection of late side effects [66–68].

1.1.4 Monitoring of cancer evolution and effectiveness of treatment with medical imaging

An effective surveillance of cancer progression and the response of the patient to therapy is a critical part of managing the disease. It allows physicians to evaluate the effectiveness of treatments, early recurrence detection, and adjust the therapy as needed.

Historically, cancer monitoring has been based on traditional imaging technologies such as CT, MRI and positron emission tomography (PET) scans. CT uses

X rays to create images that help understanding the size and extension of the tumor. MRI scans, on the other hand, rely on magnetic fields and radio waves to produce images where soft tissues are normally highlighted, giving further information about the structure and composition of the tumor. While there are multiple protocols and sequences in MRI, three stand out: T1 weighted (T1w), T2 weighted (T2w), and T1 with the suppression or saturation of fat tissues with a contrast agent such as gadolinium (T1fsgd). PET scans, lastly, combine CT technology with the detection of metabolic activity of the tumor, using radioactive tracers to identify these areas of higher metabolic activity, indicating the presence of malignant bodies. Overall, these methods allow the visualization of tumoral bodies and their surroundings, detect changes in their size, shape, composition, or presence of metastases [69–71].

Advanced technologies such as radiomics emerged as promising tools in the field, providing detailed quantitative data that can supplement traditional radiology methods and clinical information [72–74]. The features captured by radiomics data, also known as radiomics biomarkers, include information about their texture, shape, intensity, heterogeneity, and first-order statistics that analyze the distribution of individual voxel values, disregarding spatial relationships [73–75].

Radiomics biomarkers are highly rich in information, making them excellent candidates to be used for predicting treatment response in cancer patients. In this way, radiomics models are able to effectively provide supplementary knowledge to physicians when personalizing and adjusting therapies. Further, radiomics features can be useful when estimating patient prognosis by predicting tumor aggressiveness and survival probabilities, and when performing early detection of recurrence studies [76].

The advantages of radiomics biomarkers become pivotal when considering more rare types of tumors, or others with a more difficult profile. Ewing sarcomas can be found in this context: a rare malignancy found primarily in bone or soft tissues. Due to its uncommon and aggressive nature, Ewing sarcomas require close monitoring, performed via radiologist readings on the different imaging techniques. Bone scintigraphies are an alternative scanning option given the propensity of this type of tumor to infect bone tissues [77,78].

1.2 Artificial intelligence in medicine

Artificial intelligence (AI) is revolutionizing the field of medicine, providing new supplementary tools and perspectives to improve the diagnostic, treatment selection and prognostic of many diseases. The ability of Machine learning (ML) and deep learning

(DL) models to detect complex patterns from large volumes of data provides additional insights that would otherwise be unavailable to medical practitioners.

This pattern recognition ability is also used in biomedicine to develop new treatments by finding new biomarkers and predicting the drug therapy response. In this context, AI is also used in combination with genomics and proteomics data to predict safety concerns regarding drugs in development, their potential secondary effects, and unforeseen adverse effects from combining FDA / EMA-approved drugs in a treatment strategy [79,80].

Thanks to this ability to interpret vast and complex data, it is quickly integrating in multiple areas of medicine, improving the precision and efficiency of medical services and overall functions. AI models that can analyze medical images, records of patients and genomics data are useful tools for diagnosis purposes, allowing treatment to start in the early stages of a disease, when it is commonly more effective [81].

1.2.1 Evolution and state-of-the-art of AI in oncology

Oncology has been one of the fields that benefitted the most from AI integration thanks not only to the ability to interpret large volumes of clinical data, but to find patterns in complex and heterogeneous diseases such as cancer. Over the years, multiple AI models have been developed and refined for their use in oncology, each one having their strengths and weaknesses.

The first AI models developed for oncology were expert systems and simple ML algorithms. Among the former stands out the system MYCIN, used in the 70s to diagnose bacterial infections and suggest treatments. While not related to oncology, it laid out the foundation for the practical application of AI tools in medicine [82,83]. On the other hand, the first ML models used in oncology were mainly decision trees, a type of supervised learning algorithm used mostly in the 80s and in the 90s to classify types of cancer based on clinical and pathological data [84].

As computational resources became more available, the ML models developed advanced into more sophisticated ones, and began taking a more prominent role in oncology. In the 2000s, support vector machines (SVM) grew in popularity mainly for their classification capabilities and better handling of larger numbers of variables. They are used primarily to classify histological images and predict chemotherapy treatment outcomes [85,86]. On the other hand, as an evolution of decision trees, random forests (RF) have since been widely used to classify types of cancers as well. The main

advantages of RFs are their ability to handle missing values and correlated variables, making them a valuable addition in a field where incomplete data may be more common [87].

In the last decade, DL techniques have revolutionized many medical and non-medical fields with their ability to process both images and amounts of omics data on an unprecedented scale. Convolutional neural networks (CNN) have been particularly useful when analyzing mammographies (X-ray imaging), CTs and MRIs, able to detect complex patterns directly on the scans. Recurrent neural networks (RNNs), on the other hand, are primarily used to analyze time series and longitudinal data. Thus, DL techniques are currently very powerful tools to predict the response to treatments over time and follow the evolution of cancer [88,89].

Some specific models have proven to be particularly useful and effective in oncology. On the one hand, survival models play a crucial role in the prognosis and likelihood of events, with probabilities drawn usually from clinical and genomic data. One of the most common models is the semi-parametric Cox proportional hazard model [90,91], which has also led to popular extensions using more complex algorithms, such as DeepSurv, using deep neural networks [92]. On the other hand, radiomics models have become indispensable tools that can accurately address many oncology-related challenges, owing to their flexibility and the rich patterns inherent in radiomics features [45,93–97].

2 Hypotheses and aims of the thesis

Radiomics has been at the forefront of oncology research for the past years as a powerful tool to quantify information in medical imaging. This technology allows for the analysis of complex patterns, and volumetric and textural characteristics that could otherwise be overlooked. The integration of such rich information in AI models can lead to a more effective precision medicine, and to more accurate clinical decisions. Even though radiomics models have already found extensive applications in oncology owing to their versatility and predictive capabilities, there still remain many oncological challenges to be addressed.

This study aims to leverage the potential of AI and radiomics to answer some of these unsolved scientific questions regarding cancer, posed as supervised learning classification challenges. Specifically, the hypotheses of this work cover four of the most common and promising applications of radiomics models in oncology, which were introduced in the first part of the *General introduction*. Each of the working hypotheses has been addressed in the different chapters below, based on a common general objective.

1) Development and Evaluation of MR-Based Radiogenomic Models to Differentiate Atypical Lipomatous Tumors from Lipomas

The aim of this chapter was to design and implement ML and DL models to automatically distinguish ALTs and lipomas based on radiomics information extracted from MRI scans, and the presence of the genomic biomarker MDM2. These models aim to bridge the difficulty of such division, challenging for radiology residents but feasible for attendant radiologists with years of experience. Therefore, the hypothesis posed in this section is:

- Radiogenomic AI models can differentiate, with high accuracy, between biologically similar tumors such as ALTs and lipomas.

2) The importance of planning CT-based imaging features for machine learning-based prediction of pain response

This chapter intended to evaluate the capability of AI to predict the response to palliative RT treatment for pain treatment in patients with PSBM. Additionally, the most important factors for this estimation were analyzed to understand what are the most essential parameters for pain response prediction. Such capability would significantly improve optimal treatment decisions, aiming to increase the quality of life for patients in later stages of the disease. Therefore, the hypotheses posed in this section are:

- AI models can accurately predict the complete response to palliative RT treatment in PSBM patients.
- Radiomics features can capture with most precision the patterns related to complete pain response to palliative RT in PSBM patients compared to a clinical baseline.
- The spinal instability neoplastic score (SINS) can be used as a gold-standard assessment variable for the prediction of complete pain response in PSBM patients.

3) Insights from CT-based Radiomics: Predicting Breast Cancer Radiotherapy Side Effects

The goal of this chapter is to leverage the predictive power of ML models and radiomics information to anticipate adverse effects from the application of RT to breast cancer patients. A better prediction of these side effects would allow a more personalized treatment of them, with the aim to improve the quality of life of patients both short-term and long-term after therapy. Further, in this chapter the predictive influence of the total breast volume is explored, which has been reported to significantly correlate to the outcome of multiple types of radiomics models when applied to breast cancer scans [98,99]. Lastly, the extent of this correlation has been measured to evaluate the effect it has on the performance of the radiomics models. Therefore, the hypotheses formulated in this section are:

- Radiomics models can accurately predict the appearance of side effects from RT treatment of breast cancer.
- Total breast volume (TBV) influences significantly the performance of imaging radiomics models in breast cancer.

4) Can Radiomics outperform Radiologists in Predicting Ewing Sarcoma Response to Neoadjuvant Chemotherapy from Longitudinal MRI?

The motivation of this chapter is to evaluate the prognostic power of radiomics in cases of Ewing sarcoma. These rare malignancies are difficult to treat, and extracting information from their few cases is also challenging. For this reason, radiomics models can substantially improve clinical outcomes via the identification of patients not responding to the initial chemotherapy prior to surgical intervention. In addition, in this chapter the performance of radiomics models was compared to that of models trained on information extracted from radiology readings. Therefore, the hypotheses formulated in this section are:

- Radiomics models can effectively predict the histological therapy response of Ewing sarcoma patients treated with neoadjuvant chemotherapy.

- Radiomics models outperform radiology readings models when applied to the response prediction of Ewing sarcomas treated with neoadjuvant chemotherapy.

3 Overall materials and methods

While the publications contained in this thesis cover the oncological bases excellently, it is essential to highlight the rationale behind key computational steps that have been undertaken. It is often the case that some necessary processes for the correctness and reproducibility of the methodology can accidentally be overlooked or skipped. In this section, a brief description and purpose of such steps is provided.

3.1 Image acquisition to feature extraction

All CT and MRI scans, obtained from each respective center as described in the publications of the thesis, were analyzed by expert radiation oncologists to segment the appropriate volumes of interest (VOI) using 3D Slicer [100]. Preprocessing of CT and MR images included several key steps to ensure their consistency and quality before extracting the radiomics features. Intensity normalization was applied to standardize the value of all pixels in the MRI images for a better comparison. Image discretization is then used to divide the continuous pixel values into bins of discrete intervals to optimize the performance of the image analysis. Uniformity of the voxel dimension sizes was achieved via isotropic resampling using BSpline interpolation, which adjusts the spatial resolution of the images while retaining the continuity and resolution of the anatomical structures [101]. Finally, Laplacian of Gaussian filtering highlights borders and local variations in the image, helping improve the detection of structures for subsequent radiomics analysis. The steps required for each set of CT or MR images are detailed further in the methodology sections of the respective studies.

The same feature extraction process was followed for each of the image sets, using the pyRadiomics library [102] and standardized acquisition parameters according to the Imaging biomarker standardization initiative (IBSI) guidelines [103]. This initiative is used to standardize the process of extracting quantitative biomarkers from medical images, in order to increase the reproducibility and comparability of radiomics studies. To this end, IBSI establishes specific protocols for the preprocessing of images, feature extraction and data analysis, promoting homogeneity across different studies and platforms. This extraction process yields the conventional 104 to 105 variables, labelled as original, which contain the most promising and informative radiomics features [104].

The extracted radiomics features are grouped depending on the specific characteristics of the image they capture. Shape features capture the volumetric and 3D geometry details of the structures within the images. First order features aim to quantify the intensity distribution of the image voxels, which are defined by their mean, standard deviation or entropy. Additionally, there are five more groups that capture in different ways the texture features: gray level co-occurrence matrix, gray level run length matrix,

gray level size zone matrix, neighboring gray tone difference matrix, and gray level dependence matrix. These groups of texture features analyze the spatial relationships between pixels, providing information about the heterogeneity or contrast of the image by assessing variations in gray levels across the image [102].

3.2 Data preprocessing

There is a number of steps that need to be performed before the extracted features can be used to train AI models. Furthermore, their order and the stage when they are conducted is critical not just to their proper processing, but to avoid significant problems such as biased or skewed distributions, tendency to overfit, data leakage or significant performance loss.

The first step taken to preprocess the extracted radiomics features is to clean the data of missing values, and remove potential external sources of error. When the source images belong to a multicentered study, it is essential to remove the batch effects of different centers and machineries via nonparametric harmonization. This systematic effect is independent from any other corrections performed at a later stage, and should therefore be addressed first. To accomplish this, during the extraction of technical details of the scans from the digital imaging and communication in medicine (DICOM) files, the manufacturer of each CT or MR machine is recorded for their use in this process. By undertaking this step, the multicentered batch effect, which is the most significant instrumental error, is corrected.

Prior to executing computational operations with non-numeric features, which may be present in clinical or radiology feature sets, these were encoded into binary features. One hot encoding suffices for most categorical variables, though binary encoding has been employed in the cases of very sparse data. While it adds a very small amount of data leakage due to introducing a shallow correlation between the created binary variables, the sparsity of the data would otherwise transform these features into a large number of binary features, potentially harming the model's performance further.

To reduce the human error by removing the segmentation-dependent radiomics features, in all cases a number of scans have been re-segmented by a second radiation oncologist. Given that all radiomics features are numerical and continuous, features subjective to segmentation variations were identified and excluded by intraclass correlation (ICC) 3,1 and a threshold of 0.8. This specific test was chosen because the physicians were not rated as representative of a defined rater group, given their different extents of training in each case. In the case of discrete or categorical variables, present in radiology features, inter-rater variability was tested via Cohen's Kappa test with a

threshold of 0.4. Given the sparsity of the radiology data, this test was used as an alternative to Fleiss Kappa [105].

3.3 Model optimization

After addressing potential instrumental and human errors, and before splitting the data using repeated nested cross-validation (n-CV), outliers were detected and removed so that they do not affect the data distribution of the resulting resampling splits. Data normalization, feature selection and class imbalance correction were performed in the innermost fold of the n-CV in order to avoid data leakage issues, such as training data being normalized taking into account values of validation or test splits. However, datasets that come from a multicentered study, which undergo a batch harmonization step, require the data to be previously normalized. In such cases, an exception is made to meet the batch harmonization specifications.

The specific fold numbers of the n-CVs were selected based on each of the dataset sizes and their information density. In addition, for the achievement of more robust statistical significance of results, n-CV was repeated a number of times depending on available computational resources and the estimated running time of a single iteration. In some cases, it was necessary to stratify the data splits of the n-CV based on hierarchical relationships from the independent variables. For instance, in a study where patients underwent certain treatments, one patient may appear more than once due to multiple instances of the studied affliction. In such cases, all entries from the same patient have to be kept in the same fold to avoid data leakage of that patient to other folds, therefore preserving the hierarchical relationship between patients.

The last processing steps, carried out in the inner fold of the n-CV, were performed in the following order: first, feature values were transformed to a common scale using min-max scaling. This type of normalization was selected so that the same distribution of each feature is conserved in the common scale, further preserving the information within the data.

Second, after features are normalized, these were filtered using different feature selection techniques, comparing their ability to reduce redundancy, and focusing only on the most important features. To reduce the arbitrariness of such a critical step, a decision was used that combined the one in ten rule for feature selection and the estimated number of features needed to retain 95% of information by a principal component analysis (PCA) [106]. However, it is important to note that PCA can only be used as an estimate in situations where the relevance of each feature in the performance of the AI models is to be extracted. Otherwise, the nature of the transformed features by

PCA would render this task impractical. Finally, though more techniques were initially tested during research (such as variance threshold, recursive feature elimination or random forest selection), the main feature selection techniques utilized were minimum redundancy-maximum relevance (MRMR) and a two-step Spearman rank correlation coefficient, where first the most redundant features were discarded, and in a second step the most relevant features were selected [107]. MRMR was applied consistently in all four studies included in this thesis. A two-step Spearman rank correlation coefficient was also explored in all studies. However, given that the volume of results varied across the four studies, they were only incorporated in the final analyses of the third and fourth studies, where this comparison met the criteria for inclusion. The optimal number of features to be selected in each step was treated as another hyperparameter to optimize.

Third, once the data in a given fold has been normalized and features selected, class imbalance was corrected by using either of two most common techniques: by applying synthetic minority oversampling technique (SMOTE) on the minority class, combined with random undersampling of the majority class; or by using class weights. Following this order, therefore, possible generated synthetic data during class imbalance correction will not bias the feature selection process nor the scaling of the data. The right balance of SMOTE and random undersampling, and the correct class weights, have been estimated during the hyperparameter optimization process as well.

3.4 Machine learning modeling and statistical analysis

Hyperparameter optimization has been conducted using an iterative refinement approach, where optimal values are determined progressively for a finer precision in the optimization process, while decreasing running time compared to an exhaustive grid search. Several metrics were initially tested as the optimization criteria, although balanced accuracy (BA) was used as the main metric to further take into account potential class imbalances. In the case of a tie, F1 score served as tiebreaker.

An exhaustive analysis of the performance of most common ML models used with radiomics features has been performed in each of the works included in this thesis, for their respective aim. The models utilized were support vector machine (SVM), random forest (RF), least absolute shrinkage and selection operator (LASSO), a fully connected feedforward neural network (multilayer perceptron, MLP), and a logistic regression (LR) for the smaller datasets. Other models were initially considered and tested, but were not reported in the publications due to several reasons, one being the frequent overfitting caused by the use of complex models on very small datasets [108–110].

Performances have been monitored with multiple scores and plots to avoid biased conclusions, such as overoptimistic inferences in the case of imbalanced datasets when only predicting the majority class. The metrics presented are the area under the receiver-operator characteristic (AUROC), BA, F1 score, sensitivity and specificity to quantify their overall efficacy; and Matthews correlation coefficient (MCC) to assess the quality of the predictions. Whereas AUROC was used as the main metric to showcase the performance of the models, all of them were taken into account for the training and optimization process. In addition, the ROC and calibration curves are also provided. Unless otherwise specified, all scores are presented with 1.96 times the standard error (SE) for a 95% confidence interval of the results over all iterations of the results for the models. For instance, in a scenario of 50 iterations of a 5-fold n-CV, the scores have been presented as the average metric across all 250 final models ± 1.96 * SE.

Lastly, the impact that the individual features have had on the predictions has also been analyzed in all publications. Depending on the model, feature importance is reported as the weight or coefficient such a feature had (in the case of LASSO, linear kernel SVM, LR and MLP), or node impurity (in the case of RF). To account for features that are heavily relevant only in a few iterations, or features that were moderately informative but have been selected consistently, either score has been combined with the overall frequency of that particular feature being selected. This was calculated as $Score = Feature\ Importance/[(n+1)-m]$, where *Feature Importance* is the importance score corresponding to the type of model, n is the number of models, and m is the number of times the feature has been chosen.

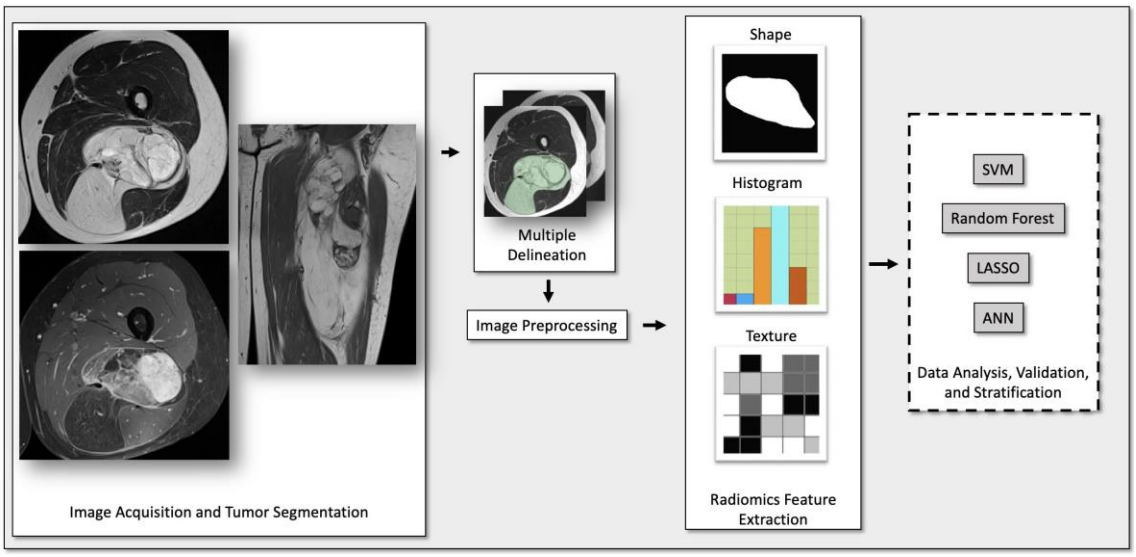


Figure 1. Graphical workflow followed for the differentiation of ALTs and lipomas (Section 4.1) [45].

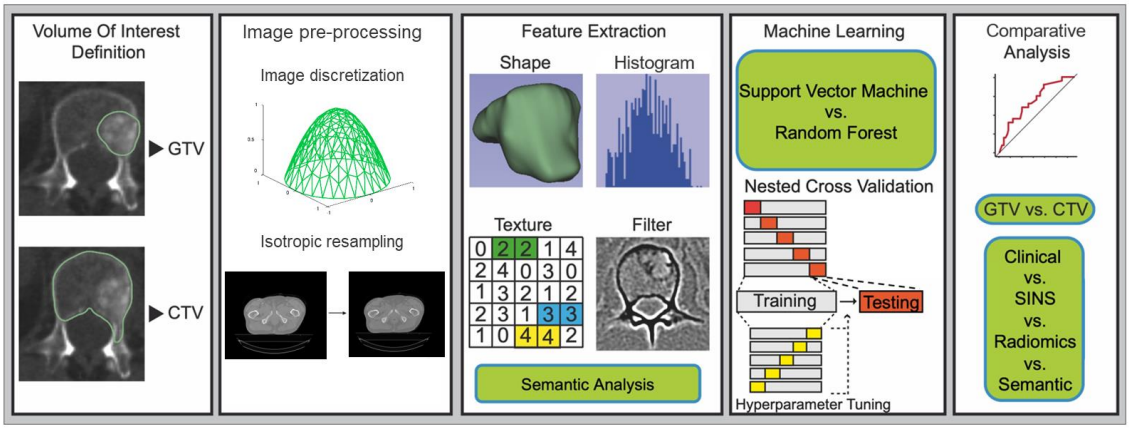


Figure 2. Graphical workflow followed for the prediction of complete pain response after palliative RT treatment in PSBM patients (Section 4.2) [93].

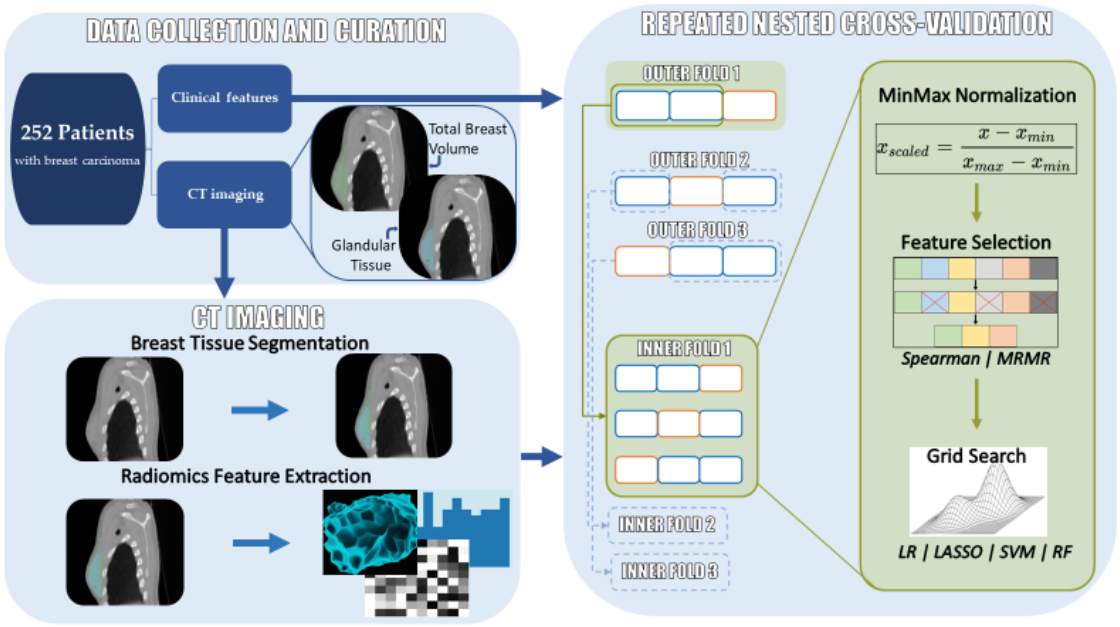


Figure 3. Graphical workflow followed for the prediction of RT side effects in breast cancer patients (Section 4.3) [111].

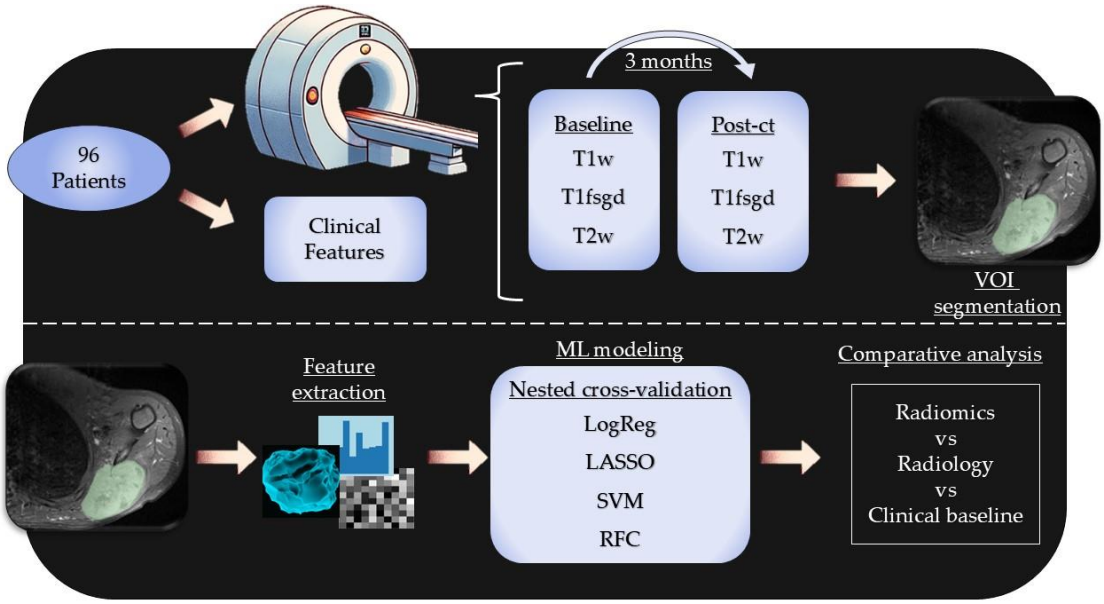


Figure 4. Graphical workflow followed for the monitoring of neoadjuvant chemotherapy treatment response to Ewing sarcoma patients (Section 4.4).

4 Results

4.1 Development and Evaluation of MR-Based Radiogenomic Models to Differentiate Atypical Lipomatous Tumors from Lipomas

cancers

MDPI

Article

Development and Evaluation of MR-Based Radiogenomic Models to Differentiate Atypical Lipomatous Tumors from Lipomas

Sarah C. Foreman [1,*,†], Oscar Llorián-Salvador [2,3,4,†], Diana E. David [3], Verena K. N. Rösner [1], Jon F. Rischewski [5], Georg C. Feuerriegel [1], Daniel W. Kramp [1], Ina Luiken [1], Ann-Kathrin Lohse [6], Jurij Kiefer [7], Carolin Mogler [8], Carolin Knebel [9], Matthias Jung [10], Miguel A. Andrade-Navarro [4], Burkhard Rost [3], Stephanie E. Combs [2], Marcus R. Makowski [1], Klaus Woertler [1], Jan C. Peeken [2,11,12,‡] and Alexandra S. Gersing [5,‡]

1 Department of Radiology, Klinikum Rechts der Isar, Technische Universität München, Ismaninger Straße 22, 81675 Munich, Germany
2 Department of Radiation Oncology, Klinikum Rechts der Isar, Technische Universität München, Ismaninger Straße 22, 81675 Munich, Germany
3 Department of Informatics, Bioinformatics and Computational Biology—i12, Technische Universität München, Boltzmannstr. 3, 85748 Munich, Germany
4 Institute of Organismic and Molecular Evolution, Johannes Gutenberg University Mainz, Hanns-Dieter-Hüsch-Weg 15, 55128 Mainz, Germany
5 Department of Diagnostic and Interventional Neuroradiology, University Hospital Munich (LMU), Marchioninistrasse 15, 81377 Munich, Germany
6 Department of Radiology, University Hospital Munich (LMU), Marchioninistrasse 15, 81377 Munich, Germany
7 Department of Plastic Surgery, University Hospital Freiburg, University of Freiburg, Hugstetterstraße 55, 79106 Freiburg im Breisgau, Germany
8 Institute of Pathology, Klinikum Rechts der Isar, Technische Universität München, Ismaninger Straße 22, 81675 Munich, Germany
9 Department of Orthopedics and Sport Orthopedics, Klinikum Rechts der Isar, Technische Universität München, Ismaninger Straße 22, 81675 Munich, Germany
10 Department of Radiology, University Hospital Freiburg, University of Freiburg, Hugstetterstraße 55, 79106 Freiburg im Breisgau, Germany
11 Helmholtz Zentrum München, Deutsches Forschungszentrum für Umwelt und Gesundheit, Institute of Radiation Medicine Neuherberg, 85764 Munich, Germany
12 Deutsches Konsortium für Translationale Krebsforschung (DKTK), Partner Site Munich, 69120 Heidelberg, Germany
* Correspondence: sarah.foreman@tum.de
† These authors contributed equally to this work.
‡ These authors contributed equally to this work.

check for updates

Citation: Foreman, S.C.; Llorián-Salvador, O.; David, D.E.; Rösner, V.K.N.; Rischewski, J.F.; Feuerriegel, G.C.; Kramp, D.W.; Luiken, I.; Lohse, A.-K.; Kiefer, J.; et al. Development and Evaluation of MR-Based Radiogenomic Models to Differentiate Atypical Lipomatous Tumors from Lipomas. *Cancers* **2023**, *15*, 2150. https://doi.org/10.3390/cancers15072150

Academic Editor: Hamid Khayyam

Received: 17 January 2023
Revised: 10 March 2023
Accepted: 27 March 2023
Published: 5 April 2023

Simple Summary: Differentiating atypical lipomatous tumors from lipomas on MR images is a challenging task due to similar imaging characteristics. Given these challenges, it would be highly beneficial to develop a reliable diagnostic tool, thereby minimizing the need for invasive diagnostic procedures. Therefore, the aim of this study was to develop and validate radiogenomic machine-learning models to predict the MDM2 gene amplification status in order to differentiate between ALTs and lipomas on preoperative MR images. The best machine-learning model was based on radiomic features from multiple MR sequences using a LASSO algorithm and showed a high discriminatory power to predict the MDM2 gene amplification. Due to the varying settings in which patients with lipomatous tumors present, this model may enhance the clinical diagnostic workup.

Abstract: Background: The aim of this study was to develop and validate radiogenomic models to predict the MDM2 gene amplification status and differentiate between ALTs and lipomas on preoperative MR images. Methods: MR images were obtained in 257 patients diagnosed with ALTs (n = 65) or lipomas (n = 192) using histology and the MDM2 gene analysis as a reference standard. The protocols included T2-, T1-, and fat-suppressed contrast-enhanced T1-weighted sequences.

Additionally, 50 patients were obtained from a different hospital for external testing. Radiomic features were selected using mRMR. Using repeated nested cross-validation, the machine-learning models were trained on radiomic features and demographic information. For comparison, the external test set was evaluated by three radiology residents and one attending radiologist. Results: A LASSO classifier trained on radiomic features from all sequences performed best, with an AUC of 0.88, 70% sensitivity, 81% specificity, and 76% accuracy. In comparison, the radiology residents achieved 60–70% accuracy, 55–80% sensitivity, and 63–77% specificity, while the attending radiologist achieved 90% accuracy, 96% sensitivity, and 87% specificity. Conclusion: A radiogenomic model combining features from multiple MR sequences showed the best performance in predicting the MDM2 gene amplification status. The model showed a higher accuracy compared to the radiology residents, though lower compared to the attending radiologist.

Keywords: radiomics; machine learning; soft-tissue sarcomas; radiology; MRI

1. Introduction

Lipomatous tumors are the most common neoplasms encountered by physicians and the most frequent soft-tissue tumors of the extremities [1]. Of these, 40 to 45% are benign adipocytic tumors (lipomas) or atypical lipomatous tumors (ALTs) [2–5]. Lipomas only require treatment if the mass effect causes symptoms such as pain or functional disorders [6]. ALTs may show locally aggressive growth and may dedifferentiate into high-grade sarcomas [7–10]. Therefore, ALTs are typically resected [11]. Histopathological differentiation relies on the detection of atypical hyperchromatic nuclei and the immunohistochemical evaluation of the molecular analysis of the mouse double minute 2 (MDM2) gene [12]. However, the detection of these atypical hyperchromatic cells can be challenging since they are frequently scattered throughout the lesion, and detection is often complicated by fibrous septa, subsequently requiring a careful analysis of the entire tumor [12–14]. Previous studies have shown that the MDM2 amplification status is the most accurate marker to differentiate ALTs and lipomas, and there is a tendency towards sampling errors if the MDM2 status is not determined [12,15–17]. Unfortunately, the majority of MR imaging studies differentiating ALTs from lipomas did not include a molecular analysis, or only performed a molecular analysis in a subset of patients [6,14,18,19].

MR imaging is the standard imaging modality for the assessment of soft-tissue tumors due to its excellent soft-tissue contrast [20–22]. Specific imaging features such as the tumor size, tumor location, presence of thick septa, and amount of contrast uptake can be used to differentiate ALTs from lipomas [6,13,18,19,23]. However, since there is a substantial overlap between these imaging features in both tumor types, differentiating ALTs from lipomas is a challenging task. Moreover, previous studies of systematic radiologic readings have reported relatively low inter-observer reproducibility, with a kappa agreement ranging from 0.17 to 0.42 [13,19,24]. Given these challenges, it would be highly beneficial to develop a reliable diagnostic tool to differentiate ALTs from lipomas on preoperative MR images, thereby minimizing the need for invasive diagnostic procedures.

Machine-learning techniques, including imaging-based radiomics, permit a non-invasive detailed analysis of a tumor phenotype by using a quantitative imaging feature analysis [25,26]. However, one of the main challenges of radiomic models includes reproducibility in different datasets [27,28]. Therefore, the aim of this study was to develop and validate radiogenomic machine-learning models based on multiparametric MR examinations to predict the MDM2 gene amplification status in order to differentiate between ALTs and lipomas on preoperative MR images. The models were evaluated using an independent external cohort for testing and were compared to the performance of radiologists.

2. Materials and Methods

The local institutional review boards approved this retrospective multi-center study (ethics committee 666/21 S) The study was performed in accordance with our institutional ethic guidelines and the 1964 Declaration of Helsinki and its later amendments. Written and informed consent was waived for this retrospective anonymized analysis.

2.1. Datasets

We retrospectively reviewed the records of all patients with lipomatous tumors in the upper or lower extremities or trunk that had surgery performed at our sarcoma referral center between 2010 and 2021 (n = 573). Of these, 424 patients had a histologically confirmed diagnosis of a lipoma or an ALT. The MDM2 amplification status, determined by fluorescence in situ hybridization (FISH) of the MDM2 gene locus, was available for n = 257 patients. Patients without an MDM2 amplification status were excluded. Therefore, in the final dataset, both the histology and the MDM2 gene amplification status were available for all patients. Two senior pathologists specializing in the analysis of soft-tissue tumors provided a final consensus diagnosis based on the MDM2 gene amplification status and histology according to the World Health Organization criteria. The patient selection process is shown in Figure 1.

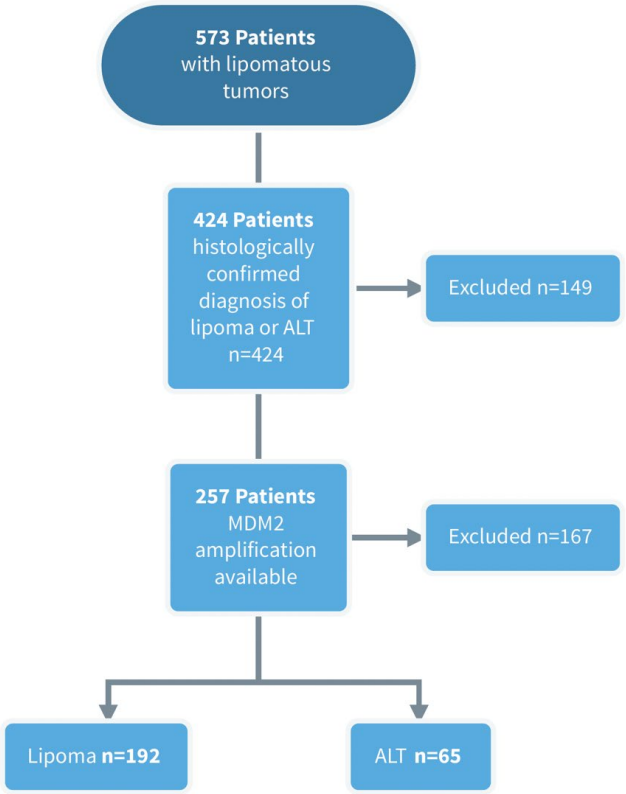


Figure 1. Subject selection flowchart. ALT = atypical lipomatous tumor; MDM2 = murine double minute.

In addition, an external test set was obtained from a further sarcoma referral center, the University Hospital of Freiburg (M1), for final independent testing and geographical validation. The external test set included patients with a diagnosis of a lipoma or an ALT confirmed by their histology and MDM2 amplification status.

2.2. MR Imaging Protocol and Image Segmentation

Pre-operative MR images were acquired using 3 or 1.5 Tesla scanners. Sequences were acquired in at least two planes that were oriented along the short and longitudinal axes of the long articulating bone(s). The protocols included a T2-w turbo spin echo (TSE) sequence (T2w), a T1-w TSE sequence (T1w), and a fat-saturated T1-w TSE sequence after the administration of a contrast agent (T1fsgd). Detailed information on the acquisition parameters is provided in Supplementary Material Table S1.

To define the volumes of interest (VOIs), tumor segmentations were performed manually by two radiology residents (S.C.F. and G.C.F.) using the open-source software 3D Slicer (3D Slicer, Version 4.8, stable release) and extracted as Neuroimaging Informatics Technology Initiative (NIfTI) label maps for further analysis. Multiple delineations were performed by S.C.F. and G.C.F. in 20 randomly selected patients to account for inter-reader variability.

2.3. Radiomic Feature Extraction and Machine-Learning Model Development

All preprocessing steps and radiomic feature extractions were conducted in accordance with the Imaging Biomarker Standardization Initiative guidelines [29] using the Python package PyRadiomics (version 2.2) implemented in Python (3.7), as previously described [30]. Image discretization was conducted using a bin width of 10 to achieve a bin count between 16 and 128, as recommended by the pyradiomics documentation [31]. Image intensity normalization was achieved via redistributing the image at the mean with a standard deviation and a scale of 100. Bspline interpolation was used to perform isotropic resampling to a voxel size of $1 \times 1 \times 1$ mm of the image and VOI mask. A total of 104 features were extracted from the original image of each sequence within the segmented label map (resulting in a total of 312 radiomic features), including first-order features, shape features, and texture features. The latter comprised "gray-level co-occurrence matrix" features, "gray-level size-zone matrix" features, "gray-level run-length matrix" features, "neighboring gray-tone difference matrix" features, and "gray-level dependence matrix" features. No features were extracted from filtered versions of the image due to a missing IBSI consensus. A detailed list of all extracted features is provided in Supplementary Material Table S2. Feature values were transformed to a common scale using min–max normalization in order to conserve their original distribution in the [0,1] range. Data normalization was performed prior to splitting the data into training and testing groups due to the batch harmonization step requirements. Nonparametric ComBatBatch harmonization was applied to account for the variability introduced by different MR scanners, as described previously [30]. Clinical features such as age, sex, and body region of the tumor (torso/head, upper extremity, or lower extremity) were also included. Categorical features were encoded into dummy numeric arrays using one hot encoder. All radiomic features susceptible to segmentation variations were excluded using a threshold intraclass correlation coefficient (ICC 3,1) of 0.8. This statistic resulted in 5, 15, and 4 radiomic features that were excluded from the T1w, T2w, and T1fsgd sequences, respectively. ICC 3,1 was chosen, as the raters were not rated as representative of a defined rater group due to their differing extents of training.

An estimate of the number of reduced features to use was calculated using a principal component analysis (PCA) with 95% of data variance: 11 to 13 features for the individual sequences (T1w, T2w, and T1fsgd) and 19 to 21 features for the combined features of all sequences. Each respective number of features was selected using minimum redundancy–maximum relevance (MRMR). Synthetic minority over-sampling and random under-sampling of the majority class were used to counteract the class imbalance. The ratios were tuned to find an optimal balance between data augmentation and data discard, with ratios of 0.5–0.6:1 after SMOTE and 0.6–0.8:1 after the random under-sampling of the majority class. The remaining class imbalance was handled by using balanced accuracy as the optimization criteria during hyperparameter optimization. Four machine-learning algorithms were implemented and compared in their performance: the support vector machine (SVM), the random forest classifier (RFC), the least absolute shrinkage and selection operator (LASSO; built from a stochastic gradient descent classifier), and a fully connected,

feedforward artificial neural network (ANN; multilayer perceptron classifier). A flow chart of the data processing and analysis of the radiomic features can be found in Supplementary Material Figure S1. For each algorithm, models were developed by (i) using demographic information only, (ii) using radiomic features for each individual sequence (T1w, T2w, or T1fsgd), (iii) using the radiomic features of all sequences, and (iv) using a combination of both the radiomic features of all sequences and demographic information. An overview of the radiomic workflow is shown in Figure 2.

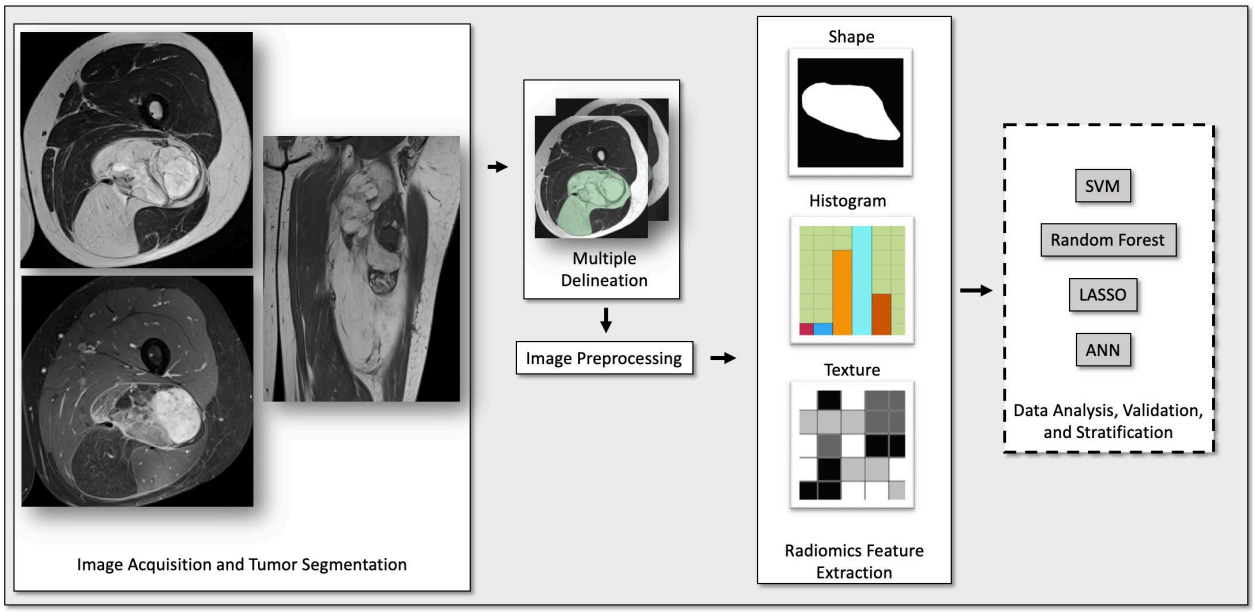


Figure 2. Radiomic workflow. Abbreviations: SVM, support vector machine; LASSO, least absolute shrinkage and selection operator; ANN, artificial neural network.

2.4. Model Optimization, Evaluation, and Statistical Analysis

Training and validation were performed using 3-fold nested cross-validation with 50 repetitions for statistical robustness, for a total of 150 averaged iterations per modeling algorithm and dataset. Hyperparameter optimization was conducted using an exhaustive grid search. This step was performed in the inner fold, after the feature selection step via MRMR, to prevent data leakage. Balanced accuracy was used as the optimization criterion to determine the best set of hyperparameters.

The performance of the models was evaluated with the area under the curve (AUC) obtained from the receiver–operator curve (ROC), plotted after averaging the yielded values. We also included the accuracy, sensitivity, and specificity as the output measures. For an unbiased evaluation, a final cross-validation step was implemented by selecting the best values obtained from the internal dataset before evaluating the performance on the external dataset. Stochastic gradient descent was used to calculate the probability of each class prediction. Calculations of model metrics were performed using scikit-learn (version 1.0.2).

For comparison, MR images of the external test set were rated independently by three radiology residents (I.L., S.C.F., and G.C.F., with 2, 3, and 5 years of experience, respectively) and one musculoskeletal imaging fellowship-trained radiologist (A.S.G., with 10 years of experience) experienced in musculoskeletal tumor imaging. All readers were blinded to all clinical and histopathological findings.

3. Results

3.1. Study Subjects

A total of 257 patients were included in the internal dataset (192 lipomas, 65 ALTs; age, 62.4 ± 14.5 years; 125 (48.6%) women). Fifty patients were included in the external dataset (30 lipomas, 20 ALTs; age, 60.6 ± 12.5 years; 22 (44%) women). All patients had a lipomatous tumor in one of the following six regions: chest, back, neck, leg, arm, hand, or foot. In both datasets, the highest number of patients had a tumor located in the leg (143/257 in the internal dataset and 27/50 in the external dataset), while the fewest number of patients had a tumor located in the foot (two in the internal dataset and none in the external dataset). Table 1 provides an overview of the subject characteristics.

Table 1. Patient characteristics.

Patient Characteristics	Internal Dataset (*n* = 257)	External Test Set (*n* = 50)
Age (years) *	62.4 ± 14.5	60.6 ± 12.5
Sex (women)	125	22
Tumor Location (Anatomical Region)		
Chest/Back	19	6
Neck	15	2
Leg	143	27
Arm	75	14
Hand	3	1
Foot	2	0
Lipomas	*n* = 192	*n* = 30
Age (years) *	62.3 ± 14.4	57.5 ± 11.1
Sex (women)	88	12
Atypical Lipomatous Tumors (ALT)	*n* = 65	*n* = 20
Age (years) *	62.5 ± 15	65.2 ± 13.5
Sex (women)	37	10

* Data are given as mean ± standard deviation.

3.2. Evaluation of the Developed Machine-Learning Models

Table 2 shows the final performance of the developed models on the external test set using demographic information only, radiomic features only (of all sequences combined), and a combination of demographic and radiomic features. The best-performing machine-learning model was based on a LASSO algorithm using a combination of all sequences, achieving an AUC of 0.88 at 70% sensitivity and 81% specificity with an accuracy of 76% on the external test set. The feature importance table, a confusion matrix, and a boxplot of the prediction probabilities from this model can be found in Supplementary Material Table S5, Supplementary Material Figure S2, and Supplementary Material Figure S3, respectively.

The AUC and accuracy for the individual sequences were lower for most models compared to models based on the radiomic parameters from all sequences combined, with a more imbalanced sensitivity/specificity. For T1w, the LASSO algorithm yielded an AUC of 0.83 at 80% sensitivity and 43% specificity with an accuracy of 58%. For T2w, the AUC was 0.82 at 42% sensitivity and 83% specificity with an accuracy of 69%. The highest AUC (0.84) was yielded for the T1fsgd sequences, though the sensitivity and specificity were highly imbalanced at 6% and 100%, respectively, with an accuracy of 60%. The performance of the developed models for the individual sequences on the external test set is shown in Supplementary Material Table S3.

Table 2. Performance of the machine-learning models on the external test set using demographic information or radiomic features only, as well as combining radiomic features and demographic information for the following model architectures: least absolute shrinkage and selection operator (LASSO), support vector machine (SVM), random forest classifier (RFC), and an artificial neural network (ANN). External performance represents the values yielded when a final cross-validation step considering only the best 150 best hyperparameter sets was implemented to predict the external test set.

