Radiomics-based machine learning classification of bone chondrosarcoma

Gitto, S.

RADIOMICS-BASED MACHINE LEARNING CLASSIFICATION OF BONE CHONDROSARCOMA

Salvatore Gitto

ISBN: 978-94-6419-438-8
Cover design by XERIOS
Printed by GILDEPRINT

This doctoral thesis was carried out jointly (co-tutelle) at the Radiology Department of the Leiden University Medical Center in The Netherlands and the Department of Biomedical Sciences for Health of the University of Milan (*Università degli Studi di Milano*) in Italy. The doctoral thesis includes studies partially supported by the 2020 Young Researchers Grant awarded by the European Society of Musculoskeletal Radiology.

Radiomics-based Machine Learning Classification of Bone Chondrosarcoma

PROEFSCHRIFT

ter verkrijging van
de graad van doctor aan de Universiteit Leiden,
op gezag van de rector magnificus prof. dr. ir. H. Bijl,
volgens besluit van het college voor promoties
te verdedigen op woensdag 16 februari 2022
klokke 15.00 uur

door

Salvatore Gitto

geboren te Messina (Italië)
in 1990

Promotores

Prof. dr. J.L. Bloem

Prof. dr. L.M. Sconfienza (University of Milan)

Leden promotiecommissie

Prof. dr. J.V.M.G. Bovée

Prof. dr. P. Erba (University of Pisa and University Medical Center Groningen)

Dr. K. van Langevelde

Prof. dr. M. Maas (Amsterdam University Medical Center)

Prof. dr. M.A.J. van de Sande

This thesis is dedicated to my parents

For their constant encouragement and invaluable support

Contents

Chapter 1

Introduction

List of abbreviations (Chapter 1)

ACT, atypical cartilaginous tumor

CT, computed tomography

CS, chondrosarcoma

MRI, magnetic resonance imaging

PET-CT, positron emission tomography-computed tomography

1.1 Bone chondrosarcoma: incidence and current definitions

Chondrosarcomas (CSs) are a heterogenous group of cartilage-forming lesions and account for 20-30% of primary bone tumors in adulthood [1]. In most cases, they arise *de novo* from the medullary cavity and are referred as primary central conventional CS. Less commonly, they are secondary to malignant transformation of benign lesions, such as osteochondroma (secondary peripheral CS) or enchondroma (secondary central CS) [2].

Based upon histopathology, conventional CSs are grouped into low to high grades, namely atypical cartilaginous tumor (ACT), CS grade I – and grades II-III [3]. The latter are malignant lesions with metastatic potential and high recurrence rates after surgery [3]. In the 2020 edition of the World Health Organization classification, the term ACT is reserved for low grade lesions located in long bones, reflecting their relatively indolent clinical behavior with unlikelihood to metastasize [4]. The incidence of ACTs has increased over the last decades mainly due to an increase in incidental findings on diagnostic imaging [5]. Cartilaginous tumors with the same histology as ACT, but located in the axial skeleton, including pelvis and skull base, are classified as CS grade I [4]. A fifth group of CS is called dedifferentiated CS, which could be seen as grade IV [3]. Two less common histotypes are mesenchymal and clear cell CSs [3], the first is highly malignant with strong tendency towards recurrence, while the second is low-grade [2]. Clinical outcome strongly depends on tumor grading, as reported 5- and 10-year overall survival rates are 87-99% and 88-95% for ACT/CS grade I, 74-99% and 58-86% for CS grade II, 31-77% and 26-55% for CS grade III, respectively [4, 5]. Overall 5-year survival rates of 7-24% have been reported in dedifferentiated or grade IV CS [4].

1.2 Bone chondrosarcoma: therapeutic strategies and diagnosis

In long bones, ACTs can be managed with marginal resection or curettage (with or without local adjuvant, such as phenol, cement, and cryotherapy) for sufficient local control [6]. The increased incidence of ACT secondary to an increase in diagnostic imaging, relative to the lack of increase in CS grades II-III, does not support the previous opinion that ACTs are at risk of dedifferentiation into high-grade lesions [5]. Also the effectiveness of curettage in preventing ACT transformation into high-grade CS has not been demonstrated. Thus, therapeutic strategy is currently shifting towards conservative approach (watchful waiting) in order to prevent overtreatment and morbidity associated with surgery [5].

Surgical excision with wide margins is the current standard of care for grade II or higher CS of long bones and for all-grade CS of the axial skeleton, pelvis and shoulder girdle [6]. CS is usually not sensitive to both radiotherapy and chemotherapy. Radiotherapy can be used for local control after incomplete resection with curative intent or for palliation if resection is not an option. Particularly, proton beam radiotherapy is used as alternative or in combination with surgery to achieve local control in CS located in skull base or sacrum. Mesenchymal CS is more sensitive to chemotherapy, which is therefore considered for adjuvant or neoadjuvant therapy [6].

As clinical management is now very different, the main challenge is to discriminate ACT from high-grade CS. Preoperative biopsy suffers from sample errors [7] and discrepancies in tumor grading even among specialized bone tumor pathologists [8]. Imaging has added substantially to our ability to differentiate between these tumors preoperatively. Particularly, magnetic resonance imaging (MRI) is the method of choice for local staging, and computed tomography (CT) and positron emission tomography-CT (PET-CT) are employed for general staging, respectively [6]. However, low reliability in tumor grading has been reported even among expert observers [9, 10]. Thus, new imaging-based methods like radiomics may improve our ability to better diagnose and grade cartilaginous bone tumors more objectively [11].

1.3 Radiomics and machine learning

The term “radiomics” derives from a combination of “radio”, which refers to medical images, and “omics”, which indicates the analysis of high amounts of data representing an entire set of some kind, like genome (genomics) and proteome (proteomics) [12]. Radiomics includes the extraction and analysis of large numbers of quantitative parameters, known as radiomic features, from medical images [13]. Radiomics has recently gained much attention in oncologic imaging and, to date, studies have focused on discriminating tumor grades and types before treatment, monitoring response to therapy, and predicting patients’ outcome [14]. The primary purpose of radiomics is to extract as much and meaningful quantitative information as possible to be used in decision support. This is also known as precision medicine, where patients that belong to different subtypes can be identified to improve their outcomes [12].

Due to its ever-growing high-dimensional nature consisting of numerous quantitative features, radiomics needs powerful analytic tools and artificial intelligence perfectly addresses this issue. Artificial intelligence comprises a broad set of systems that accurately perform inferences from large amounts of data based on computational algorithms, namely the machine learning algorithms. These algorithms learn the administered data by analyzing patterns (on the training dataset) and then make predictions on unseen dataset based on previously acquired information (test dataset) [12]. Ideally, if radiomic data could be derived from routine imaging studies, radiologists would have quantitative information that integrates qualitative assessment, helping diagnosis and outcome prediction, therefore directing clinicians towards a tailored therapeutic approach.

Despite its great potential as a non-invasive biomarker to quantify several tumor characteristics, radiomics still faces difficulties to clinical implementation [14]. Particularly, the assessment of feature reproducibility and the use of independent (external test dataset) clinical validation are two main challenges and currently lacking in most radiomic studies, as emerged in a recent systematic review we conducted on musculoskeletal sarcomas [15]. Thus, these issues need to be addressed to facilitate application and clinical transferability of radiomic models.

1.4 Objectives and outline of the thesis

The primary objective of this thesis is to determine diagnostic performance of radiomics-based machine learning in differentiating ACT from high-grade CS. The secondary objectives are: (i) to address the issue of segmentation variability and identify subsets of reproducible, robust radiomic features of cartilaginous bone tumors; and (ii) to compare the performance of radiomics-based machine learning with experienced musculoskeletal oncology radiologists.

In chapter 2, the concept of radiomics of musculoskeletal sarcomas is introduced and a systematic review on radiomic feature reproducibility and validation strategies is conducted. In chapter 3, a preliminary study is performed to investigate the performance of MRI radiomics-based machine learning in discriminating ACT from high-grade CS, using a single-center cohort, in comparison with an expert radiologist. In chapter 4, the influence of interobserver segmentation variability on the reproducibility of CT and MRI radiomic features of cartilaginous bone tumors is assessed. In chapter 5, the performance of CT

radiomics-based machine learning in discriminating ACT from high-grade CS of long bones is determined and validated using independent data from a multicenter cohort, compared to an expert radiologist. In chapter 6, the performance of MRI radiomics-based machine learning in differentiating between ACT and grade II CS of long bones is determined and validated using independent data from a multicenter cohort, in comparison with an expert radiologist. Finally, in chapter 7, the main results and implications of this thesis are summarized and discussed.

References

[1] Murphey MD, Walker EA, Wilson AJ, Kransdorf MJ, Temple HT, Gannon FH. From the archives of the AFIP: imaging of primary chondrosarcoma: radiologic-pathologic correlation. Radiographics 2003;23:1245–78.

[2] Gelderblom H, Hogendoorn PCW, Dijkstra SD, van Rijswijk CS, Krol AD, Taminiau AHM, et al. The Clinical Approach Towards Chondrosarcoma. Oncologist 2008;13:320–9.

[3] Fletcher CDM, Bridge JA, Hogendoorn PCW, Mertens F. World Health Organization Classification of Tumours of Soft Tissue and Bone. Lyon, France: International Agency for Research on Cancer Press; 2013.

[4] WHO Classification of Tumours Editorial Board. WHO Classification of Tumours: Soft Tissue and Bone Tumours. Lyon, France: International Agency for Research on Cancer Press; 2020.

[5] van Praag (Veroniek) VM, Rueten-Budde AJ, Ho V, Dijkstra PDS, Fiocco M, van de Sande MAJ, et al. Incidence, outcomes and prognostic factors during 25 years of treatment of chondrosarcomas. Surg Oncol 2018;27:402–8.

[6] Casali PG, Bielack S, Abecassis N, Aro HT, Bauer S, Biagini R, et al. Bone sarcomas: ESMO–PaedCan–EURACAN Clinical Practice Guidelines for diagnosis, treatment and follow-up. Ann Oncol 2018;29:iv79–95.

[7] Hodel S, Laux C, Farei-Campagna J, Götschi T, Bode-Lesniewska B, Müller DA. The impact of biopsy sampling errors and the quality of surgical margins on local recurrence and survival in chondrosarcoma. Cancer Manag Res 2018;10:3765–71.

[8] Eefting D, Schrage YM, Geirnaerdt MJA, Le Cessie S, Taminiau AHM, Bovée JVMG, et al. Assessment of Interobserver Variability and Histologic Parameters to Improve Reliability in Classification and Grading of Central Cartilaginous Tumors. Am J Surg Pathol 2009;33:50–7.

[9] Jones KB, Buckwalter JA, McCarthy EF, DeYoung BR, El-Khoury GY, Dolan L, et al. Reliability of Histopathologic and Radiologic Grading of Cartilaginous Neoplasms in Long Bones. J Bone Joint Surg Am 2007;89:2113–23.

[10] Zamora T, Urrutia J, Schweitzer D, Amenabar PP, Botello E. Do Orthopaedic Oncologists Agree on the Diagnosis and Treatment of Cartilage Tumors of the Appendicular Skeleton? Clin Orthop Relat Res 2017;475:2176–86.

[11] van de Sande MAJ, van der Wal RJP, Navas Cañete A, van Rijswijk CSP, Kroon HM, Dijkstra PDS, et al. Radiologic differentiation of enchondromas, atypical cartilaginous tumors, and high-grade chondrosarcomas—Improving tumor-specific treatment: A paradigm in transit? Cancer 2019;125:3288–91.

[12] Kocak B, Durmaz ES, Ates E, Kilickesmez O. Radiomics with artificial intelligence: a practical guide for beginners. Diagn Interv Radiol 2019;25:485–95.

[13] Gillies RJ, Kinahan PE, Hricak H. Radiomics: Images Are More than Pictures, They Are Data. Radiology 2016;278:563–77.

[14] Lubner MG, Smith AD, Sandrasegaran K, Sahani D V., Pickhardt PJ. CT Texture Analysis: Definitions, Applications, Biologic Correlates, and Challenges. Radiographics 2017;37:1483–503.

[15] Gitto S, Cuocolo R, Albano D, Morelli F, Pescatori LC, Messina C, et al. CT and MRI radiomics of bone and soft-tissue sarcomas: a systematic review of reproducibility and validation strategies. Insights Imaging 2021;12:68.

Chapter 2

CT and MRI radiomics of bone and soft-tissue sarcomas: a systematic review of reproducibility and validation strategies

Gitto S, Cuocolo R, Albano D, Morelli F, Pescatori LC, Messina C, Imbriaco M, Sconfienza LM

Insights Imaging 2021; 12:68
DOI: 10.1186/s13244-021-01008-3

List of abbreviations (Chapter 2)

CT, computed tomography

ICC, intraclass correlation coefficient

MRI, magnetic resonance imaging

PRISMA, Preferred Reporting Items for Systematic reviews and Meta-Analyses

Abstract

Background. Feature reproducibility and model validation are two main challenges of radiomics. This study aims to systematically review radiomic feature reproducibility and predictive model validation strategies in studies dealing with CT and MRI radiomics of bone and soft-tissue sarcomas. The ultimate goal is to promote achieving a consensus on these aspects in radiomic workflows and facilitate clinical transferability.

Results. Out of 278 identified papers, forty-nine papers published between 2008 and 2020 were included. They dealt with radiomics of bone (n=12) or soft-tissue (n=37) tumors. Eighteen (37%) studies included a feature reproducibility analysis. Inter/intra-reader segmentation variability was the theme of reproducibility analysis in 16 (33%) investigations, outnumbering the analyses focused on image acquisition or post-processing (n=2, 4%). The intraclass correlation coefficient was the most commonly used statistical method to assess reproducibility, which ranged from 0.6 and 0.9. At least one machine learning validation technique was used for model development in 25 (51%) papers and K-fold cross validation was the most commonly employed. A clinical validation of the model was reported in 19 (39%) papers. It was performed using a separate dataset from the primary institution (i.e., internal validation) in 14 (29%) studies and an independent dataset related to different scanners or from another institution (i.e., independent validation) in 5 (10%) studies.

Conclusions. The issues of radiomic feature reproducibility and model validation varied largely among the studies dealing with musculoskeletal sarcomas and should be addressed in future investigations to bring the field of radiomics from a preclinical research area to the clinical stage.

2.1 Background

Bone and soft-tissue primary malignant tumors or sarcomas are rare entities with several histological subtypes, and each has an incidence < 1/100,000/year [1, 2]. Among them, osteosarcoma is the most common sarcoma of the bone. Along with Ewing sarcoma, it has a higher incidence in the second decade of life, while chondrosarcoma is the most prevalent bone sarcoma in adulthood [1]. The most frequent soft-tissue sarcomas are liposarcoma and leiomyosarcoma [2]. Due to the rarity of these diseases, bone and soft-tissue sarcomas are managed in tertiary sarcoma centers according to current guidelines [1, 2]. Both biopsy and imaging integrate clinical data prior to the beginning of any treatment, with the former representing the reference standard for preoperative diagnosis [1, 2]. However, biopsy may be inaccurate in large, heterogenous tumors due to sampling errors and, in turn, inaccurate diagnosis may lead to inadequate treatment and subsequent need for further interventions, with increased morbidity. Additionally, the risk of biopsy tract contamination remains a concern. Imaging already plays a pivotal role in the assessment of bone and soft-tissue sarcomas. Magnetic resonance imaging (MRI) and computed tomography (CT) are employed for local and general staging, respectively [1, 2]. These modalities may certainly benefit from new imaging-based tools such as those based on radiomics, which may potentially provide additional information regarding both diagnosis and prognosis non-invasively [3].

The term “radiomics” derives from a combination of “radio”, referring to medical images, and “omics”, which indicates the analysis of high amounts of data representing an entire set of some kind, like genome (genomics) and proteome (proteomics) [3]. Therefore, “radiomics” includes extraction and analysis of large numbers of quantitative parameters, known as radiomic features, from medical images [4]. This technique has recently gained much attention in oncologic imaging as it can potentially quantify tumor heterogeneity, which can be challenging to capture by means of qualitative imaging assessment or sampling biopsies. Particularly, radiomic studies to date have focused on discriminating tumor grades and types before treatment, monitoring response to therapy and predicting outcome [5].

Despite its great potential as a non-invasive tumor biomarker, radiomics still faces challenges preventing its clinical implementation. Two main initiatives have addressed methodological issues of radiomic studies to bridge the gap between academic endeavors and real-life application. In 2017, Lambin et al. proposed the Radiomics Quality Score that

details the sequential steps to follow in radiomic pipelines and offers a tool to asses methodological rigor in their implementation [6]. In 2020, the Image Biomarkers Standardization Initiative produced and validated reference values for radiomic features, which enable verification and calibration of different software for radiomic feature extraction [7]. However, numerous challenges still remain to ensure clinical transferability of radiomics. As radiomics is essentially a two-step approach consisting of data extraction and analysis, in the first step (i.e., data extraction), the main challenge is reproducibility of radiomic features, which can be influenced by image acquisition parameters, region of interest segmentation technique and image post-processing technique [8, 9]. In the second step (i.e., data analysis), models can be built upon either conventional statistical methods or machine learning algorithms with the aim of predicting the diagnosis or outcome of interest. In either case, the main challenge is model validation [9].

The challenges of reproducibility and validation strategies in radiomics have been recently addressed in a review focusing on renal masses [10]. The aim of our study is to systematically review radiomic feature reproducibility and predictive model validation strategies in studies dealing with CT and MRI radiomics of bone and soft-tissue sarcomas. The ultimate goal is to promote and facilitate achieving a consensus on these aspects in radiomic workflows.

2.2 Methods

2.2.1 Reviewers

No Local Ethics Committee approval was needed for this systematic review. Literature search, study selection, and data extraction were performed independently by two recently-boarded radiologists with experience in musculoskeletal tumors and radiomics (S.G. and F.M.). In case of disagreement, agreement was achieved by consensus of these two readers and a third reviewer with radiology specialty and doctorate in artificial intelligence and radiomics (R.C.). The Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) guidelines [11] were followed.

2.2.2 Literature search

An electronic literature search was conducted on EMBASE (Elsevier) and PubMed (MEDLINE, U.S. National Library of Medicine and National Institutes of Health) databases

for articles published up to 31st December 2020 and dealing with CT and MRI radiomics of bone and soft-tissue sarcomas. A controlled vocabulary was adopted using medical subject headings in PubMed and the thesaurus in EMBASE. Search syntax was built by combining search terms related to two main domains, namely "musculoskeletal sarcomas" and "radiomics". The exact search query was: ("sarcoma"/exp OR "sarcoma") AND ("radiomics"/exp OR "radiomics" OR "texture"/exp OR "texture"). Studies were first screened by title and abstract, and then the full text of eligible studies was retrieved for further review. The references of selected publications were checked for additional publications to include.

2.2.3 Inclusion and exclusion criteria

Inclusion criteria were: (i) original research papers published in peer-reviewed journals; (ii) focus on CT or MRI radiomics-based characterization of sarcomas located in bone and soft-tissues for either diagnosis- or prognosis-related tasks; (iii) statement that local ethics committee approval was obtained, or ethical standards of the institutional or national research committee were followed.

Exclusion criteria were: (i) papers not dealing with mass characterization, such as those focused on computer-assisted diagnosis and detection systems; (ii) papers dealing with head and neck, retroperitoneal or visceral sarcomas; (iii) animal, cadaveric or laboratory studies; (iv) papers not written in English language.

2.2.4 Data extraction

Data were extracted to a spreadsheet with a drop-down list for each item, as defined by the first author, grouped into three main categories, namely baseline study characteristics, radiomic feature reproducibility strategies and predictive model validation strategies. Items regarding baseline study characteristics included first author's last name, year of publication, study aim, tumor type, study design, reference standard, imaging modality, database size, use of public data, segmentation process, and segmentation style. Those concerning radiomic feature reproducibility strategies included reproducibility assessment based on repeated segmentations, reproducibility assessment related to acquisition or post-processing techniques, statistical method used for reproducibility analysis, and cut-off or threshold used for reproducibility analysis. Finally, data regarding predictive model validation strategies

included the use of machine learning validation techniques, clinical validation performed on a separate internal dataset, and clinical validation performed on an external or independent dataset.

2.3 Results

2.3.1 Baseline study characteristics

A flowchart illustrating the literature search process is presented in Figure 1. After screening 278 papers and applying our eligibility criteria, 49 papers were included in this systematic review. Tables 1 and 2 detail the characteristics of papers dealing with radiomics of bone (n=12) and soft-tissue (n=37) tumors, respectively.

All studies were published between 2008 and 2020. Twenty-three out of 49 investigations (47%) were published in 2020, 14 (29%) in 2019, 4 (8%) in 2018 and 8 (16%) between 2008 and 2017. The design was prospective in 6 studies (12%) and retrospective in the remaining 43 (88%). The imaging modality of choice was MRI in 42 (86%), including one or multiple MRI sequences, and CT in 7 (14%) cases. The median size of the database was 60 patients (range, 19-226). Public data were used only in 3 (6%) studies.

The research was aimed at predicting either diagnosis or prognosis, as follows: benign vs. malignant tumor discrimination (n=14); grading (n=10); tumor histotype discrimination (n=4); proliferation index Ki67 expression (n=1); survival (n=12); response to therapy, either chemotherapy or radiotherapy (n=8); local and/or metastatic relapse (n=9). It should be noted that the aim was twofold in some studies, as detailed in Tables 1 and 2. In those focused on diagnosis-related tasks, including benign vs. malignant discrimination, grading, tumor histotype discrimination and proliferation index expression, histology was the reference standard in all cases excepting benign lesions diagnosed on the basis of stable imaging findings over time in two papers [12, 13]. In studies focused on prediction of response to chemotherapy or radiotherapy, the reference standard was histology if lesions were surgically treated, based on the percentage of viable tumor and necrosis relative to the surgical tissue specimen, or consistent imaging findings if lesions were not operated. In studies focused on prediction of tumor relapse, the diagnosis was based on histology or consistent imaging findings, as the reference standard. In studies dealing with survival prediction, survival was assessed based on follow-up.