Model Architecture	Score	Demographic Features	Combined Sequences	Combined Sequences + Demographic Features
LASSO	AUC *	0.56 (0.540.58) ± 0.07	0.88 (0.85–0.91) ± 0.07	0.72 (0.66–0.78) ± 0.15
	Accuracy	0.58	0.76	0.77
	Sensitivity	0.05	0.70	0.40
	Specificity	0.93	0.81	1.00
SVM	AUC *	0.54 (0.51–0.57) ± 0.12	0.84 (0.80–0.88) ± 0.11	0.85 (0.82–0.88) ± 0.09
	Accuracy	0.56	0.53	0.69
	Sensitivity	0.10	0.90	0.80
	Specificity	0.87	0.31	0.63
RFC	AUC *	0.63 (0.61–0.65) ± 0.06	0.87 (0.85–0.89) ± 0.05	0.87 (0.85–0.89) ± 0.05
	Accuracy	0.50	0.69	0.69
	Sensitivity	0.00	0.50	0.40
	Specificity	0.83	0.81	0.88
ANN	AUC *	0.68 (0.66–0.70) ± 0.08	0.81 (0.77–0.85) ± 0.10	0.81 (0.77–0.85) ± 0.10
	Accuracy	0.60	0.69	0.65
	Sensitivity	0.00	0.70	0.60
	Specificity	1.00	0.69	0.69

* Data are given as mean (95% confidence interval) ± standard deviation.

Interestingly, combining radiomic features and demographic information as the input for the machine-learning models did not improve the performance of the LASSO algorithm to differentiate ALTs from lipomas and resulted in a decrease in the sensitivity from 70% to 40%, though the specificity increased to 100%. The averaged nested cross-validation results of the internal dataset are shown in Supplementary Material Table S4. The training parameters and source code can be found online (https://github.com/deedeedav/alt-lipoma-radiomics (accessed on 9 March 2023)). Figure 3 shows an example of an ALT with typical imaging findings encasing the right gracilis muscle, while Figure 4 shows a typical example of a well-defined intramuscular lipoma in the right posterior thigh. Both cases were identified correctly by the machine-learning model.

3.3. Comparison with Radiologists

The results of the independent radiological readings of the external test are shown in Table 3. The radiology resident with 2 years of experience achieved an accuracy of 60%, a sensitivity of 55%, and a specificity of 63%; the resident with 3 years of experience achieved an accuracy of 70%, a sensitivity of 60%, and a specificity of 77%; and the radiology resident with 5 years of experience achieved an accuracy of 70%, a sensitivity of 80%, and a specificity of 63%. In comparison, the attending radiologist that was experienced in musculoskeletal tumor imaging achieved an accuracy of 90%, a sensitivity of 96%, and a specificity of 87%. Compared to the radiology residents, the model showed a higher accuracy and higher specificity, while the sensitivity was lower compared to the resident with 5 years of experience, but higher compared to the residents with 2 or 3 years of experience. The attending radiologist had a higher accuracy, sensitivity, and specificity. Figure 5 shows an ALT with atypical imaging findings located subcutaneously. The machine-learning model and the attending radiologist classified this tumor as an ALT, while all residents classified this tumor as a lipoma.

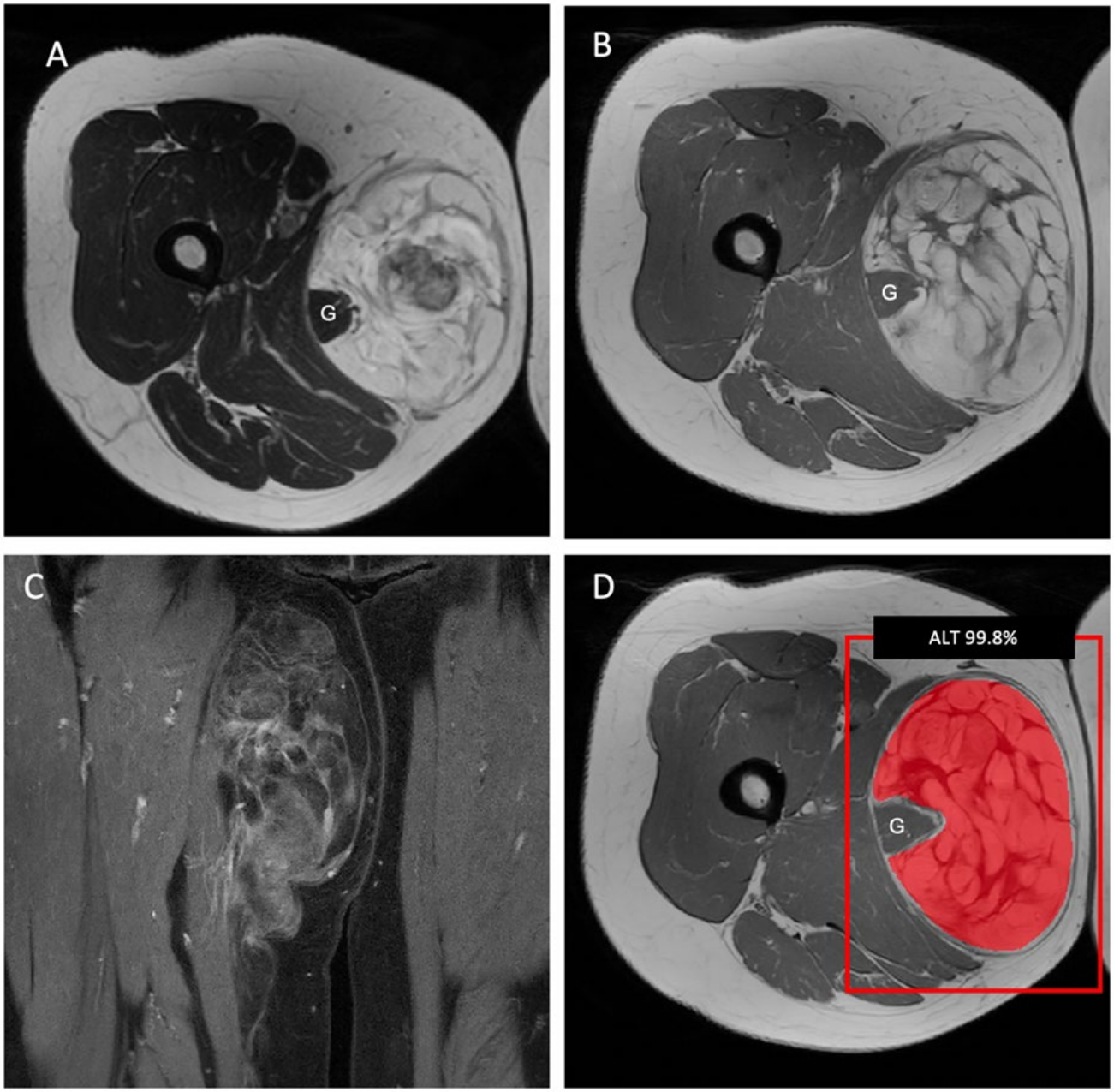


Figure 3. Lipomatous tumor in the medial right thigh, encasing the gracilis muscle (G). (**A**) The axial T2-weighted and (**B**) axial T1-weighted MR images show a large heterogeneous tumor with thick septa. (**C**) Septal contrast enhancement on the coronal T1-weighted images with fat saturation. (**D**) The machine-learning algorithm classified the tumor as an ALT with a probability of 99.8%. This diagnosis was confirmed by pathology and immunohistochemistry after surgical resection.

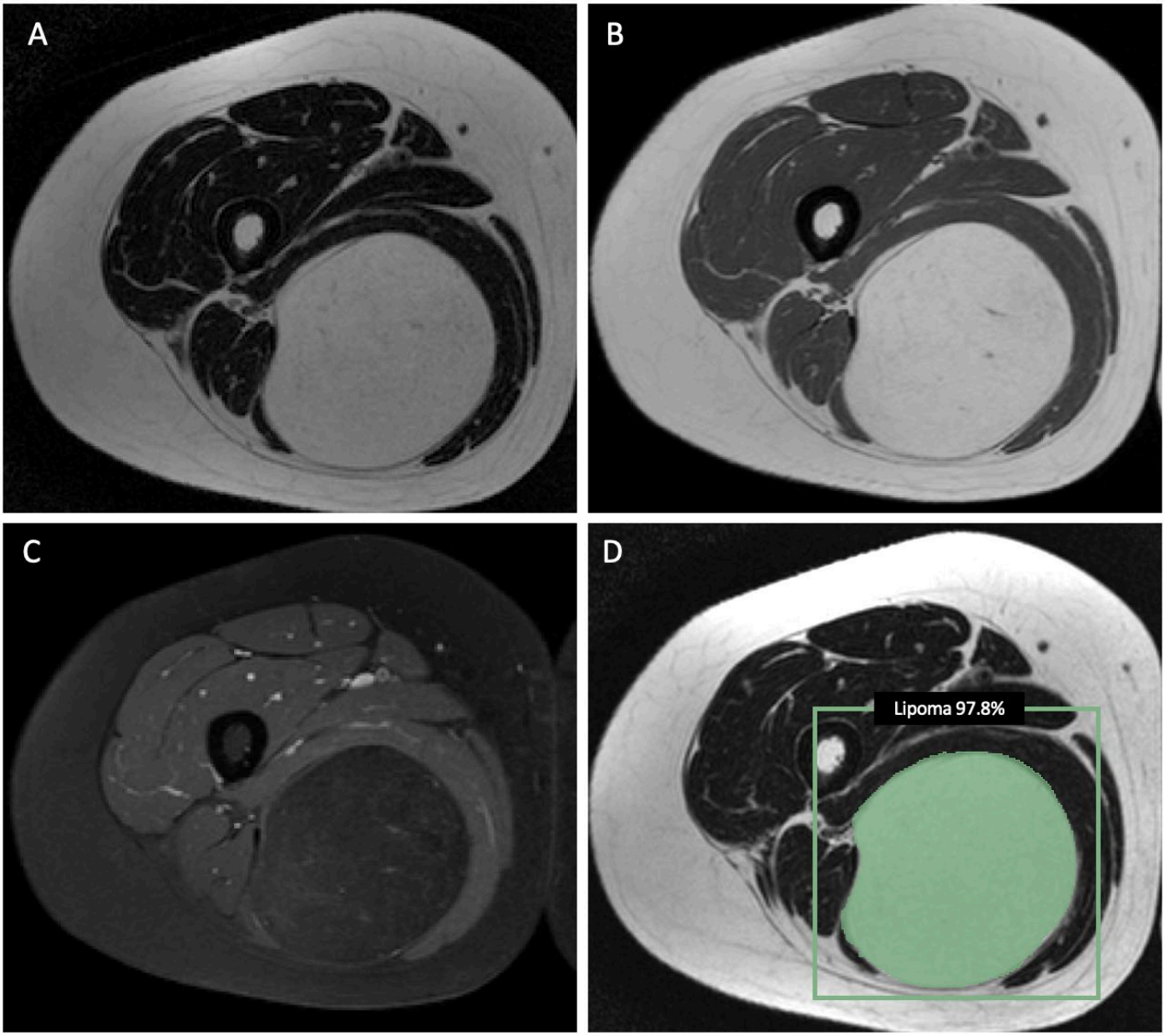


Figure 4. Axial T2-weighted (**A**) and T1-weighted (**B**) MR images showing a well-defined intramuscular lipomatous tumor (lipoma) in the right posterior thigh without significant contrast enhancement on the axial T1-weighted image with fat saturation (**C**). (**D**) The machine-learning model classified this tumor as a lipoma (probability of 97.8%). This was in accordance with the diagnosis made by the radiology residents and the attending radiologist.

Table 3. Performance of the radiology residents with 2, 3, or 5 years of experience and the fellowship-trained radiologist that was experienced in musculoskeletal tumor imaging. Readers were blinded to all clinical and histopathological findings.

Score	Radiology Resident, 2y	Radiology Resident, 3y	Radiology Resident, 5y	Fellowship-Trained Radiologist
Accuracy	0.60 (30/50)	0.70 (35/50)	0.70 (35/50)	0.90 (45/50)
Sensitivity	0.55 (11/20)	0.60 (12/20)	0.80 (16/20)	0.96 (19/20)
Specificity	0.63 (19/30)	0.77 (23/30)	0.63 (19/30)	0.87 (26/30)

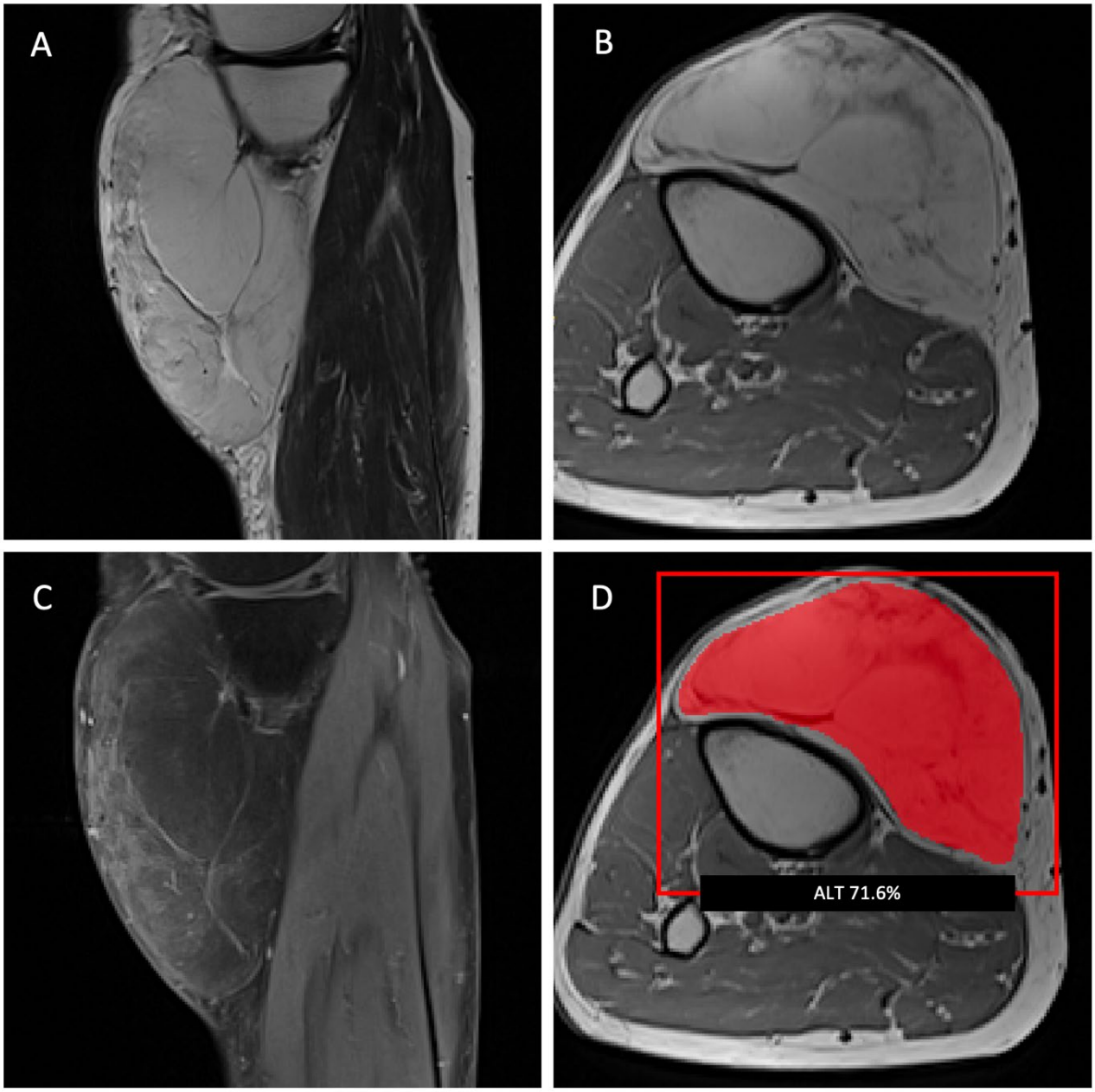


Figure 5. Sagittal T2-weighted (**A**) and axial T1-weighted (**B**) MR images of a lipomatous tumor located subcutaneously, anteromedial to the right proximal tibia. (**C**) A sagittal T1-weighted image with fat saturation shows a moderate septal contrast enhancement. All radiology residents classified this tumor as a lipoma, while the attending radiologist classified this tumor as an ALT. (**D**) The machine-learning algorithm also classified this tumor as an ALT with a probability of 71.6%. The diagnosis of an ALT was confirmed by pathology after surgical resection.

4. Discussion

In this study, machine-learning models were developed and validated to predict the amplification status of the MDM2 gene, to differentiate between atypical lipomatous tumors and lipomas on preoperative MR images, and to compare the results to the performance of radiologists using an external test set. The best-performing model was based on the combination of all MR sequences and achieved an AUC of 0.88 at 70% sensitivity and 81% specificity with an accuracy of 76%. In comparison, the accuracy of the readings by all radiology residents was lower, while the accuracy of the fellowship-trained radiologist was higher. Notably, the performance of the LASSO algorithm for each individual sequence

was lower compared to the model that included all sequences (T2w, T1w, and T1fsgd), suggesting that all sequences are required for optimal discrimination.

Radiomic models for differentiating lipomas from ALTs have previously been developed in smaller patient cohorts. Leporq et al. evaluated 2D radiomic models of 40 lipomas and 41 ALTs, including one MR image slice per patient [32]. Their best-performing model achieved an accuracy of 95% at 100% sensitivity and 90% specificity using the histology as the reference standard, though no specific information regarding the MDM2 gene amplification status was included, which may have led to a false classification of ALTs as lipomas [32]. Cay et al. evaluated 45 lipomas and 20 ALTs using histology and MDM2 amplification as the gold standards [33]. They achieved an AUC of 0.987 at 96.8% sensitivity and 93.72% specificity using 1000-fold bootstrapping [33]. However, since there was no separate test set, the algorithm was likely optimized on data used for validation in another bootstrapping iteration; therefore, these results may be inaccurately high [33]. A study by Vos et al. included 116 patients (58 lipomas and 58 ALTs) and used MDM2 amplification as the reference standard [34]. Their model performance was lower compared to our study, yielding an AUC of 0.81 at 66% sensitivity and 84% specificity with an accuracy of 75%. An important limitation of these aforementioned studies is that no external validation on an independent dataset was included. Also notably, the model performance was comparatively high in studies based on smaller patient cohorts ($n < 90$). A possible explanation may be a lack of variation in smaller datasets, which could affect the reproducibility in different datasets. However, this is not clear, since no external testing was included.

Interestingly, combining imaging parameters and clinical data did not improve the performance of most models for differentiating ALTs from lipomas, or only improved the performance marginally. While some demographic differences have been described between patients with ALTs and lipomas [23], it is likely that radiomic MR features are considerably more relevant for differentiating between these tumor types, and including parameters with less predictive power could hinder the capability of the models to identify relevant patterns. It should be noted that only a limited number of clinical features were included (age, sex, and tumor body region). Including additional clinical features may improve the predictive value of the radiomic models. Future studies could also include clinical outcome parameters to detect image-defined high-risk patients, thereby individualizing tumor treatment.

Some limitations are pertinent to this study. Since the cohort included only patients with histopathologically confirmed tumors, this potentially introduced a selection bias. Moreover, our specialized sarcoma center typically only receives larger or atypical lipomas on referral, subsequently increasing the amount of particularly challenging lipoma cases in the dataset. We also used manual segmentations as input for the models, and developing a pipeline that includes automated segmentations would be highly beneficial. In addition, more advanced sequences such as diffusion-weighted imaging or pharmacokinetic dynamic contrast-enhanced imaging were not included in the protocol. Including these sequences could potentially improve the differentiation between ALTs and lipomas. Finally, the developed models only differentiated between ALTs and lipomas, and while this is the most challenging and clinically relevant task, further studies are warranted on the ability to distinguish among all benign and malignant lipomatous tumors.

The advantages of the current study include its multicenter design, which allowed the evaluation of the models on an independent external test set, thereby reducing potential bias introduced by overfitting. Moreover, the dataset used for training was, to the best of our knowledge, the largest MRI dataset of histopathologically confirmed lipomas and ALTs. In addition, a histopathological analysis was conducted by pathologists specialized in the analysis of soft-tissue tumors and included the immunohistochemistry for the assessment of the MDM2 status in all cases. Furthermore, we excluded inter-/intra-reader segmentation-dependent features and included variability features, making the model performance more stable and reliable for other datasets.

5. Conclusions

In conclusion, radiogenomic models were developed that showed a high discriminatory power for predicting the MDM2 gene amplification status to distinguish between atypical lipomatous tumors and lipomas on preoperative MR images. The best-performing model was based on a LASSO algorithm using all MR sequences, with a higher accuracy compared to radiology residents, suggesting that these algorithms would be particularly helpful for radiologists with less experience. Due to the varying settings in which patients with lipomatous tumors present, this model may enhance the clinical diagnostic workup and improve the detection rate for atypical lipomatous tumors.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/cancers15072150/s1. Supplementary Material Table S1: Magnetic resonance imaging sequence parameters. Supplementary Material Table S2: Extracted radiomic features (n = 104). Supplementary Material Table S3: Performance of the machine-learning models on the external test set of each individual sequence (T1w, T2w, and T1fsgd) using the following model architectures: least absolute shrinkage and selection operator (LASSO), support vector machine (SVM), random forest classifier (RFC), and an artificial neural network (ANN). The external performance represents the values yielded when a final cross-validation step considering only the best 150 best hyperparameter sets was implemented. Supplementary Material Table S4: Internal performance representing the averaged values over 150 models resulting from the nested cross-validation using demographic information, radiomic features of each individual sequence (T1w, T2w, and T1fsgd), or radiomic features of all sequences combined, as well as combining radiomic features (of all sequences) and demographic information for the following model architectures: least absolute shrinkage and selection operator (LASSO), support vector machine (SVM), random forest classifier (RFC), and an artificial neural network (ANN). The metrics are given as mean ± standard deviation. Supplementary Material Table S5: Feature importance of the best-performing model (least absolute shrinkage and selection operator (LASSO) trained on features from all radiomic sequences). Supplementary Material Figure S1: Flow chart of the statistical analysis of the extracted radiomic features. Supplementary Material Figure S2: Confusion matrix of the best-performing model, a least absolute shrinkage and selection operator (LASSO) trained on all radiomic sequences. Misclassification rate: 0.23 ((FN + FP)/(N + P)). Supplementary Material Figure S3: Boxplot of the prediction probabilities made by the best-performing model (least absolute shrinkage and selection operator (LASSO) trained on features from all radiomic sequences). The probability cut-off used was 0.5.

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

Funding: This work was funded by grants from the German Society of Musculoskeletal Radiology (Deutsche Gesellschaft für muskuloskelettale Radiologie; DGMSR), the European Society of Musculoskeletal Radiology (ESSR), the Munich Clinician Scientist Program (MCSP) of the University of Munich (LMU; grant number ACS-10), and the Clinician Scientist Program (KKF) at Technische Universität München (TUM; grant number H-03).

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki, and approved by the Institutional Review Board of Klinikum Rechts der Isar, Technische Universität München (666/21 S, date of approval 24 November 2021).

Informed Consent Statement: Written informed consent was waived for this retrospective anonymized analysis.

Data Availability Statement: The training parameters and source code can be found online (https://github.com/deedeedav/alt-lipoma-radiomics (accessed on 9 March 2023)).

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

atypical lipomatous tumors	ALTs
mouse double minute 2	MDM2
fluorescence in situ hybridization	FISH
turbo spin echo	TSE
volume of interest	VOI
intraclass correlation coefficient	ICC
principal component analysis	PCA
Neuroimaging Informatics Technology Initiative	NIfTI
support vector machine	SVM
random forest classifier	RFC
least absolute shrinkage and selection operator	LASSO
artificial neural network	ANN
area under the curve	AUC
receiver–operator curve	ROC

References

1. Johnson, C.N.; Ha, A.S.; Chen, E.; Davidson, D. Lipomatous Soft-tissue Tumors. *J. Am. Acad. Orthop. Surg.* **2018**, *26*, 779–788. [CrossRef]
2. Dalal, K.M.; Antonescu, C.R.; Singer, S. Diagnosis and management of lipomatous tumors. *J. Surg. Oncol.* **2008**, *97*, 298–313. [CrossRef] [PubMed]
3. Myhre-Jensen, O.; Kaae, S.; Madsen, E.H.; Sneppen, O. Histopathological grading in soft-tissue tumours. Relation to survival in 261 surgically treated patients. *Acta Pathol. Microbiol. Immunol. Scand. A* **1983**, *91*, 145–150.
4. Rydholm, A.; Berg, N.O. Size, site and clinical incidence of lipoma. Factors in the differential diagnosis of lipoma and sarcoma. *Acta Orthop. Scand.* **1983**, *54*, 929–934. [CrossRef] [PubMed]
5. Fletcher, C.D. The evolving classification of soft tissue tumours: An update based on the new WHO classification. *Histopathology* **2006**, *48*, 3–12. [CrossRef]
6. Nagano, S.; Yokouchi, M.; Setoguchi, T.; Ishidou, Y.; Sasaki, H.; Shimada, H.; Komiya, S. Differentiation of lipoma and atypical lipomatous tumor by a scoring system: Implication of increased vascularity on pathogenesis of liposarcoma. *BMC Musculoskelet. Disord.* **2015**, *16*, 36. [CrossRef]
7. Bassett, M.D.; Schuetze, S.M.; Disteche, C.; Norwood, T.H.; Swisshelm, K.; Chen, X.; Bruckner, J.; Conrad, E.U., 3rd; Rubin, B.P. Deep-seated, well differentiated lipomatous tumors of the chest wall and extremities: The role of cytogenetics in classification and prognostication. *Cancer* **2005**, *103*, 409–416. [CrossRef]
8. Weiss, S.W.; Rao, V.K. Well-differentiated liposarcoma (atypical lipoma) of deep soft tissue of the extremities, retroperitoneum, and miscellaneous sites. A follow-up study of 92 cases with analysis of the incidence of "dedifferentiation". *Am. J. Surg. Pathol.* **1992**, *16*, 1051–1058. [CrossRef] [PubMed]
9. Bidault, F.; Vanel, D.; Terrier, P.; Jalaguier, A.; Bonvalot, S.; Pedeutour, F.; Couturier, J.M.; Dromain, C. Liposarcoma or lipoma: Does genetics change classic imaging criteria? *Eur. J. Radiol.* **2009**, *72*, 22–26. [CrossRef]
10. Evans, H.L.; Soule, E.H.; Winkelmann, R.K. Atypical lipoma, atypical intramuscular lipoma, and well differentiated retroperitoneal liposarcoma: A reappraisal of 30 cases formerly classified as well differentiated liposarcoma. *Cancer* **1979**, *43*, 574–584. [CrossRef]
11. Choi, K.Y.; Jost, E.; Mack, L.; Bouchard-Fortier, A. Surgical management of truncal and extremities atypical lipomatous tumors/well-differentiated liposarcoma: A systematic review of the literature. *Am. J. Surg.* **2020**, *219*, 823–827. [CrossRef]
12. Zhang, H.; Erickson-Johnson, M.; Wang, X.; Oliveira, J.L.; Nascimento, A.G.; Sim, F.H.; Wenger, D.E.; Zamolyi, R.Q.; Pannain, V.L.; Oliveira, A.M. Molecular testing for lipomatous tumors: Critical analysis and test recommendations based on the analysis of 405 extremity-based tumors. *Am. J. Surg. Pathol.* **2010**, *34*, 1304–1311. [CrossRef] [PubMed]
13. Brisson, M.; Kashima, T.; Delaney, D.; Tirabosco, R.; Clarke, A.; Cro, S.; Flanagan, A.M.; O'Donnell, P. MRI characteristics of lipoma and atypical lipomatous tumor/well-differentiated liposarcoma: Retrospective comparison with histology and MDM2 gene amplification. *Skelet. Radiol.* **2013**, *42*, 635–647. [CrossRef] [PubMed]
14. Ohguri, T.; Aoki, T.; Hisaoka, M.; Watanabe, H.; Nakamura, K.; Hashimoto, H.; Nakamura, T.; Nakata, H. Differential diagnosis of benign peripheral lipoma from well-differentiated liposarcoma on MR imaging: Is comparison of margins and internal characteristics useful? *AJR Am. J. Roentgenol.* **2003**, *180*, 1689–1694. [CrossRef]

15. Dei Tos, A.P.; Doglioni, C.; Piccinin, S.; Sciot, R.; Furlanetto, A.; Boiocchi, M.; Dal Cin, P.; Maestro, R.; Fletcher, C.D.; Tallini, G. Coordinated expression and amplification of the MDM2, CDK4, and HMGI-C genes in atypical lipomatous tumours. *J. Pathol.* **2000**, *190*, 531–536. [CrossRef]
16. Kulkarni, A.S.; Wojcik, J.B.; Chougule, A.; Arora, K.; Chittampalli, Y.; Kurzawa, P.; Mullen, J.T.; Chebib, I.; Nielsen, G.P.; Rivera, M.N.; et al. MDM2 RNA In Situ Hybridization for the Diagnosis of Atypical Lipomatous Tumor: A Study Evaluating DNA, RNA, and Protein Expression. *Am. J. Surg. Pathol.* **2019**, *43*, 446–454. [CrossRef] [PubMed]
17. Kashima, T.; Halai, D.; Ye, H.; Hing, S.N.; Delaney, D.; Pollock, R.; O'Donnell, P.; Tirabosco, R.; Flanagan, A.M. Sensitivity of MDM2 amplification and unexpected multiple faint alphoid 12 (alpha 12 satellite sequences) signals in atypical lipomatous tumor. *Mod. Pathol.* **2012**, *25*, 1384–1396. [CrossRef]
18. Kransdorf, M.J.; Bancroft, L.W.; Peterson, J.J.; Murphey, M.D.; Foster, W.C.; Temple, H.T. Imaging of fatty tumors: Distinction of lipoma and well-differentiated liposarcoma. *Radiology* **2002**, *224*, 99–104. [CrossRef] [PubMed]
19. Nardo, L.; Abdelhafez, Y.G.; Acquafredda, F.; Schiro, S.; Wong, A.L.; Sarohia, D.; Maroldi, R.; Darrow, M.A.; Guindani, M.; Lee, S.; et al. Qualitative evaluation of MRI features of lipoma and atypical lipomatous tumor: Results from a multicenter study. *Skelet. Radiol.* **2020**, *49*, 1005–1014. [CrossRef]
20. De Schepper, A.M.; De Beuckeleer, L.; Vandevenne, J.; Somville, J. Magnetic resonance imaging of soft tissue tumors. *Eur. Radiol.* **2000**, *10*, 213–223. [CrossRef] [PubMed]
21. Vilanova, J.C.; Woertler, K.; Narvaez, J.A.; Barcelo, J.; Martinez, S.J.; Villalon, M.; Miro, J. Soft-tissue tumors update: MR imaging features according to the WHO classification. *Eur. Radiol.* **2007**, *17*, 125–138. [CrossRef]
22. Totty, W.G.; Murphy, W.A.; Lee, J.K. Soft-tissue tumors: MR imaging. *Radiology* **1986**, *160*, 135–141. [CrossRef]
23. Knebel, C.; Neumann, J.; Schwaiger, B.J.; Karampinos, D.C.; Pfeiffer, D.; Specht, K.; Lenze, U.; von Eisenhart-Rothe, R.; Rummeny, E.J.; Woertler, K.; et al. Differentiating atypical lipomatous tumors from lipomas with magnetic resonance imaging: A comparison with MDM2 gene amplification status. *BMC Cancer* **2019**, *19*, 309. [CrossRef] [PubMed]
24. O'Donnell, P.W.; Griffin, A.M.; Eward, W.C.; Sternheim, A.; White, L.M.; Wunder, J.S.; Ferguson, P.C. Can Experienced Observers Differentiate between Lipoma and Well-Differentiated Liposarcoma Using Only MRI? *Sarcoma* **2013**, *2013*, 982784. [CrossRef]
25. Peeken, J.C.; Asadpour, R.; Specht, K.; Chen, E.Y.; Klymenko, O.; Akinkuoroye, V.; Hippe, D.S.; Spraker, M.B.; Schaub, S.K.; Dapper, H.; et al. MRI-based delta-radiomics predicts pathologic complete response in high-grade soft-tissue sarcoma patients treated with neoadjuvant therapy. *Radiother. Oncol.* **2021**, *164*, 73–82. [CrossRef]
26. Peeken, J.C.; Wiestler, B.; Combs, S.E. Image-Guided Radiooncology: The Potential of Radiomics in Clinical Application. *Recent Results Cancer Res.* **2020**, *216*, 773–794. [CrossRef]
27. Crombe, A.; Fadli, D.; Italiano, A.; Saut, O.; Buy, X.; Kind, M. Systematic review of sarcomas radiomics studies: Bridging the gap between concepts and clinical applications? *Eur. J. Radiol.* **2020**, *132*, 109283. [CrossRef] [PubMed]
28. Gitto, S.; Cuocolo, R.; Albano, D.; Morelli, F.; Pescatori, L.C.; Messina, C.; Imbriaco, M.; Sconfienza, L.M. CT and MRI radiomics of bone and soft-tissue sarcomas: A systematic review of reproducibility and validation strategies. *Insights Imaging* **2021**, *12*, 68. [CrossRef] [PubMed]
29. Zwanenburg, A.; Vallieres, M.; Abdalah, M.A.; Aerts, H.; Andrearczyk, V.; Apte, A.; Ashrafinia, S.; Bakas, S.; Beukinga, R.J.; Boellaard, R.; et al. The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping. *Radiology* **2020**, *295*, 328–338. [CrossRef] [PubMed]
30. Peeken, J.C.; Neumann, J.; Asadpour, R.; Leonhardt, Y.; Moreira, J.R.; Hippe, D.S.; Klymenko, O.; Foreman, S.C.; von Schacky, C.E.; Spraker, M.B.; et al. Prognostic Assessment in High-Grade Soft-Tissue Sarcoma Patients: A Comparison of Semantic Image Analysis and Radiomics. *Cancers* **2021**, *13*, 1929. [CrossRef] [PubMed]
31. Tixier, F.; Le Rest, C.C.; Hatt, M.; Albarghach, N.; Pradier, O.; Metges, J.P.; Corcos, L.; Visvikis, D. Intratumor heterogeneity characterized by textural features on baseline 18F-FDG PET images predicts response to concomitant radiochemotherapy in esophageal cancer. *J. Nucl. Med.* **2011**, *52*, 369–378. [CrossRef] [PubMed]
32. Leporq, B.; Bouhamama, A.; Pilleul, F.; Lame, F.; Bihane, C.; Sdika, M.; Blay, J.Y.; Beuf, O. MRI-based radiomics to predict lipomatous soft tissue tumors malignancy: A pilot study. *Cancer Imaging* **2020**, *20*, 78. [CrossRef] [PubMed]
33. Cay, N.; Mendi, B.A.R.; Batur, H.; Erdogan, F. Discrimination of lipoma from atypical lipomatous tumor/well-differentiated liposarcoma using magnetic resonance imaging radiomics combined with machine learning. *Jpn. J. Radiol.* **2022**, *40*, 951–960. [CrossRef] [PubMed]
34. Vos, M.; Starmans, M.P.A.; Timbergen, M.J.M.; van der Voort, S.R.; Padmos, G.A.; Kessels, W.; Niessen, W.J.; van Leenders, G.; Grunhagen, D.J.; Sleijfer, S.; et al. Radiomics approach to distinguish between well differentiated liposarcomas and lipomas on MRI. *Br. J. Surg.* **2019**, *106*, 1800–1809. [CrossRef] [PubMed]

4.2 The importance of planning CT-based imaging features for machine learning-based prediction of pain response

scientific reports

Check for updates

OPEN

The importance of planning CT-based imaging features for machine learning-based prediction of pain response

Óscar Llorián-Salvador[1,2,3], Joachim Akhgar[1], Steffi Pigorsch[1], Kai Borm[1], Stefan Münch[1], Denise Bernhardt[1,4,5], Burkhard Rost[2], Miguel A. Andrade-Navarro[3], Stephanie E. Combs[1,4,5] & Jan C. Peeken[1,4,5]✉

Patients suffering from painful spinal bone metastases (PSBMs) often undergo palliative radiation therapy (RT), with an efficacy of approximately two thirds of patients. In this exploratory investigation, we assessed the effectiveness of machine learning (ML) models trained on radiomics, semantic and clinical features to estimate complete pain response. Gross tumour volumes (GTV) and clinical target volumes (CTV) of 261 PSBMs were segmented on planning computed tomography (CT) scans. Radiomics, semantic and clinical features were collected for all patients. Random forest (RFC) and support vector machine (SVM) classifiers were compared using repeated nested cross-validation. The best radiomics classifier was trained on CTV with an area under the receiver-operator curve (AUROC) of 0.62 ± 0.01 (RFC; 95% confidence interval). The semantic model achieved a comparable AUROC of 0.63 ± 0.01 (RFC), significantly below the clinical model (SVM, AUROC: 0.80 ± 0.01); and slightly lower than the spinal instability neoplastic score (SINS; LR, AUROC: 0.65 ± 0.01). A combined model did not improve performance (AUROC: 0,74 ± 0,01). We could demonstrate that radiomics and semantic analyses of planning CTs allowed for limited prediction of therapy response to palliative RT. ML predictions based on established clinical parameters achieved the best results.

Bone metastasis, a common complication in oncology, poses significant difficulties in predicting pain response for patients. Machine learning (ML) techniques have been often used to address different oncological challenges, given the innovative approach they offer[1–5].

There is a significant amount of cancer research based on ML techniques, applying different ML algorithms such as support vector machines (SVMs) and random forest classifiers (RFCs)[6–8]. One field that has experienced a rapid growth over the last few years thanks to the use of ML techniques to extract information from these features is radiomics[9–13].

Radiomics data can be used for training ML models to predict clinical or biological outcomes[14–16]. Radiomics has been employed across different cancer to anticipate survival, disease prognosis, tumour response, molecular abnormalities, as well as identifying metastases or regions of invasive tumour growth[17–28].

Nonetheless, the use of radiomics feature analysis to predict non-tumour radiotherapy (RT) response hasn't been extensively explored. A few investigations have examined the projection of RT-related complications, including xerostomia, pneumonitis or proctitis[29–31]. In the context of bone metastasis, unfortunately, there remains a dearth of studies, with only a few focusing on the prediction of non-tumour RT responses[32–35]. However, there are general limitations for ML-related studies in this domain, where dataset sizes are significantly smaller than expected for the more common ML algorithms. Without the proper statistical strengthening of the resampling

[1]Department of Radiation Oncology, Klinikum Rechts der Isar, Technical University of Munich (TUM), Ismaninger Straße 22, 81675 Munich, Germany. [2]Department for Bioinformatics and Computational Biology, Informatik 12, Technical University of Munich (TUM), Boltzmannstraße 3, 85748 Garching, Germany. [3]Institute of Organismic and Molecular Evolution, Johannes Gutenberg University Mainz, Hanns-Dieter-Hüsch-Weg 15, 55128 Mainz, Germany. [4]Department of Radiation Sciences (DRS), Institute of Radiation Medicine (IRM), Helmholtz Zentrum, 85764 München, Germany. [5]Deutsches Konsortium für Translationale Krebsforschung (DKTK), Partner Site Munich, 69120 Heidelberg, Germany. ✉email: jan.peeken@tum.de

technique, this problem can potentially lead to wider error margins and, on occasions, overoptimistic results. Nevertheless, these studies underscore the importance of further research in this domain.

Painful spinal bone metastases (PSBMs) are regularly treated by palliative RT. About two thirds of the patients experience a partial or complete response in terms of pain reduction[34]. The role of biomarkers and personalised RT in PSBM cases has become increasingly prominent[36–38]. Clinical parameters, such as age, Karnofsky performance score (KPS), use of opioids or cancer histology (e.g. breast or prostate cancer), show limited predictive capabilities to identify patients that profit from palliative RT[33]. The Spinal instability neoplastic score (SINS) has been developed by the Spine Oncology Study Group to assess instability of spinal bone metastases[39]. At the same time, the SINS provides a semantic tool to predict pain response to RT[34].

In this retrospective study we sought to determine the potential of ML-based prediction of RT therapy response of PSBM. Besides clinical features, we investigated whether CT-based radiomics features and semantic features can be used to predict pain response, as well. The best strategy for the definition of volumes of interest (VOI) in regard to macroscopic or microscopic metastatic expansion was assessed for radiomics feature extraction. In order to statistically strengthen and produce more robust results, SVM, RFC and logistic regression (LR) models were trained, evaluated and compared using repeated nested cross-validation, stratifying the splits for multiple patient samples.

Materials and methods

Clinical data curation

Patient records of all (n = 491) patients treated with palliative RT for bone metastases between 2009 and 2017 at our institution were analysed. Patients with non-spinal metastases, previous interventions (e.g., surgical stabilization or kyphoplasty) or RT, haematological bone manifestations, and missing information regarding pain response were excluded (Figure S1 for a patient workflow).

Patient demographics were assessed for each patient (Table 1 for characteristics of patients, RT and metastatic disease). Clinical parameters previously shown to be associated with pain response such as KPS, age, use of opioids, and histology (breast cancer, non-small cell lung cancer (NSCLC) and others) were determined and used as input for the clinical ML models (Table S1 for the exact distribution of histologies)[33,34,40,41]. These clinical features were measured prior to RT. Histology, as the only categorical value present in the clinical data, was encoded into three dummy binary features.

Pain response was rated retrospectively on the basis of patient records following the "international consensus on palliative radiotherapy endpoints for future clinical trials in bone metastases" at the first follow-up visit 6 weeks after RT[42]: complete response: "pain score of 0 at treated site with no concomitant increase in analgesic intake", partial response: "Pain reduction of at least 2 at the treated site (scale of 0 to 10) without analgesic

Feature	Possible values
Imaging—Bone reaction	Blastic reaction
	Mixed reaction (lytic/blastic)
	Lytic reaction
Soft tissue component	Yes
	No
GTV classification	Any portion of vertebral body
	Lateralized within body
	Diffuse within body
	Body + unilateral pedicle
	Body + bilateral pedicle/transverse processBody
	Unilateral pedicle
	Unilateral lamina
	Spinous process
Posterolateral involvement of the spinal elements	Bilateral
	Unilateral
	None of the above
Vertebral body collapse	> 50% collapse
	< 50% collapse
	No collapse with > 50% body involved
	None of the above
Location	Junctional
	Mobile
	Semirigid
	Rigid

Table 1. List of semantic features.

increase, or analgesic reduction of at least 25% without pain increase", pain progression: "increase in pain score of at least 2 or increase of analgesics of at least 25%", and indeterminate response or "no response": "no response or any response not captured by the other categories". In both complete and partial responses, patient-rated worst pain measures were used.

Planning CT images acquisition parameter and orientation were performed via axial reconstruction of cross-sectional images using a Siemens Somatom Emotion 16 with 3 mm slice thickness and 0.98 mm × 0.98 mm resolution (Table S2 shows all CT image acquisition parameters). The SINS was determined by visual assessment of planning CTs following the definition of the Spine Oncology Study Group[39]. Visual assessment was performed by JA and supervised by JCP. The SINS was used for ML modelling both as a discrete variable and as a binary variable using a threshold of 7. Approval from the institutional review board of the Technical University of Munich hospital was received (reference number 466/16 s). All patients were treated after informed consent. All experiments were performed in accordance with local legal regulation allowing retrospective data analysis.

Definition of VOIs

For each metastasis, two separate VOI definitions were segmented on the planning CT scans using Eclipse 13.0 (Varian Medical Systems, Palo Alto, USA) (Table S2 for acquisition parameters). First, the visible blastic and/or lytic gross tumour volume (GTV) including any adjacent soft-tissue component was manually segmented. Secondly, a clinical target volume (CTV) considering potential microscopic spread was segmented following the International Spine Radiosurgery Consortium Consensus Guidelines for Target Volume Definition in Spinal Stereotactic Radiosurgery[43]. The segmentation process of the CTV was performed manually by HA and supervised by JCP.

Radiomics feature extraction

Pre-processing and radiomics feature extraction were performed using the pyRadiomics library (version 2.0) in Python (version 3.6.4) (Fig. 1 for study workflow)[44]. For pre-processing, a fixed bin width of 20 was used for image discretization. The intensity ranges between all patients were 218–2083 HU (Min–Max). In accordance to earlier studies, and the pyRadiomics guidelines for images with similar characteristics, a bin width of 20 was chosen in order to retain a bin number in the range of 30–130[45]. 105 radiomics features, including shape, first-order, and texture features were computed from the original image. Texture features were calculated in 3D. Gray Level Co-occurrence Matrix (GLCM) and Gray Level Run Length Matrix (GLRLM) texture features were calculated separately for each direction and then averaged. All extracted features were computed according to the "image biomarker standardization initiative" guidelines (Table S3)[46].

Semantic features extraction

Semantic features from the SINS score and other imaging descriptors were determined by an MD student (JA) and controlled by a radiation oncology resident with 3 years of experience (JCP) (see Table 1 for a complete listing). The resident trained the medical student on a per-patient basis for the first 20 patients together. Subsequently random patients were controlled and all patients with more difficult allocation to a semantic group.

Many of the semantic features are part of the SINS score (Location, Bone Reaction, Vertebral Collapse 50%, posterolateral involvement) which has been correlated with pain response[34]. For GTV classification, the extent of metastasis is part of the CTV definition recommendations and has not yet been associated with response. Soft Tissue Component was once tested in one study without showing an association with tumour response[47].

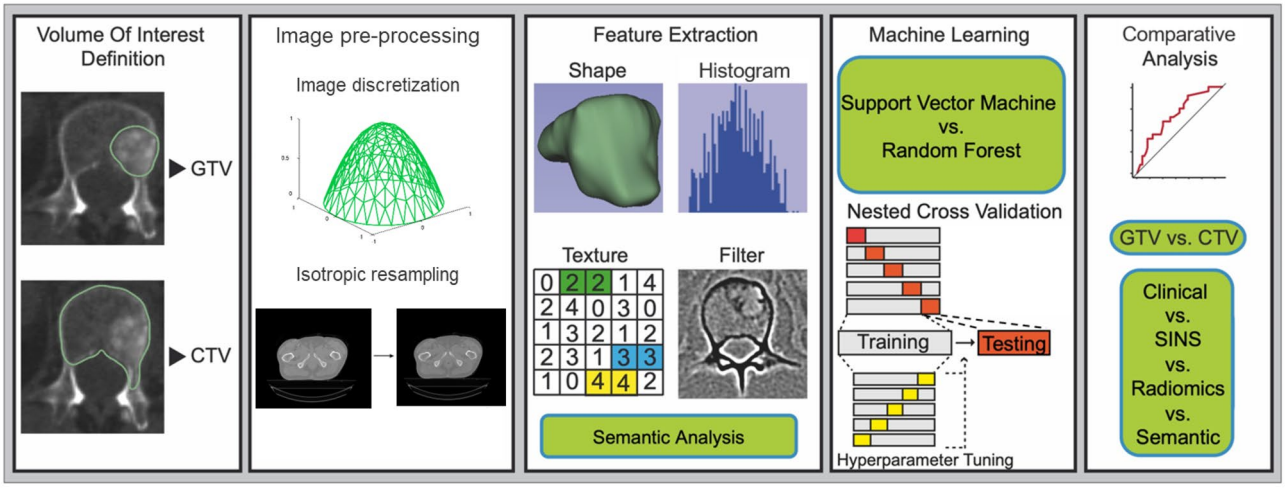


Figure 1. Workflow.

Machine learning modelling

The number of patients was filtered by removing incomplete entries, taking the intersection of patients with all CTV, GTV, semantic, clinical and SINS data, and performing outlier detection. This resulted in a dataset with 230 pre-processed PSBM with known outcome. For feature reduction, both redundancy reduction and feature correlation to the prediction target were taken into consideration with the Maximum Relevance – Minimum Redundancy (MRMR) algorithm (mrmr-selection library, version 0.2.2)[48]. For all feature sets larger than 15 throughout this study, the best 15 features were selected to be used by the respective ML algorithms, so that the number of features for every model amount to up to 10% of the number of samples.