Regarding segmentation, the process was performed manually in 45 (92%) studies and semiautomatically in 4 (8%) studies. In no case the segmentation process was fully automated. The following segmentation styles were identified: 2D without multiple sampling in 11 (23%) studies; 2D with multiple sampling in 3 (6%); 3D in 35 (71%). Of note, a single slice showing maximum tumor extension was chosen in all studies employing 2D segmentation without multiple sampling, excepting one case [14] where it was chosen based on signal intensity homogeneity.

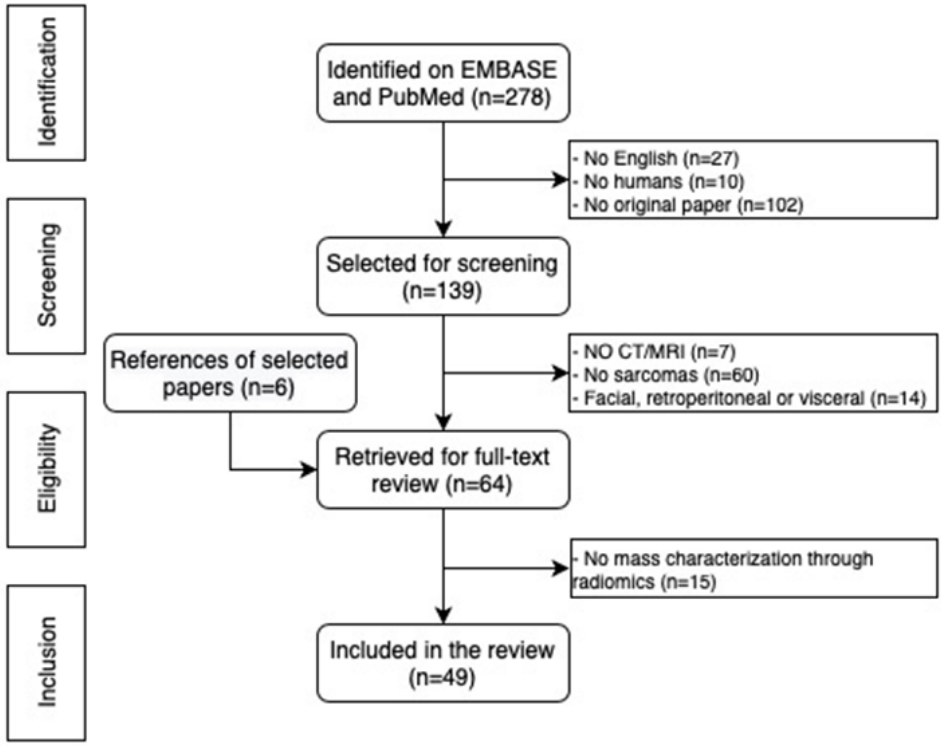


Fig. 1 PRISMA (preferred reporting items for systematic reviews and meta-analyses) flowchart of systematic identification, screening, eligibility and inclusion information from retrieved studies.

Table 1 Characteristics of the papers dealing with bone sarcomas included in the systematic review. MS, multiple sampling.

1st author	Year	Aim	Tumor	Design	Reference standard	Modality	Database size (n)	Public data	Segmentation	
									Process	Style
Baidya Kayal [50]	2020	Therapy response	Osteosarcoma	Prospective	Histology	MRI	32	No	Manual	3D
Chen [29]	2020	Local relapse Metastatic relapse	Osteosarcoma	Retrospective	Histology Imaging	MRI	93	No	Manual	2D without MS
Dai [45]	2020	Histotype	Ewing sarcoma Osteosarcoma	Retrospective	Histology	MRI	66	No	Manual	2D without MS
Fritz [13]	2018	Benign vs malignant Grading	Chondroma Chondrosarcoma	Retrospective	Histology Imaging	MRI	116	No	Manual	2D without MS
Gitto [46]	2020	Grading	Chondrosarcoma	Retrospective	Histology	MRI	58	No	Manual	2D without MS
Li [32]	2019	Histotype	Chondrosarcoma Chordoma	Retrospective	Histology	MRI	210	No	Manual	3D
Lin [15]	2020	Therapy response	Osteosarcoma	Retrospective	Histology	CT	191	No	Manual	3D
Lisson [12]	2018	Benign vs malignant	Chondroma Chondrosarcoma	Retrospective	Histology Imaging	MRI	22	No	Semiautomatic	3D
Wu [16]	2018	Survival	Osteosarcoma	Retrospective	Follow-up	CT	150	No	Manual	3D
Yin [38]	2020	Local relapse	Chondrosarcoma	Retrospective	Histology Imaging	MRI	103	No	Manual	3D
Yin [28]	2019	Histotype	Chordoma Giant cell tumor	Retrospective	Histology	CT	95	No	Manual	3D
Zhao [22]	2019	Survival	Osteosarcoma	Retrospective	Follow-up	MRI	112	No	Manual	3D

Table 2 Characteristics of the papers dealing with soft-tissue sarcomas included in the systematic review. MS, multiple sampling (continued on the next page).

1st author	Year	Aim	Tumor	Design	Reference standard	Modality	Database size (n)	Public data	Segmentation	
									Process	Style
Corino [52]	2018	Grading	Multiple sarcoma histotypes	Retrospective	Histology	MRI	19	No	Manual	3D
Crombé [39]	2020	Metastatic relapse Survival	Liposarcoma	Retrospective	Histology Follow-up	MRI	35	No	Manual	3D
Crombé [42]	2020	Metastatic relapse Survival	Multiple sarcoma histotypes	Retrospective	Histology Follow-up	MRI	50	No	Manual	3D
Crombé [30]	2019	Therapy response	Multiple sarcoma histotypes	Prospective	Histology Imaging	MRI	25	No	Manual	3D
Crombé [47]	2019	Therapy response	Multiple sarcoma histotypes	Retrospective	Histology	MRI	65	No	Manual	3D
Crombé [63]	2020	Therapy response Survival	Liposarcoma	Retrospective	Histology Follow-up	MRI	21	No	Manual	3D
Crombé [36]	2020	Therapy response Survival	Desmoid tumor	Retrospective	Imaging Follow-up	MRI	42	No	Manual	3D
Crombé [31]	2020	Metastatic relapse Survival	Multiple sarcoma histotypes	Retrospective	Histology Follow-up	MRI	70	No	Manual	3D
Gao [33]	2020	Therapy response	Multiple sarcoma histotypes	Prospective	Histology	MRI	30	No	Manual	3D
Hayano [21]	2015	Survival	Multiple sarcoma histotypes	Prospective	Follow-up	CT	20	No	Manual	2D without MS
Hong [64]	2020	Grading	Multiple sarcoma histotypes	Retrospective	Histology	MRI	42	No	Manual	3D
Juntu [48]	2010	Benign vs malignant	Multiple benign/ malignant histotypes	Retrospective	Histology	MRI	135	No	Manual	2D with MS
Kim [65]	2017	Benign vs malignant	Multiple benign/ malignant histotypes	Retrospective	Histology	MRI	40	No	Manual	3D
Leporq [19]	2020	Benign vs malignant	Lipoma Liposarcoma	Retrospective	Histology	MRI	81	No	Manual	2D without MS
Malinauskaite [23]	2020	Benign vs malignant	Lipoma Liposarcoma	Retrospective	Histology	MRI	38	No	Semiautomatic	3D
Martin-Carreras [34]	2019	Benign vs malignant	Myxoma Myxofibrosarcoma	Retrospective	Histology	MRI	56	No	Manual	3D
Mayerhoefer [57]	2008	Benign vs malignant	Multiple benign/ malignant histotypes Non-tumoral lesions	Retrospective	Histology	MRI	58	No	Manual	2D with MS
Meyer [66]	2019	Proliferation index	Multiple sarcoma histotypes	Retrospective	Histology	MRI	29	No	Manual	2D without MS

Table 2 (continued) Characteristics of the papers dealing with soft-tissue sarcomas included in the systematic review. MS, multiple sampling.

Peeken [25]	2019	Grading Survival	Multiple sarcoma histotypes	Retrospective	Histology Follow-up	MRI	225	No	Manual	3D
Peeken [27]	2019	Grading Survival	Multiple sarcoma histotypes	Retrospective	Histology Follow-up	CT	221	Yes	Manual	3D
Pressney [14]	2020	Benign vs malignant	Lipoma Liposarcoma	Retrospective	Histology	MRI	60	No	Manual	2D without MS
Spraker [51]	2019	Survival	Multiple sarcoma histotypes	Retrospective	Follow-up	MRI	226	No	Manual	3D
Tagliafico [26]	2019	Local relapse	Multiple sarcoma histotypes	Prospective	Histology Imaging	MRI	19	No	Manual	2D with MS
Thornhill [49]	2014	Benign vs malignant	Lipoma Liposarcoma	Retrospective	Histology	MRI	44	No	Semiautomatic	3D
Tian [35]	2020	Metastatic relapse	n/a	Retrospective	Histology Imaging	MRI	77	No	Manual	3D
Tian [67]	2015	Therapy response Survival	Multiple sarcoma histotypes	Prospective	Histology Follow-up	CT	20	No	Manual	2D without MS
Timbergen [20]	2020	Histotype	Desmoid tumor Multiple sarcoma histotypes	Retrospective	Histology	MRI	203	No	Manual	3D
Vallières [55]	2015	Metastatic relapse	Multiple sarcoma histotypes	Retrospective	Histology Imaging	MRI	51	Yes	Manual	3D
Vallières [68]	2017	Metastatic relapse	Multiple sarcoma histotypes	Retrospective	Histology Imaging	MRI	30	Yes	Manual	3D
Vos [40]	2019	Benign vs malignant	Lipoma Liposarcoma	Retrospective	Histology	MRI	116	No	Semiautomatic	3D
Wang [41]	2020	Grading	Multiple sarcoma histotypes	Retrospective	Histology	MRI	113	No	Manual	3D
Wang [43]	2020	Benign vs malignant	Multiple benign/ malignant histotypes	Retrospective	Histology	MRI	206	No	Manual	3D
Wang [24]	2020	Benign vs malignant	Multiple benign/ malignant histotypes	Retrospective	Histology	MRI	91	No	Manual	3D
Wu [18]	2020	Benign vs malignant	Multiple benign/ malignant histotypes	Retrospective	Histology	CT	49	No	Manual	2D without MS
Xiang [17]	2019	Grading	Multiple sarcoma histotypes	Retrospective	Histology	MRI	67	No	Manual	2D without MS
Xu [37]	2020	Grading	Multiple sarcoma histotypes	Retrospective	Histology	MRI	105	No	Manual	3D
Zhang [44]	2019	Grading	Multiple sarcoma histotypes	Retrospective	Histology	MRI	37	No	Manual	3D

2.3.2 Reproducibility strategies

Eighteen (37%) of the 49 studies included a reproducibility analysis of the radiomic features in their workflow. In 16 (33%) investigations [13, 15–29], the reproducibility of radiomic features was assessed on the basis of repeated segmentations performed by different readers and/or the same reader at different time points. Two (4%) studies presented an analysis to assess the reproducibility based on different acquisition [30] or post-processing [31] techniques. Of note, segmentations were validated by a second experienced reader in 15 studies [12, 32–45] without however addressing the issue of radiomic feature reproducibility.

The intraclass correlation coefficient (ICC) was the statistical method used in most of the papers reporting a reproducibility analysis [13, 15–18, 20, 22–25, 27–29, 31]. ICC threshold ranged between 0.6 [13] and 0.9 [22] for reproducible features. The following statistical methods were used less commonly: analysis of variance [30, 31]; Cronbach alpha statistic [26]; Pearson correlation coefficient [19] and Spearman correlation coefficient [21].

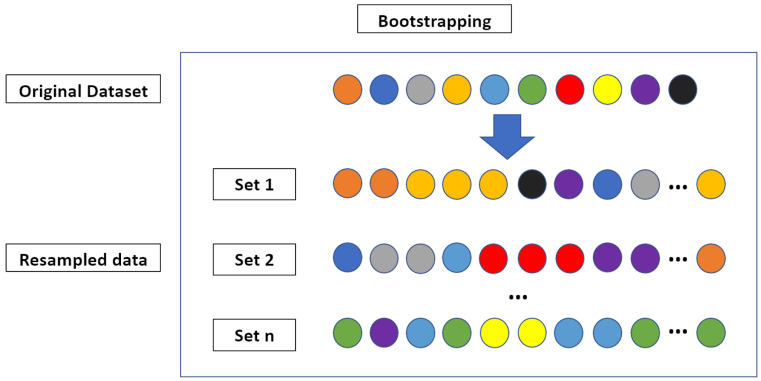
2.3.3 Validation strategies

At least one machine learning validation technique was used in 25 (51%) of the 49 papers. K-fold cross validation was used in most of the studies [13, 25, 28, 31–33, 37, 38, 40, 43, 44, 46–50]. The following machine learning validation techniques were used less commonly: bootstrapping [42, 51]; leave-one-out cross validation [34, 35, 41]; leave-p-out cross validation [52]; Monte Carlo cross validation [23]; nested cross validation [25, 27]; random-split cross validation [20]. Figure 2 provides an overview of machine learning validation techniques. Figure 3 illustrates an example of a radiomics-based machine learning pipeline.

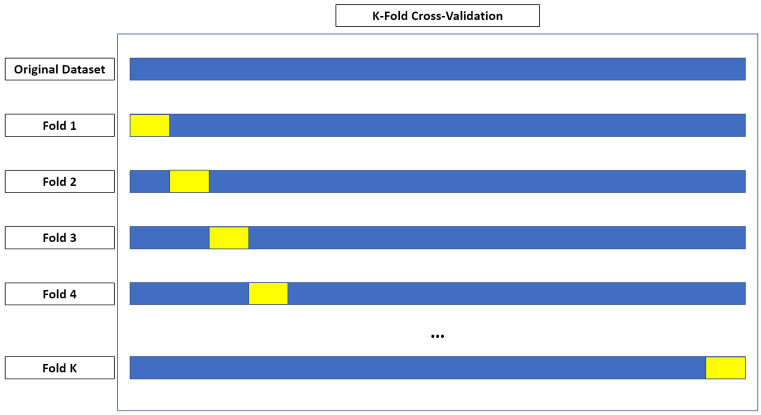
2.3.4 Clinical validation

A clinical validation of the radiomics-based prediction model was reported in 19 (39%) of the 49 papers. It was performed on a separate set of data from the primary institution, i.e. internal test set, in 14 (29%) studies [15, 16, 22, 24, 28, 31, 32, 35, 37, 38, 41, 46, 47, 52]. It was performed on an independent set of data from the primary institution (related to a different scanner) or from an external institution, i.e. external test set, in 5 (10%) studies [25, 27, 29, 43, 51].

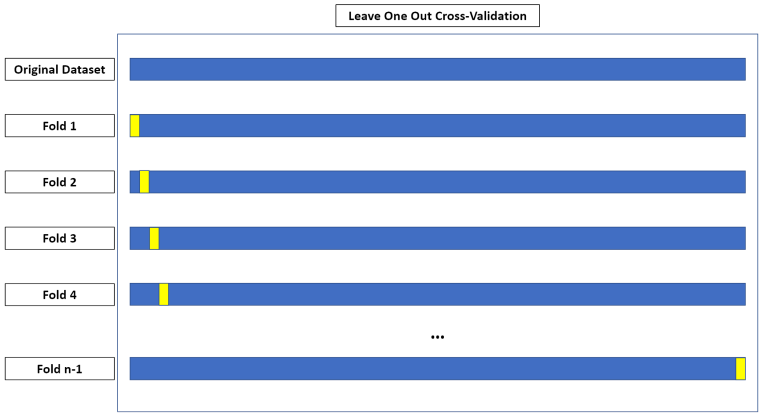
(2a)


Bootstrapping
Original Dataset
Resampled data
Set 1
Set 2
Set n

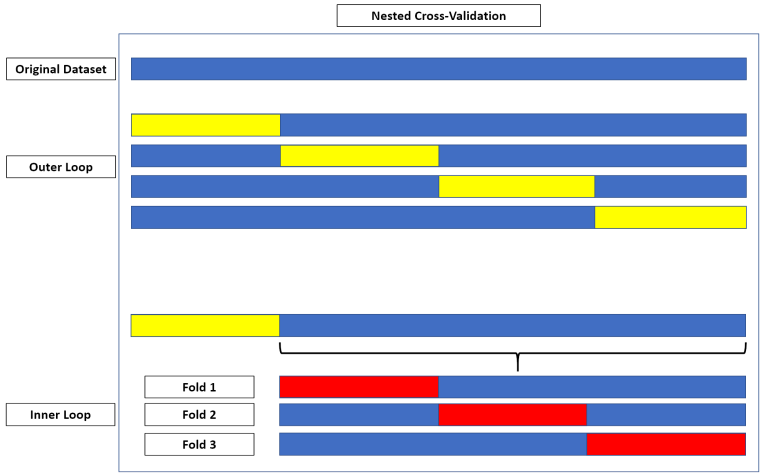
(2b)


K-Fold Cross-Validation
Original Dataset
Fold 1
Fold 2
Fold 3
Fold 4
Fold K

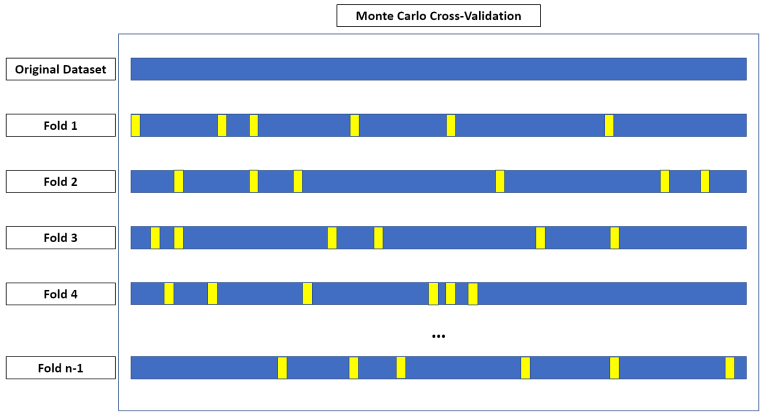
(2c)


Leave One Out Cross-Validation
Original Dataset
Fold 1
Fold 2
Fold 3
Fold 4
Fold n-1

(2d)


Nested Cross-Validation
Original Dataset
Outer Loop
Inner Loop
Fold 1
Fold 2
Fold 3

(2e)


Monte Carlo Cross-Validation
Original Dataset
Fold 1
Fold 2
Fold 3
Fold 4
...
Fold n-1

(2f)


Random Split Cross-Validation
Original Dataset
Fold
Fold 2
Fold 3
Fold 4
...
Fold n-1

(2g)

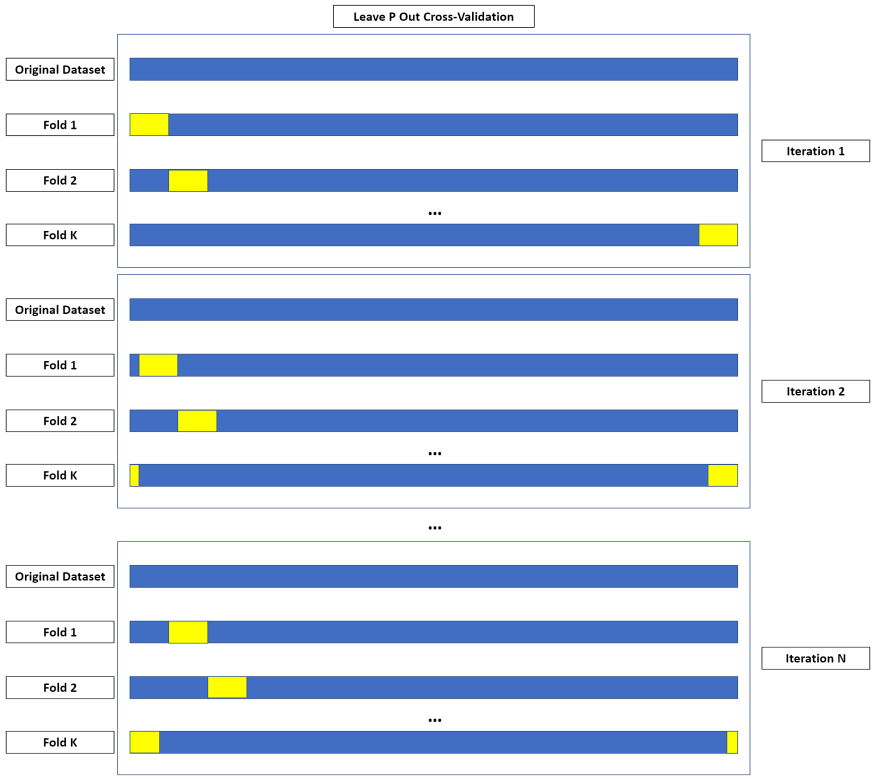


Fig. 2 Overview of machine learning validation techniques. **(a)** Bootstrapping is based on resampling with replacement, allowing to create n datasets from an original sample. These may include any number of copies of a specific instance from the original case, even none. **(b)** K-fold cross-validation is based on dividing the dataset in k parts, using each in turn as the validation set and the remaining as the training data. **(c)** In leave-one-out cross-validation, each instance in the dataset is used for model validation, using the remaining for model training. **(d)** In nested cross-validation, two loops of validation take place. The training data from each outer loop undergoes an additional K-fold cross-validation. The figure depicts a 4-fold outer loop paired with a 3-fold inner loop. In **(e)** Monte Carlo and **(f)** random split cross-validation, the folds are not made up of contiguous data but from random sampling of the entire dataset. During the first, a sample may appear in multiple folds, which is not possible in random split cross-validation. **(g)** In leave-P-out cross validation, the K-fold cross validation process is iterated to obtain all possible folding splits for the data.