Given the small dataset size and to ensure a correct hyperparameter optimization, nested fivefold cross-validation was applied to train and validate the ML models. However, multiple samples coming from the same patient, present in the same data subsample, may lead to biased and over-optimistic results. To offset this, cross-validation splits were stratified by patient ID: this way, there is an even distribution of such samples across the 5 splits in either fold. In order to correct the moderate class imbalance (negative to positive ratio of 3.11:1), Synthetic Minority Oversampling Technique (SMOTE) was used (imbalanced-learn library, version 0.8.0). To avoid overfitting by class repetition, random minority class oversampling was complemented with random majority class undersampling. The normalisation, feature selection and class imbalance correction steps were performed in the inner fold of the nested cross-validation to avoid data leakage and bias. Nested cross-validation was repeated for 50 iterations, for a total of 250 aggregated models, to increase the statistical strength of the results.

Hyperparameter optimization was performed via exhaustive grid search in the inner fold of the nested cross-validation, using balanced accuracy (BA) as the optimization criteria. SVM and RFC were used for training on the multivariable datasets i.e., both radiomic segmentations, the semantic and the clinical feature sets; and Logistic Regression (LR) was used for the analysis of SINS. For SVM, the hyperparameters optimized were C, gamma (when applicable), the degree (when applicable) and the kernel used. For RFC, the hyperparameters optimized were max_features, max_depth, min_samples_split, min_samples_leaf, bootstrap, and criterion. The only hyperparameter optimized for LR was C. A summary of the optimized hyperparameters of the best radiomics, combined and overall modelling strategies can be found in Table S12. All models come from the scikit-learn library[49] (version 0.24.2). Firstly, these models were trained on both segmentation modes: CTV and GTV to assess their predictive quality against a binary prediction target (Table 3): complete pain response (complete response vs partial response/indeterminate response/no response/pain progression). Results were compared to determine the best modelling strategy. The best model was then compared to clinical, SINS, and semantic models (Table 4). Finally, multiple combined models were devised to assess whether combined models performed better (Tables 5 and S6).

The importance given by models to their features was recorded in order to analyse the feature importance for all models developed. Since it is not possible to track the weight of features for non-linear kernels in SVM, only the percentage of feature selection was shown. For RFC models, this importance is shown as the Gini Importance or mean decrease in impurity of the nodes (the higher, the more important).

Statistical analysis

Given the small dataset size and, therefore, unclear class distribution, Min–max normalisation was performed to scale all features (scikit-learn library, version 0.24.2), while retaining the same distribution. Outlier detection is performed before the nested cross-validation (where normalization, feature selection and class balancing are conducted) to avoid extreme values from affecting the distribution of the data (scikit-learn library, version 0.24.2). All error margins are reported as standard errors with a coefficient of 1.96 for a confidence interval covering 95% of the observations. All models were evaluated, principally, using the Area Under the Receiver-Operator Curve (AUROC). In addition, BA, F1 score and Matthews Correlation Coefficient (MCC) were secondarily examined. The most important AUROC comparisons have been quantitatively evaluated with the Mann–Whitney U test to determine whether they follow the same distribution (null hypothesis), using a p-value of 0.05 for a 95% confidence interval. Given the dataset size limitation, the models trained on either radiomics segmentation did not use the intersection of all available patients but all available. This has prevented the possibility of using a DeLong test for quantitative AUROC evaluation. Statistical analysis and radiomics model building were performed using Python (version 3.7) and conducted by OL-S.

Results

Pain response to RT

A retrospective cohort of 90 patients with a total of 267 PSBM fitted the inclusion and exclusion criteria in our institution (Figure S1 for a patient workflow). Mammary carcinoma, prostate carcinoma and NSCLC were the three most frequent (63%) cancer types (Table 2 for patient characteristics and Table S1 for a distribution of all cancer histologies). There was a median of two PSBMs per patient with a total of 41 solitary PSBMs. Partial and complete pain response retrospectively assessed from patient files was achieved in 33% and 52% of patients, respectively.

Determination of the best VOI for radiomics analysis and modelling strategy

The best performing model was a RFC trained on the CTV radiomics segmentation, with the highest overall scores (AUROC: 0.62 ± 0.01) (Table 3 for outcome metrics and Fig. 2 for ROC and calibration curves). While the data was imbalanced towards the negative class (no complete pain response), it has performed better when predicting the true positive class (complete pain response), as it can be seen in the confusion matrix provided (Figure S3). This is further confirmed by a higher specificity (0.72) than sensitivity (0.44). While the RFC reached the highest performance, the SVM results were more stable. The best segmentation mode was CTV, with higher

Patient characteristic	Complete response (n = 30 p)	Partial or no response (n = 60 p)	*p* value[a]
Gender: Male	14 p (43%)	29 p (48%)	0.82
Gender: Female	16 p (57%)	31 p (52%)	
Age	m 66 (r 26–88)	m 66 (r 30–87)	0.74
Karnofsky Performance score	m 70 (r 60–100)	m 80 (r 30–90)	0.34
Opioid medication	16 p (57%)	37 p (62%)	0.50
Tumour type	Mammary/prostate carcinoma: 11 p (37%)	Mammary/prostate carcinoma: 31 p (52%)	0.30
	NSCLC: 7 p (23%)	NSCLC: 8 p (13%)	
	Others: 12 p (40%)	Others: 21 p (35%)	
Partial response	–	47 p (78%)	–
Overall survival	m 5.5 months	m 7.5 months	0.90
	(r 0.7–55.8 months)	(r 0.1–68.1 months)	
Radiotherapy			
Single dose	m 3 (r 2–8)	m 3 (r 2–8)	0.54
Total dose	m 33 (r 8–44)	m 30 (r 8–45)	0.10
Number of fractions	m 10 (r 1–22)	m 10 (r 1–19)	0.45
Bone metastases			
Number of metastases	65	196	
Number of metastases per patient	m 1.5 (r 1–6)	m 2.5 (r 1–10)	0.055
Previous RT	0 p (0%)	0 p (0%)	–
Localization	Sacrum: 8 p (12%)	Sacrum: 17 p (9%)	0.26
	Lumbar: 34 p (52%)	Lumbar: 83 p (42%)	
	Thoracic: 21 p (32%)	Thoracic: 83 p (42%)	
	Cervical: 2 p (3%)	Cervical: 13 p (7%)	
Bone reaction	Blastic: 15 p (23%)	Blastic: 56 p (29%)	0.03
	Lytic: 31p (48%)	Lytic: 28 p (14%)	
	Mixed: 19 p (29%)	Mixed: 112 p (57%)	
Soft tissue component	25 p (38%)	48 p (25%)	0.08
Extent of metastasis[b]	vertebral body: 17 p (26%)	vertebral body: 54 p (28%)	0.03
	body/pedicle: 4 p (6%)	body/pedicle: 9 p (5%)	
	`body/pedicle/transverse process: 2 p (3%)	body/pedicle/transverse process: 8 p (4%)	
	Unilateral pedicle: 23 p (35%)	Unilateral pedicle: 35 p (18%)	
	Unilateral lamina: 18 p (28%)	Unilateral lamina: 88 p (45%)	
	Spinous process: 1 p (2%)	Spinous process: 3 p (2%)	
SINS	m 7 (3–14)	m 8 (0–15)	0.02

Table 2. Characteristics of patients, radiotherapy and metastatic disease with complete information. m: median, p: patients, r: range, SINS: Spinal Instability Neoplastic Score. [a]Wilcoxon rank sum test for continuous and ordinal variables, Fisher's exact test for nominal variables, log rank test for comparison of survival times. The significance level for these tests has been Bonferroni corrected for family-wise error rate, resulting in an adjusted significance level of 3.33e-3 for an original alpha of 0.05. [b]Following the Gross Tumour Volume (GTV) classification of the International Spine Radiosurgery Consortium[1].

Segmentation	Model	AUROC	BA	F1	MCC
GTV	SVM	0.58 ± 0.01	0.54 ± 0.02	0.33 ± 0.03	0.08 ± 0.04
CTV	RFC	0.62 ± 0.01	0.58 ± 0.02	0.37 ± 0.03	0.15 ± 0.04

Table 3. AUROC, BA, F1 Score and MCC for the best modelling algorithms trained on both radiomics segmentation modes (GTV and CTV).

performance regardless of the modelling strategy. Lastly, the Mann–Whitney U test comparing the AUROC distributions of the best performing models from Table 3 had a *p* value of 4.70-e13, therefore confirming that the AUROC results are statistically different.

Comparison to clinical baseline, semantic and SINS models

The best segmentation mode among the radiomics models (CTV) was then compared to the clinical, the semantic and the SINS models (Table 4 and Fig. 2).

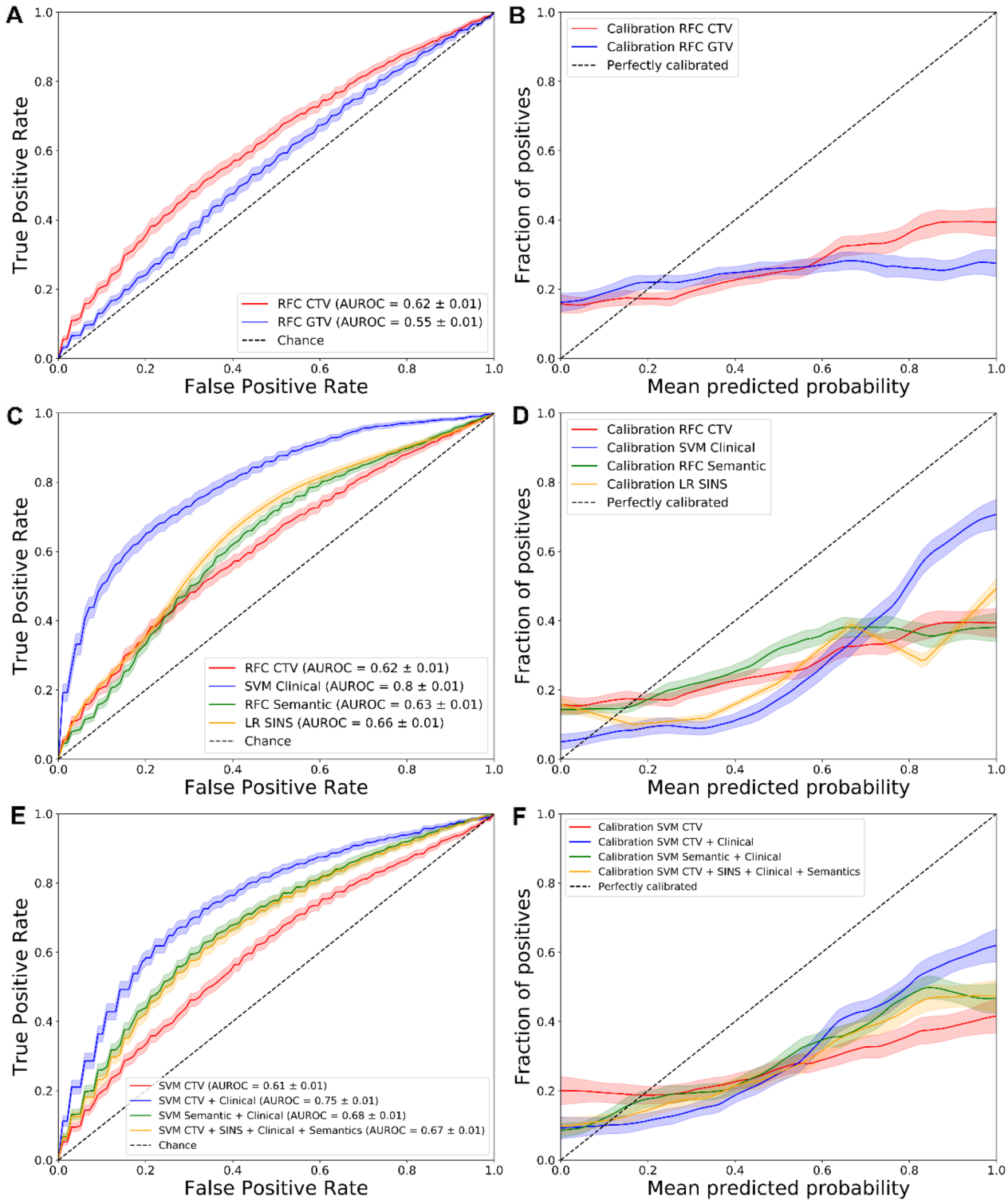


Figure 2. Receiver operator characteristic (ROC) and Calibration curves for the comparisons of different segmentation modes (**A**, **B**), clinical baseline, semantic and SINS features (**C**, **D**), and combined models (**E**, **F**).

The semantic features, on the other hand, achieved almost identical results to the best radiomics segmentation: none performed statistically better. Lastly, a LR trained only on the SINS variable achieved very different results: SINS (binarized) performed very close to random, with a poor classification quality (MCC: 0.04 ± 0.06); on the other hand, the non-binarized SINS model performed similar to the CTV-based radiomics segmentation model but higher AUROC (0.65 ± 0.01).

The clinical ML model outperformed all other models regardless of the modelling algorithm with statistical significance (Table S5 and Figure S2). The best clinical model (SVM) predicted pain response with a BA of

Data	Model	AUROC	BA	F1	MCC
CTV	RFC	0.62 ± 0.01	0.58 ± 0.02	0.37 ± 0.03	0.15 ± 0.04
Semantic	RFC	0.63 ± 0.01	0.58 ± 0.02	0.39 ± 0.03	0.16 ± 0.04
Clinical	SVM	0.80 ± 0.01	0.72 ± 0.03	0.56 ± 0.05	0.43 ± 0.06
SINS	LR	0.65 ± 0.01	0.58 ± 0.03	0.36 ± 0.05	0.16 ± 0.06
SINS (binary)		0.54 ± 0.01	0.52 ± 0.03	0.19 ± 0.05	0.04 ± 0.06

Table 4. AUROC, BA, F1 Score and MCC for the best models, comparing the best radiomics model to the semantic features, clinical baseline and SINS variable.

Data	Model	AUROC	BA	F1	MCC
CTV + SINS	SVM	0.61 ± 0.01	0.57 ± 0.02	0.36 ± 0.04	0.13 ± 0.04
CTV + Clinical		0.75 ± 0.01	0.69 ± 0.02	0.52 ± 0.03	0.35 ± 0.04
Semantic + SINS		0.62 ± 0.01	0.58 ± 0.02	0.39 ± 0.03	0.15 ± 0.04
Semantic + Clinical		0.68 ± 0.01	0.63 ± 0.02	0.45 ± 0.03	0.24 ± 0.04
CTV + SINS + Clinical + Semantic		0.67 ± 0.01	0.62 ± 0.02	0.44 ± 0.03	0.22 ± 0.04

Table 5. AUROC, BA, F1 Score and MCC for SVM models trained on the combination of radiomic, clinical, SINS and semantic features.

0.72 ± 0.03 and an AUROC of 0.80 ± 0.01. Similar to the best performing radiomics model, while the data was moderately imbalanced towards the negative class (no complete pain response), the best performing model overall has shown a better prediction quality when evaluating the true positive class (complete pain response), as it can be seen in the according confusion matrix provided (Figure S4). This is further confirmed by a higher specificity (0.82) than sensitivity (0.63). The AUROC distribution of the SVM model trained on clinical data has been compared, with a Mann–Whitney U test, to that of the other models shown in Table 4 (except to the LR model trained on SINS (binary)). The p-values were 4.21e-57, 3.02e-59 and 2.15e-49 respectively, confirming that the AUROC values of the clinical model are statistically and significantly better.

Given the limited features that the SINS and clinical datasets comprised, their respective prediction models showed a wider standard error on scores where greater variance was expected (F1 and MCC).

Benefits by combining imaging and clinical features

The SVM was evaluated on the possible performance increase by combining the best radiomics model (CTV), clinical, SINS, and semantic features (Table 4, Fig. 2 and Figure S2).

The best performance with combined models was achieved with a SVM trained on CTV and clinical data (AUROC: 0.75 ± 0.01). The addition of non-binarized SINS did not significantly affect the performance of any combined model. An SVM model trained on all data (CTV, non-binarized SINS, clinical and semantic features) outperformed one using only radiomics data; however, it was significantly worse than the best combined model. The Mann–Whitney U test comparing the AUROC distribution of the best semantic model (RFC), and a combined model of semantic and clinical data (SVM), resulted in a p-value of 1.12e-05, therefore confirming that the combined clinical models perform significantly better than semantic features alone.

Interestingly, a model trained only on semantic and clinical features achieved the same performance level as the combined model with all available features (AUROC: 0.68 ± 0.01 and 0.67 ± 0.01, respectively). None of the combined features outperformed the SVM using clinical features.

Feature importance

Feature importance was estimated for SVM and RFC trained on CTV, clinical baseline, semantic, and combined sets of data (Tables S8 to S11). None of the features from the CTV models were selected in all of the 250 cases. On the other hand, the top 15 features for both CTV models (SVM and RFC) were highly homogeneous, sharing the same top three texture features. The most important semantic features were the extent of the GTV along with features also used in the SINS score (e.g., lytic bone lesions and bilateral posterolateral involvement of the spinal element).

The mean decrease in impurity of the RFC nodes, overall, showed low values, with most being below 0.1. However, clinical features achieved a significantly higher feature importance, which is in concordance with the higher performance of those models.

The combined SVM model of CTV, Clinical, SINS and Semantic features showed the same low importance values, with almost no feature selected in 100% of the cases. The feature that was selected most often, while also retaining high importance, was the clinical feature "Tumour Type: Breast Cancer" followed by predominantly semantic and clinical features. Although the majority of all features in the combined model were radiomics (105 of 135), only four of the 12 most predictive features were radiomic, while most of them were semantic.

Discussion

In this exploratory analysis we analysed the potential of ML models to predict pain response to RT of PSBM. CT-based radiomics machine learning models predicted pain response better than random. CTV-based outperformed GTV-based models; semantic and SINS-based models outperformed random, and clinical models performed best, with SVM at the peak. The combination of radiomics features with clinical data significantly increased performance compared to the radiomics baseline. This combination, however, did not match models using only clinical features. The addition of the SINS feature neither affected the radiomics nor the combined model. The feature importance of all radiomics features showed low levels of mean impurity decrease in RFC. Texture features have proven to be the most important predictors, achieving both high percentages of feature selection and high importance scorings. Only clinical features have shown a high importance level, while they were also often consistently selected.

In our modelling approach, we compared two established ML models. Both models achieved competitive results. The best ML model, for radiomics data, was the RFC by a small but statistically significant margin (Tables 3 and S4). However, the SVMs performed better in some situations, mainly for other metrics such as the BA and F1 score. In addition, the SVM models achieved the best results when trained on clinical data, and performed better than the RFCs for the combined data models (Table 5 and S6). The SVMs achieved more consistent results when trained on radiomics features: these models had more competent performances than the RFCs when trained on features with, in principle, less useful information. Given the low importance of these features, these results indicate that the SVM is more resilient to selected features with poor importance. This is further confirmed when analysing the combined models: combined SVM models achieved consistently better performances than RFCs.

We have compared the predictive performance of multiple sets of data: two radiomics segmentation modes, clinical, semantic and SINS features. The only model that did not achieve better than random results was LR trained only on SINS (binarized; Table 4). This is to be expected: by binarizing the SINS variable, important information, that can be learnt by either model, is lost. Combined models that used clinical data had an expected performance increase compared to their respective baselines (Table 5 and S6). However, these combined models performed worse than a clinical only model: this indicates that the addition of features that are not important to the model can have a negative impact on its performance, by making it difficult for the model to identify patterns in the data. This is further confirmed by the decrease in feature importance of the clinical features when comparing them alone and in a combined model (Tables S9 and S11, respectively).

All radiomics features have shown low feature importance, which can be explained by a possible low correlation to the prediction target. This is also consistent with the fact that none were selected in any of the 250 cases. In addition, only 10 of all 105 features were selected by MRMR at least 50% of the time. This high variance when selecting features is potentially due to their low correlation towards the *complete pain response* outcome variable. On the other hand, clinical features have shown more than thrice higher feature importance towards the outcome variable, and were selected in nearly all cases when used in combined modelling (Tables S9 and S11).

Multiple previous publications have analysed factors related to pain response following RT of bone metastasis. An early retrospective study by Arcangeli et al. demonstrated that pain response depended on patients' performance status and specific histology. NSCLC patients were shown to have a worse response to RT than patients with other cancer origins[40]. This was reproduced by Nyguen et al. demonstrating a favourable response for patients with prostate and mammary carcinoma[41]. Location and pain level before therapy appeared not to influence radiation response[32,50]. These results were validated in a large prospective trial with 956 patients by Westhoff et al. Next to the aforementioned clinical factors, the use of opioids and absence of visceral metastases were positively predictive for RT response[33]. However, the multivariate model achieved only limited predictive capacity with a C-statistic of 0.56.

Van Velden et al. conducted a further prospective trial comparing the predictive performance of the SINS with clinical parameters[34]. SINS appeared to be significantly associated with complete response after adjustment for gender, tumour type and performance status. Adding SINS to the clinical parameters increased the AUROC for the prediction of complete response from 0.68 to 0.78. In our study, SINS as training data proved to perform better than random (Table 4). However, adding SINS to other datasets did not increase their performance significantly (Table 5 and S6). Combining clinical and SINS data, the overall performance was significantly better than the radiomics models (SVM and RFC AUROCS: 0.73 ± 0.01 and 0.75 ± 0.01, respectively), albeit inferior to the clinical models (Table S7). The performance difference of a combined model of clinical and SINS features between Van Velden et al. study and this exploratory analysis can be attributed to a number of reasons. Firstly, in this study, pain response was assessed retrospectively, which can potentially explain the different performance of the models towards the outcome variable. Secondly, the different proportion of metastases localization, and the presence of cervical cases, may affect the SINS, given the higher instability that some locations may entail. Thirdly, the current study employed SVMs and RFCs as models trained on clinical and SINS features, which are distinct from the multivariate logistic regression used in the previous study. Fourthly, in Van Velden et al. study it is not directly explained what resampling technique the authors have used. A difference in the resampling technique can potentially impact the prediction performances due to larger training sizes, therefore leading to over-optimistic results in some cases. Lastly, the clinical features used in both studies are significantly different, leading to different model performances (AUROC values of 0.68 and 0.80 in the previous study and ours, respectively). Therefore, the room for improvement for the SINS variable in a combined model with clinical features be substantially different.

In our study, we compared two potential modes of segmentation. Although the predictive performance was overall similar, the CTV-based segmentations were superior for both ML models. In contrast to the GTV, the CTV segmentation included vertebra compartments that are at risk of microscopic infiltration[43]. This additional

information may have improved the predictive power. Texture features were the most important radiomics features. Such features may capture texture and intensity heterogeneity that may be associated with cell density within the bone marrow. Analysis of magnetic resonance imaging data may be more suitable to quantify such changes. Recently, one other publication has analysed the potential of radiomics-based prediction of pain response[35]. The authors trained a random forest model on a single centre cohort of 69 patients using leave-one-out cross-validation. While their clinical model showed an inferior performance with an AUC of 0.70, the radiomics model was able to predict pain response with a superior AUC of 0.82. There are several reasons that may explain these differences in performance. First, the authors applied only the simplistic double-layer split into train and test set (through their leave-one-out cross-validation), instead of the more adequate triple-layer split into train, validation and test set. Consequently, the authors optimized their model on the same patients used for assessing performance, thereby opening the door to data leakage. Such leakage often leads to substantial over-estimates of performance. In contrast, our nested cross-validation results included repeated testing independent of hyperparameter optimization guaranteeing more unbiased results. Second, the authors used a different set of VOIs. Instead of a GTV or specific CTV, the authors used the spinal canal, the complete vertebra and the vertebra plus a one-centimetre margin as VOIs. So far, it remains unclear what segmentation strategy may be optimal. Third, the authors trained their model for "any pain response" instead of "complete response" which may also explain a difference in performance by having a broader prediction target. Taken together with our results, both studies could demonstrate prediction of pain response better than random albeit with different predictive power against the background of a significantly different study design.

Besides quantitative radiomics image analysis, semantic features extraction constitutes an alternative "manual" way to extract information from medical images[51]. For prediction of pain response of PSBM Mitera et al. evaluated semantic imaging features in 33 patients[47]. The authors did not find any association of semantic imaging features to pain response. Semantic features included pathological fractures, kyphosis and anatomic extent of tumour. However, the study was limited by the use of a large number of semantic features and a relatively small number of patients. For instance, the known predictive factor age did not correlate with response either. Our study has shown that, with a larger training set, it is possible to achieve better than random prediction results when training either ML algorithm with semantic data, with RFC performing best (AUROC: 0.63 ± 0.01; Table 4). It is important to note that the SINS score in itself is a score combining multiple semantic features. We used these features complemented with other additional variables. The SINS score, however, performed better than the semantic model, demonstrating that the important features are already included in the SINS score.

There are several limitations to our study. First, pain response was assessed retrospectively. Due to non-standardized or incomplete reporting of pain response determination, it may have been error prone. To allow a standardised assessment we followed the recommendation of the International Spine Radiosurgery Consortium Consensus Guidelines[43]. Patients with "indeterminate response" were excluded from analysis. This may have conferred a selection bias as missing information may be associated with confounding factors such as low KPS or early death. Secondly, in patients with multiple PBMS each metastasis was treated as a separate sample. The outcome, however, was equal between all metastases of a specific patient. Information on which specific metastases contributed to symptomatic pain remained elusive. To prevent data leakage and bias, stratified cross-validation was performed, guaranteeing that multiple samples from the same patient were evenly distributed across all splits. Thirdly, our study was of monocentric nature with a lack of an external validation set. To compensate for this, we applied nested cross-validation and repeated the process 50 times to increase the statistical strength of the results. We believe that our exploratory analysis allows the assessment of the general possibility of RT response prediction and a comparison to established factors.

Conclusions

To conclude, in this exploratory work we were able to demonstrate a predictive value of established clinical factors using machine learning for the prediction of complete pain response to palliative radiotherapy in patients with painful spinal bone metastases. CT-based radiomics and semantic machine learning models performed better than random but sub-optimally. The SINS score performed slightly better than both, and models trained on a combination of the available datasets performed even better. Using exclusively clinical features as input, however, outperformed all other models. Upon inspection of the radiomics and clinical features, their importance and selection frequency confirmed the higher predictive quality of the latter, with a more than three-fold decrease in mean impurity. Thus, CT-based radiomics features did not present supplementary value beyond models trained solely on clinical features.

Data and code availability

All data and code used in this research is available upon contact of the correspondence author (Jan C. Peeken, jan.peeken@tum.de) and in concordance to the ethics committee.

Received: 6 February 2023; Accepted: 28 September 2023
Published online: 13 October 2023

References

1. Simes, R. J. Treatment selection for cancer patients: Application of statistical decision theory to the treatment of advanced ovarian cancer. *J. Chronic. Dis.* **38**, 171–186 (1985).
2. Maclin, P. S., Dempsey, J., Brooks, J. & Rand, J. Using neural networks to diagnose cancer. *J. Med. Syst.* **15**, 11–19 (1991).
3. Cicchetti, D. V. Neural networks and diagnosis in the clinical laboratory: State of the art. *Clin. Chem.* **38**, 9–10 (1992).
4. Mitchell, T. M. *Machine Learning* 1st edn. (McGraw-Hill Inc, 1997).

5. Gupta, S. *et al.* Machine-learning prediction of cancer survival: A retrospective study using electronic administrative records and a cancer registry. *BMJ Open* **4**, e004007. https://doi.org/10.1136/bmjopen-2013-004007 (2014).
6. Peeken, J. C. *et al.* Treatment-related features improve machine learning prediction of prognosis in soft tissue sarcoma patients. *Strahlenther Onkol.* **194**, 824–834. https://doi.org/10.1007/s00066-018-1294-2 (2018).
7. Peeken, J. C. *et al.* Combining multimodal imaging and treatment features improves machine learning-based prognostic assessment in patients with glioblastoma multiforme. *Cancer Med.* **8**, 128–136. https://doi.org/10.1002/cam4.1908 (2019).
8. Peeken, J. C. *et al.* Tumor grading of soft tissue sarcomas using MRI-based radiomics. *eBioMedicine* **48**, 332–340. https://doi.org/10.1016/j.ebiom.2019.08.059 (2019).
9. Lambin, P. *et al.* Radiomics: The bridge between medical imaging and personalized medicine. *Nat. Rev. Clin. Oncol.* **14**, 749–762. https://doi.org/10.1038/nrclinonc.2017.141 (2017).
10. Kocher, M., Ruge, M. I., Galldiks, N. & Lohmann, P. Applications of radiomics and machine learning for radiotherapy of malignant brain tumors. *Strahlenther Onkol.* **196**, 856–867. https://doi.org/10.1007/s00066-020-01626-8 (2020).
11. Zhou, M. *et al.* Radiomics in brain tumor: Image assessment, quantitative feature descriptors, and machine-learning approaches. *Am. J. Neuroradiol.* **39**, 208–216. https://doi.org/10.3174/ajnr.A5391 (2018).
12. Wagner, M. W. *et al.* Radiomics, machine learning, and artificial intelligence—what the neuroradiologist needs to know. *Neuroradiology* **63**, 1957–1967. https://doi.org/10.1007/s00234-021-02813-9 (2021).
13. Peng, Z. *et al.* Application of radiomics and machine learning in head and neck cancers. *Int. J. Biol. Sci.* **17**, 475–486. https://doi.org/10.7150/ijbs.55716 (2021).
14. Peeken, J. C. *et al.* Radiomics in radiooncology - challenging the medical physicist. *Physica Medica: Eur. J. Med. Phys.* **48**, 27–36. https://doi.org/10.1016/j.ejmp.2018.03.012 (2018).
15. Lambin, P. *et al.* Radiomics: Extracting more information from medical images using advanced feature analysis. *Eur. J. Cancer* **48**, 441–446. https://doi.org/10.1016/j.ejca.2011.11.036 (2012).
16. Peeken, J. C., Wiestler, B. & Combs, S. E. Image-guided radiooncology: The potential of radiomics in clinical application. *Recent Results Cancer Res.* **216**, 773–794. https://doi.org/10.1007/978-3-030-42618-7_24 (2020).
17. Lang, D. M., Peeken, J. C., Combs, S. E., Wilkens, J. J. & Bartzsch, S. Deep learning based HPV status prediction for oropharyngeal cancer patients. *Cancers* **13**, 786. https://doi.org/10.3390/cancers13040786 (2021).
18. Navarro, F. *et al.* Development and external validation of deep-learning-based tumor grading models in soft-tissue sarcoma patients using MR imaging. *Cancers* **13**, 2866. https://doi.org/10.3390/cancers13122866 (2021).
19. Leger, S. *et al.* Comprehensive analysis of tumour sub-volumes for radiomic risk modelling in locally advanced HNSCC. *Cancers* **12**, 3047. https://doi.org/10.3390/cancers12103047 (2020).
20. Starke, S. *et al.* 2D and 3D convolutional neural networks for outcome modelling of locally advanced head and neck squamous cell carcinoma. *Sci. Rep.* **10**, 15625. https://doi.org/10.1038/s41598-020-70542-9 (2020).
21. Marr, L. *et al.* Predictive value of clinical and 18F-FDG-PET/CT derived imaging parameters in patients undergoing neoadjuvant chemoradiation for esophageal squamous cell carcinoma. *Sci. Rep.* **12**, 7148. https://doi.org/10.1038/s41598-022-11076-0 (2022).
22. Spohn, S. K. B. *et al.* The maximum standardized uptake value in patients with recurrent or persistent prostate cancer after radical prostatectomy and PSMA-PET-guided salvage radiotherapy—a multicenter retrospective analysis. *Eur. J. Nucl. Med. Mol. Imaging* https://doi.org/10.1007/s00259-022-05931-5 (2022).
23. Shahzadi, I. *et al.* Analysis of MRI and CT-based radiomics features for personalized treatment in locally advanced rectal cancer and external validation of published radiomics models. *Sci. Rep.* **12**, 10192. https://doi.org/10.1038/s41598-022-13967-8 (2022).
24. Brancato, V., Cerrone, M., Lavitrano, M., Salvatore, M. & Cavaliere, C. A systematic review of the current status and quality of radiomics for glioma differential diagnosis. *Cancers (Basel)* **14**, 2731. https://doi.org/10.3390/cancers14112731 (2022).
25. Giraud, P. *et al.* Radiomics and machine learning for radiotherapy in head and neck cancers. *Front. Oncol.* **9**, 174 (2019).
26. El Ayachy, R. *et al.* The role of radiomics in lung cancer: From screening to treatment and follow-up. *Front. Oncol.* **11**, 603595. https://doi.org/10.3389/fonc.2021.603595 (2021).
27. Kumar, A. *et al.* Machine-learning-based radiomics for classifying glioma grade from magnetic resonance images of the brain. *J. Personaliz. Med.* **13**, 920. https://doi.org/10.3390/jpm13060920 (2023).
28. Bo, L. *et al.* Differentiation of brain abscess from cystic glioma using conventional MRI based on deep transfer learning features and hand-crafted radiomics features. *Front. Med.* **8**, 748144 (2021).
29. van Dijk, L. V. *et al.* Parotid gland fat related magnetic resonance image biomarkers improve prediction of late radiation-induced xerostomia. *Radiother. Oncol.* **128**(3), 459–466. https://doi.org/10.1016/j.radonc.2018.06.012 (2018).
30. Krafft, S. P. *et al.* The utility of quantitative CT radiomics features for improved prediction of radiation pneumonitis. *Med. Phys.* **45**, 5317–5324. https://doi.org/10.1002/mp.13150 (2018).
31. Rossi, L. *et al.* Texture analysis of 3D dose distributions for predictive modelling of toxicity rates in radiotherapy. *Radiother. Oncol.* **129**, 548–553. https://doi.org/10.1016/j.radonc.2018.07.027 (2018).
32. Zeng, L. *et al.* Comparison of pain response and functional interference outcomes between spinal and non-spinal bone metastases treated with palliative radiotherapy. *Support Care Cancer* **20**, 633–639. https://doi.org/10.1007/s00520-011-1144-6 (2012).
33. Westhoff, P. G. *et al.* Quality of life in relation to pain response to radiation therapy for painful bone metastases. *Int. J. Radiat. Oncol. Biol. Phys.* **93**(3), 694–701. https://doi.org/10.1016/j.ijrobp.2015.06.024 (2015).
34. van der Velden, J. M. *et al.* Prospective evaluation of the relationship between mechanical stability and response to palliative radiotherapy for symptomatic spinal metastases. *Oncologist* **22**, 972–978. https://doi.org/10.1634/theoncologist.2016-0356 (2017).
35. Wakabayashi, K. *et al.* A predictive model for pain response following radiotherapy for treatment of spinal metastases. *Sci. Rep.* **11**, 12908. https://doi.org/10.1038/s41598-021-92363-0 (2021).
36. Sierko, E., Hempel, D., Zuzda, K. & Wojtukiewicz, M. Z. Personalized radiation therapy in cancer pain management. *Cancers* **11**, 390. https://doi.org/10.3390/cancers11030390 (2019).
37. Akezaki, Y. *et al.* Factors affecting the quality of life of patients with painful spinal bone metastases. *Healthcare* **9**, 1499. https://doi.org/10.3390/healthcare9111499 (2021).
38. Litak, J. *et al.* Biological and clinical aspects of metastatic spinal tumors. *Cancers* **14**, 4599. https://doi.org/10.3390/cancers14194599 (2022).
39. Fisher, C. G. *et al.* A novel classification system for spinal instability in neoplastic disease: An evidence-based approach and expert consensus from the spine oncology study group. *Spine* **35**, 1221–9. https://doi.org/10.1097/BRS.0b013e3181e16ae2 (2010).
40. Arcangeli, G. *et al.* Radiation therapy in the management of symptomatic bone metastases: The effect of total dose and histology on pain relief and response duration. *Int. J. Radiat. Oncol. Biol. Phys.* **42**, 1119–1126. https://doi.org/10.1016/s0360-3016(98)00264-8 (1998).
41. Nguyen, J. *et al.* Palliative response and functional interference outcomes using the brief pain inventory for spinal bony metastases treated with conventional radiotherapy. *Clin. Oncol.* **23**, 485–491. https://doi.org/10.1016/j.clon.2011.01.507 (2011).
42. Chow, E. *et al.* Update of the international consensus on palliative radiotherapy endpoints for future clinical trials in bone metastases. *Int. J. Radiat. Oncol. Biol. Phys.* **82**, 1730–1737. https://doi.org/10.1016/j.ijrobp.2011.02.008 (2012).
43. Cox, B. W. *et al.* International spine radiosurgery consortium consensus guidelines for target volume definition in spinal stereotactic radiosurgery. *Int. J. Radiat. Oncol. Biol. Phys.* **83**, e597–e605. https://doi.org/10.1016/j.ijrobp.2012.03.009 (2012).
44. van Griethuysen, J. J. M. *et al.* Computational radiomics system to decode the radiographic phenotype. *Cancer Res.* **77**, e104–e107. https://doi.org/10.1158/0008-5472.CAN-17-0339 (2017).

45. Intratumor Heterogeneity Characterized by Textural Features on Baseline 18F-FDG PET Images Predicts Response to Concomitant Radiochemotherapy in Esophageal Cancer | Journal of Nuclear Medicine Available online: https://jnm.snmjournals.org/content/52/3/369 (accessed on 23 August 2023).
46. Zwanenburg, A. *et al.* The image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping. *Radiology* **295**, 328–338. https://doi.org/10.1148/radiol.2020191145 (2020).
47. Mitera, G. *et al.* Correlation of computed tomography imaging features with pain response in patients with spine metastases after radiation therapy. *Int. J. Radiat. Oncol. Biol. Phys.* **81**, 827–830. https://doi.org/10.1016/j.ijrobp.2010.06.036 (2011).
48. Ding, C.; Peng, H. Minimum Redundancy Feature Selection From Microarray Gene Expression Data.; September 11 2003; Vol. 3, pp. 523–528.
49. Pedregosa, F. *et al.* Scikit-Learn: Machine learning in python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
50. Llorián-Salvador, O.; Akhgar, J.; Pigorsch, S.; Borm, K.; Münch, S.; Bernhardt, D.; Rost, B.; Andrade-Navarro, M.; Combs, S.; Peeken, J. Machine Learning Based Prediction of Pain Response to Palliative Radiation Therapy - Is There a Role for Planning CT-Based Radiomics and Semantic Imaging Features? 2022.
51. Peeken, J. C. *et al.* Semantic imaging features predict disease progression and survival in glioblastoma multiforme patients. *Strahlenther Onkol.* **194**, 580–590. https://doi.org/10.1007/s00066-018-1276-4 (2018).

Author contributions

Conceptualization, S.P. and J.P.; Data curation, J.A. and J.P.; Formal analysis, O.L., J.A. and J.P.; Funding acquisition, S.C. and J.P.; Investigation, O.L. and J.P.; Methodology, O.L. and J.P.; Project administration, J.P.; Resources, D.B., B.R. and S.C.; Software, O.L.; Supervision, B.R., M.A., S.C. and J.P.; Validation, O.L. and J.P.; Visualization, O.L. and J.A.; Writing – original draft, O.L. and J.P.; Writing – review & editing, J.A., S.P., K.B., S.M., D.B., B.R., M.A. and S.C.. All authors have read and agreed to the published version of the manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL. This work was funded by physician scientist programs of the medical faculty of the Technical University of Munich and the Helmholtz Zentrum Muenchen. Funding was also received from Else-Kröner-Fresenius-Stiftung.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-023-43768-6.

Correspondence and requests for materials should be addressed to J.C.P.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

4.3 CT-based radiomics for predicting breast cancer radiotherapy side effects

scientific reports

Check for updates

OPEN

CT-based radiomics for predicting breast cancer radiotherapy side effects

Óscar Llorián-Salvador[1,2,3]✉, Nora Windeler[1], Nicole Martin[2], Lucas Etzel[1,4], Miguel A. Andrade-Navarro[3], Denise Bernhardt[1,4,5], Burkhard Rost[2], Kai J. Borm[1], Stephanie E. Combs[1,4,5], Marciana N. Duma[1,6,7] & Jan C. Peeken[1,4,5]

Skin inflammation with the potential sequel of moist epitheliolysis and edema constitute the most frequent breast radiotherapy (RT) acute side effects. The aim of this study was to compare the predictive value of tissue-derived radiomics features to the total breast volume (TBV) for the moist cells epitheliolysis as a surrogate for skin inflammation, and edema. Radiomics features were extracted from computed tomography (CT) scans of 252 breast cancer patients from two volumes of interest: TBV and glandular tissue (GT). Machine learning classifiers were trained on radiomics and clinical features, which were evaluated for both side effects. The best radiomics model was a least absolute shrinkage and selection operator (LASSO) classifier, using TBV features, predicting moist cells epitheliolysis, achieving an area under the receiver operating characteristic (AUROC) of 0.74. This was comparable to TBV breast volume (AUROC of 0.75). Combined models of radiomics and clinical features did not improve performance. Exclusion of volume-correlated features slightly reduced the predictive performance (AUROC 0.71). We could demonstrate the general propensity of planning CT-based radiomics models to predict breast RT-dependent side effects. Mammary tissue was more predictive than glandular tissue. The radiomics features performance was influenced by their high correlation to TBV volume.

Keywords Radiomics, Machine learning, Breast cancer, Computed tomography, Radiotherapy, Side effects, Skin inflammation, Moist cells epitheliolysis, Edema

Breast cancer is the leading form of invasive cancer in women, accounting for the most significant proportion of cancer cases worldwide[1,2]. Approximately 14% of women are affected by breast cancer, making it a prevalent health concern[3]. Radiation therapy (RT) constitutes the standard of care after breast-conserving surgeries for most patients[4].

Radiomics, a field dedicated to extracting quantitative features from medical imaging such as computer tomography (CT) paired with machine learning (ML), shows great potential in cancer research[5]. Radiomics provides a powerful foundation for the integration of ML techniques in cancer research. By extracting quantitative features from medical images, radiomics enables the generation of high-dimensional data, which can then be utilized by computational models to create predictive models for clinical or biological endpoints[6–8].

In the context of mammary carcinoma, radiomics has been widely applied to predict survival, disease progression, treatment response, molecular aberrations, and the detection of metastases or areas of infiltrative tumor[9–16]. Nevertheless, the application of radiomics analysis for accurately predicting non-tumor response to RT remains limited. Earlier research has explored the possibility of predicting RT-related side effects, including xerostomia and pneumonitis, or pain response to palliative RT[17–19].

[1]Department of Radiation Oncology, Klinikum Rechts der Isar, Technische Universität München, Ismaninger Straße 22, 81675 Munich, Germany. [2]Department of Informatics, Bioinformatics and Computational Biology—i12, Technische Universität München, Boltzmannstr. 3, 85748 Munich, Germany. [3]Institute of Organismic and Molecular Evolution, Johannes Gutenberg University Mainz, Hanns-Dieter-Hüsch- Weg 15, 55128 Mainz, Germany. [4]Deutsches Konsortium für Translationale Krebsforschung (DKTK), Partner Site Munich, 69120 Heidelberg, Germany. [5]Department of Radiation Sciences (DRS), Institute of Radiation Medicine (IRM), Helmholtz Zentrum, 85764 München, Germany. [6]Department of Radiation Oncology, Helios Clinics of Schwerin - University Campus of MSH Medical School Hamburg, Schwerin, Germany. [7]Department for Human Medicine, MSH Medical School Hamburg, Hamburg, Germany. ✉email: oscar.llorian-salvador@tum.de

Several similar studies have investigated the use of ML and various types of imaging data to predict RT side effects in breast cancer patients. Research utilizing dosiomics features extracted from CT images managed to accurately predict acute skin toxicity[20]. Another study using electron density and biologically effective dose radiomics effectively predicted late radiation-induced subcutaneous fibrosis[21]. Additionally, a comprehensive review of ML models analyzed RT-induced complications across multiple cancer sites, including breast cancer[22]. Collectively, these studies emphasize the growing interest in using ML and imaging data to mitigate RT side effects.

The objective of this study was to develop a statistically reliable assessment of the predictive capability of radiomics features to predict the most prevalent RT side effects of moist epitheliolysis as a surrogate for skin inflammation and edema based on the total breast volume (TBV) and glandular tissue (GT).

Materials and methods

Clinical data collection and curation

The dataset consisted of 252 breast cancer patients who underwent radiotherapy between 2012 and 2016 in the Rechts der Isar university hospital of the technical university of Munich (TUM). For the patient data acquired at TUM, retrospective analysis of patient records and data is generally allowed following Article 27 of the Bavarian Hospital act (Bayerisches Krankenhausgesetz) from the Landeskrankenhausgesetz des Freistaates Bayern. Informed consent for treatment was obtained from every patient. Institutional Review Board (IRB) was acquired from the review board of TUM (reference number 466/16 s. Clinical variables were defined based on a literature review on known clinical predictors from previous publications. Moreover, variables were selected based on broad availability of data that hindered the assessment of other predictive factors[23,24]: smoker status, chemotherapy received, radiotherapy boost, the maximum prescribed radiation dose in equivalent dose at 2 Gy (EQD2, $\propto /\beta = 3$), TBV, and the two targets of prediction: (i) moist cell epitheliolysis as surrogate for common terminology criteria for adverse effects (CTCAE) grade 2 skin inflammation[25] (33 positive cases; referred henceforth simply as moist epitheliolysis); and (ii) presence of any edema (26 positive cases).

Radiomics data collection and curation

Prior to RT treatment, planning CT images of the breast were conducted. Figure S1 shows que acquisition parameters for these CT images. Exclusion criteria encompassed breast implant and mastectomy cases. Two separate volume of interest (VOI) definitions were segmented, creating two radiomics cohorts: TBV, containing radiomics information from the whole breast tissue; and glandular tissue (GT), which contained radiomics information only from this tissue. Patient outcome assessment was performed retrospectively by a medical student after thorough teaching by a radiation oncologist (JCP). All methodology has been conducted in accordance to the relevant guidelines and regulations.

Segmentation of the volumes of interest was manually performed by NW, using 3D Slicer[26]. GT was defined using the fast growcut function. BSpline interpolation was used to perform isotropic resampling to obtain a voxel size of $1 \times 1 \times 1$ mm. Image discretization was carried out with a fixed bin width of 10. Laplacian of Gaussian filtering was used for image reconstruction (Sigma values of 1.0, 2.0, 3.0, 4.0 and 5.0).

Radiomics features were extracted and filtered from the CT images and both segmentations using the Python library PyRadiomics[27] (version 3.0.1; Python version 3.8.10). A total of 104 features were obtained for each of the radiomics cohorts, which included first-order, shape, and texture features (the latter is composed of "gray-level co-occurrence matrix", "gray-level size-zone matrix", "gray-level run-length matrix", "neighboring gray-tone difference matrix", and "gray-level dependence matrix" features). Figure 1 shows a diagram of the clinical and radiomics features and side effects collection process from the patients. Further, Fig. S2 shows the distribution of patients across all clinical features and side effects measured.

Feature pre-processing and hyperparameter optimization

Repeated nested cross-validation was employed to train and validate the models. Normalization of the radiomics features was performed using min-max normalization, in order to conserve the original distribution in the [0, 1] range.

For each cohort, the most interesting features were selected and evaluated in two different ways: the first one, with a double Spearman rank correlation test, first within each dataset with a cut-off value of 0.9 to remove redundant features; and then towards each side effect prediction target, in order to keep the most relevant features. The second option was selecting features using minimum redundancy-maximum relevance (MRMR; version 1.0.2), which incorporates both tests in a single step[28]. In both cases, an estimation of the information density and, therefore, of the number of features to select, was made using Principal Component Analysis (PCA). For the TBV radiomics feature set, an average of 23 and 39 features were selected when using MRMR and a double Spearman rank correlation test, respectively. For the GT radiomics feature set, on the other hand, an average of 26 and 44 features were selected when using each of the feature selection techniques, respectively.

Before finding the optimal hyperparameter values, the class imbalance of the different side effect prediction targets was corrected depending on the level of disproportion. Moist epitheliolysis and edema had a ratio of 6.64:1 and 8.69:1 of negative to positive class sizes, respectively, and were therefore corrected using a combination of synthetic minority over-sampling technique (SMOTE; *imbalance-learn* library version 0.11.0)[29] to a ratio of 2:1, and random under-sampling of the majority class to a ratio of 1.25:1. The choice of ratios for each step was made to find a balance between avoiding excessive oversampling and losing too many samples while undersampling. Balanced accuracy (BA) was the metric used as optimization criteria for the values of the hyperparameters, capable of handling the small remainder of class imbalances. Hyperparameter optimization was conducted using an exhaustive grid search, where all combinations of hyperparameter values are tested in the validation set of the innermost fold until the optimal values are found.

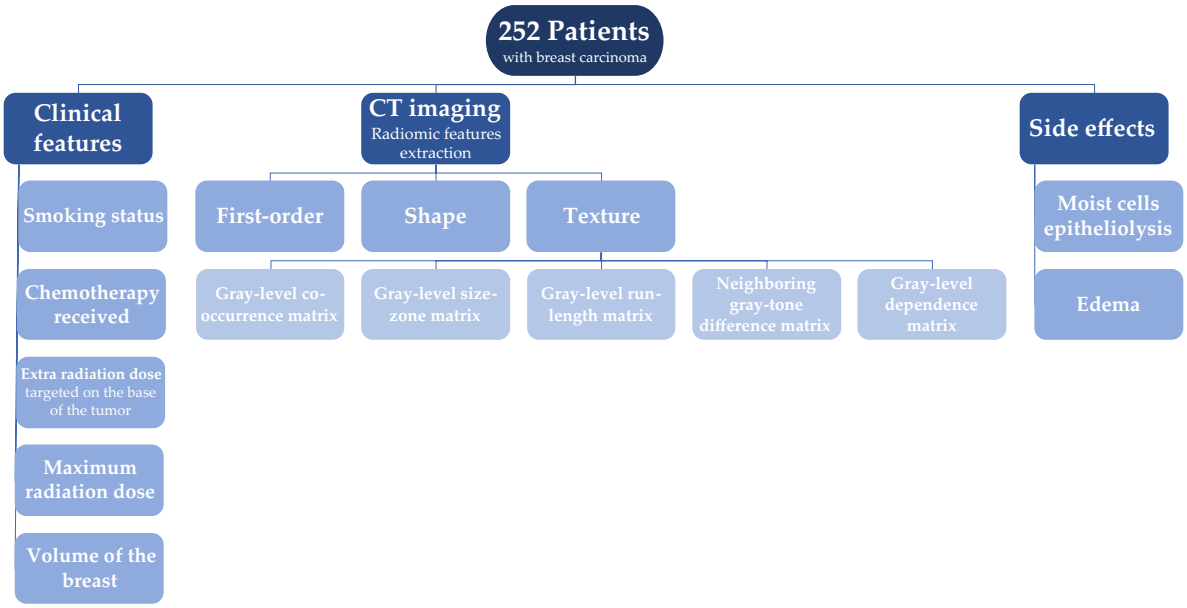


Fig. 1. Patient and data flowchart. In the left and central branches, the clinical and radiomics features can be found, respectively. The right branch shows the three RT side effects used as prediction targets.

Machine learning modeling

Four ML algorithms were implemented and evaluated: logistic regression (LR), used for its simplicity and efficiency in binary classification tasks with a low feature set dimensionality[30,31]; least absolute shrinkage and selection operator (LASSO), a variant with an optimizable regularization term that can potentially better handle imbalanced datasets[32]; support vector machine (SVM), a high flexibility algorithm thanks to the implementation of multiple kernels and explore non-linear relationships in the data[33]; and random forest classifier (RF), an ensemble learning, decision tree-based method that is more robust to overfitting effects[34]. All models were imported from the python library scikit-learn (version 1.0.2)[35]. These models were contrasted against clinical model baselines.

After comparing the four model types for each of the radiomics cohorts and feature selection types, the best models were retrained and optimized adding clinical data in order to assess whether a combined model yields a better performance in predicting the presence of any side effect. The workflow followed by the ML pipeline is shown in Fig. 2. In addition, larger reference images of the respective VOIs can be seen in Fig. S1.

Feature selection has been analyzed for all relevant models, estimating a score based on the feature importance assigned by the models and how often each feature was selected. The resulting score is calculated as $Score = Feature\ Importance/[(n+1) - m]$, where n is the number of models, and m is the number of times the feature has been chosen.

Finally, the correlation between the breast volume and the prediction probability of the best model has been analyzed to study the overall impact of the breast volume in the predictive value of radiomics features. An additional model was evaluated where radiomics features that highly correlated to the breast volume were excluded (Spearman correlation higher than 0.8), using the best performing configuration. The objective was to assess the impact of volume-correlated features on the performance of radiomics models.

Statistical analysis

Training and validation of the different models were performed using 50 repetitions of nested cross-validation (5 outer folds, 4 inner folds). This resampling technique provides additional statistical robustness, resulting in 250 final models that were aggregated to the final test results.

In order to gather more information from the radiomics features, PCA was employed as an estimation of the information density within this dataset. The variance retention by the components of PCA was used to understand the intrinsic dimensionality of our dataset. However, since the components generated by PCA are a different combination from the original features and, generally, more packed, these components should not be used as a feature selection replacement, but as an estimation. The reason behind it is the inherent added difficulty of tracing the feature importance back to the original features.

In the inner fold of the nested cross-validation normalization, feature selection and class imbalance correction were applied, in order to avoid data leakage from any training split to the validation (inner fold) or test splits (outer fold).

One of the two feature selection techniques mentioned in this study is the use of a double Spearman rank correlation test. This approach is intended to optimize feature selection by addressing redundancy and relevance in two distinct steps. First, redundancy is removed so that features that do not provide additional information are eliminated. Second, the Spearman rank correlation test is applied again comparing the dataset and the predictor, selecting instead the features that are most relevant to the prediction target.

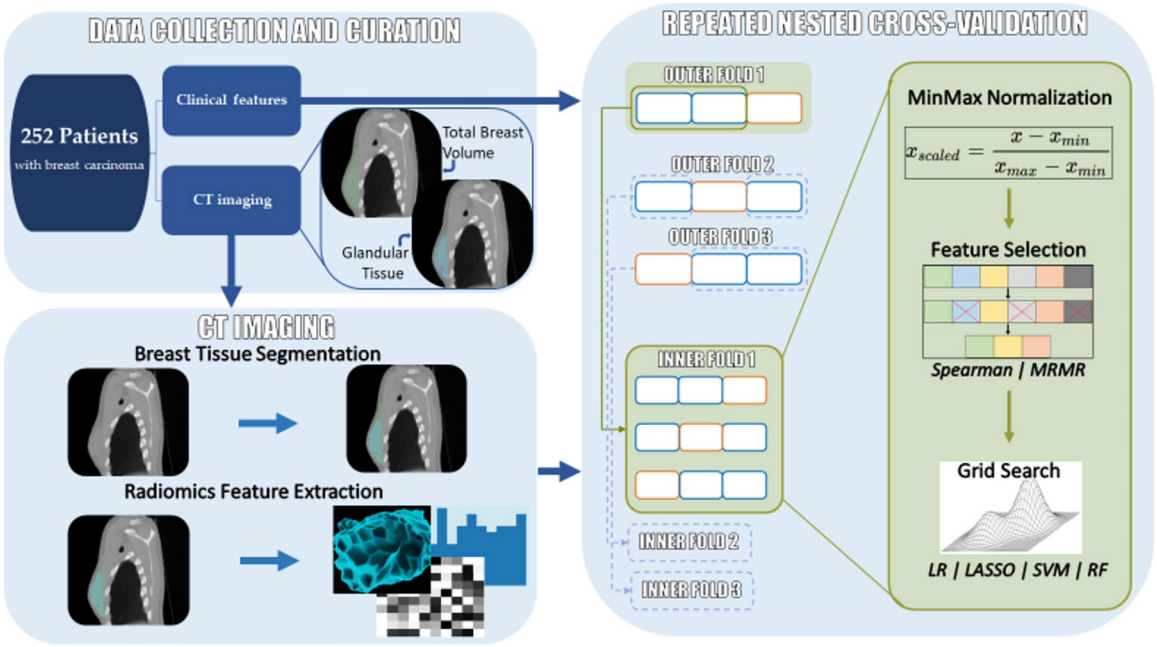


Fig. 2. Workflow of the pipeline used in the study to analyze both clinical and radiomics data. On the left half of the workflow: clinical features were obtained from all patients with CT imaging available, the respective VOIs (TBV and GT) were segmented, and the subsequent radiomics features extracted. On the right half of the workflow: for each evaluated dataset, a 50-repeat nested cross-validation was performed. Within the inner fold normalization, feature selection and an exhaustive grid search for optimal hyperparameters was performed.

The performance of the aggregated models was measured using a combination of metrics: BA, F1, precision, recall, specificity, area under the receiver-operator curve (AUROC) and Matthew's correlation coefficient (MCC). Metrics are given with 1.96 standard errors for a confidence interval of 95%. ROC curves were also used to evaluate the trade-off between the sensitivity and specificity across different decision thresholds, and to assess the discrimination power between classes of each of the models.

Results

We evaluated the possibility of predicting side effects of RT in breast cancer (moist cells epitheliolysis as a surrogate for skin inflammation and edema) based on the total breast volume (TBV), glandular tissue (GT) and using clinical features. Table 1 summarizes the results that are shown throughout this section. The feature importance was calculated for the best performing radiomics and clinical models (Table 2 and Table S8, respectively).

Side effect prediction

The ROC performance of the best trained models to predict both side effects can be seen in Fig. 3. More scores regarding the comparison of side effects as the prediction target can be seen in Table S2. In addition, the calibration curve of the best performing radiomics model is shown in Fig. S5.

While the edema models performed only slightly above random (best AUROC value of 0.55), both radiomics feature sets have shown a notable predictive value towards moist cells epitheliolysis using the LASSO classifier: an AUROC of 0.74 when using TBV, and an AUROC of 0.65 when training a RF on the GT radiomics feature set, whose features were selected using MRMR. Therefore, models trained to predict moist cells epitheliolysis perform better than predicting edema regardless of the feature selection technique, ML algorithm, or the training radiomics feature set used.

The ROC performance of both radiomics cohorts, with the clinical features as baseline, are shown in Fig. 4. The clinical model achieved an AUROC of 0.7. More scores regarding the comparison of the predictive power of each radiomics feature set and the clinical baseline can be seen in Table S3. Only the best performing ML

Side effect	TBV	GT	Clinical
Moist epitheliolysis	0.74 ± 0.01	0.65 ± 0.01	0.70 ± 0.01
Edema	0.53 ± 0.02	0.55 ± 0.01	0.53 ± 0.02

Table 1. Summary of the best AUROC performances for a given feature set used for training and side effect predicted.

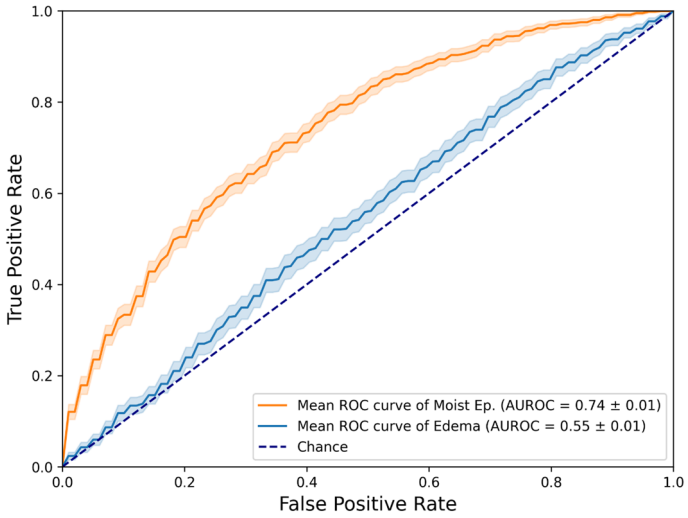


Fig. 3. Test ROC curves of the best performing models for each of the side effects predicted. Moist cells epitheliolysis: LASSO classifier trained on TBV radiomics features, selected by MRMR. Edema: LASSO classifier trained on GT radiomics features, selected by Spearman rank correlation.

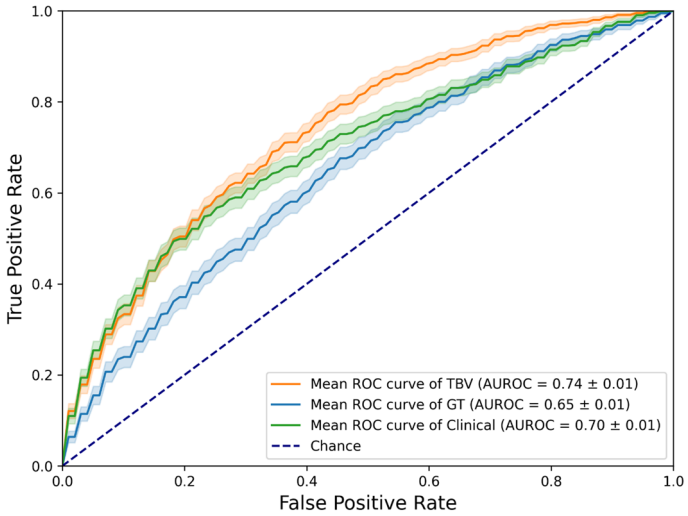


Fig. 4. Test ROC curves of the best performing models depending on the training data used for moist epitheliolysis. TBV radiomics features: LASSO classifier, with features selected by MRMR, predicting moist cells epitheliolysis. GT radiomics features: RF classifier, with features selected by MRMR, predicting moist cells epitheliolysis. Clinical baseline: LR classifier, with features selected by Spearman rank correlation, predicting moist cells epitheliolysis.

algorithms are being shown, according to the evaluation of the four types of models (shown in Table S4). An additional analysis of the best feature selection approach has been made (Table S5).

Combined modelling

Figure 5 shows the best performing models using combined datasets of either radiomics feature sets and clinical features. More scores regarding the comparison of the combinations of radiomics feature sets with clinical features, and their respective predictive performance, can be found in Table S6. When combining TBV radiomics features with clinical ones, LASSO performed best when predicting moist cells epitheliolysis (AUROC of 0.73) although without any overall improvement. RF performed best when predicting edema (AUROC of 0.53), though just above random.

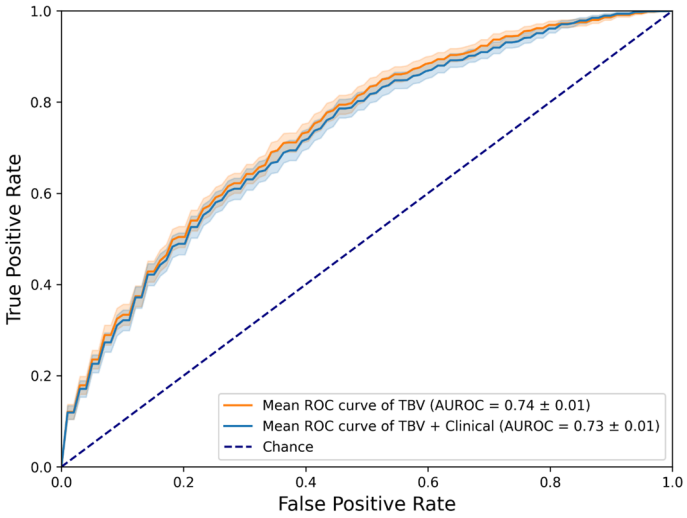


Fig. 5. Test ROC curves of the best performing model trained on a single feature set and the best performing model trained on a combined feature set. Best single feature set model: LASSO classifier trained on TBV radiomics features, selected by MRMR, predicting moist cells epitheliolysis. Best combined feature set model: LASSO classifier trained on a combination of TBV and clinical features, selected by MRMR, predicting moist cells epitheliolysis.

Feature type - name	% chosen	Importance	Score
Shape - Maximum D Diameter Column	99.6	4.30	4.29
Shape - Least Axis Length	100	2.39	2.39
Glcm - Imc	96.4	1.59	1.53
Shape - Surface Area	95.2	1.39	1.32
Shape - Flatness	88.8	1.44	1.28
Glszm - Gray Level Non-Uniformity	98.4	1.05	1.03
Glrlm - Run Length Non-Uniformity	99.6	1.00	0.99
Shape - Maximum D Diameter	53.6	1.83	0.98
Glszm - Size Zone Non-Uniformity	66.4	1.47	0.97
Glrlm - Gray Level Non-Uniformity	93.6	1.01	0.94
Shape - Major Axis Length	77.6	1.20	0.93
Shape - Maximum D Diameter Slice	35.6	2.33	0.83
Firstorder - Energy	96.4	0.85	0.82
Gldm - Dependence Variance	84	0.96	0.81
Shape - Maximum D Diameter Row	65.2	1.23	0.80

Table 2. Top 15 feature importance report of the best performing model: a LASSO classifier trained on TBV radiomics features, selected by MRMR, and predicting moist cells epitheliolysis as a surrogate for skin inflammation. This report shows how often a feature has been chosen out of the 250 iterations (% chosen), the feature importance value given by the model (importance; LASSO coefficients), and a score encompassing the feature importance value and how often that feature was selected (product of these two values).

Feature importance

Table 2 shows the feature importance scores for the best models. To account for both the importance score and the frequency by which a feature was chosen, we computed a score that was the product of these values and ranked the features accordingly.

From the list of the 15 most predictive features, more than half of them belong to the shape type, confirming that planar and volumetric information has a significant influence on the performance of oncological ML models[36–38].

Predictive influence of the total breast volume

The influence of the volume of the whole breast on the prediction quality of the models has been further analyzed. A logistic regression model has been trained only on TBV breast volume, with an AUROC of 0.75 ± 0.01, performing similarly to the best model trained on all TBV radiomics features.

Over all 250 runs, there was a median Spearman correlation coefficient of 0.82 between TBV breast volume and the best radiomics model. Figure S3 shows the Spearman's correlation distribution between the TBV breast volume and the prediction probabilities of the best performing model. The distribution of the respective p-values can be seen in Fig. S4. The p-values of their correlation to the prediction probabilities of the model were significant ($p < 0.05$) in 243 runs.

The predictive influence of the breast volume has been evaluated by retraining the best performing model, but excluding all features with a Spearman correlation coefficient higher than 0.8. These results can be seen in Table S7. With an AUROC of 0.71, performance has slightly but significantly decreased (from an AUROC of 0.74), confirming an effect of the breast volume on the performance of these radiomics models.

Discussion

In this study, we analyzed the relevance of CT-based radiomics to predict two common RT side effects: epitheliolysis of moist cells as a surrogate for skin inflammation; and edema, using a statistically robust pipeline. The best prediction model was a LASSO classifier that was trained on radiomics features from the TBV and selected using MRMR, predicting moist cells epitheliolysis. This model achieved a moderate discriminatory power with an AUROC of 0.74. Clinical features alone or in combination with radiomics did not significantly improve predictive performances.

In contrast, edema was more difficult to predict with a performance level just above random (AUROC score of 0.55 for the best model). The best radiomics model for moist cell epitheliolysis was largely correlated to the TBV volume which itself showed the same reasonable predictive performance with an AUROC of 0.74.

These results have uncovered the previously known fact that radiomics features are largely correlated with the size of the VOI[39,40]. Eliminating volume-correlated features slightly mitigated the performance of the radiomics model (AUROC of 0.71 from 0.74). As consequence, radiomic features do carry relevant information for the prediction of radiotherapy side effects. However, these features are less predictive than TBV volume.

The analysis of the importance of other features revealed several logical patterns. First, shape features appeared to be the most influential ones, indicating that geometrical features play a dominant role in predicting RT-dependent side effects. Maximum D Diameter Column being the most influential feature supports this idea, implying that larger tumors or more irregular tumors may cause more adverse effects to RT due to how the dose distribution is made, and how it affects the neighboring tissue. Further, the presence of multiple gray level types of features suggests that the heterogeneity of the tumor tissue is another significant factor, possibly due to how different types of tissues may react to RT, and the side effects that appear as a cause of this non-uniformity[41,42].

Naturally, the given radiation dose is a decisive factor for development or RT-dependent side effects. The dose was part of the clinical prediction model achieving a decent predictive performance albeit inferior to the TBV volume. In fact, the radiation doses given were largely similar, yielding low variability and thus predictive value. Moreover, this cohort was solely treated with normofractionated RT (conventional RT dose fractionation schedule). The START B trial, however, could also demonstrate the predictive performance of breast size on physician-assessed normal tissue effects in the breast[43].

While LASSO yielded the overall best results, all other ML algorithms have proven to be on a similar level. Only SVM has performed slightly but statistically worse, with an AUROC of 0.69 on the best configuration (compared to LASSO: AUROC of 0.74). The choice of algorithm is relevant but does not affect the performance of the model, as long as the model is optimized and properly trained. The choice of the feature selection technique had a small impact on the overall performance, managing to reduce the data dimensionality without losing much information.

This study is subject to two main limitations. The first one stems from the retrospective nature of our side effect data, deriving from past patient records, which presents a challenge to data quality. To this end, we decided to predict moist cells epitheliolysis as it constitutes a binarized endpoint describing more aggravated skin inflammation. On the other hand, the detection and extent of edema was completely dependent on the subjective physician assessment. The second limitation regards the absence of an external validation cohort for an unbiased estimation of the performance of our models. To compensate for this and have a more reliable and unswayed model performance assessment, we decided to apply a more robust resampling technique, in this case a 50-repeat nested cross-validation.

Conclusions

To conclude, the radiomics models developed in this study have shown a reasonable prediction power towards the epitheliolysis of moist cells side effect, while clinical features yielded intermediate albeit competitive results. Adding information from the whole breast tissue, instead of just glandular tissue, achieved better results overall. The radiomics prediction probabilities were largely correlated to breast volume which remained the most predictive feature, though this correlation only affected to a small extent the prediction power of radiomics features in general. These findings, however, should be further validated on larger, more diverse and multi-centered datasets. Future studies should investigate the potential variations in RT side effects prediction using radiomics information depending on the subtype and stage of breast cancer.

Data availability

All data and code used in this research is available upon contact of the correspondence author (Óscar Llorián-Salvador, oscar.llorian-salvador@tum.de) and in concordance to the ethics committee.

Received: 26 May 2024; Accepted: 20 August 2024
Published online: 29 August 2024

References

1. 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. https://doi.org/10.3322/caac.21492 (2018).
2. Siegel, R. L., MIller, K. D., Wagle, N. S. & Jemal, A. Cancer statistics. *CA Cancer J. Clin.* **73**, 17–48. https://doi.org/10.3322/caac.21763 (2023).
3. Lin, L. *et al.* Regional, and national cancer incidence and death for 29 cancer groups in 2019 and trends analysis of the global cancer burden, 1990–2019. *J. Hematol. Oncol.* **14**, 197. https://doi.org/10.1186/s13045-021-01213-z (2021).
4. Shah, C., Al-Hilli, Z. & Vicini, F. Advances in breast cancer radiotherapy: implications for current and future practice. *JCO Oncol. Pract.* **17**, 697–706. https://doi.org/10.1200/OP.21.00635 (2021).
5. Peeken, J. C., Wiestler, B., Combs, S. E., Image-Guided, & Radiooncology,. The potential of radiomics in clinical application. *Recent. Results Cancer Res.* **216**, 773–794. https://doi.org/10.1007/978-3-030-42618-7_24 (2020).
6. Kumar, V. *et al.* Radiomics: The process and the challenges. *Magn. Reson. Imaging* **30**, 1234–1248. https://doi.org/10.1016/j.mri.2012.06.010 (2012).
7. Desideri, I. *et al.* Application of radiomics for the prediction of radiation-induced toxicity in the IMRT era: Current state-of-the-art. *Front. Oncol.* **10**, 1708 (2020).
8. Peeken, J. C. *et al.* Prognostic assessment in high-grade soft-tissue sarcoma patients: A comparison of semantic image analysis and radiomics. *Cancers* **13**, 1929. https://doi.org/10.3390/cancers13081929 (2021).
9. Bi, W. L. *et al.* Artificial intelligence in cancer imaging: Clinical challenges and applications. *Cancer J. Clin.* **69**, 127–157. https://doi.org/10.3322/caac.21552 (2019).
10. Bera, K., Braman, N., Gupta, A., Velcheti, V. & Madabhushi, A. Predicting cancer outcomes with radiomics and artificial intelligence in radiology. *Nat. Rev. Clin. Oncol.* **19**, 132–146. https://doi.org/10.1038/s41571-021-00560-7 (2022).
11. Peeken, J. C., Nusslin, F. & Combs, S. E. Radio-oncomics: The potential of radiomics in radiation oncology. *Strahlenther. Onkol.* **193**, 767–779. https://doi.org/10.1007/s00066-017-1175-0 (2017).
12. Fox, M. J., Gibbs, P. & Pickles, M. D. Minkowski functionals: An MRI texture analysis tool for determination of the aggressiveness of breast cancer. *J. Magn. Reson. Imaging* **43**, 903–910. https://doi.org/10.1002/jmri.25057 (2016).
13. Feng, Q., Hu, Q., Liu, Y., Yang, T. & Yin, Z. Diagnosis of triple negative breast cancer based on radiomics signatures extracted from preoperative contrast-enhanced chest computed tomography. *BMC Cancer* **20**, 579. https://doi.org/10.1186/s12885-020-07053-3 (2020).
14. Aristei, C. *et al.* Personalization in modern radiation oncology: Methods, results and pitfalls. Personalized interventions and breast cancer. *Front. Oncol.* **11**, 616042 (2021).
15. Hacking, S. M., Yakirevich, E. & Wang, Y. From immunohistochemistry to new digital ecosystems: A state-of-the-art biomarker review for precision breast cancer medicine. *Cancer* **14**, 3469. https://doi.org/10.3390/cancers14143469 (2022).
16. Yamamoto, S., Maki, D. D., Korn, R. L. & Kuo, M. D. Radiogenomic analysis of breast cancer using MRI: A preliminary study to define the landscape. *Am. J. Roentgenol.* **199**, 654–663. https://doi.org/10.2214/AJR.11.7824 (2012).
17. Dijk, L. V. *et al.* Parotid gland fat related magnetic resonance image biomarkers improve prediction of late radiation-induced xerostomia. *Radiother. Oncol.* **128**, 459–466. https://doi.org/10.1016/j.radonc.2018.06.012 (2018).
18. Llorián-Salvador, Ó. *et al.* The importance of planning ct-based imaging features for machine learning-based prediction of pain response. *Sci. Rep.* **13**, 17427. https://doi.org/10.1038/s41598-023-43768-6 (2023).
19. Kraus, K. M., Oreshko, M., Bernhardt, D., Combs, S. E. & Peeken, J. C. Dosiomics and radiomics to predict pneumonitis after thoracic stereotactic body radiotherapy and immune checkpoint inhibition. *Front. Oncol.* **13**, 1124592 (2023).
20. Saadatmand, P. *et al.* A dosiomics model for prediction of radiation-induced acute skin toxicity in breast cancer patients: Machine learning-based study for a closed bore Linac. *Eur. J. Med. Res.* **29**, 282. https://doi.org/10.1186/s40001-024-01855-y (2024).
21. Avanzo, M. *et al.* Electron density and biologically effective dose (BED) radiomics-based machine learning models to predict late radiation-induced subcutaneous fibrosis. *Front. Oncol.* https://doi.org/10.3389/fonc.2020.00490 (2020).
22. Isaksson, L. J. et al. Machine Learning-Based Models for Prediction of Toxicity Outcomes in Radiotherapy. *Frontiers in Oncology 10*, 790 (2020).
23. Lilla, C. *et al.* Predictive factors for late normal tissue complications following radiotherapy for breast cancer. *Breast Cancer Res. Treat.* **106**, 143–150. https://doi.org/10.1007/s10549-006-9480-9 (2007).
24. Kole, A. J., Kole, L. & Moran, M. S. Acute radiation dermatitis in breast cancer patients: Challenges and solutions. *Breast Cancer (Dove Med. Press)* **9**, 313–323. https://doi.org/10.2147/BCTT.S109763 (2017).
25. Huang, C. J. *et al.* RTOG, CTCAE and WHO criteria for acute radiation dermatitis correlate with cutaneous blood flow measurements. *Breast* **24**, 230–236. https://doi.org/10.1016/j.breast.2015.01.008 (2015).
26. Fedorov, A. *et al.* 3D slicer as an image computing platform for the quantitative imaging network. *Magn. Reson. Imaging* **30**, 1323–1341. https://doi.org/10.1016/j.mri.2012.05.001 (2012).
27. van Griethuysen, J. J. M. *et al.* Computational radiomics system to decode the radiographic phenotype. *Cancer Res.* **77**, e104–e107. https://doi.org/10.1158/0008-5472.CAN-17-0339 (2017).
28. Peng, H., Long, F. & Ding, C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.* **27**, 1226–1238. https://doi.org/10.1109/TPAMI.2005.159 (2005).
29. Lemaître, G., Nogueira, F. & Aridas, C. K. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *J. Mach. Learn. Res.* **18**, 1–5 (2017).
30. Jr, D. W. H. & Lemeshow, S. Applied Logistic Regression. (Wiley, , UK, 2004).
31. Brancato, V., Cerrone, M., Lavitrano, M., Salvatore, M. & Cavaliere, C. A systematic review of the current status and quality of radiomics for glioma differential diagnosis. *Cancers (Basel)* **14**, 2731. https://doi.org/10.3390/cancers14112731 (2022).
32. Tibshirani, R. Regression shrinkage and selection via the Lasso. *J. Royal Stat. Soc. Ser. B (Methodological)* **58**, 267–288 (1996).
33. Cortes, C. & Vapnik, V. Support-vector networks. *Mach. Learn.* **20**, 273–297. https://doi.org/10.1007/BF00994018 (1995).
34. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32. https://doi.org/10.1023/A:1010933404324 (2001).
35. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
36. Ludwig, C. G., Lauric, A., Malek, J. A., Mulligan, R. & Malek, A. M. Performance of radiomics derived morphological features for prediction of aneurysm rupture status. *J. NeuroInterventional Surg.* **13**, 755–761. https://doi.org/10.1136/neurintsurg-2020-016808 (2021).
37. Trinh, D. L., Kim, S. H., Yang, H. J. & Lee, G. S. The efficacy of shape radiomics and deep features for glioblastoma survival prediction by deep learning. *Electronics* **11**, 1038. https://doi.org/10.3390/electronics11071038 (2022).
38. Yap, F. Y. *et al.* Shape and texture-based radiomics signature on CT effectively discriminates benign from malignant renal masses. *Eur. Radiol.* **31**, 1011–1021. https://doi.org/10.1007/s00330-020-07158-0 (2021).
39. Hatt, M. *et al.* 18F-FDG PET uptake characterization through texture analysis: Investigating the complementary nature of heterogeneity and functional tumor volume in a multi–cancer site patient cohort. *J. Nucl. Med.* **56**, 38–44. https://doi.org/10.2967/jnumed.114.144055 (2015).
40. Welch, M. L. *et al.* Vulnerabilities of radiomic signature development: The need for safeguards. *Radiother. Oncol.* **130**, 2–9. https://doi.org/10.1016/j.radonc.2018.10.027 (2019).

41. van Timmeren, J. E., Cester, D., Tanadini-Lang, S., Alkadhi, H. & Baessler, B. Radiomics in medical imaging—How-to guide and critical reflection. *Insights Imaging* **11**, 91. https://doi.org/10.1186/s13244-020-00887-2 (2020).
42. Zhang, W., Guo, Y. & Jin, Q. Radiomics and its feature selection: A review. *Symmetry* **15**, 1834. https://doi.org/10.3390/sym15101834 (2023).
43. Haviland, J. S. *et al.* The UK standardisation of breast radiotherapy (START) trials of radiotherapy hypofractionation for treatment of early breast cancer: 10-year follow-up results of two randomised controlled trials. *Lancet Oncol.* **14**, 1086–1094. https://doi.org/10.1016/S1470-2045(13)70386-3 (2013).

Author contributions

Conceptualization, O.L.-S, N.M., L.E., M.A.A.-N., B.R., S.E.C., K.J.B., M.N.D., and J.C.P.; Data curation, M.N.D., N.W., J.C.P.; Formal analysis, O.L.-S., N.M., and J.C.P.; Funding acquisition, J.C.P.; Investigation, O.L.-S., N.M. and J.C.P.; Methodology, O.L.-S., N.M. and J.C.P.; Project administration, J.C.P.; Resources, J.C.P. and S.E.C.; Software, O.L.-S. and N.M.; Supervision, M.A.A.-N., B.R., S.E.C. and J.C.P.; Validation, O.L.-S. and J.C.P.; Visualization, O.L.-S. and N.M.; Writing – original draft, O.L.-S.; All authors reviewed and edited the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL. This work was funded by physician scientist programs of the medical faculty of the Technical University of Munich and the Helmholtz Zentrum Muenchen.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-70723-w.

Correspondence and requests for materials should be addressed to Ó.L.-S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

4.4 Can Radiomics outperform Radiologists in Predicting Ewing Sarcoma Response to Neoadjuvant Chemotherapy from longitudinal MRI?

Can Radiomics outperform Radiologists in Predicting Ewing Sarcoma Response to Neoadjuvant Chemotherapy from longitudinal MRI?

Óscar Llorián-Salvador [1,2,3,*,†], Daniel Rusche [1,†], Johanna Luitjens [4,5], Nicole Martin [2], Felix Gassert [6], Florian Gassert [6], Stefan Weissinger, Miguel A. Andrade-Navarro [3], Stephanie E. Combs [1,7,8], Burkhard Rost [2], Alexandra S. Gersing [5,‡], Jan C. Peeken [1,7,8,‡]

[1] Department of Radiation Oncology, Klinikum Rechts der Isar, Technische Universität München, Ismaninger Straße 22, 81675 Munich, Germany
[2] Department of Informatics, Bioinformatics and Computational Biology—i12, Technische Universität München, Boltzmannstr. 3, 85748 Munich, Germany
[3] Institute of Organismic and Molecular Evolution, Johannes Gutenberg University Mainz, Hanns-Dieter-Hüsch-Weg 15, 55128 Mainz, Germany
[4] Department of Radiology and Biomedical Imaging, University of California, San Francisco, CA, USA
[5] Department of Diagnostic and Interventional Neuroradiology, University Hospital Munich (LMU), Marchioninistrasse 15, 81377 Munich, Germany
[6] Department of Diagnostic and Interventional Radiology, School of Medicine and Klinikum Rechts der Isar, Technical University of Munich, Munich, Germany
[7] German Cancer Consortium (DKTK), Partner Site Munich and German Cancer Research Center (DKFZ), Munich, Germany
[8] Helmholtz Zentrum München (HMGU) GmbH, German Research Center for Environmental Health, Institute of Radiation Medicine (IRM), Munich, Germany
* Corresponding author
† These authors contributed equally to this work
‡ These authors contributed equally to this work

Abstract (max. 300 words)

Background: Ewing sarcoma is a common bone cancer in children and adolescents, typically treated with neoadjuvant chemotherapy followed by surgery. Evaluating the response to initial neoadjuvant treatment is crucial, as inadequate response may warrant adapted therapy. However, definitive histological assessment can only be obtained during surgery, leading to delays for non-responders and, therefore, a potential metastatic spread. This study aimed to assess the predictive potential of radiomics features from MRI to evaluate the histological response to neoadjuvant chemotherapy in Ewing sarcoma patients, and compare it to the predictive value of features extracted from radiology readings.

Methods: Radiomics features were extracted from pre- and post-neoadjuvant chemotherapy MRI scans of Ewing sarcoma patients. Several machine learning models were trained to predict the histological response based on the radiomics data. The performance of these radiomics models was compared to others trained on features extracted from expert radiology readings.

Results: Radiomics features, particularly those related to low gray level values and their distribution, performed best when considering their ratio of change in T1 fat-saturated via contrast agent imaging modalities from pre- to post-neoadjuvant therapy (logistic regression with an area of the receiver-operator curve, AUROC, of 0.62 ± 0.03). They outperformed the best model based on the ratio of change of radiology features, based exclusively on the maximum diameter of extratumoral soft tissue components (logistic regression with an AUROC of 0.58 ± 0.02). Radiomics models trained solely on pre-neoadjuvant therapy data also

performed comparatively well and outperformed the clinical baseline with an AUROC of 0.61 ± 0.03.

Conclusion: Quantitative radiomics analysis of MRI data can provide valuable supplementary insights to guide adaptive treatment strategies for Ewing sarcoma patients, potentially improving clinical outcomes by identifying non-responders to initial neoadjuvant chemotherapy earlier.

Keywords: radiology; radiomics; machine learning; Ewing sarcoma; MRI; neoadjuvant therapy; histological response.

---

1. Introduction (max 400 words)

Ewing sarcoma is the second most frequent bone sarcoma in children and adolescents after osteosarcoma [1]. The most common therapy procedure consists of neoadjuvant chemotherapy, surgical resection and adjuvant chemotherapy. To monitor the effect of neoadjuvant therapies, magnetic resonance imaging (MRI) is performed before neoadjuvant therapy as baseline examination and directly afterwards at the first follow-up. With insufficient response at follow-up, an adapted chemotherapy protocol can be administered additionally, as well as radiation therapy.

Unfortunately, a definitive histological reference for the response to chemotherapy can only be obtained during surgical resection. As a result, non-responders spend valuable time during the extended neoadjuvant treatment before receiving resection, even risking metastatic spread. Therefore, a prediction of the response to the first neoadjuvant treatment based on imaging characteristics would be clinically relevant and can potentially improve the clinical outcome.

Recent studies show that MRIs can be used to evaluate the response to neoadjuvant chemotherapy [2] and machine learning methods can reliably aid in predicting therapy outcomes in patients with sarcomas [3].

The aim of this study is to assess the predictive power of radiomics features obtained from MRIs with several machine learning classifier models, and applied to the therapeutic response to neoadjuvant chemotherapy in reference to the histological response of patients with Ewing sarcoma. In addition, we evaluated the previously developed models against the expert radiological readings from the same MRIs with the best configuration of the machine learning approach. Finally, we analyzed the predictive value of the most important radiomics features, as well as those extracted from radiological readings.

2. Materials and methods

2.1. Dataset collection and preparation

Patient databases from Ludwig-Maximilians-University (LMU) hospital and Technical University of Munich Hospital rechts der Isar (TUM) were retrospectively searched for bone tumor patients treated with VIDA scheme neoadjuvant chemotherapy between 2004 and 2020.

Patients treated with neoadjuvant chemoradiotherapy or radiotherapy were excluded, as well as patients with histological diagnosis other than Ewing sarcoma and insufficient imaging (see Figure 1 for a more detailed patient workflow). In total, n = 96 patients met the clinical criteria.

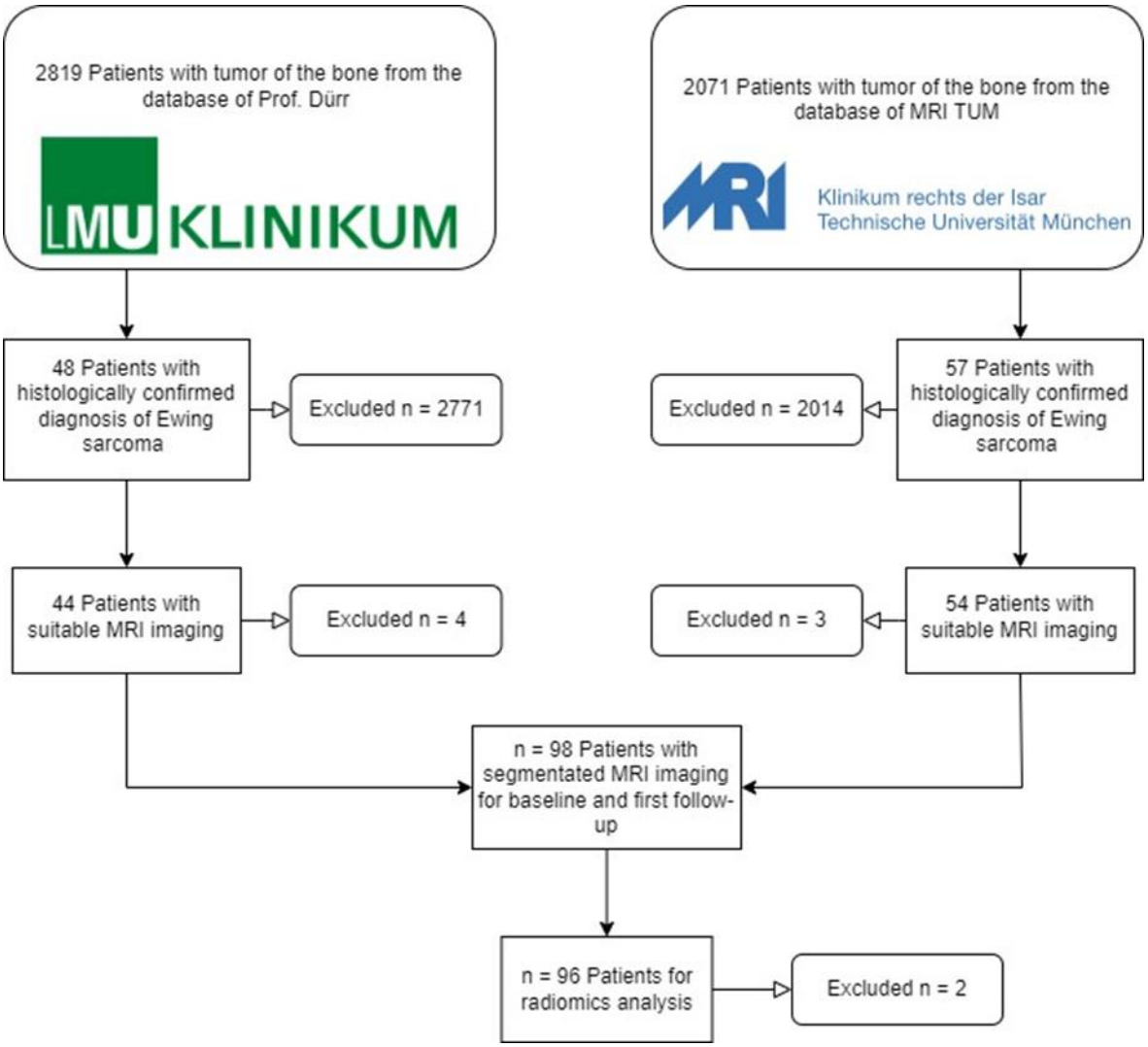


Figure 1. Patient cohort and dataset creation. Histological samples taken after the first follow-up were evaluated according to the Salzer-Kuntschik score for histological response [4]. Patients with a score of up to III were considered responders and larger than III as non-responders. For appropriate imaging, MRI before (baseline) and after (follow-up) were mandatory. Due to technical issues during the radiomics analysis, two more patients had to be excluded.

For each MRI modality, the delta of the baseline and follow-up radiomics features were calculated to further evaluate the treatment effect on the histological response and its evolution.

2.2. MR Imaging protocol and image segmentation

Imaging data consisted of pre-therapeutic baseline and follow-up MRI using 1.0, 1.5 and 3.0 Tesla scanners. We considered three protocols: T1 native (T1w), T1 fat-saturated with the administration of a contrast agent (T1fsgd) and T2 native (T2w). More details regarding the MRI acquisition parameters can be seen in Table S1. The available Digital Imaging and Communications in Medicine (DICOM) files were exported and the volumes of interest (VOI) were segmented (3D Slicer, version 4.11) [5]. The definition of the tumor volume was

supervised by radiology experts (J. L. and F. & F. G.) with two and three years of training and conducted manually by a doctoral student specifically trained for that task. Segmentation label maps were extracted as neuroimaging informatics technology initiative (NIfTI) files. For 20 cases segmentation was performed twice by J. L. and S. W. for later computation of inter-rater variability.

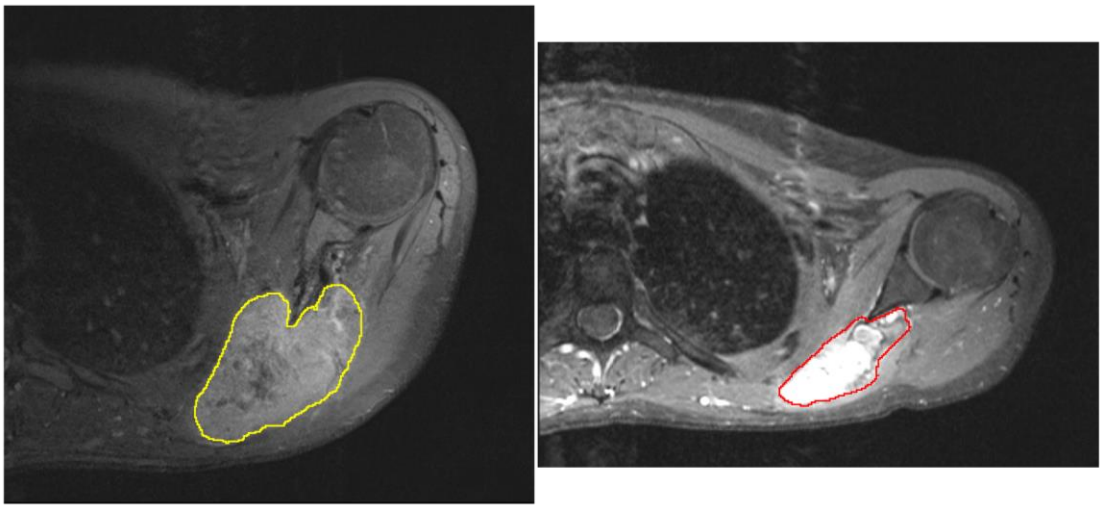


Figure 2. Segmentation of gross tumor volume of an Ewing Sarcoma located in the scapula region at baseline (left) and follow-up (right).

2.3. *Feature sets extraction*

Feature extraction and the implementation of the machine learning pipeline in this project were based on Python3 (version 3.8.10) [6] and a selection of its open-source Python libraries. We extracted a total of 104 features for each sequence at baseline and follow-up respectively (SimpleITK, version 2.2.0 [7]; Pyradiomics, version 3.0.1 [8]). Included were first order, shape and texture features. A clinical feature set contained sex, age, and presence of metastasis at baseline and follow-up. Besides the response evaluation criteria in solid tumors (RECIST) [9] derived from T2w, the mesh volume of the tumor was also considered. All extracted radiomics features are listed in Table S2. To account for inter-rater variability, we calculated the Intraclass Correlation (3,1) (ICC) (Pingouin, version 0.5.3) [10]. Features with an ICC < 0.8 were excluded, reducing the feature space of the radiomics sets to 57.

Pooled features from baseline and follow-up of each sequence were transformed to the standard normal distribution using min-max normalization (Scikit-learn library version 1.0.2) [11]. Feature normalization was performed before splitting the data for training due to the batch harmonization requirements. With the MRI scanner names extracted from the DICOM files with Pydicom (version 2.3.0) [12] we performed harmonization of the combined feature sets from one sequence at baseline and follow-up using Combat (version 0.2.1) [13] to compensate for variances introduced by different brands of MRI scanners. Delta (follow-up - baseline) and ratio of change (delta / baseline) time points for the feature sets were also calculated to explore the predictive quality of feature evolution.

Principal component analysis (PCA) estimated the count of components for each feature as an estimation of the information density of each feature set for the subsequent feature selection step.

Trained radiologists (J. L., and F. G. and F. G., xx and xx years of experience; F. G. and F. G. performed one reading together) determined two sets of 19 features for each MRI scan. Continuous radiology features susceptible to segmentation variations were also excluded using the ICC 3,1 statistic with a threshold of 0.8. Discrete features, given their high sparsity, were tested for inter-rater variability using the Cohen's Kappa test with a threshold of 0.4. Categorical features, prior to this test, were transformed using binary encoding. A list of the radiology features is provided in Table S3. The remaining features were used for training a radiology model that followed the same configuration as the best performing radiomics model.

2.4. Machine Learning modeling

The radiomics and clinical features sets were evaluated for their predictive value for the positive evolution of histological response using four different ML algorithms: logistic regression (LogReg), least absolute shrinkage and selection operator (LASSO), support vector machine (SVM) and random forest (RFC). Data was split and resampled using a 3-fold repeated nested cross-validation for 50 iterations, resulting in the aggregate of 150 models for each set of features and algorithm. The general workflow followed throughout this study is shown in Figure 3. In the inner fold, feature selection and imbalance correction were conducted to respectively reduce redundancy of the data and offset the imbalance of classes in the histological response.

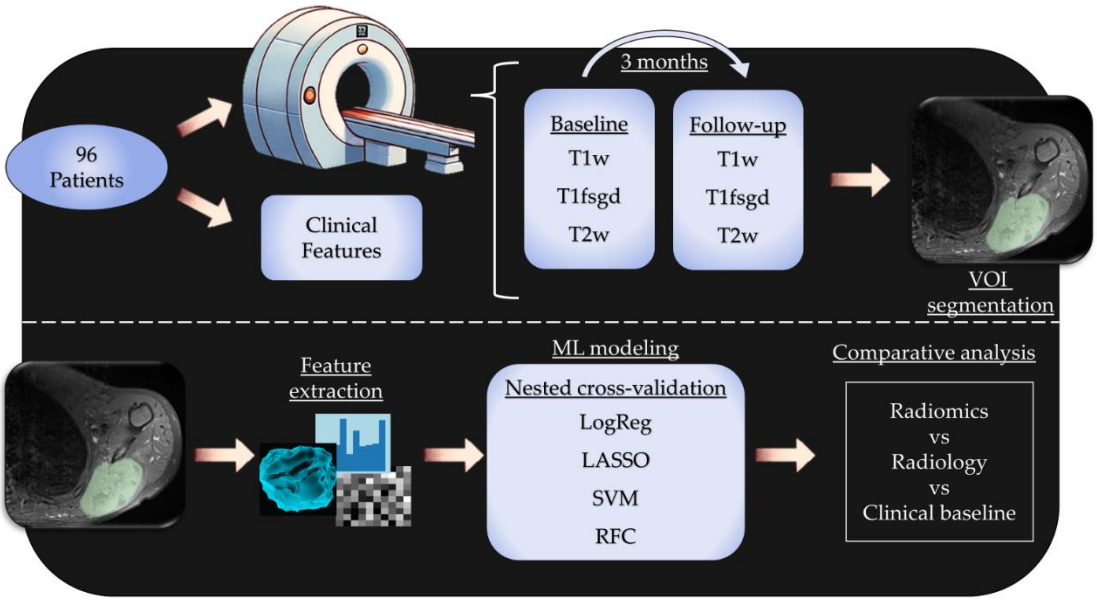


Figure 3. Pipeline followed by the features analyzed in this study.