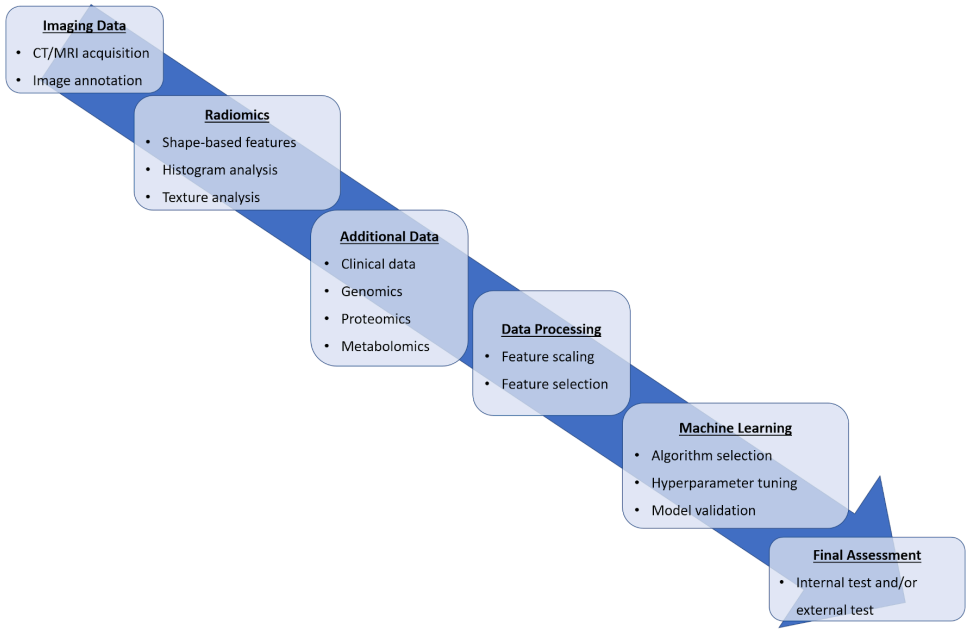


Fig. 3 Example of a radiomics-based machine learning pipeline, listing the most commonly employed steps in an ideal order of execution.

2.4 Discussion

This systematic review focused on the radiomics literature regarding MRI and CT of bone and soft-tissue sarcomas with particular emphasis on reproducibility and validation strategies. The number of papers reporting the assessment of radiomic feature reproducibility and the use of independent or external clinical validation was relatively small. This finding is in line with recent literature reviews showing that the quality of sarcoma radiomics studies is low [53, 54], which may hamper performance generalizability of radiomic models on independent cohorts and, consequently, their practical application [53]. Thus, these issues need to be addressed in the radiomic workflow of future studies to facilitate clinical transferability.

2.4.1 Baseline study characteristics

MRI and CT radiomics of bone and soft-tissue sarcomas has progressively gained attention in musculoskeletal and oncologic imaging. The number of papers has rapidly increased over the recent years, and almost half of those included in our review (47%) was published in 2020. Radiomics was used in attempt to answer clinical questions related to both

diagnosis and prognosis of musculoskeletal sarcomas. Most studies (88%) were retrospective in nature, as this design allowed including relatively large number of patients with imaging data already available and bone or soft-tissue sarcomas, which are rare diseases. A prospective analysis, while not strictly necessary in radiomic studies [5], may however have advantages for controlling data gathering in reproducibility assessment and matching certain patient or imaging characteristics in independent datasets. Public data were used in no study regarding bone sarcomas and in a small proportion of the studies (6%) concerning soft-tissue sarcomas. A public database [55] available on The Cancer Imaging Archive (https://www.cancerimagingarchive.net) was used in all these studies. Public databases afford opportunities for researchers who do not have sufficient data at their institution and allow research groups from around the world to test and compare new radiomic methods using common data. Thus, research employing radiomics in this field would certainly be enhanced if further imaging databases are made publicly available in the near future.

Regarding segmentation, the process was performed manually in most of the studies (92%) and semiautomatically in the remaining, both requiring human intervention to some extent. Even though the influence of inter-observer and/or intra-observer variability on the reproducibility of radiomic features can be assessed as part of the radiomic workflow, fully automated segmentation algorithms would ideally achieve higher reliability and deserve future investigation. Annotations included the entire lesion volume (3D segmentation) in most of the studies (71%) and a single slice (2D), without multiple sampling, in the remaining (23%). However, to date no study has compared the outcome of 2D and 3D segmentations in musculoskeletal sarcomas. As 2D annotations are time saving and have recently proven higher performance than 3D segmentation in oropharyngeal cancers [56], this should represent another area of research in the near future. Of note, a limited number of studies (6%) used a 2D segmentation style with multiple sampling as a data augmentation technique to increase the number of labeled slices [26, 48, 57]. This practice can be useful for an uncommon entity as musculoskeletal sarcomas but should be employed with care to avoid the introduction of bias in the final model. The inclusion of samples from the same case in both the training and test sets could lead to overly optimistic results.

2.4.2 Reproducibility strategies

A great variability in radiomic features has emerged as a major issue across studies and attributed to different segmentation, image acquisition and post-processing approaches [4]. Therefore, methodological analyses are advisable prior to conducting radiomic studies in order to assess feature robustness and avoid biases due to non-reproducible, noisy features. This concept is in line with recent literature emphasizing the importance of reproducibility in artificial intelligence and radiology [58]. In our review, we noted that about one third of the included papers described a reproducibility analysis in their workflow. In this subgroup of papers, inter- and/or intra-reader segmentation variability was the main focus of the reproducibility analysis. Segmentation variability-related analyses outnumbered those addressing reproducibility issues due to image acquisition or post-processing differences, which were reported in one paper per each [30, 31]. This finding underlines that further research should deal with dependencies of radiomic features on image acquisition and post-processing. While these analyses may already be performed in retrospective series, when patients underwent more than one study in a short interval, prospective studies could facilitate the identification of reliable radiomic features within this domain. Finally, ICC was the statistical method used in most of the papers evaluating radiomic feature reproducibility. Of note, guidelines for performing and assessing ICC are available and can be followed to achieve consensus on the cut-off and threshold values [59].

2.4.3 Validation strategies

Proper validation of radiomic models is highly desirable to bridge the gap between concepts and clinical application [53]. Machine learning validation techniques are employed to avoid any information leak from the test to the training set during model development [60]. Resampling strategies can be extremely useful, especially with relatively limited samples of data, which may not be truly representative for the population of interest, with the aim of reducing overfitting and better estimating the performance of the radiomics-based predictive model on new data (i.e., the test set) [61, 62]. K-fold cross validation was the most commonly used technique for this task in the studies included in this review.

Ideally, in both prospective and retrospective studies, a clinical validation of the model is performed against completely independent sets of data, i.e., the external or independent test set [4]. We found that clinical validation was performed against an

independent dataset from the primary institution (using different scanners) or from a different institution only in a small number of studies (10%) included in this systematic review. More studies (29%) validated the model using a separate set of data from the primary institution, i.e., an internal test set. Therefore, future studies should be carried out in more than one institution and include external testing of the model with large and independent sets of data.

2.5 Limitations and conclusions

This study is limited to a systematic review of the literature, and no meta-analysis was performed due to the lack of homogeneity between studies in terms of objectives and subgroups of sarcoma with a rather limited number of papers per each objective and subgroup. Different metrics were also used, preventing us from providing an estimation of model performance for each objective. Furthermore, it was outside of the scope of the review to perform a formal assessment of the quality of each included study, as our focus was on reporting methodological data that can be in and of themselves quality indicators. Limitations notwithstanding, we reviewed the radiomics literature regarding bone and soft-tissue sarcomas with emphasis on the methodologic issues of feature reproducibility and predictive model validation. They varied largely among the included studies, and, in particular, no reproducibility analysis was provided in more than half the papers. Additionally, less than half the studies included a clinical validation and only 10% used an independent dataset for this purpose. Thus, in order to bring the field of radiomics from a preclinical research area to the clinical stage, both these issues should be addressed in future studies dealing with musculoskeletal sarcomas.

References

1. Casali PG, Bielack S, Abecassis N, et al (2018) Bone sarcomas: ESMO–PaedCan–EURACAN Clinical Practice Guidelines for diagnosis, treatment and follow-up. Ann Oncol 29:iv79–iv95. https://doi.org/10.1093/annonc/mdy310
2. Casali PG, Abecassis N, Bauer S, et al (2018) Soft tissue and visceral sarcomas: ESMO–EURACAN Clinical Practice Guidelines for diagnosis, treatment and follow-up. Ann Oncol 29:iv51–iv67. https://doi.org/10.1093/annonc/mdy096
3. Kocak B, Durmaz ES, Ates E, Kilickesmez O (2019) Radiomics with artificial intelligence: a practical guide for beginners. Diagn Interv Radiol 25:485–495. https://doi.org/10.5152/dir.2019.19321
4. Gillies RJ, Kinahan PE, Hricak H (2016) Radiomics: Images Are More than Pictures, They Are Data. Radiology 278:563–577. https://doi.org/10.1148/radiol.2015151169
5. Lubner MG, Smith AD, Sandrasegaran K, et al (2017) CT Texture Analysis: Definitions, Applications, Biologic Correlates, and Challenges. Radiographics 37:1483–1503. https://doi.org/10.1148/rg.2017170056
6. Lambin P, Leijenaar RTH, Deist TM, et al (2017) Radiomics: the bridge between medical imaging and personalized medicine. Nat Rev Clin Oncol 14:749–762. https://doi.org/10.1038/nrclinonc.2017.141
7. Zwanenburg A, Vallières M, Abdalah MA, et al (2020) The image biomarker standardization initiative: Standardized quantitative radiomics for high-throughput image-based phenotyping. Radiology. https://doi.org/10.1148/radiol.2020191145
8. Varghese BA, Cen SY, Hwang DH, Duddalwar VA (2019) Texture Analysis of Imaging: What Radiologists Need to Know. AJR Am J Roentgenol 212:520–528. https://doi.org/10.2214/AJR.18.20624
9. Traverso A, Wee L, Dekker A, Gillies R (2018) Repeatability and Reproducibility of Radiomic Features: A Systematic Review. Int J Radiat Oncol Biol Phys 102:1143–1158. https://doi.org/10.1016/j.ijrobp.2018.05.053
10. Kocak B, Durmaz ES, Erdim C, et al (2020) Radiomics of Renal Masses: Systematic Review of Reproducibility and Validation Strategies. AJR Am J Roentgenol 214:129–136. https://doi.org/10.2214/AJR.19.21709
11. Moher D, Liberati A, Tetzlaff J, Altman DG (2009) Preferred Reporting Items for

Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Med 6:e1000097. https://doi.org/10.1371/journal.pmed.1000097