Two different feature selection techniques were tested to reduce the redundancy of the features and focus on the most predictive ones: a two-step Spearman's rank correlation coefficient, to first exclude redundant features and then exclude low relevance features; and minimum redundancy-maximum relevance (MRMR). The number of selected features was approximated using a principal component analysis (PCA) with 90% and 95% of data variance.

To offset the various degrees of class imbalance of the training set (from 1.42:1 to 2.06:1), synthetic minority oversampling technique (SMOTE; imbalanced-learn library version 0.8.0) [14], random undersampling of the majority class were used. Due to the low size of training sets, wherever minority class oversampling could not be applied, class weights were used instead. To account for a possible remainder class imbalance, balanced accuracy (BA) was the metric used to optimize the hyperparameters of the models.

2.5. Statistical Analysis

After feature normalization and batch harmonization, outlier detection was conducted before nested cross-validation to prevent extreme values to influence the distribution of the data (Scikit-learn library version 1.0.2).

All error margins of the mean observations are reported as 1.96 standard errors for a 95% confidence interval. The models were compared primarily with the area under the receiver-operator characteristic curve (AUROC). Other metrics taken into consideration were BA, F1 score, recall, specificity and Matthews correlation coefficient (MCC).

Feature importance was reported as a score encompassing the frequency of a feature being selected and the weight value that each models computes i.e., coefficient for LogReg, LASSO and SVM, reduction of node impurity for RFC: $Score = Feature\ Importance * m$, where m is the percentage of times that feature has been chosen.

3. Results

3.1. Study Subjects

In this study, 96 patients with Ewing sarcoma were included (age 27.7 ± 12.5; 41 women). The properties of the final imaging datasets after curation and cleaning are depicted in Table 1. Most Ewing sarcomas, as shown in Table S4, were located in the lower limb (thigh, lower leg and foot, n = 46).

Table 1. Number of available MRI sequences, and the subset with available histological response after surgical resection. T1w: T1 weighted MRI sequence; T1fsgd: T1 weighted MRI sequence with fat saturation and contrast agent, T2w: T2 weighted MRI sequence.

Imaging modality	Baseline	Follow-up
T1w	81 (55)	74 (49)
T1fsgd	73 (46)	81 (53)
T2w	58 (37)	66 (42)

3.2. Determination of the best radiomics feature set for histological response detection

The detection of positive histological response was evaluated for all combinations of imaging modalities and scanning times, using all modeling strategies. The clinical feature set served as a baseline for comparison. Table 2 shows the performance of the models for which the AUROC score is statistically significant with respect to the clinical baseline.

Table 2. Test performance of the radiomics models that were statistically better than the clinical baseline, as well as the optimal feature selection technique.

	Ratio of change T1fsgd	Baseline T2	Follow-up T1fsgd	Clinical

Radiomics Score	*LogReg | Spearman*	*RFC | MRMR*	*RFC | MRMR*	*RFC*
AUROC	0.62 ± 0.03	0.61 ± 0.03	0.58 ± 0.02	0.53 ± 0.02
BA	0.57	0.60	0.57	0.53
F1	0.48	0.49	0.43	0.39
MCC	0.15	0.22	0.14	0.06
Recall	0.47	0.48	0.42	0.40
Specificity	0.66	0.72	0.71	0.65

* Data is given as mean ± 1.96 standard errors for a 95% confidence interval.

Three radiomics models performed better than the clinical baseline, which reached an AUROC of 0.53 ± 0.02. The ratio of change of the T1fsgd imaging modality performed best with an AUROC of 0.62 ± 0.03 when trained with LogReg, and using a two-step Spearman's correlation coefficient to filter the number of radiomics features. Besides this model, RFC reached the highest performance for the best baseline, follow-up and clinical baseline models, with respective AUROCs of 0.61 ± 0.03, 0.58 ± 0.02 and 0.53 ± 0.02. The same conclusion can be drawn for the best feature selection technique, where the 2-step Spearman's correlation coefficient performed best overall with the ratio delta T1fsgd radiomics set, but MRMR performed better for the baseline and follow-up models. Interestingly, there is one model per scanning time configuration (baseline, follow-up, and a delta / ratio estimation) that is statistically better than the clinical baseline.

None of the models working with T1w image features could outperform the clinical-based models. The same is valid for all the models working on T1 fsgd at baseline and T2w at follow-up. An AUROC overview from best models per MRI sequence and time point can be found in Figure S1.

3.3. Comparison with Radiologists

The majority of discrete features from the radiology readings, including the encoded categorical features, exhibited minimal change between the baseline and follow-up time points. As a result, these radiology features were largely uninformative for the delta and ratio of change time points. Additionally, all discrete features showed high redundancy, and were therefore filtered out during feature selection for all time points. Two of the continuous features were also consistently filtered out during feature selection, leaving only the extratumoral soft tissue component's maximum diameter as the most informative predictor. The configuration for the best radiomics model was used to train models with the radiology readings features for each of the time points shown in Table 3.

Table 3. Test performance of the features extracted from radiology readings for the ratio of change, baseline and follow-up time points. In all cases a LogReg was used with a two-step Spearman's correlation coefficient as the feature selection technique.

Score	**Ratio of change**	**Baseline**	**Follow-up**
AUROC	0.58 ± 0.02	0.53 ± 0.02	0.49 ± 0.02
BA	0.53	0.52	0.51
F1	0.02	0.31	0.35
MCC	0.09	0.07	0.01
Recall	0.15	0.03	0.42

Specificity	0.09	0.75	0.59

* Data is given as mean ± 1.96 standard errors for a 95% confidence interval.

The best performing time point for features from radiology readings was the ratio of change, yielding an AUROC of 0.58 ± 0.02 at a BA of 0.53. Clinically important metrics performed insufficiently with an F1 score of 0.02, sensitivity of 0.15 and specificity of 0.09. The predictive performance of baseline and follow-up time points for the radiology features scored near-random at 0.53 ± 0.02 and 0.49 ± 0.02 AUROC, respectively.

3.4. Feature importance

To gain more insight into the radiomics features contributing to the highest predictive performance, we conducted a feature importance analysis. Table 4 highlights a list of the 10 most predictive radiomics features for the best performing model, as well as their selection frequency, average importance, and a score product of the average feature importance and the selection frequency.

Table 4. Feature importance report of the top 10 radiomics features from the best performing model: a LogReg trained on the ratio of change T1fsgd features, selected with a two-step Spearman's correlation coefficient.

Feature Type - Name	**% Chosen**	**Importance**	**Score**
Gldm - Small Dependence Low Gray Level Emphasis	80.7	0.11	0.09
Glrlm - Long Run Low Gray Level Emphasis	17.3	0.23	0.04
Glszm - Size Zone Non Uniformity	25.3	0.16	0.04
Glszm - Gray Level Non Uniformity Normalized	28.7	0.05	0.01
Glcm - Joint Entropy	16.7	0.08	0.01
Firstorder - Uniformity	8.7	0.12	0.01
Glcm - Difference Entropy	5.3	0.18	0.01
Shape - Minor Axis Length	14.7	0.06	0.01
Gldm - Gray Level Non Uniformity	10.0	0.07	0.01
Shape - Maximum D Diameter Row	9.3	0.06	0.01

Considering the 10 most important features for the best performing model, 7 of them were based on gray level values and their distribution. The most predictive feature was *Small Dependence Low Gray Level Emphasis* with an occurrence in 80.7% of the cases and an average importance of 0.11, for a final score of 0.09: more than double than the score of the second and third most influential predictors, and 6 times higher than the remainder of top radiomics features. This suggests that a high proportion of small, low gray-level pixel dependencies play a relevant role in the classification criteria of the model.

A patient-wise comparison of the 6 most predictive features between baseline and follow-up time points is illustrated in Figure 3. The radiomics feature evolution was further tested with a Wilcoxon signed-rank test [15] to test the null hypothesis that there is no evolution between the baseline and follow-up distributions.

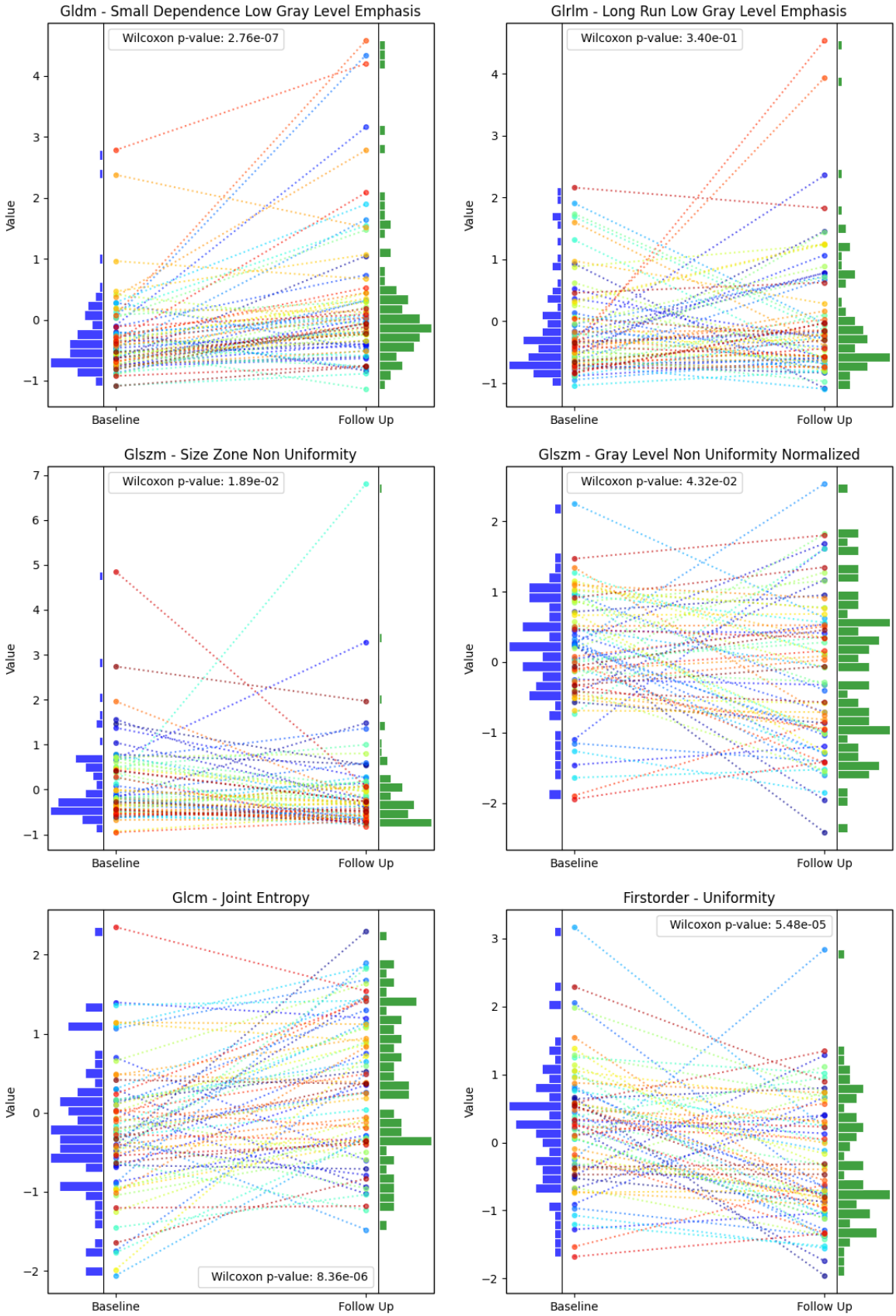


Figure 4. Evolution from baseline to follow-up of all patients for the 6 most predictive features in the best performing model: a LogReg trained on ratio of change T1fsgd features selected by a two-step Spearman's correlation coefficient. Each radiomics feature is accompanied by the p-value of the Wilcoxon signed-rank test.

In all cases the p-values were lower than 0.05, confirming with a 95% confidence interval that there is an evolution between the baseline and follow-up distributions, showing a noticeable response to treatment in the histological response.

4. Discussion

In this study, we analyzed the capabilities of MRI radiomics features compared to extracted features from radiology readings to predict the outcome of neoadjuvant chemotherapy on Ewing Sarcoma. The best performing model was a LogReg based on the ratio of change of radiomics features from T1fsgd with an AUROC of 0.62 ± 0.03 at a BA of 0.57, outperforming the radiology readings at an AUROC of 0.58 ± 0.02.

The superior results of the ratio of change T1fsgd features, when considered within the radiomics analysis, appears logical taking into consideration the effects of chemotherapy on vital tissue. As with an effective neoadjuvant treatment the vital tumor tissue is reduced, the amount of tissue with a significant uptake of contrast agent declines as well. This decline in contrast enhanced tissue is a feature best portrayed in the ratio of change of T1fsgd radiomics features. This impact of changes in contrast enhancement on the prediction of the pathological response of STS to neoadjuvant treatment was also pointed out by Huang et al. [16].

Considering contrast agent uptake on the gray level value scale, the impact of the change is also depicted within the feature importance. The two highest scoring features are both based on gray level values, albeit on low gray level values. This might contrast the proposed importance of high gray level values within tissue with uptake of contrast agent. However, it merely emphasizes the properties of non-tumor-tissue, as with a vanishing tumor the gray levels of the surrounding tissues gain in importance, a trend that can be verified in Figure 3.

This hypothesis is further backed by the insight acquired from the radiology readings. Here, the remaining feature after pre-processing, redundancy reduction and filtering for feature relevance was the extratumoral soft tissue component's maximum diameter. Other features did not show any change between baseline and follow-up to reflect any meaningful evolution in the ratio of change, nor did they convey enough information to be kept after feature selection. Therefore, both methods appear to register the change in soft tissue amount most likely caused by a successful chemotherapy regime. However, features from radiology readings were less predictive compared to the ratio of change information of the radiomics features extracted from T1fsgd MR images.

As shown in Tables 2 and 3, baseline radiomics features from the T2w MRI modality also outperform the radiology readings at an AUROC of 0.61 ± 0.03. Similar to what other studies have shown for STS in general, this analysis of the largest Ewing sarcoma dataset so far emphasizes the trend of radiomics as a diagnostic staging feature able to aid radiologists and oncologists in therapeutic decisions [17].

5. Conclusion

In this study, we posed the question of whether radiomics can outperform radiologists in predicting Ewing sarcoma response to neoadjuvant chemotherapy from longitudinal MRI data. The MRI ratio of change radiomics based prediction of the response to neoadjuvant chemotherapy of Ewing sarcoma outperforms radiology readings. Moreover, radiomics

baseline information performs comparably well, posing an early supplementary tool for the prognosis of Ewing sarcoma patients. Finally, the analysis of radiology and MRI radiomics features suggest that both can monitor the response to neoadjuvant chemotherapy through the change in soft tissue amount, either directly or via measuring the contrast agent uptake effect on the gray level value scale.

Author contributions: Conceptualization, O.L.-S., J.L., N.M., S.W., M.A.A.-N., S.E.C., B.R., A.S.G. and J.C.P.; Data curation, J.L., F.G., F.G., S.W. and J.C.P.; Formal analysis, O.L.-S., N.M. and J.C.P.; Funding acquisition, A.S.G. and J.C.P.; Investigation, O.L.-S., J.L., N.M., F.G., F.G., S.W. and J.C.P.; Methodology, O.L.-S., J.L., N.M., F.G., F.G., S.W., A.S.G. and J.C.P.; Project administration, O.L.-S., A.S.G. and J.C.P.; Resources, S.E.C., B.R., A.S.G. and J.C.P.; Software, O.L.-S. and N.M.; Supervision, O.L.-S., M.A.A.-N., S.E.C., B.R., A.S.G. and J.C.P.; Validation, O.L.-S., N.M. and J.C.P.; Visualization, O.L.-S. and N.M.; Writing – original draft, O.L.-S., D.R. and N.M.; Writing – review & editing, O.L.-S., D.R., J.L., N.M., F.G., F.G., S.W., M.A.A-N., S.E.C., B.R., A.S.G. and J.C.P. All authors have read and agreed to the published version of the manuscript.

Data and code availability: All data and code used in this research is available upon contact of the correspondence author (Óscar Llorián-Salvador, oscar.llorian-salvador@tum.de) and in concordance to the ethics committee.

Ethics vote: The local institutional review board approved this retrospective multi-center study (ethics committee 21-0282) The study was performed in accordance with our institutional ethic guidelines and the 1964 Declaration of Helsinki and its later amendments. Written and informed consent was waived for this retrospective anonymized analysis.

Funding: This work was funded by physician scientist programs of the medical faculty of the Technical University of Munich and the Helmholtz Zentrum Muenchen.

Figure and Table legends: Figure 1. Patient cohort and dataset creation. Histological samples taken after the first follow-up were evaluated according to the Salzer-Kuntschik score for histological response [4]. Patients with a score of up to III were considered responders and larger than III as non-responders. For appropriate imaging, MRI before (baseline) and after (follow-up) were mandatory. Due to technical issues during the radiomics analysis, two more patients had to be excluded.; Figure 2. Segmentation of gross tumor volume of an Ewing Sarcoma located in the scapula region at baseline (left) and follow up (right).; Figure 3. Pipeline followed by the features analyzed in this study. Figure 4. Evolution from baseline to follow-up of all patients for the 6 most predictive features in the best performing model: a LogReg trained on ratio of change T1fsgd features selected by a two-step Spearman's correlation coefficient. Each radiomics feature is accompanied by the p-value of the Wilcoxon signed-rank test.; Table 1. Number of available MRI sequences, and the subset with available histological response after surgical resection. T1w: T1 weighted MRI sequence; T1fsgd: T1 weighted MRI sequence with fat saturation and contrast agent, T2w: T2 weighted MRI sequence.; Table 2. Test performance of the radiomics models that were statistically better than the clinical baseline, as well as the optimal feature selection technique.; Table 3. Test performance of the features extracted from radiology readings for the ratio of change, baseline and follow-up time points. In all cases a LogReg was used with a two-step Spearman's correlation coefficient as the feature selection technique.; Table 4. Feature importance report of the top 10 radiomics features from the best performing model: a LogReg trained on the ratio of change T1fsgd features, selected with a two-step Spearman's correlation coefficient.; Table S1. Acquisition parameters extracted from the DICOM metadata files.; Table S2. Radiomics features extracted from the MRI scans. All extracted features were computed according to the "image biomarker standardization initiative" (IBSI) guidelines [1]. The pyRadiomics package (version 2.0) implemented in python (version 3.6.4) was used for feature extraction [2].; Table S3. Radiology features extracted at baseline and follow-up times.; Table S4. Distribution of the general body locations where Ewing sarcomas were found.; Figure 1. Clustered barplot of the AUROC scores achieved by the best models for each combination of MRI modality (T1w, T1fsgd, T2w) and time point (at baseline, follow-up, delta of change and ratio of change).

References

1. Strauss, S.J.; Frezza, A.M.; Abecassis, N.; Bajpai, J.; Bauer, S.; Biagini, R.; Bielack, S.; Blay, J.Y.; Bolle, S.; Bonvalot, S.; et al. Bone Sarcomas: ESMO-EURACAN-GENTURIS-ERN PaedCan Clinical Practice Guideline for Diagnosis, Treatment and Follow-Up. *Ann Oncol* **2021**, *32*, 1520–1536, doi:10.1016/j.annonc.2021.08.1995.
2. Wunder, J.S.; Paulian, G.; Huvos, A.G.; Heller, G.; Meyers, P.A.; Healey, J.H. The Histological Response to Chemotherapy as a Predictor of the Oncological Outcome of Operative Treatment of Ewing Sarcoma*. *JBJS* **1998**, *80*, 1020.

3. Fanciullo, C.; Gitto, S.; Carlicchi, E.; Albano, D.; Messina, C.; Sconfienza, L.M. Radiomics of Musculoskeletal Sarcomas: A Narrative Review. *J Imaging* **2022**, *8*, 45, doi:10.3390/jimaging8020045.
4. Salzer-Kuntschik, M.; Brand, G.; Delling, G. [Determination of the degree of morphological regression following chemotherapy in malignant bone tumors]. *Pathologe* **1983**, *4*, 135–141.
5. Fedorov, A.; Beichel, R.; Kalpathy-Cramer, J.; Finet, J.; Fillion-Robin, J.-C.; Pujol, S.; Bauer, C.; Jennings, D.; Fennessy, F.; Sonka, M.; et al. 3D Slicer as an Image Computing Platform for the Quantitative Imaging Network. *Magnetic Resonance Imaging* **2012**, *30*, 1323–1341, doi:10.1016/j.mri.2012.05.001.
6. Van Rossum, G.; Drake, F.L. *Python 3 Reference Manual*; CreateSpace: Scotts Valley, CA, 2009; ISBN 978-1-4414-1269-0.
7. Yaniv, Z.; Lowekamp, B.C.; Johnson, H.J.; Beare, R. SimpleITK Image-Analysis Notebooks: A Collaborative Environment for Education and Reproducible Research. *J Digit Imaging* **2018**, *31*, 290–303, doi:10.1007/s10278-017-0037-8.
8. 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.J.W.L. Computational Radiomics System to Decode the Radiographic Phenotype. *Cancer Research* **2017**, *77*, e104–e107, doi:10.1158/0008-5472.CAN-17-0339.
9. Eisenhauer, E.A.; Therasse, P.; Bogaerts, J.; Schwartz, L.H.; Sargent, D.; Ford, R.; Dancey, J.; Arbuck, S.; Gwyther, S.; Mooney, M.; et al. New Response Evaluation Criteria in Solid Tumours: Revised RECIST Guideline (Version 1.1). *Eur J Cancer* **2009**, *45*, 228–247, doi:10.1016/j.ejca.2008.10.026.
10. Vallat, R. Pingouin: Statistics in Python. *Journal of Open Source Software* **2018**, *3*, 1026, doi:10.21105/joss.01026.
11. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-Learn: Machine Learning in Python. *Journal of Machine Learning Research* **2011**, *12*, 2825–2830.
12. Mason, D. SU-E-T-33: Pydicom: An Open Source DICOM Library. *Medical Physics* **2011**, *38*, 3493–3493, doi:10.1118/1.3611983.
13. Behdenna, A.; Colange, M.; Haziza, J.; Gema, A.; Appé, G.; Azencott, C.-A.; Nordor, A. pyComBat, a Python Tool for Batch Effects Correction in High-Throughput Molecular Data Using Empirical Bayes Methods. *BMC Bioinformatics* **2023**, *24*, 459, doi:10.1186/s12859-023-05578-5.
14. Lemaître, G.; Nogueira, F.; Aridas, C.K. Imbalanced-Learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. *Journal of Machine Learning Research* **2017**, *18*, 1–5.
15. Woolson, R.F. Wilcoxon Signed-Rank Test. In *Wiley Encyclopedia of Clinical Trials*; John Wiley & Sons, Ltd, 2008; pp. 1–3 ISBN 978-0-471-46242-2.
16. Huang, W.; Beckett, B.R.; Tudorica, A.; Meyer, J.M.; Afzal, A.; Chen, Y.; Mansoor, A.; Hayden, J.B.; Doung, Y.-C.; Hung, A.Y.; et al. Evaluation of Soft Tissue Sarcoma Response to Preoperative Chemoradiotherapy Using Dynamic Contrast-Enhanced Magnetic Resonance Imaging. *Tomography* **2016**, *2*, 308–316, doi:10.18383/j.tom.2016.00202.
17. Peeken, J.C.; Asadpour, R.; Specht, K.; Chen, E.Y.; Klymenko, O.; Akinkuoroye, V.; Hippe, D.S.; Spraker, M.B.; Schaub, S.K.; Dapper, H.; et al. MRI-Based Delta-Radiomics Predicts Pathologic Complete Response in High-Grade Soft-Tissue Sarcoma Patients Treated with Neoadjuvant Therapy. *Radiother Oncol* **2021**, *164*, 73–82, doi:10.1016/j.radonc.2021.08.023.

5 General discussion

This thesis was dedicated to the study of AI models mainly applied to radiomics information and their predictive potential. Different models were developed for predominant challenges in oncology that had remained to be addressed. Their predictive significance provided insights in multiple areas, as well as confirmed established knowledge in the forefront of ML and radiomics.

In light of the results presented in this thesis, the subsequent discussion will address the working hypotheses in detail. Furthermore, other topics of interest that have emerged throughout the course of this research will be addressed. This includes a consideration of the many commonalities observed across all conclusions of the studies encompassed by this work.

5.1 Evaluation of working hypotheses

- **Radiogenomic AI models can differentiate, with high accuracy, between biologically similar tumors such as ALTs and lipomas.**

In the first study presented in the results of this thesis, ML models were developed and validated, using MRI-based radiomics features to predict the MDM2 gene amplification status: a key biomarker for the differentiation between ALTs and lipomas. The best model tasked to perform this classification was a LASSO classifier trained on the combination of the radiomics features from all three available MRI sequences, with an AUROC of 0.88, and a sensitivity, specificity and accuracy of 0.70, 0.81 and 0.76, respectively. Interestingly, while adding demographic features to this model did not improve the overall performance of the LASSO model (AUROC of 0.72 at 0.77 accuracy and 0.40 sensitivity), it reached a perfect classification of all lipomas as such, with 1.00 sensitivity.

Another conclusion of this study was that the addition of clinical features to the best performing radiomics model did not help differentiating ALTs and lipomas further. Even though in previous studies some demographic differences were noted between ALT and lipoma cases [112], it is possible that MRI-based radiomics features are considerably more important, and therefore demographic features are either discarded during the feature selection process, or given a very low weight by the models. Coupled with this, it is important to mention that this study only included the age, sex and anatomical region of the tumor. Including more diverse demographic features could still increase the predictive potential of radiogenomic AI models. Future studies may benefit from combining the existing MRI-based radiomics features with clinical features such as

additional clinical treatments or further physical conditions, especially those more closely related to the location of type of soft tissue tumor.

So far, published studies in this direction had several limitations that needed to be overcome in order to have a higher significance of the conclusions [113–115]. The first and an important shortcoming is the lack of an external test set (also referred to as external validation cohort in the first study presented in this thesis, and in other similar studies). While proper data handling and resampling techniques can lower overoptimistic results due to data leakage, there is a potential inherent relationship of the data when training and testing patients belonging to the same center, which can be solved with an external test set. Moreover, previous studies were based on smaller patient samples [113–115]. Given that smaller sample sizes, especially of patients, have a propensity to lack variability, this could affect the significance of results in two ways. First, lower variability can also lead to overoptimistic results, particularly when tied to the lack of an external test set. Second, smaller patient sets can lead to results that are not reproducible on other sets of patients given their lack of variability.

Therefore, it can be confirmed to a great extent, with the results discussed from the first study, that: radiogenomic AI models can differentiate, with high accuracy, between biologically similar tumors such as ALTs and lipomas. The models developed significantly outperformed radiology residents, with an average accuracy of 0.65, sensitivity of 0.68 and specificity of 0.70; despite they did not improve the performance of the attending radiologist (approximately 10 years of experience), with an accuracy of 0.90, sensitivity of 0.96 and sensitivity of 0.87. Therefore, these models have a great potential as a supplementary tool in the learning experience of radiology residents.

- **AI models can accurately predict the complete response to palliative RT treatment in PSBM patients.**

The second study included in this thesis presented the predictability of complete pain response from patients of PSBM under palliative RT treatment, using ML tools a combination of radiomics, semantics and clinical features. The best models trained with each of the feature sets managed an above-random performance: both radiomics segmentations, gross tumor volume (GTV) and clinical target volume (CTV), had a respective AUROC of 0.58 (SVM trained on GTV radiomics features) and 0.62 (RFC trained on CTV radiomics features). Semantic and the spinal instability neoplastic score (SINS) models performed similarly, with AUROCs of 0.63 and 0.65 respectively. On the other hand, demographic-based models performed best with an AUROC of 0.80 (SVM).

Models trained on combinations of different feature sets were also compared. None of the models outperformed the clinical-only SVM: the best combined model was an SVM trained on CTV radiomics and clinical features, reaching an AUROC of 0.75. A possible explanation, backed by the feature importance reports provided, is that clinical information was the most predictive type: therefore, the addition of other features would only hide the most insightful patterns of the data with less predictive inputs.

Another interesting conclusion from this work was the overall low correlation and predictive power of the radiomics features: none of them were consistently selected across all radiomics models; and only 10 of them were chosen in at least half of the feature selection processes. The nature of PSBMs as distant metastases can explain such results, being more heterogeneous than in other types of cancer, therefore hindering the ability of radiomics of capturing such patterns for the prediction of complete pain response to palliative RT. In contrast, clinical features have proven to be more predictive for distant metastases and more heterogeneous types of cancer. Previous works focused on the monitoring of pain response after RT also support this observation, particularly for features related to the performance status of the patient, their specific histology, use of opioids, and absence of visceral metastases [50,52,53,65,93,116].

Therefore, it can be confirmed to a great extent, with the results provided by the second study, that AI models can accurately predict the complete response to palliative RT treatment in PSBM patients. Although radiomics data was not as informative as other types of data studied, all of them performed significantly above random, with clinical features achieving the maximum efficacy. The addition of more demographic features, especially others supported by previous studies and other works in this thesis, suggests a potential further enhancement of AI models towards the prediction of complete pain response in these circumstances of the disease [50–53].

- **Radiomics features can capture with most precision the patterns related to complete pain response to palliative RT in PSBM patients compared to a clinical baseline.**

According to the findings mentioned from the second study, while radiomics ML models achieved significant efficacies with AUROCs of 0.58 (GTV segmentation) and 0.62 (CTV segmentation), they did not outperform clinical-based models. In light of these results, it cannot be confirmed that radiomics features can capture with most precision the patterns related to complete pain response to palliative RT in PSBM patients compared to a clinical baseline. The heterogeneous nature of PSBMs can explain the predictive difficulty of radiomics features; whereas clinical information is, for the most part, more independent or takes into consideration factors such as the type or location of the primary cancer.

- **The spinal instability neoplastic score (SINS) can be used as a gold-standard assessment variable for the prediction of complete pain response in PSBM patients.**

Previous works already conducted an analysis of the predictive performance of the SINS, particularly in comparison with clinical features [51,117]. In the second study presented in this thesis, SINS appeared to be significantly informative regarding complete pain response after adjustment for gender, tumor type and performance status. However, ML models trained exclusively on these clinical features outperformed the SINS models. As expected, training a LR with a binarized SINS decreased the efficacy from an AUROC of 0.65 to 0.54: by binarizing this score, patterns that can be picked up by the model are bound to be lost.

Therefore, it can be confirmed to a limited extent, with the results provided by the second study, that the SINS can be used as a gold-standard assessment variable for the prediction of complete pain response in PSBM patients. While it is significantly predictive on its own, and more insightful than radiomics models, it was outperformed by clinical-only models and combined models including clinical features.

- **Radiomics models can accurately predict the appearance of side effects from RT treatment of breast cancer.**

In the third study presented in this thesis, the appearance of two of the most common acute side effects to RT treatment in breast cancer patients was explored by using CT-based radiomics models. The best performing model found was a LASSO classifier trained with radiomics features extracted from the total breast volume segmentation (TBV), and predicting the epitheliolysis of moist cells as a surrogate for skin inflammation (AUROC of 0.74). In this case, the combination of clinical features, or such features on their own, did not manage to improve the performance of the best model. On the other hand, the best model trained to predict epitheliolysis of moist cells using radiomics features from the glandular tissue segmentation (GT) had a considerably lower efficacy (AUROC of 0.65). This drop in performance is expected, considering that the TBV segmentation encompasses information from multiple tissues, whereas GT is a single tissue segmentation.

The other side effect considered, edema, was considerably harder to predict. The highest efficacy was achieved by another LASSO classifier, but trained only on radiomics features from the glandular tissue segmentation (GT; AUROC of 0.55). The difference with the TBV-based radiomics model was minimal and non-significant: a RF trained on TBV radiomics features performed with an AUROC of 0.53. However, given that these

models perform only slightly above random, conclusions regarding the edema side effects are limited.

Findings regarding the importance of specific features included the high importance of shape-based features for the prediction of side effects after RT for treating best cancer, an observation that was found previously in literature and in other studies of this thesis [118–120]. Furthermore, the presence of high-importance gray-level texture features indicates the relevance of the heterogeneity of the tumor tissue. This can be explained by individual compositions of the anatomy (for instance, dense or less dense breast tissue) that may lead to different reactions to RT.

Taking into consideration the conclusions from the study, it can be confirmed with different degrees that radiomics models can accurately predict the appearance of side effects from RT treatment of breast cancer. Regarding the epitheliolysis of moist cells as a surrogate for skin inflammation, high levels of significance were achieved when trained with TBV radiomics features, with slight variations depending on the ML algorithm. This finding can be extended, with slightly lower levels of efficacy, to models trained with GT radiomics features, clinical features, and volumetric-excluded radiomics features. However, models trained to predict edema, albeit significant, performed significantly worse, indicating that radiomics and clinical information varies in efficacy depending on the type of side effect.

- **Total breast volume (TBV) influences significantly the performance of breast cancer imaging radiomics models.**

The conclusions from this study support a previously observed finding: in this type of cancer, radiomics features are largely influenced by the volume of the breast [98,99]. In this case, the Spearman correlation coefficient found between the TBV and the prediction probabilities of the best model was 0.82, with only 7 of the 250 runs having a non-significant p-value (threshold of 0.05 for a 95% confidence interval). Interestingly, even though there is a high correlation between them, another LASSO classifier trained without the volumetric-dependent features still had a similar performance (AUROC of 0.71).

In light of this, it can be confirmed to a limited extent that the TBV correlates highly to the performance of breast cancer imaging radiomics models, affecting their performance significantly. While a strong correlation was observed between TBV and the performance of the models, its impact on the performance was found to be limited as seen by radiomics models trained without volumetric-dependent features.

- **Radiomics models can effectively predict the prognostic outcome of Ewing sarcoma patients treated with neoadjuvant chemotherapy.**

The last study presented in this thesis, analyzed the ability of MRI-based radiomics features to predict the therapeutic outcome of neoadjuvant chemotherapy for Ewing sarcoma patients. Additionally, these radiomics features were captured before treatment (at baseline) and 3 months after treatment (after the first cycle of chemotherapy, referred as post-ct), being able to estimate an absolute and a relative delta in the radiomics features. The modeling strategy that could best capture the therapeutic outcome via histological response was a LR trained on the relative delta of radiomics features from the T1fsgd MRI sequence (AUROC of 0.62).

It seems logical that T1fsgd is the sequence that best captures this evolution: with an effective neoadjuvant chemotherapy treatment, there is a large change in vital tumoral tissue volume between baseline and post-ct. As noted in this study, this volume change can also be observed in the feature importances from the models trained on absolute or relative delta radiomics information, specifically in the features portraying low gray level values: an effective neoadjuvant chemotherapy treatment reduces the vital tumoral tissue volume (high gray level values). The larger absence of high gray level areas emphasizes, by contrast, low gray level features, which may portray larger proportions of necrotic tissue visible in post-ct MRI scans.

The prognostic potential of radiomics models trained on the relative delta of T1fsgd features is limited by the necessity post-ct data. However, it was possible to achieve comparable results using only baseline information: a LR trained only on baseline T2w radiomics information reached an AUROC of 0.61.

Finally, it is essential to mention that, as a very rare malignancy, it is considerably harder for ML models to adequately recognize patterns due to smaller sample sizes due to the scarcity of cases, and the heterogeneity bound to smaller sample sizes. Nevertheless, the last study included in this thesis analyses the largest, to date, Ewing sarcoma dataset. This underscores the need of supplementary AI tools to help radiologists and oncologists with therapeutic decisions, in cases where experience is difficult to gain.

In view of these findings, it can be confirmed that radiomics models can predict the prognostic outcome of Ewing sarcoma patients treated with neoadjuvant chemotherapy to a certain extent. Multiple radiomics ML models managed to achieve significant

performances even with only baseline information. Their predictive power, lower than other radiomics models trained in this thesis for other oncological challenges, is likely limited by the rarity of this type of malignancy and, therefore, lower number of cases.

- **Radiomics models outperform radiology readings models when applied to the prognosis of Ewing sarcomas treated with neoadjuvant chemotherapy.**

The best performance achieved by ML models trained on features extracted from radiology readings was from a LR trained on their relative delta, with an AUROC of 0.58. The only radiology feature that was found to be predictive was the maximum diameter of the extratumoral soft tissue component. This further supports the possible reason why the best radiomics model was the one trained with the relative delta of T1fsgd features, where features related to the relevance of low gray value analysis were the most informative. Other models trained on information from radiology readings performed random or significantly worse.

Therefore, with the results provided by the last study from this thesis, it can be confirmed to a limited extent that radiomics models outperform radiology readings models when applied to the prognosis of Ewing sarcomas treated with neoadjuvant chemotherapy. The best radiology-based ML model was outperformed by several radiomics-based models, both including relative delta and only baseline features. However, this difference was not statistically significant given the wider 95% confidence intervals from having a smaller sample size.

5.2 Synthesis and integration of key findings

5.2.1 Technical optimization

The pipelines used for the development and validation of ML models in the different studies included in this thesis had flexibility in several of their specific procedures. This flexibility allowed for an exploratory analysis in search for the best option in each step. Here, the most impactful ones are discussed.

The first step considered is the correction of the batch effect correction. This data harmonization step requires data to be normalized. However, normalizing all data together introduces some degree of data leakage: when scaling future testing values, future training samples are taken into consideration as well. In the first study from this thesis, the option of forgoing batch harmonization in order to avoid data leakage was studied. The best performing model in that study, a LASSO classifier trained on the combined radiomics features from all three MRI sequences, was replicated using the same but non-harmonized, non-normalized data. While the original LASSO scored an

AUROC of 0.88, a balanced accuracy (BA) of 0.75 and an MCC of 0.51, the same model trained on non-harmonized, non-normalized data scored an AUROC of 0.87, BA of 0.50, and an MCC of 0.01, confirming how essential it is to perform batch harmonization: the introduction of batch effects usually leads to biased, lower performing models due to instrument noise.

The feature selection technique used was an important decision, which varied in each of the studies. The main techniques compared were a two-step Spearman rank correlation coefficient test and MRMR. Each technique performed better in different scenarios, without one performing best in all cases. MRMR was the best technique when selecting radiomics features from breast cancer cases for the prediction of common acute side effects: the LASSO classifier trained on TBV radiomics features to predict epitheliolysis of moist cells used MRMR, and achieved the AUROC of 0.74. A two-step Spearman rank correlation coefficient test worked best when predicting the prognosis of Ewing sarcoma patients after neoadjuvant chemotherapy, with a LR trained on the relative delta of T1fsgd radiomics features reaching the best performance of an AUROC of 0.62. However, in both cases the alternative technique achieved respective performance of two and three points lower (AUROCs of 0.72 and 0.59), concluding that the feature selection technique utilized in each scenario has to be compared for achieving optimal performances.

Other feature selection techniques were also tested: variance threshold, random forest selection and a single Spearman rank correlation coefficient test for redundancy reduction, all yielded consistent and significantly lower performances. A possible explanation for the lower performance of these feature selection techniques is that they rely mainly only on one feature reduction mechanism (redundancy reduction or relevance maximization). By taking into account only one of the two mechanisms, harsher filtering thresholds are needed to achieve the same final number of selected features.

Lastly, the choice and balancing type of class imbalance correction can have critical effects on the performance of the models, especially in small datasets where model generalization is majorly hindered by overfitting. The two main class imbalance correction measures compared were a combination of SMOTE and random undersampling of the majority class and using class weights. In most cases, the former has proven to be a better alternative thanks to a higher customization of how class imbalance is corrected. The addition of too much synthetic data, in datasets where class imbalance is larger, can also lead to potential overfitting. Therefore, finding the best proportion of synthetic oversampling and random undersampling was in all cases treated as another hyperparameter to optimize. The most extreme case of class imbalance

was the second study, where the complete pain response prediction to palliative RT of PSBM patients was analyzed. With an initial ratio of 10:1 of non-responders to positive response, the optimal ratio found was to correct 60% of the total class imbalance via SMOTE, 20% via random undersampling, and accounting for the remainder by using BA as the optimization criteria score.

One special example of the use of class imbalance correction was the last study of this thesis: even though the available Ewing sarcomas dataset is the largest to this date with 96 patients available for analysis, histological assessment could only be obtained during surgery, which was only available to a subset of the cases: ranging from 37 to 55 depending on the MRI sequence and at baseline or post-ct. Although class imbalance was not as steep in this study (ranging from 1.42:1 to 2.06:1), the small sample size in some datasets made impossible to apply a reasonable SMOTE or random undersampling balancing. In such cases, balanced class weights were used.

5.2.2 Dataset fingerprint impact

The unique characteristics and patterns of a dataset, often referred as fingerprint, have been a deciding factor in most stages of the analysis process. The respective sample sizes of each of the four studies included in this thesis were 257, 261, 252, and 96. However, multiple factors such as missing data, outliers and medical requirements have lowered these numbers to different extents. While always a critical factor, having a clean and curated dataset becomes even more important when the sample size is so small. In these cases, the outcome of each prediction becomes proportionally more important, and one misprediction can more significantly impact the final performance of a ML model.

It is this aspect of the dataset fingerprint that made the main impact on deciding the n-CV fold numbers. With a lower number of folds, each validation split will have more samples, therefore decreasing the impact of each misprediction. Further, given that smaller datasets are prone to overfit more due to the generalization difficulty, a smaller fold number will also increase each training split at the cost of a larger standard deviation. This finding is further backed by the initial testing of leave-one-out cross validation (LOOCV) as the resampling technique for the second study. However, as observed in the profiles of learning curves, none of the models trained managed to generalize sufficiently.

An additional effect of having a smaller sample size in datasets where patient data is analyzed is the higher variance of the samples, translated into wider error margins, enhanced by having lowered the n-CV fold number as well. These two reasons are the main rationale behind the use a more complex resampling technique such as n-CV, and

increasing the number of iterations as much as it was computationally possible. The final fold numbers, ranging from 3 to 5, worked rather well when paired with 50 iterations: averaging 150 to 250 final models in each case have substantially narrowed error margins for the 95% confidence interval that is shown by default. Nevertheless, the impact of the overall dataset size is present in the final error margins: the respective AUROCs of the best performing models from each study, including a 95% confidence interval error margin, were 0.88 ± 0.01 (LASSO trained on combined MRI-based radiomics features to differentiate ALTs and lipomas), 0.80 ± 0.01 (SVM trained on clinical features to predict complete pain response), 0.74 ± 0.01 (LASSO trained on CT-based TBV radiomics features to predict epitheliolysis of moist cells), and 0.62 ± 0.03 (LR trained on the relative delta of T1fsgd radiomics features to predict prognosis on Ewing sarcomas). Even though, there are more factors at play that may affect the final variance of the predictions, the dataset sizes and the heterogeneous nature of this type of data have shown to be significantly impactful. An estimation of the correlation between the dataset sizes from the studies included in this thesis, and their respective error margins, results in a Spearman correlation of -0.78. This assessment, which suggests a strong negative correlation between the dataset sizes and the error margins, supports the previously mentioned variance observations.

Another aspect to consider regarding the dataset fingerprint are the radiomics feature types analyzed in the studies. As multiple previous studies suggested, shape-related radiomics features have been found to be the most predictive [118–120]. The feature importance reports from the best performing radiomics models in each of the studies can be respectively found in Table S5 (first study), Table S8 (second study), Table 2 (third study) and Table 4 (fourth study). The most significant features were determined as $Score = Feature\ Importance/[(n + 1) - m]$, where *Feature Importance* is the specific metric used by each model (such as coefficients or node impurities), n is the number of models, and m is the number of times the feature has been chosen. Out of the seven types of radiomics features, shape was the most common overall: out of the 10 most significant radiomics features from each of the studies, 6, 2, 5 and 2 of them were only shape-based, while the rest were a combination of the rest of types. The absolute majority in two of the studies, and major presence in the others, supports the findings from previous radiomics works regarding the importance of volumetric radiomics features.

However, while shape-based features emerged as the overall most significant type, it is important to acknowledge that the relevance of radiomics features mostly depend on the specific use case. Although shape features may play a more prominent role, particularly in the distinction between tumor types (as seen in the first study), texture and intensity features often provide key information that can be equally important. This is especially when assessing heterogeneity or predicting treatment outcome from histological response (as seen in the fourth study). This also aligns with findings from

previous studies, which emphasize that no single category of radiomics features universally dominates, and the optimal selected features is often task-dependent [75,121].

We make a final observation regarding the clinical feature importances, and how well clinical models performed depending on the type of cancer and available clinical information. In the first study, where clinical features included age, sex and body region of the tumor, their performance was significantly above random, but far from the performance of radiomics models (Table S4). The best clinical model was a RF, with an AUROC of 0.63. In the second study, clinical features included age, Karnofsky performance scale (KPS), use of opioids, and type of primary tumor that evolved into PSBM. This clinical information was the most predictive overall, with an SVM achieving an AUROC of 0.80. In the third study, the available clinical features were the smoker status, chemotherapy received, RT boost, and maximum radiation dose. Clinical-based models managed to predict the epitheliolysis of moist cells with an AUROC of 0.70 and edema with an AUROC of 0.53. Finally, in the last study the clinical features contained age, sex and presence of metastasis at baseline and post-ct. In this case, clinical-based models performed with an AUROC of 0.53 as well. The main finding from analyzing the clinical features and the models trained on them is that the best performing clinical models (from the second and third study) contained more varied and insightful demographic features. In the case of prediction of complete pain response to palliative RT therapy on PSBM patients, where data can be more heterogeneous given the source of the primary tumor, taking this location into consideration as a clinical feature has proven to be not just insightful, but the most important clinical feature as observed in Table S9. In the third study, where common acute side effects from RT treatment in breast cancer cases are predicted, clinical features were also varied, and have already been reported to be associated to RT side effects [122,123].

5.2.3 Comparative model efficacy

The choice of ML algorithm is arguably one of the most important decisions in the framework. Out of the four models that were considered in all studies (SVM, RF, LASSO and LR), none performed best in all circumstances. Table 1 shows the AUROC of the best performing models, using the same training data, for each algorithm.

Table 1. AUROC scores for the best performing models trained on the same data for each algorithm. Each column corresponds to the studies included in the results of this thesis. Each training data corresponds to the training data of the best performing model overall. In the second study, only SVM and RF models had been considered.

AUROC	Study 1	Study 2	Study 3	Study 4
SVM	0.84 ± 0.01	0.8 ± 0.01	0.69 ± 0.02	0.59 ± 0.03
RF	0.87 ± 0.01	0.78 ± 0.01	0.72 ± 0.01	0.50 ± 0.03
LASSO	0.88 ± 0.01	-	0.74 ± 0.01	0.62 ± 0.03

LR	0.86 ± 0.01	-	0.73 ± 0.01	0.62 ± 0.03

Considering their relative best efficacies, LASSO shows the highest average performance across the 3 studies where it has been used. However, such conclusion is not statistically significant. Moreover, LASSO and LR were initially considered and discarded for the study of complete pain response in PSBM patients after palliative RT therapy, but given their low performance they were discarded for further modeling and final results in order to adhere to the scope of the project. For this reason, SVM also stands out as the most reliable model, with consistent efficacy in all studies regardless of whether the training data is clinical, radiomic, radiogenomic, semantic or otherwise.

The use of combined models, generally by combining radiomics and clinical features, did not increase the predictive power of models. In the first study, the best performing model, which encompassed radiomics features from all MRI sequences, did not improve the performance of an AUROC of 0.88 with the addition of clinical features (AUROC of 0.72). Although most scores decreased, it still held some radiological interest by being able to perfectly identify all lipomas (specificity of 1.00), while still having a competent generalization capability. In the rest of studies, combined models did not outperform the best existing models in any case, regardless of the types of combined features.

A final factor to evaluate is the running time for each of the models during the hyperparameter optimization and training process. Here, the complexity of the model and the number and depth of hyperparameters play a pivotal role. RF was the model that required the most time, given its higher relative complexity and customization of hyperparameters to accommodate and generalize many different types of datasets. On the other hand, LR had the lowest running time, with LASSO being slightly slower, possibly due to the incorporation of the extra regularization parameter to optimize. Finally, SVM had a relative intermediate running time, though it was highly dependent on the complexity of the kernel utilized.

6 General conclusion and outlook

The present work offers insight into the fields of artificial intelligence, oncology and radiomics: different radiomics AI models have been developed and evaluated on their ability to solve a range of common and recurrent cancer-related challenges.

1. Radiogenomic models can differentiate between atypical lipomatous tumors and lipomas via the detection of the MDM2 gene biomarker, with a performance significantly higher than that of radiology residents. The use of all MRI radiomics features combined in a single LASSO model achieved this goal with the highest efficacy. These models, still with a lower performance than attending radiologists, can improve the clinical diagnostic workup and serve as a supplementary tool in the learning process of medical residents. Further works in this area could enhance the interpretability and understandability of AI tools oriented to physicians for their ease of understanding and use.

2. Clinical factors can accurately predict the complete pain response of painful spinal bone metastasis patients after radiotherapy (RT). CT-based radiomics and semantics models can also predict complete pain response, though to a more limited extent. Spinal instability neoplastic score is predictive enough to offer significant insight on its own, albeit still outperformed by clinical-based models. Given the clinical efficacy in cases of distant metastases or heterogeneous types of cancer, future studies may benefit from a more extensive demographic collection of information, especially tailored to the type of malignancy.

3. Radiomics models can reliably anticipate the presence of common acute RT-related side effects after treatment for breast cancer, especially epitheliolysis of moist cells as a surrogate for skin inflammation. The predictability of edema, on the other hand, was limited. The radiomics predictions were largely correlated to breast volume; however, its impact on the predictive performance was minimal, as it was observed in a radiomics model after excluding volumetric-dependent features. Provided the comparatively high performance of TBV radiomics and clinical-based models, subsequent research work should consider expanding the range of side effects to analyze, possibly also including late RT side effects.

4. The absolute and relative delta radiomics features extracted from longitudinal MRI scans can, to a considerable degree, predict the histological response after neoadjuvant chemotherapy of Ewing sarcoma patients. Baseline-only models perform comparatively, adding the value of relying exclusively on early, preoperative information. Radiology readings models reached a lower albeit significant efficacy. Both types of models can track the response to neoadjuvant chemotherapy, respectively, via contrast agent uptake shift on gray level feature values and the volume change of soft tissue. Granted the rarity of this malignancy and its generalization difficulty, follow-up investigations should try to acquire a

more complete, curated dataset to ensure more reliable pattern recognition. A possible alternative would be to monitor the treatment response using a more accessible indicator. This approach is particularly relevant for the discarded patients that did not undergo surgery, and for whom histological response could not be acquired.

A. Curriculum vitae

Óscar Llorián-Salvador
Glückstraße, 3; Germering, Bavaria, Germany
82110

oscar.llorian-salvador@tum.de | ollorian@hotmail.com

Education and training

2021-Current	Doctoral program *Johannes Gutenberg University, Mainz, Germany* *Technical University of Munich, Munich, Germany*
2015-2017	M. Sc. Bioinformatics *Wageningen University, Wageningen, The Netherlands*
2010-2014	B. Sc. Biotechnology *Oviedo University, Oviedo, Spain*
2008-2010	Scientific and Technologic secondary school title *Santa Teresa de Jesús School, Oviedo, Spain*

Work experience

01.2021 – Current	Doctoral program in the Computational Biology and Data Mining group (Johannes Gutenberg University, Mainz), and collaborating with the research groups Rostlab (Bioinformatics) and Radiation Oncology (Klinikum Rechts der Isar) in TUM, Munich.
03.2018 – 12.2020	Predoctoral research assistant in the Rostlab group at the Bioinformatics department in collaboration with the Radiomics group at the Radiation Oncology department (TUM, Munich).
03.2017 – 10.2017	Bioinformatics internship in the Rostlab group at the Bioinformatics department in the Technical University of Munich (TUM) in Garching, Germany.
05.2016 – 03.2017	Master's Bioinformatics scholarship researcher at the Bioinformatics department in the Wageningen University, Wageningen, The Netherlands.
03.2015 – 05.2015	Software development project member at the Wageningen University, Wageningen, The Netherlands.

04.2014 – 11.2014	Bachelor's Lab scholarship researcher in a Physical Chemistry project at the Oviedo University, Oviedo, Spain.
01.2014 – 02.2014	Laboratory intern at Reny Picot's® microbiological and quality control laboratory, Navia, Spain.
09.2012 – 06.2013	Member in FEBiotec's project "FEBiotec Divulga", Oviedo, Spain.

Skills

Languages	Spanish (mother tongue) English (C1) French (B1) German (A2)
Programming languages	Python Matlab, R, SQL (advanced) Java, Awk, Javascript, Perl, HTML, C++, NoSQL (intermediate)
Soft skills	SPSS, SPQR, BLAST, Augustus, BiNGO, Galaxy (advanced) OS: Windows, Linux (advanced) Stata, SPAdes, GENSCAN (intermediate) OS: macOS (intermediate)

Publications and projects

- Doctoral Publication: Foreman, S.C.†; Llorián-Salvador, O.†; David, D.E.; Rösner, V.K.N.; Rischewski, J.F.; Feuerriegel, G.C.; Kramp, D.W.; Luiken, I.; Lohse, A.-K.; Kiefer, J.; et al. Development and Evaluation of MR-Based Radiogenomic Models to Differentiate Atypical Lipomatous Tumors from Lipomas. Cancers 2023, 15, 2150. https://doi.org/10.3390/cancers15072150.
- Doctoral Publication: Llorián-Salvador, Ó., Akhgar, J., Pigorsch, S. et al. The importance of planning CT-based imaging features for machine learning-based prediction of pain response. Sci Rep 13, 17427 (2023). https://doi.org/10.1038/s41598-023-43768-6.
- Doctoral Publication: Llorián-Salvador, Ó., Windeler, N., Martin, N. et al. CT-based radiomics for predicting breast cancer radiotherapy side effects. Sci Rep 14, 20051 (2024). https://doi.org/10.1038/s41598-024-70723-w
- Doctoral Publication: Llorián-Salvador, Ó.†; Rusche, D.†; Dürr, H.R.; Martin, N.; Luitjens, J.; Klein, A.; Gassert, F.; Gassert, Fl.; Weissinger, S.; Andrade-Navarro, M.A.; et al. Can Radiomics outperform Radiologists in Predicting Ewing Sarcoma Response to Neoadjuvant Chemotherapy from longitudinal MRI? (Manuscript submitted and pending approval).
- Publication: Oscar Llorian-Salvador; Michael Bernhofer; Yannick Mahlich; Burkhard Rost An Exhaustive Analysis of Single Amino Acid Variants in Helical Transmembrane Proteins. bioRxiv 2019, 2019.12.18.881318, doi:10.1101/2019.12.18.881318.
- Publication: Erdur, A.C.; Rusche, D.; Scholz, D.; Kiechle, J.; Fischer, S.; Llorián-Salvador, O.; Buchner, J.A.; Nguyen, M.Q.; Etzel, L.; Weidner, J.; et al. Deep learning for autosegmentation for radiotherapy treatment planning: State-of-the-art and novel perspectives. Strahlenther Onkol (2024). https://doi.org/10.1007/s00066-024-02262-2

- Publication: Jan Zaucha, Michael Heinzinger, A Kulandaisamy, Evans Kataka, Óscar Llorian Salvador, Petr Popov, Burkhard Rost, M Michael Gromiha, Boris S Zhorov, Dmitrij Frishman; Mutations in transmembrane proteins: diseases, evolutionary insights, prediction and comparison with globular proteins, Briefings in Bioinformatics, Volume 22, Issue 3, May 2021, bbaa132.
- Publication: Ayhan C. Erdur, Daniel Rusche, Daniel Scholz, Johannes Kiechle, Stefan Fischer, Óscar Llorián-Salvador, Josef A. Buchner, Mai Q. Nguyen, Lucas Etzel, Jonas Weidner, Marie-Christin Metz, Benedikt Wiestler, Julia Schnabel, Daniel Rueckert, Stephanie E. Combs; Deep Learning autosegmentation for radiotherapy treatment planning: State-of-the-art and novel perspectives. (Manuscript approved and waiting for publication).
- Doctoral project: Search for clinically relevant imaging biomarkers in a multicentred study of anal squamous cell carcinoma (ASCC) patients.
- Doctoral project: Pain response prediction to palliative radiation therapy for bone metastasis patients.
- Doctoral project: Prediction of side effects after radiotherapy for breast cancer patients using CT-based radiomics.
- Doctoral project: Differentiating atypical lipomatous tumours (ALTs) from lipomas using MRI-based radiomics.
- Doctoral project: Prediction of therapeutic outcome of Ewing-sarcomas using MRI-based radiomics.
- Research Assistant Project: Migration of a transmembrane protein topology predictor tool (TMSEG) from Java to Python (TMSEG).
- Master thesis project: Design and application of synthetic yeast promoters using ML tools.
- Bachelor thesis Project: A Machine Learning approach to gene expression prediction.
- Project: Recreation and thermometric analysis of an OCR (Oscillating Chemical Reaction).
- Project: Dungeons & Dragons game tool development for Desktop and Android platforms.

B. Acknowledgements

Many people have helped me during the years of this thesis, and I could fill a book with the virtues of each one of them. I feel fortunate to be surrounded by such great individuals, brilliant in their respective fields and as human beings.

I would like to thank my doctoral supervisors. Miguel Andrade has been a diligent guide, offering his knowledge and time, counseling beyond what he was required to do. I would also like to thank Jan Peeken, for offering me this opportunity and showing me how rewarding this field can be, and how to help people with my work, even if indirectly. I am also deeply grateful for the guidance offered by Burkhard Rost. One walk at a time, he has taught me how to grow as a scientist and a professional, with advice and experience that goes beyond fields of expertise. Any input from either of the three is priceless in its own way, and I will honor them by not forgetting a single lesson.

To all my colleagues: thank you. All of you have helped me increase my knowledge, share the burden of a doctoral thesis, and advance through the toughest of all: German bureaucracy. I am especially grateful to those in the Rostlab, many of whom I am fortunate to also call my close friends. You are good people that I know I will keep for life, regardless of where we are in the world, sharing stories and tabletop games.

I would like to thank some miscellaneous people in my life who, in one way or another, have helped me be the person I am today. To Quique Martín Yáñez, for his passion in science mad of him an excellent teacher, and it was contagious enough to decide to pursue science and teaching myself. To Esther Méndez, who has brought evolution and stability in a time of rut and chaos. To Tim Karl and Inga Weisse, unsung heroes of the Rostlab for a long time, without whom not just my work, but that of many others wouldn't have come to fruition.

To my closest friends. No matter how much time it passes, or where each one of us are: I know our friendship is imperishable, and you have sticked with me through thick and thin many times. You know me inside and out, and I can only hope you consider me half as good a friend as you all are to me. And to all other friends: you know who you are, we share great moments even online. You help me a great deal just by being there.

To Sandra, my soul mate. No words are needed between us, and nobody knows me like you do. Anything I could write here would fall short of expressing all I want to say and how grateful I am. Fortunately, we have all the time in the world for each other.

Last but not least to my family, who have raised me and have also helped me be the person I am today, and taught me how to find honorable principles to live life by. Especial thanks to my sister María, who has helped me walking through life by my side, knowing our problems and solutions back-to-back. You perform such a titanic task seamlessly, and I will never thank enough for that.

Finally, one last thought: Journey before destination. The lessons, experiences and growth along the way are what truly shape us, and are what truly matters.

C. List of publications

The results presented in this thesis have been published previously or submitted and pending approval. Each publication corresponds, in order, to the four results sections.

1. Foreman, S.C. †; Llorián-Salvador, O. †; David, D.E.; Rösner, V.K.N.; Rischewski, J.F.; Feuerriegel, G.C.; Kramp, D.W.; Luiken, I.; Lohse, A.-K.; Kiefer, J.; et al. Development and Evaluation of MR-Based Radiogenomic Models to Differentiate Atypical Lipomatous Tumors from Lipomas. Cancers 2023, 15, 2150, doi:10.3390/cancers15072150.

2. Llorián-Salvador, Ó.; Akhgar, J.; Pigorsch, S.; Borm, K.; Münch, S.; Bernhardt, D.; Rost, B.; Andrade-Navarro, M.A.; Combs, S.E.; Peeken, J.C. The Importance of Planning CT-Based Imaging Features for Machine Learning-Based Prediction of Pain Response. *Sci Rep* **2023**, *13*, 17427, doi:10.1038/s41598-023-43768-6.

3. Llorián-Salvador, Ó.; Windeler, N.; Martin, N.; Etzel, L.; Andrade-Navarro, M.A.; Bernhardt, D.; Rost, B.; Borm, K.J.; Combs, S.E.; Duma, M.N.; et al. CT-Based Radiomics for Predicting Breast Cancer Radiotherapy Side Effects. Sci Rep 2024, 14, 20051, doi:10.1038/s41598-024-70723-w.

4. Llorián-Salvador, Ó.†; Rusche, D.†; Dürr, H.R.; Martin, N.; Luitjens, J.; Klein, A.; Gassert, F.; Gassert, Fl.; Weissinger, S.; Andrade-Navarro, M.A.; et al. Can Radiomics outperform Radiologists in Predicting Ewing Sarcoma Response to Neoadjuvant Chemotherapy from longitudinal MRI? (Manuscript submitted and pending approval).

† First authors contributed equally to this work.

D. Statutory declaration

I hereby declare that I wrote the contents of this dissertation with the topic

"**Development and application of AI tools to improve diagnostic and prognostic capabilities of medical imaging data**"

Independently and that I have used no other aids than those cited. In each individual case, I have clearly identified the source of the passages that are taken word for word or paraphrased from other works. Other contributions to this work in terms of collaboration and co-authors are clearly indicated and acknowledged in the "Author contribution" section from each respective study.

I hereby declare that this thesis has not been presented for any other academic degree at another institution.

I hereby declare that I have not failed any other PhD, doctoral, or equivalent graduation procedure in the same subjects as this thesis.

I hereby declare that I have followed good academic practices and that I did not receive paid help from third parties for the dissertation

---

Óscar Llorián-Salvador

E. Appendix

E.1 Supplemental material for chapter 4.1

Supplementary Material Table 1: Magnetic Resonance imaging sequence parameters

Sequence	T_1		T_2		T_1-FS-GD	
Field strength (T)	1.5	3.0	1.5	3.0	1.5	3.0
Repetition time (ms)	537-557	855-1100	4020-6110	4260-5600	500-595	722-1050
Echo time (ms)	15-16	12	85-104	78-94	13-16	12-13
Flip angle (°)	90-172	175-180	180	180	90-180	180-136
In-plane resolution (mm)	0.2-0.5x0.2-0.5	0.5-0.8x0.4-0.7	0.4-0.6x0.4-0.6	0.5-0.7x0.5-0.7	0.3-0.6x0.3-0.6	0.6-0.7x0.6-0.7
Slice thickness (mm)	3-4	3-5	3-5	4-5	3-5	4-5
Gap (%)	10-25	10-50	20-60	10-40	20-60	10-40
Bandwidth (Hz/pixel)	85-128	160-172	109-119	200-203	85-130	150-160
Echo train length (n)	118	162-262	14-58	15-37	118	116-288

GD: gadolinium-enhanced; FS: fat-saturated
Field of view was limited to the area of interest according to the respective tumor volume to maximize anatomical detail
Sequence parameters were adjusted in accordance with the optimal protocol of the anatomical tumor region

Supplementary Material Table 2: Extracted radiomics features (n=104)

All extracted features were computed according the “image biomarker standardization initiative” (IBSI) guidelines [1].

	Shape Features
1	Mesh Volume
2	Voxel Volume
3	Surface Area
4	Surface Volume Ratio
5	Sphericity
6	Maximum 3D Diameter
7	Maximum 2D Diameter Slice
8	Maximum 2D Diameter Column
9	Maximum 2D Diameter Row
10	Major Axis
11	Minor Axis
12	Least Axis
13	Elongation
14	Flatness
	First Order Features
1	Energy
2	Intensity Histogram Entropy
3	Minimum
4	10th Percentile
5	90th Percentile
6	Maximum
7	Mean
8	Median
9	Interquartile Range
10	Range
11	Mean Absolute Deviation (MAD)
12	Root Mean Squared (RMS)
13	Skewness
14	Excess Kurtosis
15	Variance
16	Intensity Histogram Uniformity
	Gray Level Co-occurrence Matrix (GLCM) Features
1	Autocorrelation
2	Joint Average
3	Cluster Prominence
4	Cluster Shade
5	Cluster Tendency
6	Contrast
7	Correlation
8	Difference Average
9	Difference Entropy
10	Difference Variance

11	Joint Energy (IBSI: Angular Second Moment)
12	Joint Entropy
13	Informal Measure of Correlation (IMC) 1
14	Informal Measure of Correlation (IMC) 2
15	Inverse Difference Moment (IDM)
16	Inverse Difference Moment Normalized (IDMN)
17	Inverse Difference (ID)
18	Inverse Difference Normalized (IDN)
19	Inverse Variance
20	Maximum Probability (IBSI: Joint maximum)
21	Sum Entropy
22	Sum of Squares (IBSI: Sum of Squares)
23	Maximal Correlation Coefficient (MCC)
	Gray Level Size Zone Matrix (GLSZM) Features
1	Small Area Emphasis (SAE)
2	Large Area Emphasis (LAE)
3	Gray Level Non-Uniformity (GLN)
4	Gray Level Non-Uniformity Normalized (GLNN)
5	Size-Zone Non-Uniformity (SZN)
6	Size-Zone Non-Uniformity Normalized (SZNN)
7	Zone Percentage (ZP)
8	Gray Level Variance (GLV)
9	Zone Variance (ZV)
10	Zone Entropy (ZE)
11	Low Gray Level Zone Emphasis (LGLZE)
12	High Gray Level Zone Emphasis (HGLZE)
13	Small Area Low Gray Level Emphasis (SALGLE)
14	Small Area High Gray Level Emphasis (SAHGLE)
15	Large Area Low Gray Level Emphasis (LALGLE)
16	Large Area High Gray Level Emphasis (LAHGLE)
	Gray Level Run Length Matrix (GLRLM) Features
1	Short Run Emphasis (SRE)
2	Long Run Emphasis (LRE)
3	Gray Level Non-Uniformity (GLN)
4	Gray Level Non-Uniformity Normalized (GLNN)
5	Run Length Non-Uniformity (RLN)
6	Run Length Non-Uniformity Normalized (RLNN)
7	Run Percentage (RP)
8	Gray Level Variance (GLV)
9	Run Variance (RV)
10	Run Entropy (RE)
11	Low Gray Level Run Emphasis (LGLRE)
12	High Gray Level Run Emphasis (HGLRE)
13	Short Run Low Gray Level Emphasis (SRLGLE)
14	Short Run High Gray Level Emphasis (SRHGLE)
15	Long Run Low Gray Level Emphasis (LRLGLE)
16	Long Run High Gray Level Emphasis (LRHGLE)
	Neighbouring Gray Tone Difference Matrix (NGTDM) Features

1	Coarseness
2	Contrast
3	Busyness
4	Complexity
5	Strength
	Gray Level Dependence Matrix (GLDM) Features
1	Small Dependence Emphasis (SDE)
2	Large Dependence Emphasis (LDE)
3	Gray Level Non-Uniformity (GLN)
4	Dependence Non-Uniformity (DN)
5	Dependence Non-Uniformity Normalized (DNN)
6	Gray Level Variance (GLV)
7	Dependence Variance (DV)
8	Dependence Entropy (DE)
9	Low Gray Level Emphasis (LGLE)
10	High Gray Level Emphasis (HGLE)
11	Small Dependence Low Gray Level Emphasis (SDLGLE)
12	Small Dependence High Gray Level Emphasis (SDHGLE)
13	Large Dependence Low Gray Level Emphasis (LDLGLE)
14	Large Dependence High Gray Level Emphasis (LDHGLE)

[1] 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 2020:191145.

Supplementary Material Table 3: Performance the machine learning models on the external test set of each individual sequence T1w, T2w, and T1fsgd using the following model architectures: least absolute shrinkage and selection operator (LASSO), support vector machine (SVM), random forest classifier (RFC), and an artificial neural network (ANN). The external performance represents the values yielded when a final cross-validation step considering only the best 150 best hyperparameter sets was implemented.

Model Architecture	Score	T1w	T2w	T1fsgd
LASSO	**AUC***	0.83 (0.82-0.84) ± 0.02	0.82 (0.81-0.83) ± 0.04	0.84 (0.83-0.85) ± 0.03
	Accuracy	0.58	0.69	0.60
	Sensitivity	0.80	0.42	0.06
	Specificity	0.43	0.83	1.00
SVM	**AUC***	0.78 (0.76-0.80) ± 0.07	0.81 (0.78-0.84) ± 0.08	0.78 (0.74-0.82) ± 0.12
	Accuracy	0.64	0.63	0.65
	Sensitivity	0.30	0.17	0.24
	Specificity	0.89	0.88	0.96
RFC	**AUC***	0.80 (0.79-0.81) ± 0.02	0.80 (0.79-0.81) ± 0.04	0.82 (0.81-0.83) ± 0.04
	Accuracy	0.70	0.66	0.63
	Sensitivity	0.35	0.33	0.18
	Specificity	0.96	0.83	0.96
ANN	**AUC***	0.77 (0.75-0.79) ± 0.08	0.79 (0.76-0.82) ± 0.08	0.79 (0.77-0.81) ± 0.08
	Accuracy	0.71	0.58	0.70
	Sensitivity	0.45	0.08	0.29
	Specificity	0.89	0.83	1.00

* Data is given as mean (95% confidence interval) ± standard deviation

Supplementary Material Table 4: Internal performance representing the averaged values over 150 models resulting from the nested cross-validation using demographic information, radiomic features of each individual sequence T1w, T2w, and T1fsgd or radiomic features of all sequences combined, as well as combining radiomic features (of all sequences) and demographic information for the following model architectures: least absolute shrinkage and selection operator (LASSO), support vector machine (SVM), random forest classifier (RFC), and an artificial neural network (ANN). The metrics are given as mean ± standard deviation.

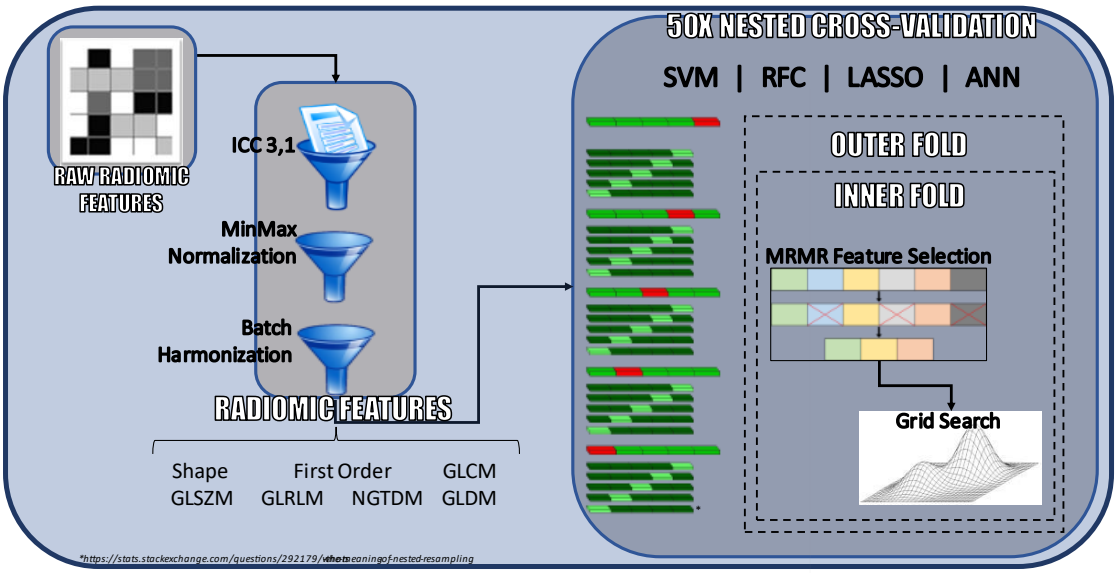
Model Architecture	Score	Demographic Features	T1w	T2w	T1fsgd	Combined Sequences	Combined Sequences + Demographic Features
LASSO	**AUC***	0.56 (0.55-0.57) ± 0.06	0.83 (0.83-0.83) ± 0.04	0.80 (0.79-0.81) ± 0.05	0.82 (0.81-0.83) ± 0.05	0.88 (0.87-0.89) ± 0.05	0.88 (0.87-0.89) ± 0.05
	Accuracy*	0.67 (0.65-0.69) ± 0.15	0.79 (0.79-0.79) ± 0.04	0.78 (0.77-0.79) ± 0.05	0.79 (0.78-0.80) ± 0.04	0.85 (0.84-0.86) ± 0.04	0.85 (0.84-0.86) ± 0.05
	Sensitivity	0.14	0.41	0.34	0.34	0.51	0.49
	Specificity	0.85	0.92	0.91	0.92	0.93	0.94
SVM	**AUC***	0.56 (0.54-0.58) ± 0.12	0.80 (0.79-0.81) ± 0.09	0.78 (0.77-0.79) ± 0.08	0.75 (0.73-0.77) ± 0.13	0.84 (0.83-0.85) ± 0.09	0.85 (0.84-0.86) ± 0.08
	Accuracy*	0.73 (0.73-0.73) ± 0.03	0.78 (0.78-0.78) ± 0.04	0.79 (0.78-0.80) ± 0.05	0.79 (0.78-0.80) ± 0.05	0.84 (0.83-0.85) ± 0.04	0.84 (0.83-0.85) ± 0.05
	Sensitivity	0.04	0.38	0.37	0.34	0.50	0.45
	Specificity	0.97	0.93	0.92	0.91	0.93	0.93
RFC	**AUC***	0.63 (0.62-0.64) ± 0.06	0.85 (0.84-0.86) ± 0.05	0.79 (0.78-0.80) ± 0.05	0.79 (0.78-0.80) ± 0.06	0.86 (0.85-0.87) ± 0.05	0.87 (0.86-0.88) ± 0.05
	Accuracy*	0.69 (0.69-0.69) ± 0.04	0.81 (0.81-0.81) ± 0.03	0.80 (0.80-0.80) ± 0.03	0.80 (0.80-0.80) ± 0.03	0.86 (0.86-0.86) ± 0.03	0.86 (0.86-0.86) ± 0.03
	Sensitivity	0.23	0.50	0.39	0.36	0.49	0.51
	Specificity	0.85	0.92	0.92	0.92	0.95	0.95
ANN	**AUC***	0.68 (0.67-0.69) ± 0.08	0.78 (0.77-0.79) ± 0.08	0.76 (0.75-0.77) ± 0.09	0.77 (0.76-0.78) ± 0.08	0.83 (0.82-0.84) ± 0.08	0.79 (0.78-0.80) ± 0.09
	Accuracy*	0.73 (0.73-0.73) ± 0.03	0.78 (0.77-0.79) ± 0.05	0.77 (0.76-0.78) ± 0.05	0.78 (0.77-0.79) ± 0.05	0.83 (0.82-0.84) ± 0.04	0.84 (0.83-0.85) ± 0.04
	Sensitivity	0.07	0.49	0.45	0.41	0.53	0.44
	Specificity	0.95	0.89	0.86	0.88	0.90	0.94

* Data is given as mean (95% confidence interval) ± standard deviation

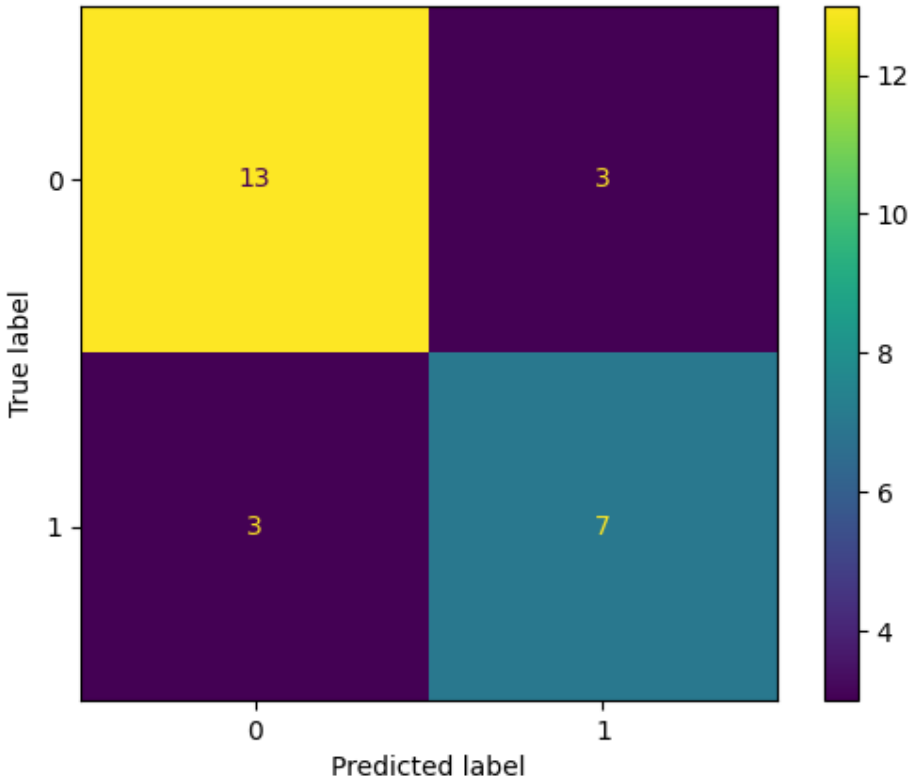
Supplementary Material Table 5: Feature Importance of the best performing model (least absolute shrinkage and selection operator (LASSO) trained on features from all radiomic sequences).

Feature Name	Score
T1fs_original_glszm_ZoneEntropy	22,30362659
T1_original_shape_Elongation	21,75575404
T2_original_glszm_ZoneEntropy	19,1399285
T1fs_original_shape_Flatness	14,13490208

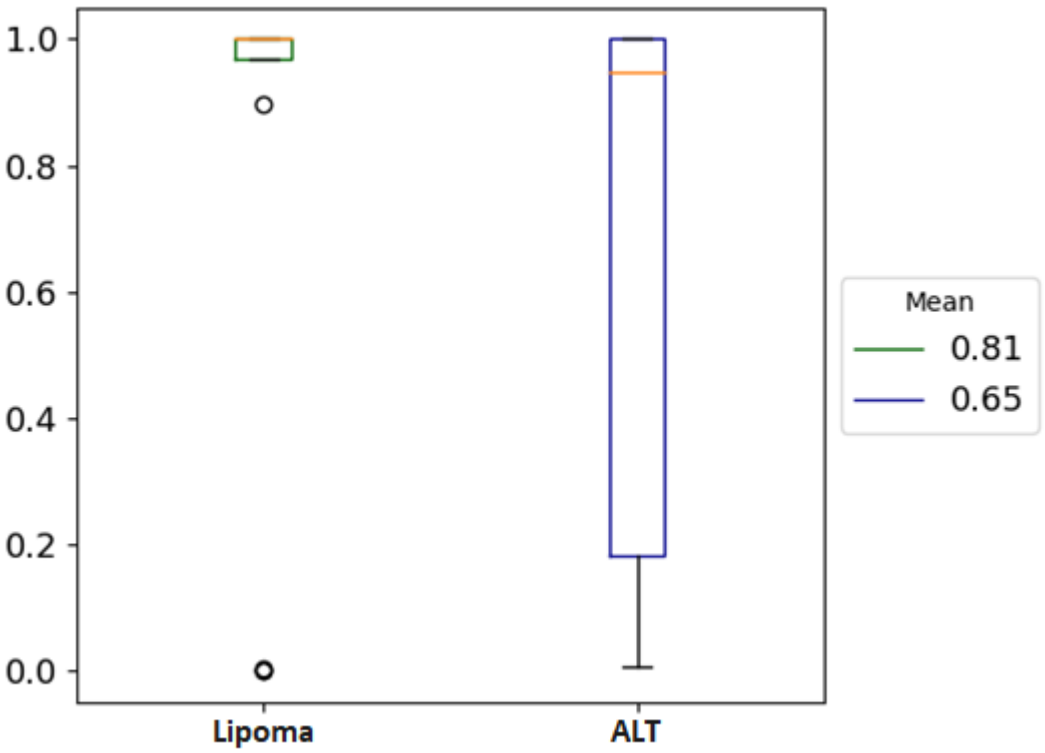
T1_original_shape_Maximum2DDiameterRow	14,06067563
T2_original_shape_Maximum2DDiameterRow	12,80611959
T1_original_ngtdm_Busyness	10,64687961
T2_original_shape_Maximum2DDiameterColumn	10,5844085
T1fs_original_firstorder_InterquartileRange	9,39608846
T2_original_shape_MajorAxisLength	9,064467759
T2_original_shape_Maximum3DDiameter	8,953856721
T1fs_original_firstorder_Energy	7,286263865
T1_original_shape_MajorAxisLength	6,747650604
T2_original_ngtdm_Complexity	2,597293753
T1_original_shape_Maximum2DDiameterSlice	2,586796401
T1fs_original_glcm_DifferenceEntropy	1,213683594
T1_original_shape_Maximum3DDiameter	1,118527794
T1fs_original_glcm_JointEntropy	0,253434882
T1_original_glszm_LargeAreaLowGrayLevelEmphasis	0,195501056
T1fs_original_glcm_JointAverage	0,11638087



Supplementary Material Figure 1: Flow chart of the statistical analysis of the extracted radiomic features.



Supplementary Material Figure 2: Confusion matrix of the best performing model, a least absolute shrinkage and selection operator (LASSO) trained on all radiomic sequences. Misclassification rate: 0.23 ((FN + FP) / (N + P))



Supplementary Material Figure 3: Boxplot of the prediction probabilities made by the best performing model (least absolute shrinkage and selection operator (LASSO) trained on features from all radiomic sequences). The probability cut-off used was 0.5.

E.2 Supplemental material for chapter 4.2

Supplemental Material

Supplemental Tables

TABLE S1
Histology distribution

Relative proportion of all histology types ranked by their frequency.

Histology	Proportion of patients
Mammary carcinoma	25 (28 %)
Prostate carcinoma	17 (19 %)
Non-Small Cell Lung Cancer	15 (16 %)
Urothelium cancer	6 (7 %)
Pancreatic carcinoma	3 (3 %)
Renal cell carcinoma	9 (10 %)
Cancer of unknown primary	1 (1 %)
Rectal carcinoma	1 (1 %)
Cholangiocellular carcinoma	1 (1 %)
Parotic carcinoma	1 (1 %)
Esophageal carcinoma	2 (2 %)
Thymoma	1 (1 %)
Hepatocellular carcinoma	1 (1 %)
Gliosarcoma	1 (1 %)
Small Cell Lung Cancer	1 (1 %)
Gastric cancer	2 (2 %)
Thyroid cancer	1 (1 %)
Adrenal gland carcinoma	1 (1 %)
Hypopharyngeal carcinoma	1 (1 %)

TABLE S2
CT acquisition Parameters

Parameter	
Scanner type	Somatom Emotion 16 Siemens (Erlangen, Germany)
Axial scan dimensions	512 x 512 pixel
Pixel Spacing	0.98 mm x 0.98 mm
Slice thickness	3 mm
Kernel	B31s
Voltage	130 KVP
X ray tube Current	398.5 (94-650)

Abbreviations: CT: computed tomography

TABLE S3
Extracted radiomics features

All extracted features were computed according to the “image biomarker standardization initiative” (IBSI) guidelines [1]. The pyRadiomics package (version 2.0) implemented in python (version 3.6.4) was used for feature extraction [2]. For pre-processing, a fixed bin width of 20 was used for image discretization. Isotropic resampling was performed to a voxel size of 1x1x1 mm using Bspline interpolation.

	Shape Features
1.)	Volume
2.)	Surface Area
3.)	Surface Volume Area
4.)	Sphericity
5.)	Spherical Disproportion
6.)	Maximum 3D Diameter
7.)	Maximum 2D Diameter Slice
8.)	Maximum 2D Diameter Column
9.)	Maximum 2D Diameter Row
10.)	Major Axis
11.)	Minor Axis
12.)	Least Axis
13.)	Elongation
14.)	Flatness
	First Order Features
1.)	Energy
2.)	Intensity Histogram Entropy
3.)	Minimum
4.)	10th Percentile
5.)	90th Percentile
6.)	Maximum
7.)	Mean
8.)	Median
9.)	Interquartile Range

10.)	Range
11.)	Mean Absolute Deviation (MAD)
12.)	Robust Mean Absolute Deviation (rMAD)
13.)	Root Mean Squared (RMS)
14.)	Skewness
15.)	Excess Kurtosis
16.)	Variance
17.)	Intensity Histogram Uniformity
	Local Binary pattern (LBP) Features-m1
1.)	Energy
2.)	Intensity Histogram Entropy
3.)	Minimum
4.)	10th Percentile
5.)	90th Percentile
6.)	Maximum
7.)	Mean
8.)	Median
9.)	Interquartile Range
10.)	Range
11.)	Mean Absolute Deviation (MAD)
12.)	Robust Mean Absolute Deviation (rMAD)
13.)	Root Mean Squared (RMS)
14.)	Skewness
15.)	Excess Kurtosis
16.)	Variance
17.)	Intensity Histogram Uniformity
	Local Binary pattern (LBP) Features-m2
1.)	Energy
2.)	Intensity Histogram Entropy
3.)	Minimum

4.)	10th Percentile
5.)	90th Percentile
6.)	Maximum
7.)	Mean
8.)	Median
9.)	Interquartile Range
10.)	Range
11.)	Mean Absolute Deviation (MAD)
12.)	Robust Mean Absolute Deviation (rMAD)
13.)	Root Mean Squared (RMS)
14.)	Skewness
15.)	Excess Kurtosis
16.)	Variance
17.)	Intensity Histogram Uniformity
	Local Binary pattern (LBP) Features-kurtosis
1.)	Energy
2.)	Intensity Histogram Entropy
3.)	Minimum
4.)	10th Percentile
5.)	90th Percentile
6.)	Maximum
7.)	Mean
8.)	Median
9.)	Interquartile Range
10.)	Range
11.)	Mean Absolute Deviation (MAD)
12.)	Robust Mean Absolute Deviation (rMAD)
13.)	Root Mean Squared (RMS)
14.)	Skewness
15.)	Excess Kurtosis

16.)	Variance
17.)	Intensity Histogram Uniformity
	Gray Level Co-occurrence Matrix (GLCM) Features
1.)	Autocorrelation
2.)	Joint Average
3.)	Cluster Prominence
4.)	Cluster Shade
5.)	Cluster Tendency
6.)	Contrast
7.)	Correlation
8.)	Difference Average
9.)	Difference Entropy
10.)	Difference Variance
11.)	Joint Energy (IBSI: Angular Second Moment)
12.)	Joint Entropy
13.)	Informal Measure of Correlation (IMC) 1
14.)	Informal Measure of Correlation (IMC) 2
15.)	Inverse Difference Moment (IDM)
16.)	Inverse Difference Moment Normalized (IDMN)
17.)	Inverse Difference (ID)
18.)	Inverse Difference Normalized (IDN)
19.)	Inverse Variance
20.)	Maximum Probability (IBSI: Joint maximum)
21.)	Sum Entropy
22.)	Sum of Squares (IBSI: Sum of Squares)
23.)	Maximal Correlation Coefficient (MCC)
	Gray Level Size Zone Matrix (GLSZM) Features
1.)	Small Area Emphasis (SAE)
2.)	Large Area Emphasis (LAE)
3.)	Gray Level Non-Uniformity (GLN)

4.)	Gray Level Non-Uniformity Normalized (GLNN)
5.)	Size-Zone Non-Uniformity (SZN)
6.)	Size-Zone Non-Uniformity Normalized (SZNN)
7.)	Zone Percentage (ZP)
8.)	Gray Level Variance (GLV)
9.)	Zone Variance (ZV)
10.)	Zone Entropy (ZE)
11.)	Low Gray Level Zone Emphasis (LGLZE)
12.)	High Gray Level Zone Emphasis (HGLZE)
13.)	Small Area Low Gray Level Emphasis (SALGLE)
14.)	Small Area High Gray Level Emphasis (SAHGLE)
15.)	Large Area Low Gray Level Emphasis (LALGLE)
16.)	Large Area High Gray Level Emphasis (LAHGLE)
	Gray Level Run Length Matrix (GLRLM) Features
1.)	Short Run Emphasis (SRE)
2.)	Long Run Emphasis (LRE)
3.)	Gray Level Non-Uniformity (GLN)
4.)	Gray Level Non-Uniformity Normalized (GLNN)
5.)	Run Length Non-Uniformity (RLN)
6.)	Run Length Non-Uniformity Normalized (RLNN)
7.)	Run Percentage (RP)
8.)	Gray Level Variance (GLV)
9.)	Run Variance (RV)
10.)	Run Entropy (RE)
11.)	Low Gray Level Run Emphasis (LGLRE)
12.)	High Gray Level Run Emphasis (HGLRE)
13.)	Short Run Low Gray Level Emphasis (SRLGLE)
14.)	Short Run High Gray Level Emphasis (SRHGLE)
15.)	Long Run Low Gray Level Emphasis (LRLGLE)
16.)	Long Run High Gray Level Emphasis (LRHGLE)

	Neighbouring Gray Tone Difference Matrix (NGTDM) Features
1.)	Coarseness
2.)	Contrast
3.)	Busyness
4.)	Complexity
5.)	Strength
	Gray Level Dependence Matrix (GLDM) Features
1.)	Small Dependence Emphasis (SDE)
2.)	Large Dependence Emphasis (LDE)
3.)	Gray Level Non-Uniformity (GLN)
4.)	Dependence Non-Uniformity (DN)
5.)	Dependence Non-Uniformity Normalized (DNN)
6.)	Gray Level Variance (GLV)
7.)	Dependence Variance (DV)
8.)	Dependence Entropy (DE)
9.)	Low Gray Level Emphasis (LGLE)
10.)	High Gray Level Emphasis (HGLE)
11.)	Small Dependence Low Gray Level Emphasis (SDLGLE)
12.)	Small Dependence High Gray Level Emphasis (SDHGLE)
13.)	Large Dependence Low Gray Level Emphasis (LDLGLE)
14.)	Large Dependence High Gray Level Emphasis (LDHGLE)

TABLE S4
Extension of Table 3

AUROC, BA, F1 Score and MCC for the Support Vector Machine (SVM) and Random Forest Classifier (RFC) models trained on both segmentation modes, Gross Tumour Volume (GTV) and Clinical Target Volume (CTV).

Model	Segmentation	AUC	BA	F1	MCC
SVM	GTV	0.58 ± 0.01	0.54 ± 0.02	0.33 ± 0.03	0.08 ± 0.04
	CTV	0.61 ± 0.01	0.57 ± 0.02	0.36 ± 0.03	0.13 ± 0.04
RFC	GTV	0.55 ± 0.01	0.52 ± 0.02	0.29 ± 0.03	0.05 ± 0.04
	CTV	0.62 ± 0.01	0.58 ± 0.02	0.37 ± 0.03	0.15 ± 0.04

* Data is given as mean ± 1.96 standard errors for a 95% confidence interval.

TABLE S5
Extension of Table 4

AUROC, BA, F1 Score and MCC for the SVM, RFC and Logistic Regression (LR) models trained on semantic, clinical, and SINS features.

Model	Data	AUROC	BA	F1	MCC
SVM	CTV	0.61 ± 0.01	0.57 ± 0.02	0.36 ± 0.03	0.13 ± 0.04
RFC		0.62 ± 0.01	0.58 ± 0.02	0.37 ± 0.03	0.15 ± 0.04
SVM	Semantic	0.61 ± 0.01	0.57 ± 0.02	0.38 ± 0.03	0.13 ± 0.04
RFC		0.63 ± 0.01	0.58 ± 0.02	0.39 ± 0.03	0.16 ± 0.04
SVM	Clinical	0.80 ± 0.01	0.72 ± 0.03	0.56 ± 0.05	0.43 ± 0.06
RFC		0.79 ± 0.01	0.73 ± 0.03	0.58 ± 0.05	0.44 ± 0.06
LR	SINS	0.65 ± 0.01	0.58 ± 0.03	0.36 ± 0.05	0.16 ± 0.06
	SINS (binary)	0.54 ± 0.01	0.52 ± 0.03	0.19 ± 0.05	0.04 ± 0.06

* Data is given as mean ± 1.96 standard errors for a 95% confidence interval.

TABLE S6
Extension of Table 5

AUROC, BA, F1 Score and MCC for the SVM and RFC models trained on the different combined datasets.

Model	Data	AUROC	BA	F1	MCC
SVM	CTV + SINS	0.61 ± 0.01	0.57 ± 0.02	0.36 ± 0.04	0.13 ± 0.04
	CTV + Clinical	0.75 ± 0.01	0.69 ± 0.02	0.52 ± 0.03	0.35 ± 0.04
	Semantic + SINS	0.62 ± 0.01	0.58 ± 0.02	0.39 ± 0.03	0.15 ± 0.04
	Semantic + Clinical	0.68 ± 0.01	0.63 ± 0.02	0.45 ± 0.03	0.24 ± 0.04
	CTV + SINS + Clinical	0.74 ± 0.01	0.68 ± 0.02	0.50 ± 0.03	0.33 ± 0.05
	CTV + SINS + Clinical + Semantic	0.67 ± 0.01	0.62 ± 0.02	0.44 ± 0.03	0.22 ± 0.04
RFC	CTV + SINS	0.60 ± 0.01	0.57 ± 0.02	0.36 ± 0.04	0.13 ± 0.04
	CTV + Clinical	0.68 ± 0.01	0.61 ± 0.02	0.42 ± 0.03	0.21 ± 0.04
	Semantic + SINS	0.65 ± 0.01	0.59 ± 0.02	0.40 ± 0.03	0.17 ± 0.04
	Semantic + Clinical	0.72 ± 0.01	0.64 ± 0.02	0.48 ± 0.03	0.27 ± 0.04
	CTV + SINS + Clinical	0.67 ± 0.01	0.61 ± 0.02	0.42 ± 0.03	0.21 ± 0.04

	CTV + SINS + Clinical + Semantic	0.67 ± 0.01	0.61 ± 0.02	0.44 ± 0.03	0.20 ± 0.04

* Data is given as mean ± 1.96 standard errors for a 95% confidence interval.

TABLE S7

AUROC, BA, F1 Score and MCC for the SVM and RFC models trained on clinical and SINS features.

Model	Data	AUROC	BA	F1	MCC
SVM	Clinical + SINS	0.73 ± 0.01	0.68 ± 0.03	0.52 ± 0.04	0.32 ± 0.05
RFC		0.75 ± 0.01	0.68 ± 0.03	0.53 ± 0.04	0.35 ± 0.05

* Data is given as mean ± 1.96 standard errors for a 95% confidence interval.

TABLE S8

Feature importance table for SVM and RFC models trained on CTV features.

Importance in RFC is measured by the mean decrease in impurity. Feature importance in SVM is not possible to estimate given that most kernels were non-linear. Score is calculated by multiplying the importance of a feature by their frequency.

SVM Feature Name	% Chosen		RFC Feature Name	% Chosen	Importance	Score
GLSZM - LAHGLE	94.8		GLCM - DE	94.4	0.071	0.067
GLCM - DE	92		GLCM - Cluster Shade	90.8	0.0699	0.0635
GLCM - Cluster Shade	90		GLSZM - LAHGLE	96.8	0.0648	0.0627
GLRLM - SRE	81.6		Shape - Maximum 2D Diameter Row	78.8	0.0751	0.0592
Shape - Maximum 2D Diameter Row	76.8		GLRLM - SRE	87.2	0.0626	0.0546
GLRLM - LRE	63.6		GLRLM - LRE	64.8	0.0669	0.0434
Shape - Elongation	54.4		GLCM - IDMN	54	0.0675	0.0365
GLSZM - SZNN	54		Shape - Elongation	51.2	0.0704	0.036
GLCM - IDMN	53.6		GLRLM - RLNN	52.8	0.0632	0.0334
GLRLM - RLNN	51.6		GLSZM - SZNN	52.4	0.0633	0.0332
GLSZM - SAE	45.6		Shape - Surface Volume Ratio	39.2	0.0713	0.0279
GLDM - LDLGLE	40.8		GLSZM - SAE	44	0.059	0.026
GLDM - LDHGLE	37.2		GLDM - LDLGLE	42	0.0616	0.0259
GLSZM - GLNN	35.2		GLDM - LDE	35.6	0.0697	0.0248
GLCM - Cluster Prominence	34.4		GLCM - Cluster Prominence	35.6	0.0692	0.0246

TABLE S9

Feature importance table for SVM and RFC models trained on clinical features.