12. Lisson CS, Lisson CG, Flosdorf K, et al (2018) Diagnostic value of MRI-based 3D texture analysis for tissue characterisation and discrimination of low-grade chondrosarcoma from enchondroma: a pilot study. Eur Radiol 28:468–477. https://doi.org/10.1007/s00330-017-5014-6
13. Fritz B, Müller DA, Sutter R, et al (2018) Magnetic Resonance Imaging–Based Grading of Cartilaginous Bone Tumors. Invest Radiol 53:663–672. https://doi.org/10.1097/RLI.0000000000000486
14. Pressney I, Khoo M, Endozo R, et al (2020) Pilot study to differentiate lipoma from atypical lipomatous tumour/well-differentiated liposarcoma using MR radiomics-based texture analysis. Skeletal Radiol 49:1719–1729. https://doi.org/10.1007/s00256-020-03454-4
15. Lin P, Yang P-F, Chen S, et al (2020) A Delta-radiomics model for preoperative evaluation of Neoadjuvant chemotherapy response in high-grade osteosarcoma. Cancer Imaging 20:7. https://doi.org/10.1186/s40644-019-0283-8
16. Wu Y, Xu L, Yang P, et al (2018) Survival Prediction in High-grade Osteosarcoma Using Radiomics of Diagnostic Computed Tomography. EBioMedicine 34:27–34. https://doi.org/10.1016/j.ebiom.2018.07.006
17. Xiang P, Zhang X, Liu D, et al (2019) Distinguishing soft tissue sarcomas of different histologic grades based on quantitative MR assessment of intratumoral heterogeneity. Eur J Radiol 118:194–199. https://doi.org/10.1016/j.ejrad.2019.07.028
18. Wu G, Xie R, Li Y, et al (2020) Histogram analysis with computed tomography angiography for discriminating soft tissue sarcoma from benign soft tissue tumor. Medicine (Baltimore) 99:e18742. https://doi.org/10.1097/MD.0000000000018742
19. Leporq B, Bouhamama A, Pilleul F, et al (2020) MRI-based radiomics to predict lipomatous soft tissue tumors malignancy: a pilot study. Cancer Imaging 20:78. https://doi.org/10.1186/s40644-020-00354-7
20. Timbergen MJM, Starmans MPA, Padmos GA, et al (2020) Differential diagnosis and mutation stratification of desmoid-type fibromatosis on MRI using radiomics. Eur J Radiol 131:109266. https://doi.org/10.1016/j.ejrad.2020.109266

21. Hayano K, Tian F, Kambadakone AR, et al (2015) Texture Analysis of Non–Contrast-Enhanced Computed Tomography for Assessing Angiogenesis and Survival of Soft Tissue Sarcoma. J Comput Assist Tomogr 39:607–612. https://doi.org/10.1097/RCT.0000000000000239
22. Zhao S, Su Y, Duan J, et al (2019) Radiomics signature extracted from diffusion-weighted magnetic resonance imaging predicts outcomes in osteosarcoma. J Bone Oncol 19:100263. https://doi.org/10.1016/j.jbo.2019.100263
23. Malinauskaite I, Hofmeister J, Burgermeister S, et al (2020) Radiomics and Machine Learning Differentiate Soft-Tissue Lipoma and Liposarcoma Better than Musculoskeletal Radiologists. Sarcoma 2020:1–9. https://doi.org/10.1155/2020/7163453
24. Wang H, Nie P, Wang Y, et al (2020) Radiomics nomogram for differentiating between benign and malignant soft-tissue masses of the extremities. J Magn Reson Imaging 51:155–163. https://doi.org/10.1002/jmri.26818
25. Peeken JC, Spraker MB, Knebel C, et al (2019) Tumor grading of soft tissue sarcomas using MRI-based radiomics. EBioMedicine 48:332–340. https://doi.org/10.1016/j.ebiom.2019.08.059
26. Tagliafico AS, Bignotti B, Rossi F, et al (2019) Local recurrence of soft tissue sarcoma: a radiomic analysis. Radiol Oncol 53:300–306. https://doi.org/10.2478/raon-2019-0041
27. Peeken JC, Bernhofer M, Spraker MB, et al (2019) CT-based radiomic features predict tumor grading and have prognostic value in patients with soft tissue sarcomas treated with neoadjuvant radiation therapy. Radiother Oncol 135:187–196. https://doi.org/10.1016/j.radonc.2019.01.004
28. Yin P, Mao N, Zhao C, et al (2019) Comparison of radiomics machine-learning classifiers and feature selection for differentiation of sacral chordoma and sacral giant cell tumour based on 3D computed tomography features. Eur Radiol 29:1841–1847. https://doi.org/10.1007/s00330-018-5730-6
29. Chen H, Liu J, Cheng Z, et al (2020) Development and external validation of an MRI-based radiomics nomogram for pretreatment prediction for early relapse in osteosarcoma: A retrospective multicenter study. Eur J Radiol 129:109066. https://doi.org/10.1016/j.ejrad.2020.109066

30. Crombé A, Saut O, Guigui J, et al (2019) Influence of temporal parameters of DCE-MRI on the quantification of heterogeneity in tumor vascularization. J Magn Reson Imaging 50:1773–1788. https://doi.org/10.1002/jmri.26753
31. Crombé A, Kind M, Fadli D, et al (2020) Intensity harmonization techniques influence radiomics features and radiomics-based predictions in sarcoma patients. Sci Rep 10:15496. https://doi.org/10.1038/s41598-020-72535-0
32. Li L, Wang K, Ma X, et al (2019) Radiomic analysis of multiparametric magnetic resonance imaging for differentiating skull base chordoma and chondrosarcoma. Eur J Radiol 118:81–87. https://doi.org/10.1016/j.ejrad.2019.07.006
33. Gao Y, Kalbasi A, Hsu W, et al (2020) Treatment effect prediction for sarcoma patients treated with preoperative radiotherapy using radiomics features from longitudinal diffusion-weighted MRIs. Phys Med Biol 65:175006. https://doi.org/10.1088/1361-6560/ab9e58
34. Martin-Carreras T, Li H, Cooper K, et al (2019) Radiomic features from MRI distinguish myxomas from myxofibrosarcomas. BMC Med Imaging 19:67. https://doi.org/10.1186/s12880-019-0366-9
35. Tian L, Zhang D, Bao S, et al (2021) Radiomics-based machine-learning method for prediction of distant metastasis from soft-tissue sarcomas. Clin Radiol 76:158.e19-158.e25. https://doi.org/10.1016/j.crad.2020.08.038
36. Crombé A, Kind M, Ray-Coquard I, et al (2020) Progressive Desmoid Tumor: Radiomics Compared With Conventional Response Criteria for Predicting Progression During Systemic Therapy—A Multicenter Study by the French Sarcoma Group. AJR Am J Roentgenol 215:1539–1548. https://doi.org/10.2214/AJR.19.22635
37. Xu W, Hao D, Hou F, et al (2020) Soft Tissue Sarcoma: Preoperative MRI-Based Radiomics and Machine Learning May Be Accurate Predictors of Histopathologic Grade. AJR Am J Roentgenol 215:963–969. https://doi.org/10.2214/AJR.19.22147
38. Yin P, Mao N, Liu X, et al (2020) Can clinical radiomics nomogram based on 3D multiparametric MRI features and clinical characteristics estimate early recurrence of pelvic chondrosarcoma? J Magn Reson Imaging 51:435–445. https://doi.org/10.1002/jmri.26834
39. Crombé A, Le Loarer F, Sitbon M, et al (2020) Can radiomics improve the

prediction of metastatic relapse of myxoid/round cell liposarcomas? Eur Radiol 30:2413–2424. https://doi.org/10.1007/s00330-019-06562-5

40. Vos M, Starmans MPA, Timbergen MJM, et al (2019) Radiomics approach to distinguish between well differentiated liposarcomas and lipomas on MRI. Br J Surg 106:1800–1809. https://doi.org/10.1002/bjs.11410
41. Wang H, Chen H, Duan S, et al (2020) Radiomics and Machine Learning With Multiparametric Preoperative MRI May Accurately Predict the Histopathological Grades of Soft Tissue Sarcomas. J Magn Reson Imaging 51:791–797. https://doi.org/10.1002/jmri.26901
42. Crombé A, Fadli D, Buy X, et al (2020) High-Grade Soft-Tissue Sarcomas: Can Optimizing Dynamic Contrast-Enhanced MRI Postprocessing Improve Prognostic Radiomics Models? J Magn Reson Imaging 52:282–297. https://doi.org/10.1002/jmri.27040
43. Wang H, Zhang J, Bao S, et al (2020) Preoperative MRI-Based Radiomic Machine-Learning Nomogram May Accurately Distinguish Between Benign and Malignant Soft-Tissue Lesions: A Two-Center Study. J Magn Reson Imaging 52:873–882. https://doi.org/10.1002/jmri.27111
44. Zhang Y, Zhu Y, Shi X, et al (2019) Soft Tissue Sarcomas: Preoperative Predictive Histopathological Grading Based on Radiomics of MRI. Acad Radiol 26:1262–1268. https://doi.org/10.1016/j.acra.2018.09.025
45. Dai Y, Yin P, Mao N, et al (2020) Differentiation of Pelvic Osteosarcoma and Ewing Sarcoma Using Radiomic Analysis Based on T2-Weighted Images and Contrast-Enhanced T1-Weighted Images. Biomed Res Int 2020:9078603. https://doi.org/10.1155/2020/9078603
46. Gitto S, Cuocolo R, Albano D, et al (2020) MRI radiomics-based machine-learning classification of bone chondrosarcoma. Eur J Radiol 128:109043. https://doi.org/10.1016/j.ejrad.2020.109043
47. Crombé A, Périer C, Kind M, et al (2019) T 2 -based MRI Delta-radiomics improve response prediction in soft-tissue sarcomas treated by neoadjuvant chemotherapy. J Magn Reson Imaging 50:497–510. https://doi.org/10.1002/jmri.26589
48. Juntu J, Sijbers J, De Backer S, et al (2010) Machine learning study of several

classifiers trained with texture analysis features to differentiate benign from malignant soft-tissue tumors in T1-MRI images. J Magn Reson Imaging 31:680–689. https://doi.org/10.1002/jmri.22095