SVM Feature Name	% Chosen		RFC Feature Name	% Chosen	Importance	Score
Age	100		Tumour Type: Others	100	0.2229	0.2229
KPS	100		Tumour Type: Breast Cancer	100	0.2218	0.2218
Opiate Medication	100		Tumour Type: NSCLC	100	0.1969	0.1969

Tumour Type: Breast Cancer	100		Opiate Medication	100	0.1508	0.1508
Tumour Type NSCLC	100		KPS	100	0.1235	0.1235
Tumour Type: Others	100		Age	100	0.0841	0.0841

TABLE S10
Feature importance table for SVM and RFC models trained on semantic features.

SVM Feature Name	% Chosen		RFC Feature Name	% Chosen	Importance	Score
Vertebral body collapse: <50% collapse	100		GTV - Classification: Body + bilateral pedicle/transverse processBody	98.4	0.0775	0.0763
Posterolateral involvement of the spinal elements: None of the above	100		Imaging - Bone reaction: Lytic	98	0.0743	0.0728
Location: Mobile	99.2		Posterolateral involvement of the spinal elements: Bilateral	94.4	0.0753	0.0711
Vertebral body collapse: >50% collapse	98.8		GTV - Classification: Body + bilateral pedicle/transverse	93.2	0.0729	0.0679
Imaging - Bone reaction: Lytic	97.6		Soft tissue component: Yes	84.4	0.0778	0.0656
GTV - Classification: Body + unilateral pedicle	97.6		Vertebral body collapse: >50% collapse	99.2	0.0653	0.0647
Posterolateral involvement of the spinal elements: Bilateral	96		Posterolateral involvement of the spinal elements: None of the above	100	0.0644	0.0644
GTV - Classification: Body + bilateral pedicle/transverse processBody	94.4		Soft tissue component: No	82.8	0.0774	0.0641
Location: Semirigid	88.4		Imaging - Bone reaction: Mixed	49.2	0.1105	0.0544
Soft tissue component: Yes	86		Vertebral body collapse: <50% collapse	100	0.0536	0.0536
Soft tissue component: No	84.8		Posterolateral involvement of the spinal elements: Unilateral	71.2	0.0708	0.0504
Posterolateral involvement of the spinal elements: Unilateral	64		Imaging - Bone reaction: Blastic	39.2	0.1115	0.0437
Vertebral body collapse: No collapse with >50% body involved	52.8		Location: Mobile	98	0.0426	0.0418
GTV - Classification: Unilateral pedicle	51.6		GTV - Classification: Unilateral pedicle	47.6	0.0747	0.0356

Vertebral body collapse: None of the above	48		Location: Semirigid	88.4	0.0397	0.0351

TABLE S11
Feature importance table for SVM and RFC models trained on CTV, clinical, SINS and semantic features.

SVM Feature Name	% Chosen		RFC Feature Name	% Chosen	Importance	Score
Tumour Type: Breast Cancer	100		Posterolateral involvement of the spinal elements: None of the above	94	0.0714	0.0671
Tumour Type: Others	93.6		Tumour Type: Breast Cancer	99.2	0.0598	0.0594
Posterolateral involvement of the spinal elements: None of the above	90.8		GLSZM - LAHGLE	89.6	0.0653	0.0585
GLSZM - LAHGLE	88.4		Tumour Type: Others	93.2	0.061	0.0569
GTV - Classification: Unilateral pedicle	84		Vertebral body collapse: <50% collapse	79.2	0.0716	0.0567
Vertebral body collapse: <50% collapse	78.4		GLCM - Cluster Shade	78	0.0721	0.0562
GLCM - Cluster Shade	76		GTV - Classification: Unilateral pedicle	83.2	0.0627	0.0521
Location: Mobile	70.4		Location: Mobile	70	0.0721	0.0505
GLDM - LDHGLE	68		GLDM - LDHGLE	72.4	0.067	0.0485
GLCM - DE	58		Imaging - Bone reaction: Lytic	61.6	0.0699	0.0431
Imaging - Bone reaction: Lytic	57.6		Location: Semirigid	50	0.0663	0.0331
Location: Semirigid	46.8		GLCM - DE	48.8	0.0668	0.0326
GTV - Classification: Body + bilateral pedicle/transverse processBody	34		GTV - Classification: Body + bilateral pedicle/transverse processBody	36.4	0.0677	0.0246
Posterolateral involvement of the spinal elements: Bilateral	31.6		Posterolateral involvement of the spinal elements: Bilateral	34.4	0.0652	0.0224
Vertebral body collapse: >50% collapse	30.4		Vertebral body collapse: >50% collapse	27.2	0.0727	0.0198

TABLE S12
Averaged optimal hyperparameter values for the three best models.

The values shown have been averaged across the 250 final models. In the cases where an average is not possible (e.g., categorical parameters), the mode has been used. In the cases where a parameter does not make sense or alters the model in any way (e.g., *gamma* or *degree* for a linear kernel in SVM), it has been accordingly excluded.

Model	max_features	max_depth	min_samples_split	min_samples_leaf	Bootstrap	Criterion

CTV radiomics model (RFC)	'auto'	503,22	5,21	2,92	TRUE	'gini'
Clinical model (SVM)	**C**	**Kernel**	**Degree**	**Gamma**		
	0,5	poly'	4,54	8,62		
Combined CTV and clinical model (SVM)	**C**	**Kernel**	**Degree**	**Gamma**		
	0,43	poly'	3,57	4,73		

SUPPLEMENTAL FIGURES

FIGURE S1
Patients workflow

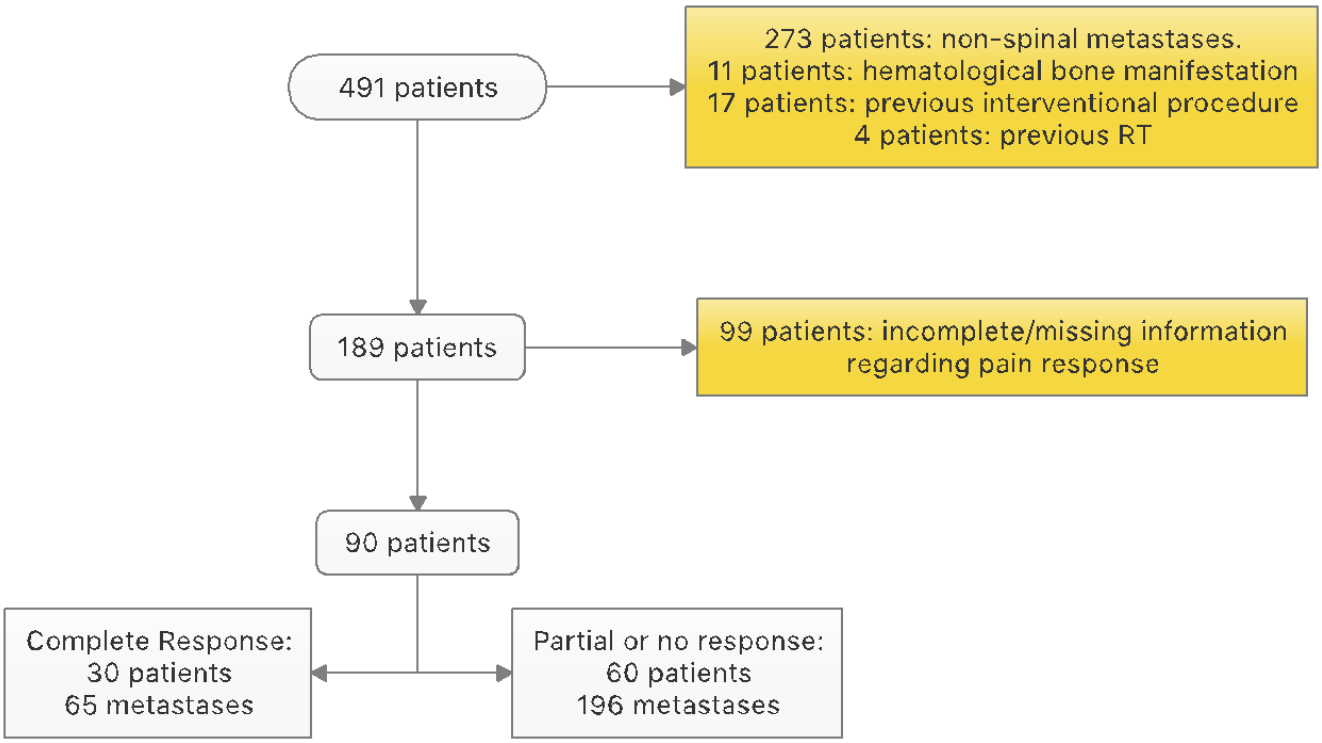


FIGURE S2
ROC and Calibration Curves: Extension of Figure 2

Receiver operator characteristic (ROC) and Calibration curves for the comparisons of the remaining models, from Table 3 and Table 5, that are not shown in Figure 2.

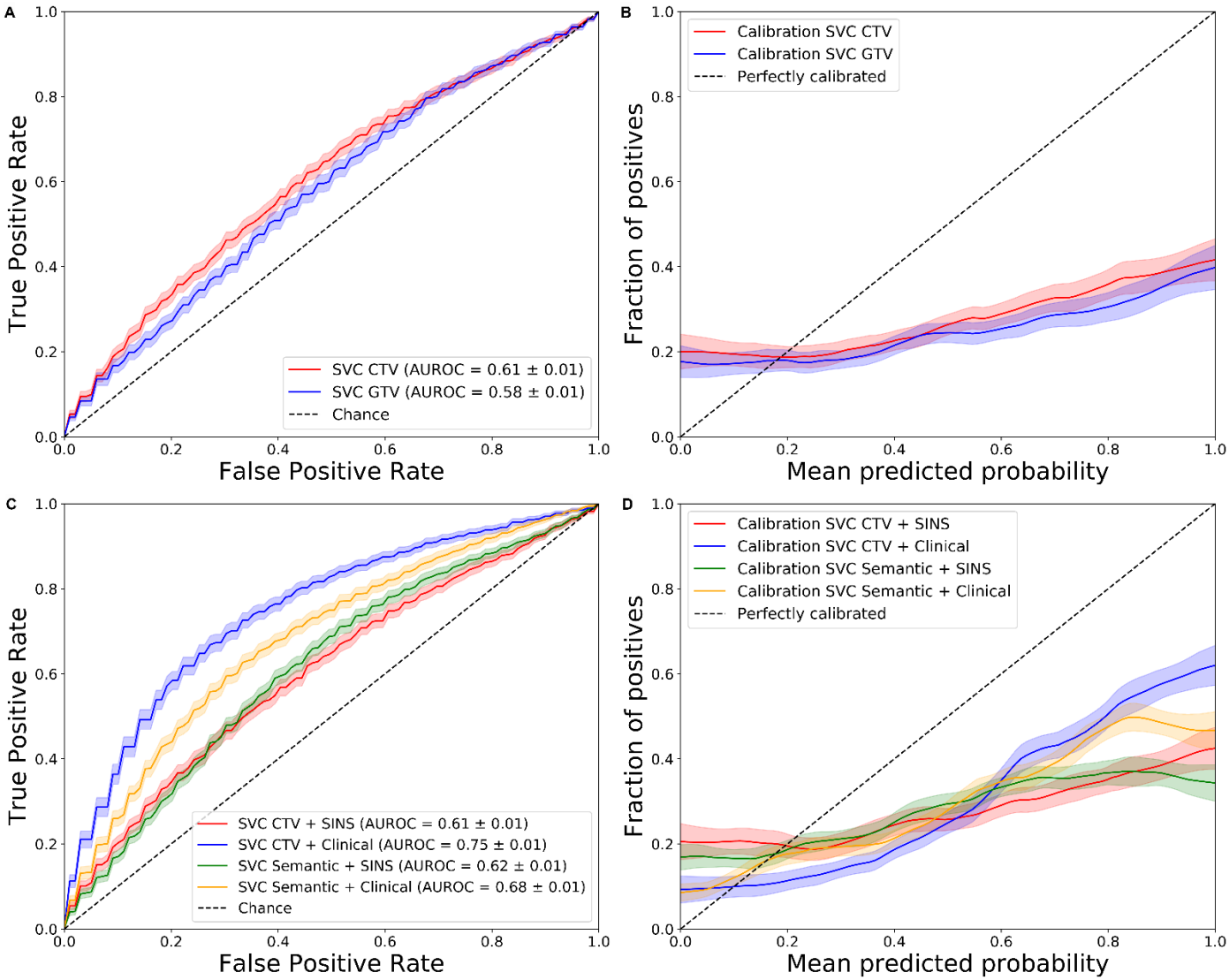

A
True Positive Rate
False Positive Rate
SVC CTV (AUROC = 0.61 ± 0.01)
SVC GTV (AUROC = 0.58 ± 0.01)
Chance
B
Fraction of positives
Mean predicted probability
Calibration SVC CTV
Calibration SVC GTV
Perfectly calibrated
C
True Positive Rate
False Positive Rate
SVC CTV + SINS (AUROC = 0.61 ± 0.01)
SVC CTV + Clinical (AUROC = 0.75 ± 0.01)
SVC Semantic + SINS (AUROC = 0.62 ± 0.01)
SVC Semantic + Clinical (AUROC = 0.68 ± 0.01)
Chance
D
Fraction of positives
Mean predicted probability
Calibration SVC CTV + SINS
Calibration SVC CTV + Clinical
Calibration SVC Semantic + SINS
Calibration SVC Semantic + Clinical
Perfectly calibrated

FIGURE S3
Averaged Confusion Matrix for the best performing radiomics model

Confusion Matrix that averages the classification performances of the 250 best radiomics models (RFC trained on CTV radiomics features).

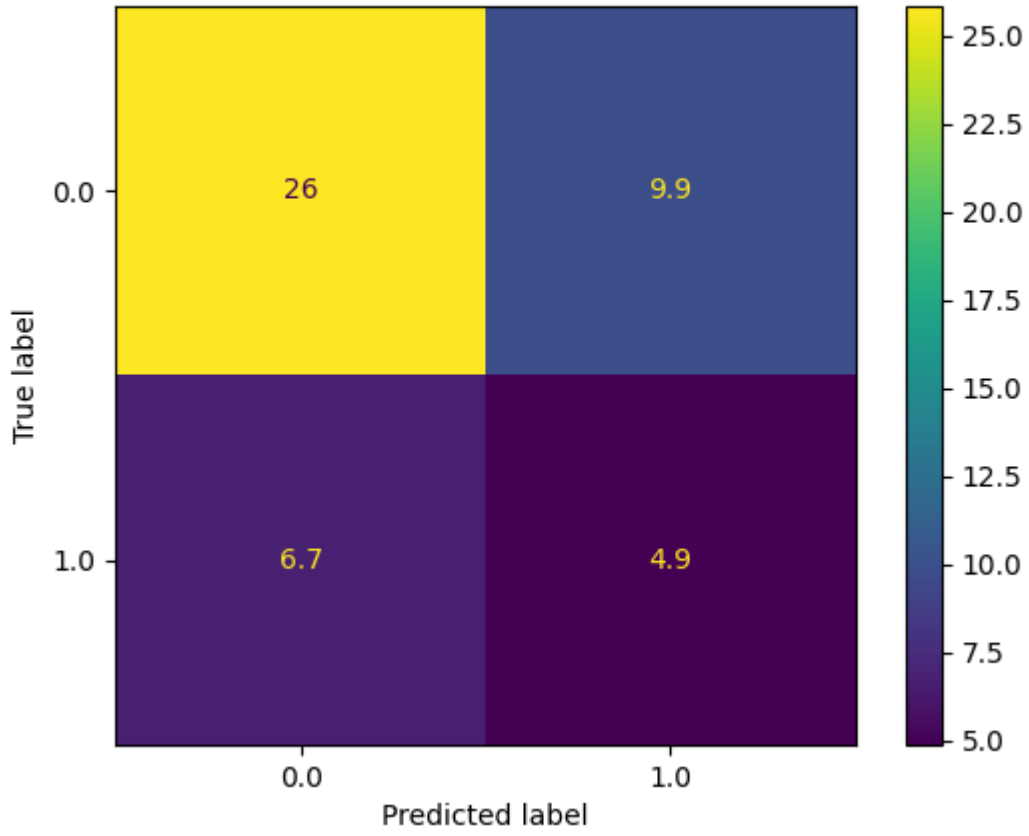
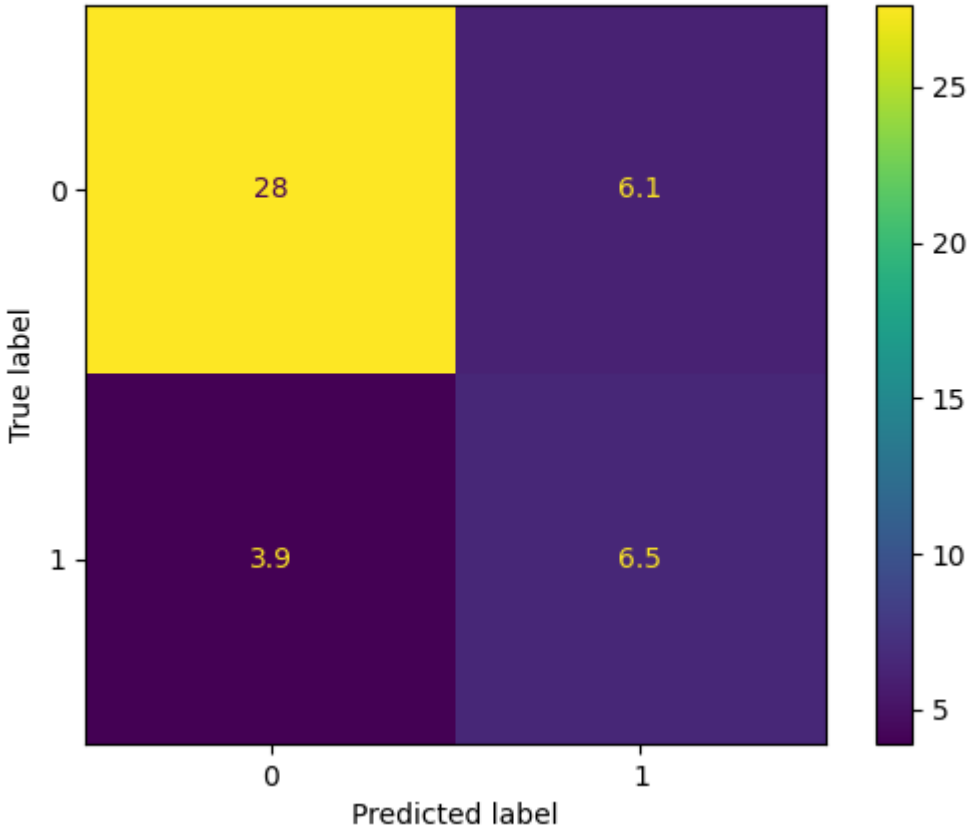


FIGURE S4
Averaged Confusion Matrix for the best performing model overall

Confusion Matrix that averages the classification performances of the 250 best models overall (RFC trained on CTV radiomics features).



References

1. Zwanenburg, A.; Vallières, M.; Abdalah, M.A.; Aerts, H.J.W.L.; Andrearczyk, V.; Apte, A.; Ashrafinia, S.; Bakas, S.; Beukinga, R.J.; Boellaard, R.; et al. The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping. *Radiology* **2020**, 191145, doi:10.1148/radiol.2020191145.
2. 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.J.W.L. Computational Radiomics System to Decode the Radiographic Phenotype. *Cancer Res.* **2017**, *77*, e104-e107, doi:10.1158/0008-5472.CAN-17-0339.

E.3 Supplemental material for chapter 4.3

Supplemental Material

Supplemental Tables

TABLE S1
Side effect prediction extended scores

Table S1. Acquisition parameters extracted from the DICOM metadata files of the TUM center.

Acquisition parameters	Matrix (pixel)	Pixel Spacing (mm)	Slice Thickness (mm)	Kernel	Tube Current (kV)	CT Scanner
TUM	512 x 512	0.97 x 0.97	3	B31s	130	Siemens Somatom Emotion 16

TABLE S2
Side effect prediction extended scores

Table S2. Test scores of the best performing models for each of the side effects and their configuration. For instance, the best model that predicted skin inflammation was a RF trained on the TBV cohort, which used Spearman's correlation as the feature selection technique.

Metric	Moist Cells Epitheliolysis TBV, LASSO, MRMR	Edema GT, LASSO, Spearman
AUROC*	0.74 ± 0.01	0.55 ± 0.01
Balanced Accuracy	0.65	0.52
F1	0.35	0.15
Sensitivity	0.56	0.31
Specificity	0.75	0.73
MCC	0.25	0.03

* Data is given as mean ± 1.96 standard errors for a 95% confidence interval

TABLE S3
Best radiomics cohort extended scores

Table S3. Test scores of the best performing models depending on the training data. For instance, for the radiomics cohort TBV, a LASSO classifier performed best when using MRMR as the feature selection technique, and predicting moist cells epitheliolysis.

Metric	TBV Moist ep., LASSO, MRMR	GT Moist ep., RF, MRMR	Clinical Features Moist ep., LR, Spearman
AUROC	0.74 ± 0.01	0.65 ± 0.01	0.70 ± 0.01
Balanced Accuracy	0.65	0.59	0.65
F1	0.35	0.27	0.34
Sensitivity	0.56	0.46	0.57
Specificity	0.75	0.71	0.71
MCC	0.25	0.14	0.23

* Data is given as mean ± 1.96 standard errors for a 95% confidence interval

TABLE S4
Best modelling strategy

The four ML algorithms have been compared, and their best configurations of prediction target, radiomics feature set and feature selection technique are shown. The best performing ML algorithm was a LASSO classifier trained on TBV radiomics features to predict moist epitheliolysis (AUROC of 0.74), albeit not by a statistically significant margin: LR and RF are still within 1.96 standard errors at 0.73 and 0.72, respectively. Regardless of the algorithm selected, the configuration that has proven to perform best was using TBV as the training radiomics features, and MRMR as the feature selection technique. Moist epitheliolysis as the prediction target has, again, proven to yield the best results.

Table S4. Test scores of the best performing model configurations for each of the four ML algorithms. For instance, the best LR performance was achieved when trained on volume A radiomics features, using MRMR as the feature selection technique, and predicting moist cells epitheliolysis.

Metric	LR TBV, Moist ep., MRMR	LASSO TBV, Moist ep., MRMR	SVM TBV, Moist ep., MRMR	RF TBV, Moist ep., MRMR
AUROC	0.73 ± 0.01	0.74 ± 0.01	0.69 ± 0.02	0.72 ± 0.01
Balanced Accuracy	0.65	0.65	0.63	0.64
F1	0.34	0.35	0.32	0.33
Sensitivity	0.56	0.56	0.47	0.5
Specificity	0.74	0.75	0.78	0.77
MCC	0.23	0.25	0.21	0.22

* Data is given as mean ± 1.96 standard errors for a 95% confidence interval

TABLE S5
Best feature selection approach

The performance of the two different feature selection techniques is shown. Using either technique, LASSO performed best when training on the selected TBV features, and predicting moist cells epitheliolysis.

The impact of the feature selection technique is minimal, albeit noticeable, when compared to other modelling configurations, such as the prediction target or the ML algorithm. This can be seen not only for the best model overall (LASSO classifier, trained on the selected TBV radiomics features, and predicting moist cells epitheliolysis), but in many other instances. While MRMR has performed slightly better on the most optimal models, Spearman leads to marginally better results on just above random performing models.

Table S5. Test scores of the best performing models depending on the feature selection technique utilized. For instance, when using MRMR to select the best volume A radiomics features, LASSO performed best when predicting moist cells epitheliolysis.

Metric	MRMR TBV, Moist ep., LASSO	Spearman TBV, Moist ep., LASSO
AUROC	0.74 ± 0.01	0.72 ± 0.01
Balanced Accuracy	0.65	0.64

F1	0.35	0.33
Sensitivity	0.56	0.5
Specificity	0.75	0.77
MCC	0.25	0.23

* Data is given as mean ± 1.96 standard errors for a 95% confidence interval

TABLE S6
Combined modelling extended results

Table S6. Test scores of the best performing combined models for each of the radiomics cohorts. For instance, when using TBV radiomics features together with clinical features as training data, a RF classifier performed best when predicting skin inflammation.

Metric	TBV + Clinical Features		GT + Clinical Features	
	Moist ep.	**Edema**	**Moist ep.**	**Edema**
	LASSO, MRMR	RF, Spearman	RF, MRMR	LASSO, Spearman
AUROC	0.73 ± 0.01	0.53 ± 0.02	0.67 ± 0.01	0.55 ± 0.01
Balanced Accuracy	0.65	0.51	0.6	0.52
F1	0.34	0.12	0.29	0.15
Sensitivity	0.55	0.17	0.49	0.33
Specificity	0.74	0.86	0.71	0.71
MCC	0.23	0.03	0.16	0.02

* Data is given as mean ± 1.96 standard errors for a 95% confidence interval

TABLE S7
Best performing TBV model excluding volume-correlated features

Table S7. Test scores of the best performing TBV model, excluding features with a Spearman correlation coefficient larger than 0.8 towards breast volume. The configuration was a LASSO classifier, using MRMR as the feature selection technique for further refinement, and predicting moist epitheliolysis.

Metric	TBV excluding volume features
	Moist ep., LASSO, MRMR
AUROC	0.71 ± 0.01
Balanced Accuracy	0.63
F1	0.32
Sensitivity	0.51
Specificity	0.75
MCC	0.21

* Data is given as mean ± 1.96 standard errors for a 95% confidence interval

TABLE S8
Feature importance report of the best performing clinical model

Table S8. Feature importance report of the best performing clinical model: a LR classifier trained on clinical features, selected with a double Spearman rank correlation test, and predicting moist cells epitheliolysis. Score is calculated by multiplying the average importance of a feature by their selection frequency. The best 15 features are shown.

Feature	% Selected	Average Importance	Score
TBV Volume	100	3.29	3.29
RT Boost	100	0.42	0.42
EQD2 Max Radiation Dose	100	0.31	0.31
Smoker Status	100	0.28	0.28
Chemotherapy	100	0.26	0.26

Supplemental Figures

FIGURE S1
VOI references of CT scans for TBV (left) and GT (right)

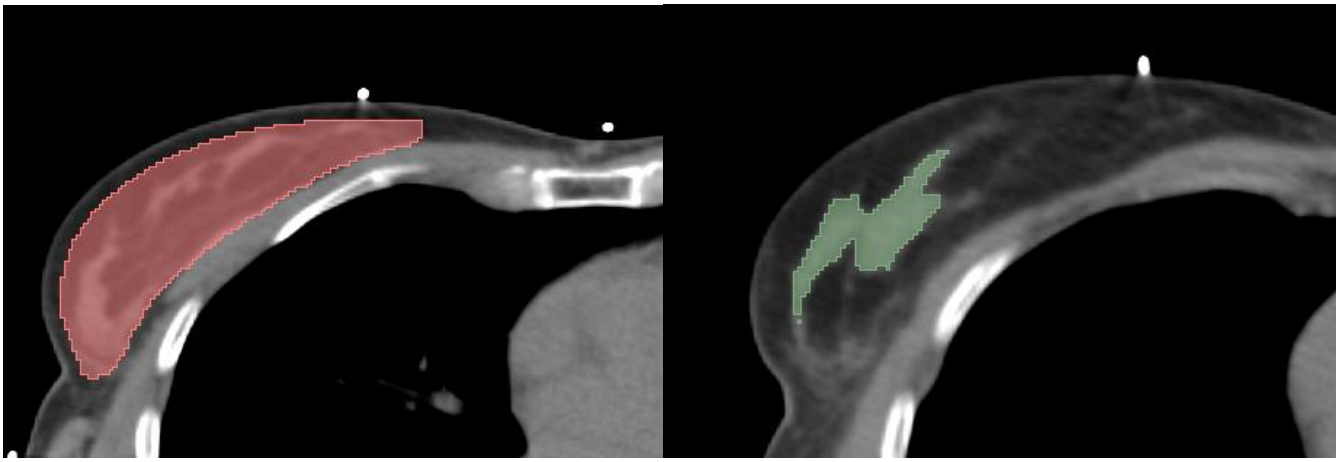


Figure S1. CT scans highlighting both VOIs studied in this research: TBV (left) and GT (right).

FIGURE S2
Patient workflow of clinical features and side effects

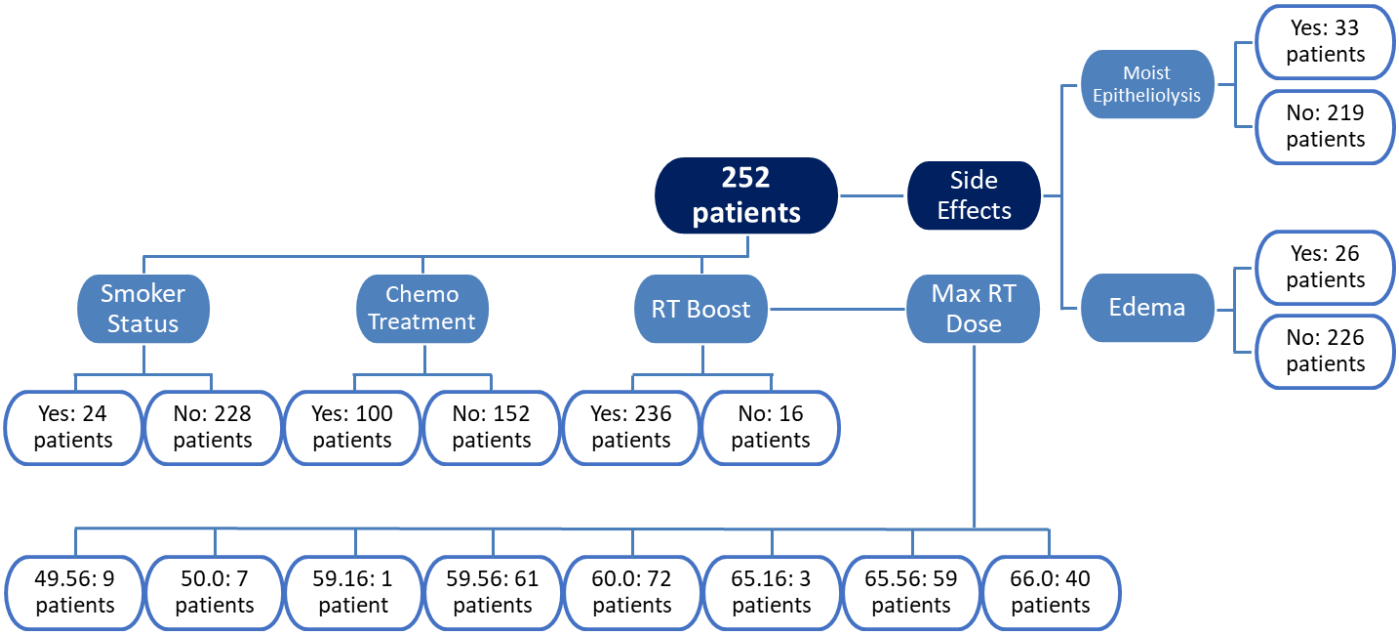


Figure S2. Patient workflow summarizing their clinical features and side effects distributions.

FIGURE S3
Extended predictive influence of the breast volume (correlation scores)

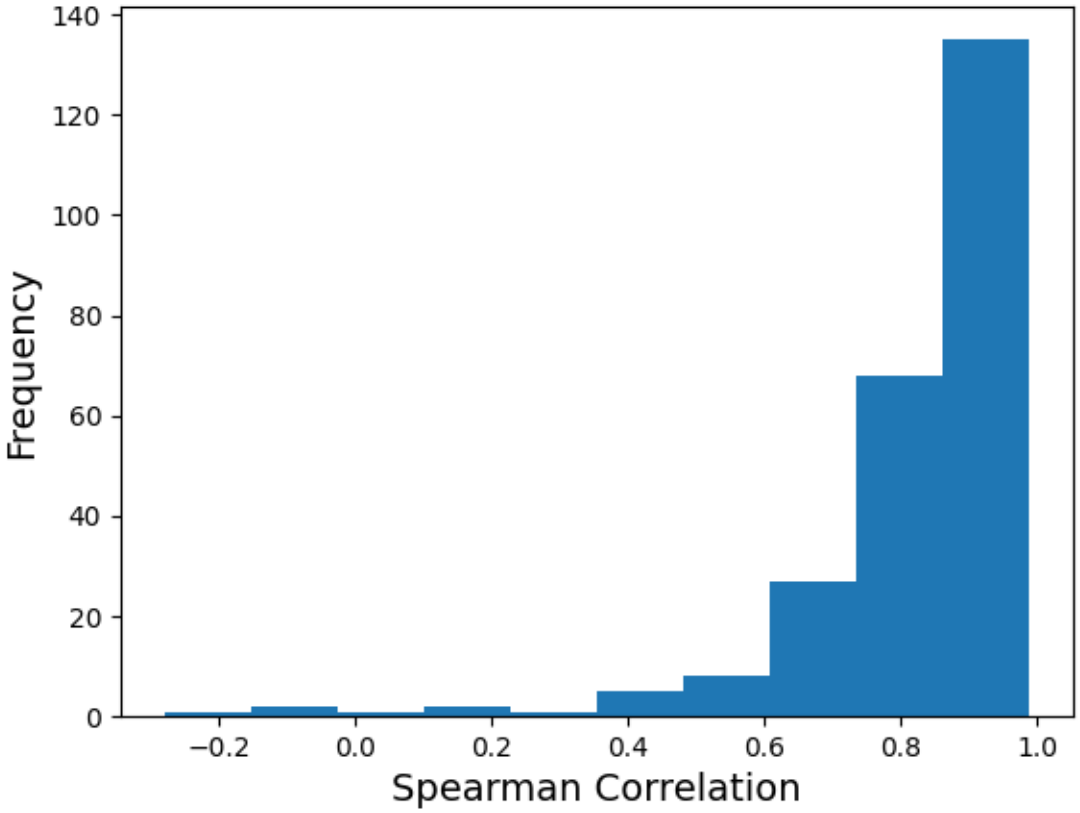


Figure S3. Spearman's correlation scores between the volume of the whole breast and the prediction probabilities of the best performing model, a LASSO classifier trained on TBV radiomics features, selected by MRMR, and used to predict moist cells epitheliolysis as a surrogate for skin inflammation side effect.

FIGURE S4
Extended predictive influence of the breast volume (correlation p-values)

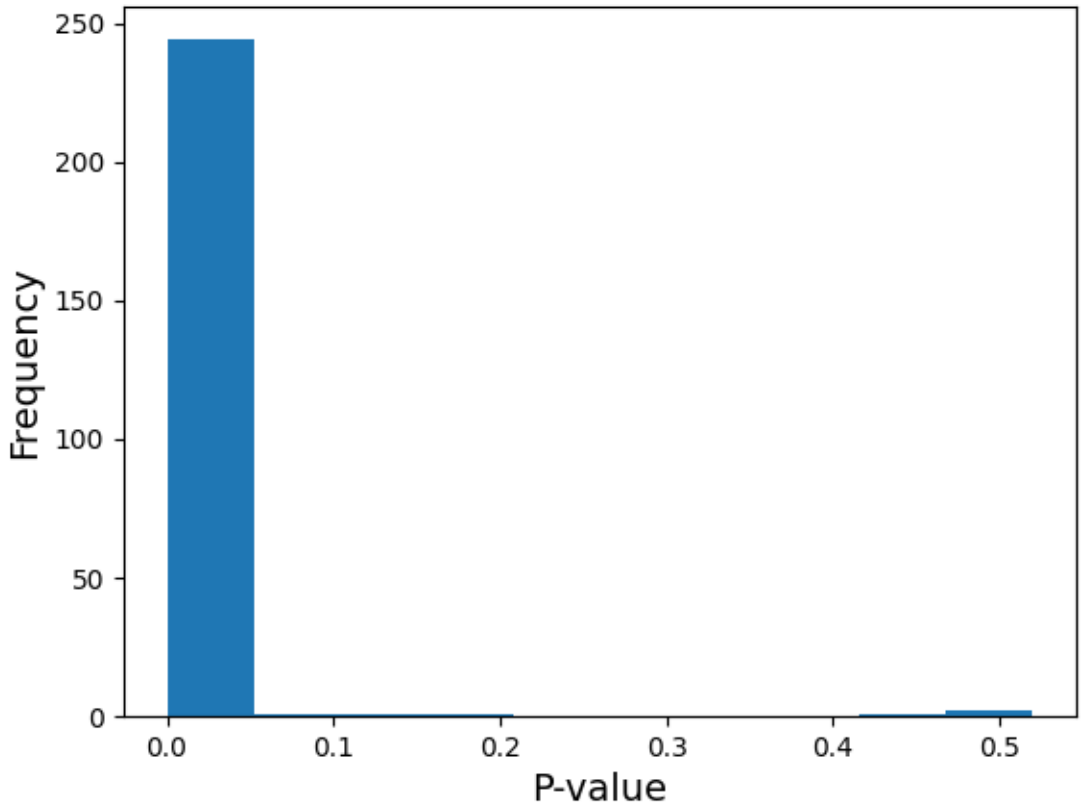


Figure S4. P-values of the Spearman's correlations between the volume of the whole breast and the prediction probabilities of the best performing model, a LASSO classifier trained on TBV radiomics features, selected by MRMR, and used to predict moist cells epitheliolysis as a surrogate for skin inflammation side effect.

FIGURE S5
Calibration curve of the best performing radiomics model

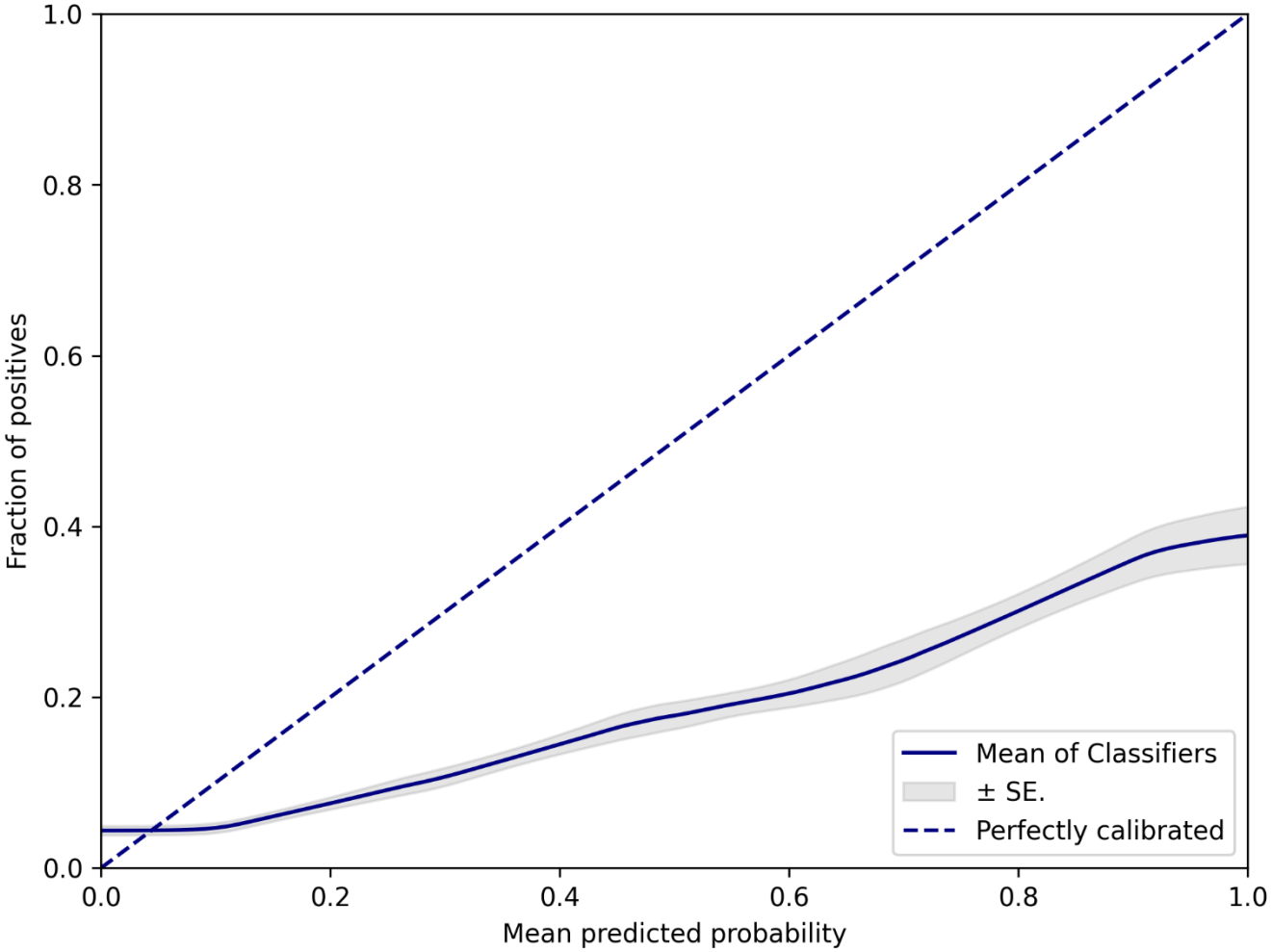


Figure S5. Calibration curve of the best performing radiomics model: a LASSO classifier trained on TBV radiomics features, selected with MRMR, and predicting moist cells epitheliolysis.

E.4 Supplemental material for chapter 4.4

Supplemental Material

Supplemental Tables

TABLE S1
MRI acquisition parameters

Table S1. Acquisition parameters extracted from the DICOM metadata files.

Acquisition Parameters	Magnetic Field Strength (T)			Repetition Time (ms) *	Echo Time (ms) *	Slice Thickness (mm) *	Flip Angle (°) *	Pixel Spacing (mm) *
	1.0	1.5	3.0					
Baseline	3	63	30	1559 - 2603.2	23.9 - 40.1	4.4 - 5.2	101.8 - 232.2	0.66 - 0.82 x 0.66 - 0.82
Follow Up	6	55	34	1875.5 - 2959.3	26.8 - 39.6	4.7 - 5.3	111.6 - 131.8	0.75 - 0.91 x 0.75 - 0.91

* Data is given as the 95% confidence interval boundary values.

TABLE S2
Extracted radiomics features

Table S2. Radiomics features extracted from the MRI scans. All extracted features were computed according to the "image biomarker standardization initiative" (IBSI) guidelines [1]. The pyRadiomics package (version 2.0) implemented in Python (version 3.6.4) was used for feature extraction [2].

Shape Features	
Volume	Surface Area
Surface Volume Area	Sphericity
Spherical Disproportion	Maximum 3D Diameter
Maximum 2D Diameter Slice	Maximum 2D Diameter Column
Maximum 2D Diameter Row	Major Axis
Minor Axis	Least Axis
Elongation	Flatness
First Order Features	
Energy	Intensity Histogram Entropy
Minimum	10th Percentile
90th Percentile	Maximum
Mean	Median
Interquartile Range	Range
Mean Absolute Deviation (MAD)	Robust Mean Absolute Deviation (rMAD)
Root Mean Squared (RMS)	Skewness
Excess Kurtosis	Variance
Intensity Histogram Uniformity	
Gray Level Co-occurrence Matrix (GLCM) Features	
Autocorrelation	Joint Average
Cluster Prominence	Cluster Shade
Cluster Tendency	Contrast
Correlation	Difference Average
Difference Entropy	Difference Variance
Joint Energy (IBSI: Angular Second Moment)	Joint Entropy
Informal Measure of Correlation (IMC) 1	Informal Measure of Correlation (IMC) 2

Inverse Difference Moment (IDM)	Inverse Difference Moment Normalized (IDMN)
Inverse Difference (ID)	Inverse Difference Normalized (IDN)
Inverse Variance	Maximum Probability (IBSI: Joint maximum)
Sum Entropy	Sum of Squares (IBSI: Sum of Squares)
Maximal Correlation Coefficient (MCC)	
Gray Level Size Zone Matrix (GLSZM) Features	
Small Area Emphasis (SAE)	Large Area Emphasis (LAE)
Gray Level Non-Uniformity (GLN)	Gray Level Non-Uniformity Normalized (GLNN)
Size-Zone Non-Uniformity (SZN)	Size-Zone Non-Uniformity Normalized (SZNN)
Zone Percentage (ZP)	Gray Level Variance (GLV)
Zone Variance (ZV)	Zone Entropy (ZE)
Low Gray Level Zone Emphasis (LGLZE)	High Gray Level Zone Emphasis (HGLZE)
Small Area Low Gray Level Emphasis (SALGLE)	Small Area High Gray Level Emphasis (SAHGLE)
Large Area Low Gray Level Emphasis (LALGLE)	Large Area High Gray Level Emphasis (LAHGLE)
Gray Level Run Length Matrix (GLRLM) Features	
Short Run Emphasis (SRE)	Long Run Emphasis (LRE)
Gray Level Non-Uniformity (GLN)	Gray Level Non-Uniformity Normalized (GLNN)
Run Length Non-Uniformity (RLN)	Run Length Non-Uniformity Normalized (RLNN)
Run Percentage (RP)	Gray Level Variance (GLV)
Run Variance (RV)	Run Entropy (RE)
Low Gray Level Run Emphasis (LGLRE)	High Gray Level Run Emphasis (HGLRE)
Short Run Low Gray Level Emphasis (SRLGLE)	Short Run High Gray Level Emphasis (SRHGLE)
Long Run Low Gray Level Emphasis (LRLGLE)	Long Run High Gray Level Emphasis (LRHGLE)
Neighbouring Gray Tone Difference Matrix (NGTDM) Features	
Coarseness	Contrast
Busyness	Complexity
Strength	
Gray Level Dependence Matrix (GLDM) Features	
Small Dependence Emphasis (SDE)	Large Dependence Emphasis (LDE)
Gray Level Non-Uniformity (GLN)	Dependence Non-Uniformity (DN)
Dependence Non-Uniformity Normalized (DNN)	Gray Level Variance (GLV)
Dependence Variance (DV)	Dependence Entropy (DE)
Low Gray Level Emphasis (LGLE)	High Gray Level Emphasis (HGLE)
Small Dependence Low Gray Level Emphasis (SDLGLE)	Small Dependence High Gray Level Emphasis (SDHGLE)
Large Dependence Low Gray Level Emphasis (LDLGLE)	Large Dependence High Gray Level Emphasis (LDHGLE)

Supplemental Figures

FIGURE S1
Patient workflow

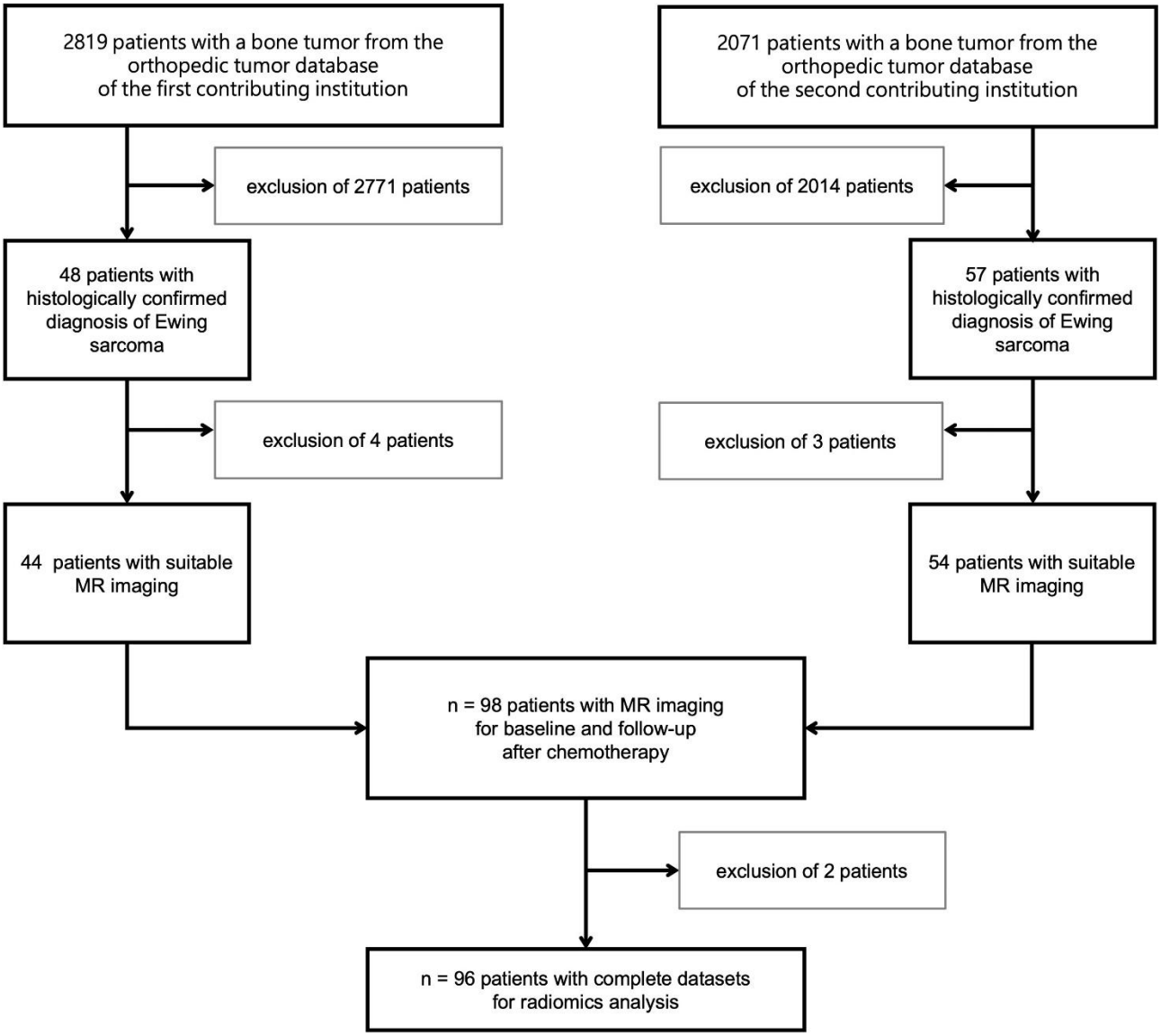


Figure S1. Patient cohort and dataset creation. Histological samples taken after the first chemotherapy cycle were evaluated according to the Salzer-Kuntschik score for histological response [3]. For appropriate imaging, MR imaging scans before (baseline) and after (post-ct) were mandatory. Due to technical issues during the radiomics analysis, two more patients had to be excluded.

FIGURE S2
AUROC score comparison of the best radiomics and radiology models

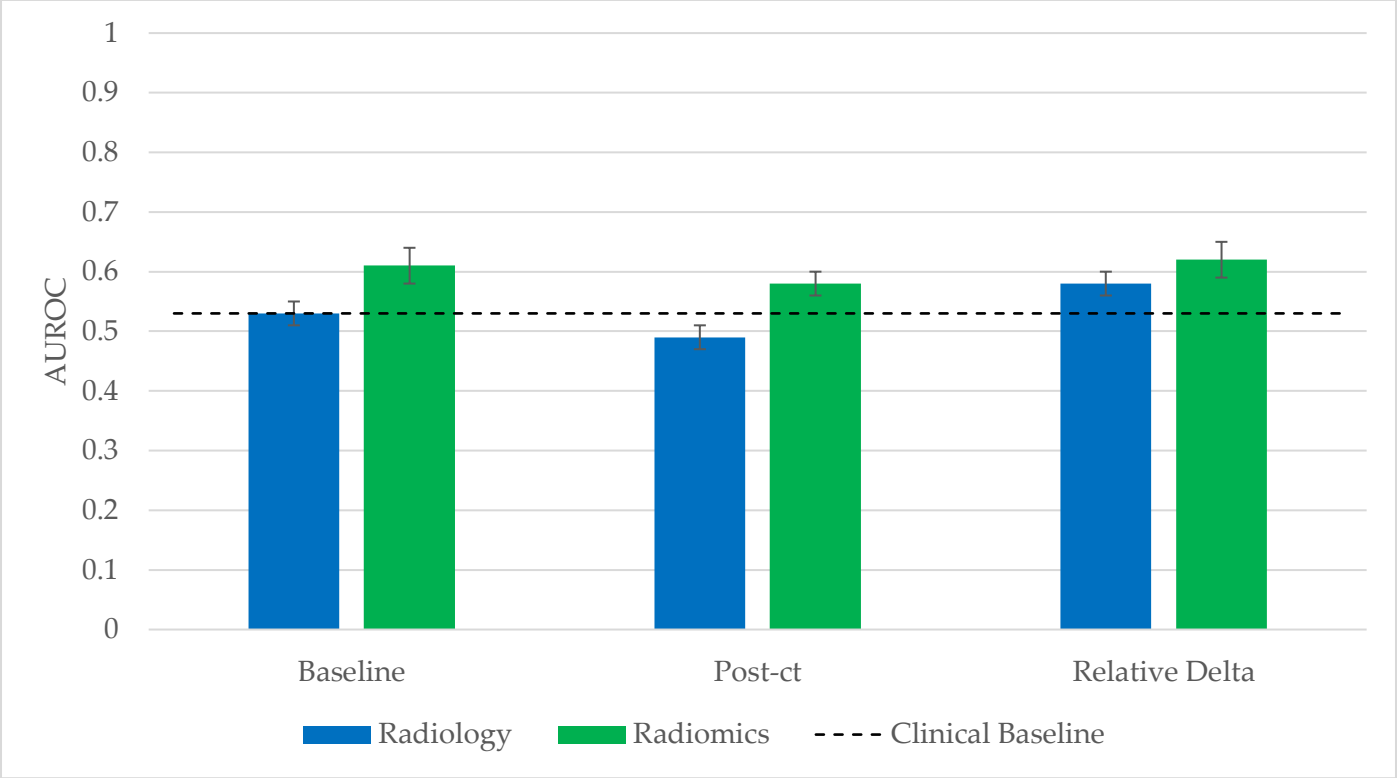


Figure S2. AUROC values of baseline, post-ct and relative delta time points of the best performing radiology (blue) and radiomics (green; T1fsgd MRI modality) models.

FIGURE S3
Patient-wise comparison of the 6 most predictive features between baseline and post-ct

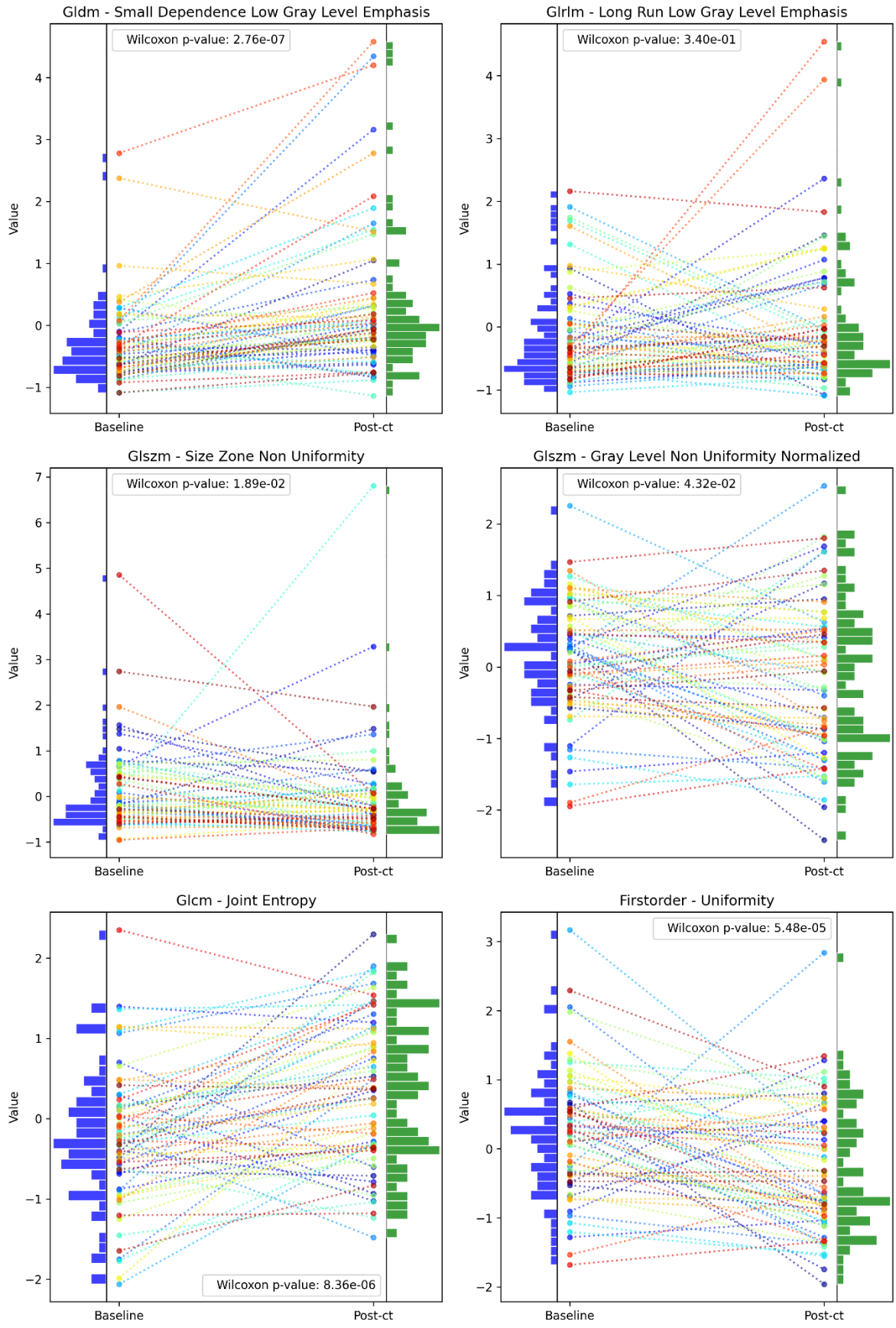


Figure 3. Evolution from baseline to post-ct of all patients for the 6 most predictive features in the best performing model: a LogReg trained on the relative delta of T1fsgd features selected by a two-step Spearman's correlation coefficient. Each radiomics feature is accompanied by the p-value of the Wilcoxon signed-rank test.

REFERENCES

1. Zwanenburg, A.; Vallières, M.; Abdalah, M.A.; Aerts, H.J.W.L.; Andrearczyk, V.; Apte, A.; Ashrafinia, S.; Bakas, S.; Beukinga, R.J.; Boellaard, R.; et al. The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-Based Phenotyping. *Radiology* **2020**, *295*, 328–338, doi:10.1148/radiol.2020191145.
2. 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.J.W.L. Computational Radiomics System to Decode the Radiographic Phenotype. *Cancer Research* **2017**, *77*, e104–e107, doi:10.1158/0008-5472.CAN-17-0339.
3. Salzer-Kuntschik, M.; Brand, G.; Delling, G. Determination of the Degree of Morphological Regression Following Chemotherapy in Malignant Bone Tumors. *Der Pathologe* **1983**, *4*, 135–141.

F. Bibliography

1. Menezes, M.-R. The Biology of Cancer. *The Yale Journal of Biology and Medicine* **2015**, *88*, 199–200.
2. Sung, H.; Ferlay, J.; Siegel, R.L.; Laversanne, M.; Soerjomataram, I.; Jemal, A.; Bray, F. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J Clin* **2021**, *71*, 209–249, doi:10.3322/caac.21660.
3. Hanahan, D.; Weinberg, R.A. Hallmarks of Cancer: The next Generation. *Cell* **2011**, *144*, 646–674, doi:10.1016/j.cell.2011.02.013.
4. Vogelstein, B.; Kinzler, K.W. Cancer Genes and the Pathways They Control. *Nat Med* **2004**, *10*, 789–799, doi:10.1038/nm1087.
5. Stratton, M.R.; Campbell, P.J.; Futreal, P.A. The Cancer Genome. *Nature* **2009**, *458*, 719–724, doi:10.1038/nature07943.
6. Fearon, E.R.; Vogelstein, B. A Genetic Model for Colorectal Tumorigenesis. *Cell* **1990**, *61*, 759–767, doi:10.1016/0092-8674(90)90186-i.
7. Hanahan, D.; Weinberg, R.A. The Hallmarks of Cancer. *Cell* **2000**, *100*, 57–70, doi:10.1016/s0092-8674(00)81683-9.
8. Ferlay, J.; Colombet, M.; Soerjomataram, I.; Mathers, C.; Parkin, D.M.; Piñeros, M.; Znaor, A.; Bray, F. Estimating the Global Cancer Incidence and Mortality in 2018: GLOBOCAN Sources and Methods. *Int J Cancer* **2019**, *144*, 1941–1953, doi:10.1002/ijc.31937.
9. Anand, P.; Kunnumakkara, A.B.; Sundaram, C.; Harikumar, K.B.; Tharakan, S.T.; Lai, O.S.; Sung, B.; Aggarwal, B.B. Cancer Is a Preventable Disease That Requires Major Lifestyle Changes. *Pharm Res* **2008**, *25*, 2097–2116, doi:10.1007/s11095-008-9661-9.
10. Peto, J. Cancer Epidemiology in the Last Century and the next Decade. *Nature* **2001**, *411*, 390–395, doi:10.1038/35077256.
11. Colditz, G.A.; Wei, E.K. Preventability of Cancer: The Relative Contributions of Biologic and Social and Physical Environmental Determinants of Cancer Mortality. *Annu Rev Public Health* **2012**, *33*, 137–156, doi:10.1146/annurev-publhealth-031811-124627.
12. Clapp, R.W.; Jacobs, M.M.; Loechler, E.L. Environmental and Occupational Causes of Cancer: New Evidence 2005-2007. *Rev Environ Health* **2008**, *23*, 1–37, doi:10.1515/reveh.2008.23.1.1.
13. Parkin, D.M. The Global Health Burden of Infection-Associated Cancers in the Year 2002. *Int J Cancer* **2006**, *118*, 3030–3044, doi:10.1002/ijc.21731.
14. Plummer, M.; de Martel, C.; Vignat, J.; Ferlay, J.; Bray, F.; Franceschi, S. Global Burden of Cancers Attributable to Infections in 2012: A Synthetic Analysis. *Lancet Glob Health* **2016**, *4*, e609-616, doi:10.1016/S2214-109X(16)30143-7.
15. Humans, I.W.G. on the E. of C.R. to *Arsenic, Metals, Fibres and Dusts*; International Agency for Research on Cancer, 2012; ISBN 978-92-832-1320-8.
16. LaDou, J. The Asbestos Cancer Epidemic. *Environ Health Perspect* **2004**, *112*, 285–290.
17. Grosse, Y.; Loomis, D.; Guyton, K.Z.; El Ghissassi, F.; Bouvard, V.; Benbrahim-Tallaa, L.; Mattock, H.; Straif, K.; International Agency for Research on Cancer Monograph Working Group Carcinogenicity of Some Industrial Chemicals. *Lancet Oncol* **2016**, *17*, 419–420, doi:10.1016/S1470-2045(16)00137-6.
18. Bray, F.; Laversanne, M.; Sung, H.; Ferlay, J.; Siegel, R.L.; Soerjomataram, I.; Jemal, A. Global Cancer Statistics 2022: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: A Cancer Journal for Clinicians* **2024**, *74*, 229–263, doi:10.3322/caac.21834.
19. Preventing Cancer Available online: https://www.who.int/activities/preventing-cancer (accessed on 14 June 2024).

20. Kushi, L.H.; Doyle, C.; McCullough, M.; Rock, C.L.; Demark-Wahnefried, W.; Bandera, E.V.; Gapstur, S.; Patel, A.V.; Andrews, K.; Gansler, T.; et al. American Cancer Society Guidelines on Nutrition and Physical Activity for Cancer Prevention: Reducing the Risk of Cancer with Healthy Food Choices and Physical Activity. *CA Cancer J Clin* **2012**, *62*, 30–67, doi:10.3322/caac.20140.
21. Clinton, S.K.; Giovannucci, E.L.; Hursting, S.D. The World Cancer Research Fund/American Institute for Cancer Research Third Expert Report on Diet, Nutrition, Physical Activity, and Cancer: Impact and Future Directions. *J Nutr* **2020**, *150*, 663–671, doi:10.1093/jn/nxz268.
22. Chhabra, N.; Kennedy, J. A Review of Cancer Immunotherapy Toxicity: Immune Checkpoint Inhibitors. *J Med Toxicol* **2021**, *17*, 411–424, doi:10.1007/s13181-021-00833-8.
23. Sharma, P.; Allison, J.P. The Future of Immune Checkpoint Therapy. *Science* **2015**, *348*, 56–61, doi:10.1126/science.aaa8172.
24. Hodi, F.S.; Chiarion-Sileni, V.; Gonzalez, R.; Grob, J.-J.; Rutkowski, P.; Cowey, C.L.; Lao, C.D.; Schadendorf, D.; Wagstaff, J.; Dummer, R.; et al. Nivolumab plus Ipilimumab or Nivolumab Alone versus Ipilimumab Alone in Advanced Melanoma (CheckMate 067): 4-Year Outcomes of a Multicentre, Randomised, Phase 3 Trial. *Lancet Oncol* **2018**, *19*, 1480–1492, doi:10.1016/S1470-2045(18)30700-9.
25. Rini, B.I.; Plimack, E.R.; Stus, V.; Gafanov, R.; Hawkins, R.; Nosov, D.; Pouliot, F.; Alekseev, B.; Soulières, D.; Melichar, B.; et al. Pembrolizumab plus Axitinib versus Sunitinib for Advanced Renal-Cell Carcinoma. *N Engl J Med* **2019**, *380*, 1116–1127, doi:10.1056/NEJMoa1816714.
26. Weinstein, I.B.; Joe, A.K. Mechanisms of Disease: Oncogene Addiction—a Rationale for Molecular Targeting in Cancer Therapy. *Nat Rev Clin Oncol* **2006**, *3*, 448–457, doi:10.1038/ncponc0558.
27. Basch, E.; Deal, A.M.; Kris, M.G.; Scher, H.I.; Hudis, C.A.; Sabbatini, P.; Rogak, L.; Bennett, A.V.; Dueck, A.C.; Atkinson, T.M.; et al. Symptom Monitoring With Patient-Reported Outcomes During Routine Cancer Treatment: A Randomized Controlled Trial. *J Clin Oncol* **2016**, *34*, 557–565, doi:10.1200/JCO.2015.63.0830.
28. Temel, J.S.; Greer, J.A.; Muzikansky, A.; Gallagher, E.R.; Admane, S.; Jackson, V.A.; Dahlin, C.M.; Blinderman, C.D.; Jacobsen, J.; Pirl, W.F.; et al. Early Palliative Care for Patients with Metastatic Non-Small-Cell Lung Cancer. *N Engl J Med* **2010**, *363*, 733–742, doi:10.1056/NEJMoa1000678.
29. Smith, R.A.; Andrews, K.S.; Brooks, D.; Fedewa, S.A.; Manassaram-Baptiste, D.; Saslow, D.; Brawley, O.W.; Wender, R.C. Cancer Screening in the United States, 2018: A Review of Current American Cancer Society Guidelines and Current Issues in Cancer Screening. *CA Cancer J Clin* **2018**, *68*, 297–316, doi:10.3322/caac.21446.
30. Siegel, R.L.; Miller, K.D.; Jemal, A. Cancer Statistics, 2019. *CA Cancer J Clin* **2019**, *69*, 7–34, doi:10.3322/caac.21551.
31. Widhe, B.; Widhe, T. Initial Symptoms and Clinical Features in Osteosarcoma and Ewing Sarcoma. *J Bone Joint Surg Am* **2000**, *82*, 667–674.
32. Dalal, K.M.; Antonescu, C.R.; Singer, S. Diagnosis and Management of Lipomatous Tumors. *J Surg Oncol* **2008**, *97*, 298–313, doi:10.1002/jso.20975.
33. Rydholm, A.; Berg, N.O. Size, Site and Clinical Incidence of Lipoma. Factors in the Differential Diagnosis of Lipoma and Sarcoma. *Acta Orthop Scand* **1983**, *54*, 929–934, doi:10.3109/17453678308992936.
34. Fletcher, C.D.M. The Evolving Classification of Soft Tissue Tumours: An Update Based on the New WHO Classification. *Histopathology* **2006**, *48*, 3–12, doi:10.1111/j.1365-2559.2005.02284.x.

35. Myhre-Jensen, O.; Kaae, S.; Madsen, E.H.; Sneppen, O. Histopathological Grading in Soft-Tissue Tumours. Relation to Survival in 261 Surgically Treated Patients. *Acta Pathol Microbiol Immunol Scand A* **1983**, *91*, 145–150.
36. Bassett, M.D.; Schuetze, S.M.; Disteche, C.; Norwood, T.H.; Swisshelm, K.; Chen, X.; Bruckner, J.; Conrad III, E.U.; Rubin, B.P. Deep-Seated, Well Differentiated Lipomatous Tumors of the Chest Wall and Extremities. *Cancer* **2005**, *103*, 409–416, doi:10.1002/cncr.20779.
37. Weiss, S.W.; Rao, V.K. Well-Differentiated Liposarcoma (Atypical Lipoma) of Deep Soft Tissue of the Extremities, Retroperitoneum, and Miscellaneous Sites. A Follow-up Study of 92 Cases with Analysis of the Incidence of "Dedifferentiation." *Am J Surg Pathol* **1992**, *16*, 1051–1058, doi:10.1097/00000478-199211000-00003.
38. Bidault, F.; Vanel, D.; Terrier, P.; Jalaguier, A.; Bonvalot, S.; Pedeutour, F.; Couturier, J.M.; Dromain, C. Liposarcoma or Lipoma: Does Genetics Change Classic Imaging Criteria? *European Journal of Radiology* **2009**, *72*, 22–26, doi:10.1016/j.ejrad.2009.05.025.
39. Evans, H.L.; Soule, E.H.; Winkelmann, R.K. Atypical Lipoma, Atypical Intramuscular Lipoma, and Well Differentiated Retroperitoneal Liposarcoma: A Reappraisal of 30 Cases Formerly Classified as Well Differentiated Liposarcoma. *Cancer* **1979**, *43*, 574–584, doi:10.1002/1097-0142(197902)43:2<574::aid-cncr2820430226>3.0.co;2-7.
40. De Schepper, A.M.; De Beuckeleer, L.; Vandevenne, J.; Somville, J. Magnetic Resonance Imaging of Soft Tissue Tumors. *Eur Radiol* **2000**, *10*, 213–223, doi:10.1007/s003300050037.
41. Vilanova, J.C.; Woertler, K.; Narváez, J.A.; Barceló, J.; Martínez, S.J.; Villalón, M.; Miró, J. Soft-Tissue Tumors Update: MR Imaging Features According to the WHO Classification. *Eur Radiol* **2007**, *17*, 125–138, doi:10.1007/s00330-005-0130-0.
42. Totty, W.G.; Murphy, W.A.; Lee, J.K. Soft-Tissue Tumors: MR Imaging. *Radiology* **1986**, *160*, 135–141, doi:10.1148/radiology.160.1.3715024.
43. Enzinger and Weiss's Soft Tissue Tumors - NLM Catalog - NCBI Available online: https://www.ncbi.nlm.nih.gov/nlmcatalog/101604149 (accessed on 14 June 2024).
44. Kransdorf, M.J.; Bancroft, L.W.; Peterson, J.J.; Murphey, M.D.; Foster, W.C.; Temple, H.T. Imaging of Fatty Tumors: Distinction of Lipoma and Well-Differentiated Liposarcoma. *Radiology* **2002**, *224*, 99–104, doi:10.1148/radiol.2241011113.
45. Foreman, S.C.; Llorián-Salvador, O.; David, D.E.; Rösner, V.K.N.; Rischewski, J.F.; Feuerriegel, G.C.; Kramp, D.W.; Luiken, I.; Lohse, A.-K.; Kiefer, J.; et al. Development and Evaluation of MR-Based Radiogenomic Models to Differentiate Atypical Lipomatous Tumors from Lipomas. *Cancers* **2023**, *15*, 2150, doi:10.3390/cancers15072150.
46. Hacking, S.M.; Yakirevich, E.; Wang, Y. From Immunohistochemistry to New Digital Ecosystems: A State-of-the-Art Biomarker Review for Precision Breast Cancer Medicine. *Cancers* **2022**, *14*, 3469, doi:10.3390/cancers14143469.
47. Fidler, I.J. The Pathogenesis of Cancer Metastasis: The "seed and Soil" Hypothesis Revisited. *Nat Rev Cancer* **2003**, *3*, 453–458, doi:10.1038/nrc1098.
48. Gupta, G.P.; Massagué, J. Cancer Metastasis: Building a Framework. *Cell* **2006**, *127*, 679–695, doi:10.1016/j.cell.2006.11.001.
49. Steeg, P.S. Tumor Metastasis: Mechanistic Insights and Clinical Challenges. *Nat Med* **2006**, *12*, 895–904, doi:10.1038/nm1469.
50. Westhoff, P.G.; Graeff, A. de; Monninkhof, E.M.; Pomp, J.; Vulpen, M. van; Leer, J.W.H.; Marijnen, C.A.M.; Linden, Y.M. van der Quality of Life in Relation to Pain Response to Radiation Therapy for Painful Bone Metastases. *International Journal of Radiation Oncology, Biology, Physics* **2015**, *93*, 694–701, doi:10.1016/j.ijrobp.2015.06.024.
51. van der Velden, J.M.; Versteeg, A.L.; Verkooijen, H.M.; Fisher, C.G.; Chow, E.; Oner, F.C.; van Vulpen, M.; Weir, L.; Verlaan, J. Prospective Evaluation of the Relationship Between Mechanical Stability and Response to Palliative Radiotherapy for Symptomatic Spinal Metastases. *Oncologist* **2017**, *22*, 972–978, doi:10.1634/theoncologist.2016-0356.

52. Arcangeli, G.; Giovinazzo, G.; Saracino, B.; D'Angelo, L.; Giannarelli, D.; Arcangeli, G.; Micheli, A. Radiation Therapy in the Management of Symptomatic Bone Metastases: The Effect of Total Dose and Histology on Pain Relief and Response Duration. *Int J Radiat Oncol Biol Phys* **1998**, *42*, 1119–1126, doi:10.1016/s0360-3016(98)00264-8.
53. Nguyen, J.; Chow, E.; Zeng, L.; Zhang, L.; Culleton, S.; Holden, L.; Mitera, G.; Tsao, M.; Barnes, E.; Danjoux, C.; et al. Palliative Response and Functional Interference Outcomes Using the Brief Pain Inventory for Spinal Bony Metastases Treated with Conventional Radiotherapy. *Clinical Oncology* **2011**, *23*, 485–491, doi:10.1016/j.clon.2011.01.507.
54. El Ayachy, R.; Giraud, N.; Giraud, P.; Durdux, C.; Giraud, P.; Burgun, A.; Bibault, J.E. The Role of Radiomics in Lung Cancer: From Screening to Treatment and Follow-Up. *Front Oncol* **2021**, *11*, 603595, doi:10.3389/fonc.2021.603595.
55. Chow, E.; Hoskin, P.; Mitera, G.; Zeng, L.; Lutz, S.; Roos, D.; Hahn, C.; van der Linden, Y.; Hartsell, W.; Kumar, E.; et al. Update of the International Consensus on Palliative Radiotherapy Endpoints for Future Clinical Trials in Bone Metastases. *Int J Radiat Oncol Biol Phys* **2012**, *82*, 1730–1737, doi:10.1016/j.ijrobp.2011.02.008.
56. Hartsell, W.F.; Scott, C.B.; Bruner, D.W.; Scarantino, C.W.; Ivker, R.A.; Roach, M.; Suh, J.H.; Demas, W.F.; Movsas, B.; Petersen, I.A.; et al. Randomized Trial of Short- versus Long-Course Radiotherapy for Palliation of Painful Bone Metastases. *J Natl Cancer Inst* **2005**, *97*, 798–804, doi:10.1093/jnci/dji139.
57. Akezaki, Y.; Nakata, E.; Kikuuchi, M.; Sugihara, S.; Katayama, Y.; Katayama, H.; Hamada, M.; Ozaki, T. Factors Affecting the Quality of Life of Patients with Painful Spinal Bone Metastases. *Healthcare* **2021**, *9*, 1499, doi:10.3390/healthcare9111499.
58. Fu, M.R. Breast Cancer-Related Lymphedema: Symptoms, Diagnosis, Risk Reduction, and Management. *World J Clin Oncol* **2014**, *5*, 241–247, doi:10.5306/wjco.v5.i3.241.
59. Early Breast Cancer Trialists' Collaborative Group (EBCTCG); Darby, S.; McGale, P.; Correa, C.; Taylor, C.; Arriagada, R.; Clarke, M.; Cutter, D.; Davies, C.; Ewertz, M.; et al. Effect of Radiotherapy after Breast-Conserving Surgery on 10-Year Recurrence and 15-Year Breast Cancer Death: Meta-Analysis of Individual Patient Data for 10,801 Women in 17 Randomised Trials. *Lancet* **2011**, *378*, 1707–1716, doi:10.1016/S0140-6736(11)61629-2.
60. Darby, S.C.; Ewertz, M.; McGale, P.; Bennet, A.M.; Blom-Goldman, U.; Brønnum, D.; Correa, C.; Cutter, D.; Gagliardi, G.; Gigante, B.; et al. Risk of Ischemic Heart Disease in Women after Radiotherapy for Breast Cancer. *N Engl J Med* **2013**, *368*, 987–998, doi:10.1056/NEJMoa1209825.
61. Heins, M.J.; de Ligt, K.M.; Verloop, J.; Siesling, S.; Korevaar, J.C.; PSCCR group Adverse Health Effects after Breast Cancer up to 14 Years after Diagnosis. *Breast* **2022**, *61*, 22–28, doi:10.1016/j.breast.2021.12.001.
62. Lovelace, D.L.; McDaniel, L.R.; Golden, D. Long-Term Effects of Breast Cancer Surgery, Treatment, and Survivor Care. *J Midwifery Womens Health* **2019**, *64*, 713–724, doi:10.1111/jmwh.13012.
63. Suresh, R.; Raffi, J.; Yuen, F.; Murase, J.E. Treatment of Moist Desquamation for Patients Undergoing Radiotherapy. *Int J Womens Dermatol* **2019**, *5*, 124–125, doi:10.1016/j.ijwd.2018.12.002.
64. Verbelen, H.; Tjalma, W.; Dombrecht, D.; Gebruers, N. Breast Edema, from Diagnosis to Treatment: State of the Art. *Arch Physiother* **2021**, *11*, 8, doi:10.1186/s40945-021-00103-4.
65. Young-Afat, D.A.; Gregorowitsch, M.L.; van den Bongard, D.H.; Burgmans, I.; van der Pol, C.C.; Witkamp, A.J.; Bijlsma, R.M.; Koelemij, R.; Schoenmaeckers, E.J.; Jonasse, Y.; et al. Breast Edema Following Breast-Conserving Surgery and Radiotherapy: Patient-Reported Prevalence, Determinants, and Effect on Health-Related Quality of Life. *JNCI Cancer Spectr* **2019**, *3*, pkz011, doi:10.1093/jncics/pkz011.

66. Kenyon, M.; Mayer, D.K.; Owens, A.K. Late and Long-Term Effects of Breast Cancer Treatment and Surveillance Management for the General Practitioner. *J Obstet Gynecol Neonatal Nurs* **2014**, *43*, 382–398, doi:10.1111/1552-6909.12300.
67. Caplan, L. Delay in Breast Cancer: Implications for Stage at Diagnosis and Survival. *Front Public Health* **2014**, *2*, 87, doi:10.3389/fpubh.2014.00087.
68. Bentzen, S.M. Preventing or Reducing Late Side Effects of Radiation Therapy: Radiobiology Meets Molecular Pathology. *Nat Rev Cancer* **2006**, *6*, 702–713, doi:10.1038/nrc1950.
69. Kalra, M.K.; Maher, M.M.; Toth, T.L.; Hamberg, L.M.; Blake, M.A.; Shepard, J.-A.; Saini, S. Strategies for CT Radiation Dose Optimization. *Radiology* **2004**, *230*, 619–628, doi:10.1148/radiol.2303021726.
70. Mcrobbie, D. *Essentials of MRI Safety*; 2020; ISBN 978-1-119-55717-3.
71. Mayerhoefer, M.E.; Prosch, H.; Beer, L.; Tamandl, D.; Beyer, T.; Hoeller, C.; Berzaczy, D.; Raderer, M.; Preusser, M.; Hochmair, M.; et al. PET/MRI versus PET/CT in Oncology: A Prospective Single-Center Study of 330 Examinations Focusing on Implications for Patient Management and Cost Considerations. *Eur J Nucl Med Mol Imaging* **2020**, *47*, 51–60, doi:10.1007/s00259-019-04452-y.
72. Kumar, V.; Gu, Y.; Basu, S.; Berglund, A.; Eschrich, S.A.; Schabath, M.B.; Forster, K.; Aerts, H.J.; Dekker, A.; Fenstermacher, D.; et al. Radiomics: The Process and the Challenges. *Magn Reson Imaging* **2012**, *30*, 1234–1248, doi:10.1016/j.mri.2012.06.010.
73. Lambin, P.; Rios-Velazquez, E.; Leijenaar, R.; Carvalho, S.; van Stiphout, R.G.P.M.; Granton, P.; Zegers, C.M.L.; Gillies, R.; Boellard, R.; Dekker, A.; et al. Radiomics: Extracting More Information from Medical Images Using Advanced Feature Analysis. *Eur J Cancer* **2012**, *48*, 441–446, doi:10.1016/j.ejca.2011.11.036.
74. 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, doi:10.1038/nrclinonc.2017.141.
75. Zhang, W.; Guo, Y.; Jin, Q. Radiomics and Its Feature Selection: A Review. *Symmetry* **2023**, *15*, 1834, doi:10.3390/sym15101834.
76. Wu, Q.; Wang, S.; Chen, X.; Wang, Y.; Dong, L.; Liu, Z.; Tian, J.; Wang, M. Radiomics Analysis of Magnetic Resonance Imaging Improves Diagnostic Performance of Lymph Node Metastasis in Patients with Cervical Cancer. *Radiother Oncol* **2019**, *138*, 141–148, doi:10.1016/j.radonc.2019.04.035.
77. Bernstein, M.; Kovar, H.; Paulussen, M.; Randall, R.L.; Schuck, A.; Teot, L.A.; Juergens, H. Ewing's Sarcoma Family of Tumors: Current Management. *Oncologist* **2006**, *11*, 503–519, doi:10.1634/theoncologist.11-5-503.
78. Gielen, J.L.M.A.; De Schepper, A.M.; Vanhoenacker, F.; Parizel, P.M.; Wang, X.L.; Sciot, R.; Weyler, J. Accuracy of MRI in Characterization of Soft Tissue Tumors and Tumor-like Lesions. A Prospective Study in 548 Patients. *Eur Radiol* **2004**, *14*, 2320–2330, doi:10.1007/s00330-004-2431-0.
79. Riechelmann, R.P.; Krzyzanowska, M.K. Drug Interactions and Oncological Outcomes: A Hidden Adversary. *Ecancermedicalscience* **2019**, *13*, ed88, doi:10.3332/ecancer.2019.ed88.
80. Akbar, Z.; Rehman, S.; Khan, A.; Khan, A.; Atif, M.; Ahmad, N. Potential Drug–Drug Interactions in Patients with Cardiovascular Diseases: Findings from a Prospective Observational Study. *J Pharm Policy Pract* **2021**, *14*, 63, doi:10.1186/s40545-021-00348-1.
81. Hunter, B.; Hindocha, S.; Lee, R.W. The Role of Artificial Intelligence in Early Cancer Diagnosis. *Cancers (Basel)* **2022**, *14*, 1524, doi:10.3390/cancers14061524.

82. Yu, V.L.; Buchanan, B.G.; Shortliffe, E.H.; Wraith, S.M.; Davis, R.; Scott, A.C.; Cohen, S.N. Evaluating the Performance of a Computer-Based Consultant. *Comput Programs Biomed* **1979**, *9*, 95–102, doi:10.1016/0010-468x(79)90022-9.
83. Shortliffe, E. Computer-Based Medical Consultations: MYCIN. *Artificial Intelligence - AI* **1976**, *388*, doi:10.1097/00004669-197610000-00011.
84. Quinlan, J.R. Induction of Decision Trees. *Mach Learn* **1986**, *1*, 81–106, doi:10.1007/BF00116251.
85. Cortes, C.; Vapnik, V. Support-Vector Networks. *Mach Learn* **1995**, *20*, 273–297, doi:10.1007/BF00994018.
86. Statistical Learning Theory | Wiley Available online: https://www.wiley.com/en-us/Statistical+Learning+Theory-p-9780471030034 (accessed on 14 June 2024).
87. Breiman, L. Random Forests. *Machine Learning* **2001**, *45*, 5–32, doi:10.1023/A:1010933404324.
88. LeCun, Y.; Bengio, Y.; Hinton, G. Deep Learning. *Nature* **2015**, *521*, 436–444, doi:10.1038/nature14539.
89. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. *Commun. ACM* **2017**, *60*, 84–90, doi:10.1145/3065386.
90. Cox, D.R. Regression Models and Life-Tables. *Journal of the Royal Statistical Society: Series B (Methodological)* **1972**, *34*, 187–202, doi:10.1111/j.2517-6161.1972.tb00899.x.
91. Aerts, H.J.W.L.; Velazquez, E.R.; Leijenaar, R.T.H.; 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, doi:10.1038/ncomms5006.
92. Katzman, J.L.; Shaham, U.; Cloninger, A.; Bates, J.; Jiang, T.; Kluger, Y. DeepSurv: Personalized Treatment Recommender System Using a Cox Proportional Hazards Deep Neural Network. *BMC Medical Research Methodology* **2018**, *18*, 24, doi:10.1186/s12874-018-0482-1.
93. Llorián-Salvador, O.; Akhgar, J.; Pigorsch, S.; Borm, K.; Münch, S.; Bernhardt, D.; Rost, B.; Andrade-Navarro, M.; Combs, S.; Peeken, J. Machine Learning Based Prediction of Pain Response to Palliative Radiation Therapy - Is There a Role for Planning CT-Based Radiomics and Semantic Imaging Features? 2022.
94. Peeken, J.C.; Bernhofer, M.; Wiestler, B.; Goldberg, T.; Cremers, D.; Rost, B.; Wilkens, J.J.; Combs, S.E.; Nüsslin, F. Radiomics in Radiooncology - Challenging the Medical Physicist. *Physica Medica: European Journal of Medical Physics* **2018**, *48*, 27–36, doi:10.1016/j.ejmp.2018.03.012.
95. Peeken, J.C.; Spraker, M.B.; Knebel, C.; Dapper, H.; Pfeiffer, D.; Devecka, M.; Thamer, A.; Shouman, M.A.; Ott, A.; Eisenhart-Rothe, R. von; et al. Tumor Grading of Soft Tissue Sarcomas Using MRI-Based Radiomics. *eBioMedicine* **2019**, *48*, 332–340, doi:10.1016/j.ebiom.2019.08.059.
96. Peeken, J.C.; Asadpour, R.; Specht, K.; Chen, E.Y.; Klymenko, O.; Akinkuoroye, V.; Hippe, D.S.; Spraker, M.B.; Schaub, S.K.; Dapper, H.; et al. MRI-Based Delta-Radiomics Predicts Pathologic Complete Response in High-Grade Soft-Tissue Sarcoma Patients Treated with Neoadjuvant Therapy. *Radiotherapy and Oncology* **2021**, *164*, 73–82, doi:10.1016/j.radonc.2021.08.023.
97. Peeken, J.C.; Bernhofer, M.; Spraker, M.B.; Pfeiffer, D.; Devecka, M.; Thamer, A.; Shouman, M.A.; Ott, A.; Nüsslin, F.; Mayr, N.A.; et al. CT-Based Radiomic Features Predict Tumor Grading and Have Prognostic Value in Patients with Soft Tissue Sarcomas Treated with Neoadjuvant Radiation Therapy. *Radiotherapy and Oncology* **2019**, *135*, 187–196, doi:10.1016/j.radonc.2019.01.004.
98. Hatt, M.; Majdoub, M.; Vallières, M.; Tixier, F.; Rest, C.C.L.; Groheux, D.; Hindié, E.; Martineau, A.; Pradier, O.; Hustinx, R.; et al. 18F-FDG PET Uptake Characterization Through Texture Analysis: Investigating the Complementary Nature of Heterogeneity

and Functional Tumor Volume in a Multi–Cancer Site Patient Cohort. *Journal of Nuclear Medicine* **2015**, *56*, 38–44, doi:10.2967/jnumed.114.144055.
99. Welch, M.L.; McIntosh, C.; Haibe-Kains, B.; Milosevic, M.F.; Wee, L.; Dekker, A.; Huang, S.H.; Purdie, T.G.; O'Sullivan, B.; Aerts, H.J.W.L.; et al. Vulnerabilities of Radiomic Signature Development: The Need for Safeguards. *Radiotherapy and Oncology* **2019**, *130*, 2–9, doi:10.1016/j.radonc.2018.10.027.
100. Fedorov, A.; Beichel, R.; Kalpathy-Cramer, J.; Finet, J.; Fillion-Robin, J.-C.; Pujol, S.; Bauer, C.; Jennings, D.; Fennessy, F.; Sonka, M.; et al. 3D Slicer as an Image Computing Platform for the Quantitative Imaging Network. *Magnetic Resonance Imaging* **2012**, *30*, 1323–1341, doi:10.1016/j.mri.2012.05.001.
101. *A Practical Guide to Splines*;
102. 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.J.W.L. Computational Radiomics System to Decode the Radiographic Phenotype. *Cancer Research* **2017**, *77*, e104–e107, doi:10.1158/0008-5472.CAN-17-0339.
103. Zwanenburg, A.; Vallières, M.; Abdalah, M.A.; Aerts, H.J.W.L.; Andrearczyk, V.; Apte, A.; Ashrafinia, S.; Bakas, S.; Beukinga, R.J.; Boellaard, R.; et al. The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-Based Phenotyping. *Radiology* **2020**, *295*, 328–338, doi:10.1148/radiol.2020191145.
104. Gillies, R.J.; Kinahan, P.E.; Hricak, H. Radiomics: Images Are More than Pictures, They Are Data. *Radiology* **2016**, *278*, 563–577, doi:10.1148/radiol.2015151169.
105. Kappa Statistics for Attribute Agreement Analysis Available online: https://support.minitab.com/en-us/minitab/help-and-how-to/quality-and-process-improvement/measurement-system-analysis/how-to/attribute-agreement-analysis/attribute-agreement-analysis/interpret-the-results/all-statistics-and-graphs/kappa-statistics/ (accessed on 14 June 2024).
106. Freedman Professor, D.A.; Freedman Professor, D.A. A Note on Screening Regression Equations. *The American Statistician* **1983**, *37*, 152–155, doi:10.1080/00031305.1983.10482729.
107. Ramírez-Gallego, S.; Lastra, I.; Martínez-Rego, D.; Bolón-Canedo, V.; Benítez, J.M.; Herrera, F.; Alonso-Betanzos, A. Fast-mRMR: Fast Minimum Redundancy Maximum Relevance Algorithm for High-Dimensional Big Data: FAST-mRMR ALGORITHM FOR BIG DATA. *Int. J. Intell. Syst.* **2017**, *32*, 134–152, doi:10.1002/int.21833.
108. Hawkins, D.M. The Problem of Overfitting. *J Chem Inf Comput Sci* **2004**, *44*, 1–12, doi:10.1021/ci0342472.
109. Ying, X. An Overview of Overfitting and Its Solutions. *J. Phys.: Conf. Ser.* **2019**, *1168*, 022022, doi:10.1088/1742-6596/1168/2/022022.
110. Chawla, N.V. Data Mining for Imbalanced Datasets: An Overview. In *Data Mining and Knowledge Discovery Handbook*; Maimon, O., Rokach, L., Eds.; Springer US: Boston, MA, 2005; pp. 853–867 ISBN 978-0-387-25465-4.
111. Llorián-Salvador, Ó.; Martin, N.; Etzel, L.; Windeler, N.; Andrade, M.A.; Rost, B.; Combs, S.E.; Peeken, J.C. Insights from Radiomics: Predicting Breast Cancer Radiotherapy Side Effects.
112. Knebel, C.; Neumann, J.; Schwaiger, B.J.; Karampinos, D.C.; Pfeiffer, D.; Specht, K.; Lenze, U.; von Eisenhart-Rothe, R.; Rummeny, E.J.; Woertler, K.; et al. Differentiating Atypical Lipomatous Tumors from Lipomas with Magnetic Resonance Imaging: A Comparison with MDM2 Gene Amplification Status. *BMC Cancer* **2019**, *19*, 309, doi:10.1186/s12885-019-5524-5.
113. Leporq, B.; Bouhamama, A.; Pilleul, F.; Lame, F.; Bihane, C.; Sdika, M.; Blay, J.-Y.; Beuf, O. MRI-Based Radiomics to Predict Lipomatous Soft Tissue Tumors Malignancy: A Pilot Study. *Cancer Imaging* **2020**, *20*, 78, doi:10.1186/s40644-020-00354-7.

114. Cay, N.; Mendi, B.A.R.; Batur, H.; Erdogan, F. Discrimination of Lipoma from Atypical Lipomatous Tumor/Well-Differentiated Liposarcoma Using Magnetic Resonance Imaging Radiomics Combined with Machine Learning. *Jpn J Radiol* **2022**, *40*, 951–960, doi:10.1007/s11604-022-01278-x.
115. Vos, M.; Starmans, M.P.A.; Timbergen, M.J.M.; van der Voort, S.R.; Padmos, G.A.; Kessels, W.; Niessen, W.J.; van Leenders, G.J.L.H.; Grünhagen, D.J.; Sleijfer, S.; et al. Radiomics Approach to Distinguish between Well Differentiated Liposarcomas and Lipomas on MRI. *Br J Surg* **2019**, *106*, 1800–1809, doi:10.1002/bjs.11410.
116. Zeng, L.; Chow, E.; Zhang, L.; Culleton, S.; Holden, L.; Jon, F.; Khan, L.; Tsao, M.; Barnes, E.; Danjoux, C.; et al. Comparison of Pain Response and Functional Interference Outcomes between Spinal and Non-Spinal Bone Metastases Treated with Palliative Radiotherapy. *Support Care Cancer* **2012**, *20*, 633–639, doi:10.1007/s00520-011-1144-6.
117. Fisher, C.G.; Schouten, R.; Versteeg, A.L.; Boriani, S.; Varga, P.P.; Rhines, L.D.; Kawahara, N.; Fourney, D.; Weir, L.; Reynolds, J.J.; et al. Reliability of the Spinal Instability Neoplastic Score (SINS) among Radiation Oncologists: An Assessment of Instability Secondary to Spinal Metastases. *Radiat Oncol* **2014**, *9*, 69, doi:10.1186/1748-717X-9-69.
118. Trinh, D.-L.; Kim, S.-H.; Yang, H.-J.; Lee, G.-S. The Efficacy of Shape Radiomics and Deep Features for Glioblastoma Survival Prediction by Deep Learning. *Electronics* **2022**, *11*, 1038, doi:10.3390/electronics11071038.
119. Yap, F.Y.; Varghese, B.A.; Cen, S.Y.; Hwang, D.H.; Lei, X.; Desai, B.; Lau, C.; Yang, L.L.; Fullenkamp, A.J.; Hajian, S.; et al. Shape and Texture-Based Radiomics Signature on CT Effectively Discriminates Benign from Malignant Renal Masses. *Eur Radiol* **2021**, *31*, 1011–1021, doi:10.1007/s00330-020-07158-0.
120. Ludwig, C.G.; Lauric, A.; Malek, J.A.; Mulligan, R.; Malek, A.M. Performance of Radiomics Derived Morphological Features for Prediction of Aneurysm Rupture Status. *Journal of NeuroInterventional Surgery* **2021**, *13*, 755–761, doi:10.1136/neurintsurg-2020-016808.
121. Zhang, X.; Zhang, Y.; Zhang, G.; Qiu, X.; Tan, W.; Yin, X.; Liao, L. Deep Learning With Radiomics for Disease Diagnosis and Treatment: Challenges and Potential. *Front. Oncol.* **2022**, *12*, doi:10.3389/fonc.2022.773840.
122. Gentili, C.; Sanfilippo, O.; Silvestrini, R. Cell Proliferation and Its Relationship to Clinical Features and Relapse in Breast Cancers. *Cancer* **1981**, *48*, 974–979, doi:10.1002/1097-0142(19810815)48:4<974::AID-CNCR2820480420>3.0.CO;2-#.
123. Dent, R.; Trudeau, M.; Pritchard, K.I.; Hanna, W.M.; Kahn, H.K.; Sawka, C.A.; Lickley, L.A.; Rawlinson, E.; Sun, P.; Narod, S.A. Triple-Negative Breast Cancer: Clinical Features and Patterns of Recurrence. *Clinical Cancer Research* **2007**, *13*, 4429–4434, doi:10.1158/1078-0432.CCR-06-3045.