49. Thornhill RE, Golfam M, Sheikh A, et al (2014) Differentiation of Lipoma From Liposarcoma on MRI Using Texture and Shape Analysis. Acad Radiol 21:1185–1194. https://doi.org/10.1016/j.acra.2014.04.005
50. Baidya Kayal E, Kandasamy D, Khare K, et al (2021) Texture analysis for chemotherapy response evaluation in osteosarcoma using MR imaging. NMR Biomed 34:1–17. https://doi.org/10.1002/nbm.4426
51. Spraker MB, Wootton LS, Hippe DS, et al (2019) MRI Radiomic Features Are Independently Associated With Overall Survival in Soft Tissue Sarcoma. Adv Radiat Oncol 4:413–421. https://doi.org/10.1016/j.adro.2019.02.003
52. Corino VDA, Montin E, Messina A, et al (2018) Radiomic analysis of soft tissues sarcomas can distinguish intermediate from high-grade lesions. J Magn Reson Imaging 47:829–840. https://doi.org/10.1002/jmri.25791
53. Crombé A, Fadli D, Italiano A, et al (2020) Systematic review of sarcomas radiomics studies: Bridging the gap between concepts and clinical applications? Eur J Radiol 132:109283. https://doi.org/10.1016/j.ejrad.2020.109283
54. Zhong J, Hu Y, Si L, et al (2021) A systematic review of radiomics in osteosarcoma: utilizing radiomics quality score as a tool promoting clinical translation. Eur Radiol 31:1526–1535. https://doi.org/10.1007/s00330-020-07221-w
55. Vallières M, Freeman CR, Skamene SR, El Naqa I (2015) A radiomics model from joint FDG-PET and MRI texture features for the prediction of lung metastases in soft-tissue sarcomas of the extremities. Phys Med Biol 60:5471–5496. https://doi.org/10.1088/0031-9155/60/14/5471
56. Ren J, Yuan Y, Qi M, Tao X (2020) Machine learning–based CT texture analysis to predict HPV status in oropharyngeal squamous cell carcinoma: comparison of 2D and 3D segmentation. Eur Radiol 30:6858–6866. https://doi.org/10.1007/s00330-020-07011-4
57. Mayerhoefer ME, Breitenseher M, Amann G, Dominkus M (2008) Are signal intensity and homogeneity useful parameters for distinguishing between benign and

malignant soft tissue masses on MR images? Magn Reson Imaging 26:1316–1322. https://doi.org/10.1016/j.mri.2008.02.013

58. Mongan J, Moy L, Kahn CE (2020) Checklist for Artificial Intelligence in Medical Imaging (CLAIM): A Guide for Authors and Reviewers. Radiol Artif Intell 2:e200029. https://doi.org/10.1148/ryai.2020200029
59. Koo TK, Li MY (2016) A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. J Chiropr Med 15:155–163. https://doi.org/10.1016/j.jcm.2016.02.012
60. Chianca V, Cuocolo R, Gitto S, et al (2021) Radiomic Machine Learning Classifiers in Spine Bone Tumors: A Multi-Software, Multi-Scanner Study. Eur J Radiol. https://doi.org/10.1016/j.ejrad.2021.109586
61. Cuocolo R, Caruso M, Perillo T, et al (2020) Machine Learning in oncology: A clinical appraisal. Cancer Lett 481:55–62. https://doi.org/10.1016/j.canlet.2020.03.032
62. Parmar C, Barry JD, Hosny A, et al (2018) Data Analysis Strategies in Medical Imaging. Clin Cancer Res 24:3492–3499. https://doi.org/10.1158/1078-0432.CCR-18-0385
63. Crombe A, Sitbon M, Stoeckle E, et al (2020) Magnetic resonance imaging assessment of chemotherapy-related adipocytic maturation in myxoid/round cell liposarcomas: specificity and prognostic value. Br J Radiol 93:20190794. https://doi.org/10.1259/bjr.20190794
64. Hong JH, Jee W-H, Jung C-K, Chung Y-G (2020) Tumor grade in soft-tissue sarcoma. Medicine (Baltimore) 99:e20880. https://doi.org/10.1097/MD.0000000000020880
65. Kim HS, Kim J-H, Yoon YC, Choe BK (2017) Tumor spatial heterogeneity in myxoid-containing soft tissue using texture analysis of diffusion-weighted MRI. PLoS One 12:e0181339. https://doi.org/10.1371/journal.pone.0181339
66. Meyer H-J, Renatus K, Höhn AK, et al (2019) Texture analysis parameters derived from T1-and T2-weighted magnetic resonance images can reflect Ki67 index in soft tissue sarcoma. Surg Oncol 30:92–97. https://doi.org/10.1016/j.suronc.2019.06.006
67. Tian F, Hayano K, Kambadakone AR, Sahani D V. (2015) Response assessment to

neoadjuvant therapy in soft tissue sarcomas: using CT texture analysis in comparison to tumor size, density, and perfusion. Abdom Imaging 40:1705–1712. https://doi.org/10.1007/s00261-014-0318-3

68. Vallières M, Laberge S, Diamant A, El Naqa I (2017) Enhancement of multimodality texture-based prediction models via optimization of PET and MR image acquisition protocols: a proof of concept. Phys Med Biol 62:8536–8565. https://doi.org/10.1088/1361-6560/aa8a49

Chapter 3

MRI radiomics-based machine-learning classification of bone chondrosarcoma

Gitto S, Cuocolo R, Albano D, Chianca V, Messina C, Gambino A, Ugga L, Cortese MC, Lazzara A, Ricci D, Spairani R, Zanchetta E, Luzzati A, Brunetti A, Parafioriti A, Sconfienza LM

Eur J Radiol 2020; 128:109043
DOI: 10.1016/j.ejrad.2020.109043

List of abbreviations (Chapter 3)

MRI, magnetic resonance imaging

ROC, receiver operator characteristic

ROI, region of interest

Abstract

Purpose. To evaluate the diagnostic performance of machine learning for discrimination between low-grade and high-grade cartilaginous bone tumors based on radiomic parameters extracted from unenhanced Magnetic Resonance Imaging (MRI).

Methods. We retrospectively enrolled 58 patients with histologically-proven low-grade/atypical cartilaginous tumor of the appendicular skeleton (n=26) or higher-grade chondrosarcoma (n=32, including 16 appendicular and 16 axial lesions). They were randomly divided into training (n=42) and test (n=16) groups for model tuning and testing, respectively. All tumors were manually segmented on T1-weighted and T2-weighted images by drawing bidimensional regions of interest, which were used for first order and texture feature extraction. A Random Forest wrapper was employed for feature selection. The resulting dataset was used to train a locally weighted ensemble classifier (AdaboostM1). Its performance was assessed via 10-fold cross-validation on the training data and then on the previously unseen test set. Thereafter, an experienced musculoskeletal radiologist blinded to histological and radiomic data qualitatively evaluated the cartilaginous tumors in the test group.

Results. After feature selection, the dataset was reduced to 4 features extracted from T1-weighted images. AdaboostM1 correctly classified 85.7% and 75% of the lesions in the training and test groups, respectively. The corresponding areas under the Receiver Operating Characteristic curve were 0.85 and 0.78. The radiologist correctly graded 81.3% of the lesions. There was no significant difference in performance between the radiologist and machine learning classifier (P=0.453).

Conclusions. Our machine learning approach showed good diagnostic performance for classification of low-to-high grade cartilaginous bone tumors and could prove a valuable aid in preoperative tumor characterization.

3.1 Introduction

Chondrosarcoma accounts for approximately 25% of primary bone malignant tumors and, excluding hematopoietic malignancies of bone marrow origin, is exceeded in frequency only by osteosarcoma [1]. It shows an increased incidence after the fourth decade [2]. Chondrosarcoma can be either primary and develop de novo [3], or secondary and superimpose on a preexisting benign cartilage-forming tumor [4]. The clinical outcome depends on the histological grading, as the 10-year overall survival decreases from 88% for low-grade/atypical cartilaginous tumor to 62% and 26% for grade II and grade III chondrosarcoma, respectively [5]. Treatment changes substantially and consists of curettage or even watchful waiting for low-grade lesions in the extremities, and resection with wide margins for axial skeleton tumors and higher-grade lesions in the extremities [6]. A reliable preoperative diagnosis is thus crucial, and relies on a combination of clinical presentation, imaging and biopsy [6–8]. However, discrepancies in tumor grading are common even among expert radiologists and pathologists due to overlapping imaging and histological findings [9,10], advocating the need for more accurate diagnostic tools.

Texture analysis is an emerging post-processing method for quantification of tumor heterogeneity, a key feature of malignancy that reflects adverse tumor biology but is hard to capture using conventional imaging tools or sampling biopsies [11,12]. It belongs to the growing field of radiomics, which includes extraction, analysis and interpretation of large numbers of quantitative data from medical images [13]. Their interpretation can be aided by data mining techniques and machine learning algorithms to identify the best subset of texture features and create usable predictive model for the diagnosis of interest [14].

To date, texture analysis has been applied to imaging studies in combination to classical univariate and multivariate statistical analyses with the aim of discriminating tumor grades and types before treatment, monitoring response to therapy and predicting outcome [15]. The purpose of this study is to evaluate the diagnostic accuracy of machine learning for discrimination between low-grade and high-grade cartilaginous bone tumors based on radiomic parameters extracted from unenhanced magnetic resonance imaging (MRI).

3.2 Material and Methods

3.2.1 Study design and population

Institutional Review Board approval and a waiver for informed consent were obtained. This retrospective study included patients with low-to-high grade cartilaginous tumors of the bone who underwent MRI, between 2015 and 2018, at a tertiary bone tumor center. Information was retrieved through electronic surgical records. Inclusion criteria were: (i) primary low-to-high grade cartilaginous tumor that underwent surgery, such as intralesional curettage or resection; (ii) definitive histological diagnosis based on the assessment of the surgical specimen and considered as the reference standard; (iii) 1.5-T MRI performed within 3 months before surgery; and (iv) available T1-weighted and T2-weighted sequences. Patients with pathological fractures or images affected by artifacts were excluded.

Overall, a total of 58 patients (53 ± 16 [mean ± standard deviation] years of age) were enrolled, including:

- 26 low-grade/atypical cartilaginous tumors of the appendicular skeleton. These lesions were located in the femur (n=12), fibula (n=2), humerus (n=11) and tibia (n=1);
- 32 high-grade tumors, such as conventional G2 (n=18), conventional G3 (n=11), dedifferentiated (n=2) or mesenchymal (n=1) chondrosarcomas. These lesions were located in the femur (n=7), fibula (n=1), humerus (n=7), pelvis (n=5), scapula (n=4), spine (n=7) and tibia (n=1).

3.2.2 Quantitative image analysis

Patients were randomly divided into training (n=42) and test (n=16) groups for model tuning and testing, respectively. A last-year radiology resident with 12 months of supervised experience in musculoskeletal imaging performed image segmentation blinded to information about histological diagnosis, course of the disease and any other imaging study. The T1-weighted and T2-weighted images showing the largest tumor area were selected, on axial sequences as first choice and coronal or sagittal sequences as second choice, and then imported into ITK-SNAP (v3.6) [16], an open-source software for medical image segmentation. Next, the reader manually segmented each tumor by drawing a bidimensional polygonal region of interest (ROI) along its borders.

ROIs were used for first order and texture feature extraction using PyRadiomics (v2.2.0) [17], an open-source Python software. Image pre-processing consisted in resampling to a 2x2x2 isotropic voxel, intensity normalization and discretization with a fixed bin width of 2. A detailed description of the implementation of these steps and radiomic features extracted by the software is available in the official documentation (https://pyradiomics.readthedocs.io/en/latest/features.html).

Data mining and machine learning analysis were performed using the Weka data mining platform (v3.8.3) [18]. A Random Forest feature subset selection wrapper and an ensemble meta-algorithm (AdaboostM1) were employed for feature selection and for final model development, respectively, as they have shown good performance on similar datasets in literature [19,20]. Feature subset selection is a technique that aims to identify the optimal set of parameters while taking into consideration both the relevance of features in relation to the class of interest and their redundancy [21]. Using the wrapper approach, feature relevance evaluation is performed via a black box induction algorithm (random forest in our case) and cross-validation within the training set. Random forest was selected for this task as it is an ensemble algorithm, as the AdaboostM1 used for developing the final model, often employed to compute feature importance [22]. Final model performance was first assessed through stratified 10-fold cross-validation in the training cohort, and then the test cohort was used to confirm our findings on previously unseen cases. Our radiomics workflow pipeline is shown in Fig. 1.

3.2.3 Qualitative image analysis

A musculoskeletal radiologist with 10 years of experience performed qualitative MRI analysis on the test set cases independently and blinded to information regarding histological diagnosis, course of the disease and any other imaging study. Non-contrast MRI sequences were available for review, including T2-weighted with and without fat suppression and T1-weighted sequences. The reader was asked to predict tumor grade based on the following features: tumor location; maximum diameter; adjacent bone marrow edema; bone expansion; cortical thickening; cortical breakthrough; periosteal reaction; soft-tissue mass edema; and soft-tissue mass.

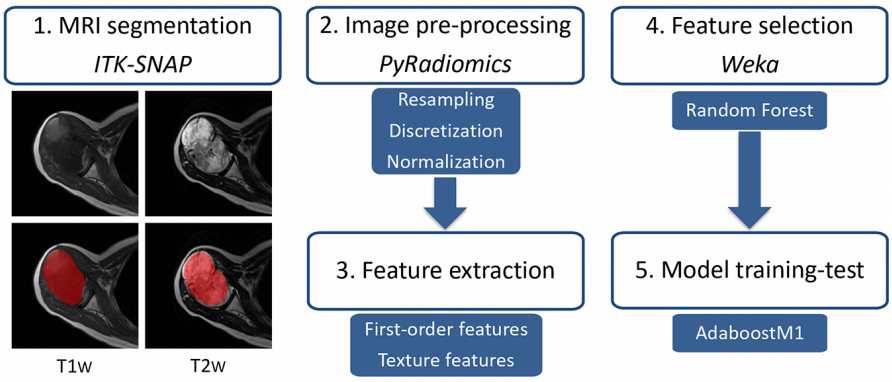


Fig. 1 Radiomics workflow pipeline. Manual segmentation of the lesions was performed on T1-weighted and T2-weighted images. Image pre-processing included resampling to a 2x2x2 isotropic voxel, intensity normalization and discretization with a fixed bin width of 2. First-order and texture features were extracted and then selected using a Random Forest wrapper. Finally, an ensemble meta-algorithm (AdaboostM1) was employed to automatically classify the lesions with a training-test approach.

3.2.4 Statistical analysis

Categorical variables were reported as absolute value and percentage; continuous variables were reported as mean ± standard deviation. Accuracy measures of the machine-learning classifier performance included, among others:

- F-score, i.e. the harmonic average of the precision (also known as positive predictive value) and recall (also known as sensitivity), ranging from 0 to 1 (perfect accuracy);
- Matthews correlation coefficient, i.e. a measure of the quality of binary classifications in machine learning, ranging from + 1 (perfect prediction) to 0 (average random prediction) and − 1 (inverse prediction);
- Area under the precision-recall curve, i.e. an alternative to the area under the receiver operator characteristic (ROC) curve, which is more informative for imbalanced classes.

Data analysis was performed using IBM SPSS Statistics (IBM Corp., Armonk, NY, USA). The diagnostic performances of the machine learning classifier and the radiologist were compared using McNemar's test. A p-value < 0.05 indicated statistical significance.

3.3 Results

A total of 172 radiomic features were extracted from each patient, 86 for each MRI sequence, shown in their pairwise correlation cluster map (Supplementary file 1). Among these, the Random Forest wrapper selected the 4 most informative. These were all derived from T1-weighted images and included: Energy derived from first order histogram analysis; Joint Average derived from Gray Level Co-occurrence Matrix; Large Dependence High Gray Level Emphasis derived from Gray Level Dependence Matrix; and Gray Level Non-Uniformity derived from Gray Level Size Zone Matrix. Table 1 details the characteristics of each feature and relative class. Figure 2 shows their univariate and bivariate distribution in our population based on lesion grade.

The final model was a locally weighted (k=3) boosted (AdaboostM1, 10 iterations) decision stump ensemble algorithm. AdaboostM1 algorithm code is provided as supplementary material (Supplementary file 2). Overall, its accuracy was 85.7% in the 10-fold stratified cross-validation performed in the training cohort (36/42 correctly classified lesions), and 75% in the test one (12/16 correctly classified lesions). The corresponding areas under the ROC curve were 0.85 and 0.78 (Supplementary file 3). Specifically, model accuracy for the identification of high-grade and low-grade tumors was 86.4% and 85% in the training, and 70% and 83.3% in the test cohort, respectively. Other evaluation metrics are reported in Table 2, derived from the confusion matrix presented in Table 3.

The musculoskeletal radiologist correctly graded 81.3% (13/16) of cartilaginous tumors from the test cohort. Specifically, his accuracy was 90% (9/10) for the detection of high-grade and 66.7% (4/6) for that of low-grade tumors. Maximum tumor diameter was 9 ± 6 cm. There was no statistical difference in terms of diagnostic performance between the machine learning classifier and the radiologist (P=0.453).

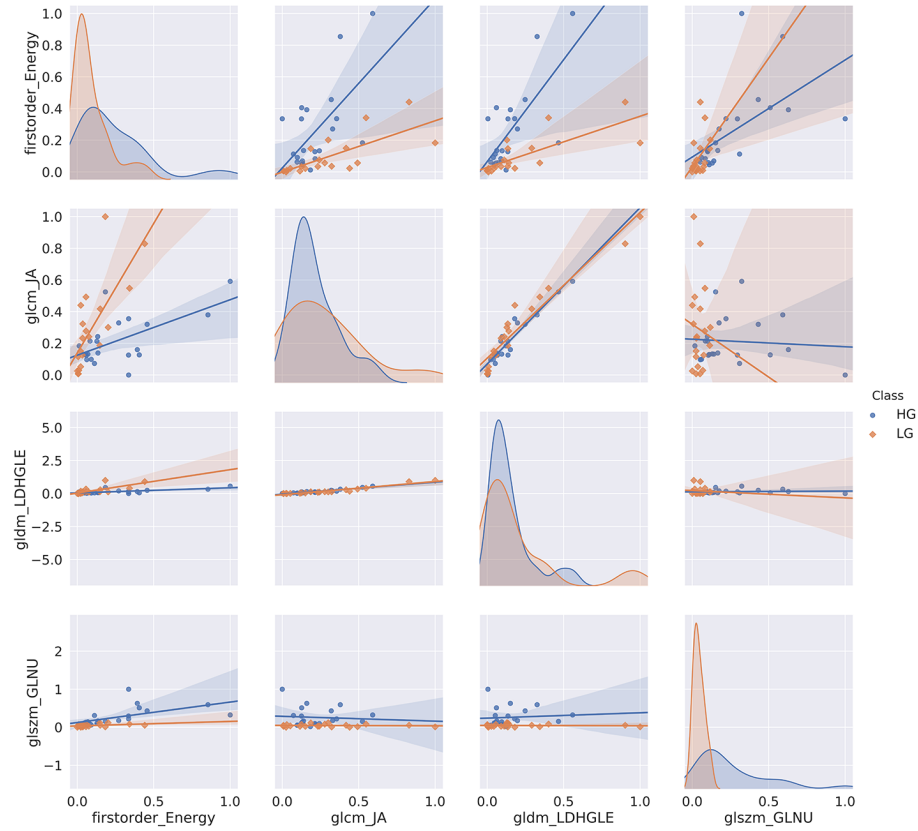


Fig. 2 Univariate and bivariate distribution with regression lines for the selected feature subset in relation to the tumor grade. Firstorder_energy, Energy derived from first order histogram analysis; glcm_JA, Joint Average derived from Gray Level Co-occurrence Matrix; gldm_LDHGLE, Large Dependence High Gray Level Emphasis derived from Gray Level Dependence Matrix; glszm_GLNU, Gray Level Non-Uniformity derived from Gray Level Size Zone Matrix; HG, high-grade tumor; LG, low-grade/atypical cartilaginous tumor.

Table 1 Characteristics of each selected feature and relative class according to PyRadiomics official documentation (https://pyradiomics.readthedocs.io/en/latest/features.html). All features were extracted from T1-weighted images.

Feature	Feature characteristics	Feature class	Class characteristics
Energy	Measures the magnitude of voxel values	First order	Describes the distribution of voxel intensities
Joint Average	Returns the mean gray level intensity	Gray Level Co-occurrence Matrix	Quantifies how often pairs of pixels with specific values occur in a specified spatial range
Large Dependence High Gray Level Emphasis	Measures the joint distribution of large dependence with higher gray-level values	Gray Level Dependence Matrix	Quantifies gray level dependencies, i.e. the number of connected voxels within a set distance that are dependent on the center voxel
Gray Level Non-Uniformity	Measures the variability of gray-level intensity values	Gray Level Size Zone Matrix	Quantifies gray level zones, i.e. the number of connected voxels sharing the same intensity value

Table 2 Classifier accuracy metrics weighted average and by class in both the training and test cohorts. FP, false positive; MCC, Matthews correlation coefficient; PRC, precision-recall curve; ROC, receiver operator curve; TP, true positive.

Cohort	Class	TP rate	FP rate	Precision	F-score	MCC	ROC area	PRC area
Training	**High-grade**	0.864	0.150	0.864	0.864	0.714	0.850	0.812
	Low-grade	0.850	0.136	0.850	0.850	0.714	0.853	0.794
	Weighted average	0.857	0.144	0.857	0.857	0.714	0.852	0.803
Test	**High-grade**	0.700	0.167	0.875	0.778	0.516	0.775	0.813
	Low-grade	0.833	0.300	0.625	0.714	0.516	0.775	0.588
	Weighted average	0.750	0.217	0.781	0.754	0.516	0.775	0.728

Table 3 Confusion matrix for the training and test cohorts.

			Actual class	
			High-grade	**Low-grade**
Predicted class	**Training cohort**	**High-grade**	19	3
		Low-grade	3	17
	Test cohort	**High-grade**	7	1
		Low-grade	3	5

3.4 Discussion

The main finding of this study is that our machine learning approach showed good diagnostic performance for classification of low-to-high grade cartilaginous bone tumors based on radiomic features extracted from unenhanced MRI, which was not significantly different compared to an experienced musculoskeletal radiologist.

The accurate grading of cartilaginous bone tumors is highly desired to select the most appropriate treatment, which ranges from conservative-to-aggressive for low-to-high grade lesions [6]. However, preoperative biopsy may erroneously lead to down-grading of chondrosarcoma as only small tumor areas are sampled [23], and interobserver variability in tumor grading has been seen even among specialized bone pathologists [9,10]. In turn, inaccurate preoperative grading may result in an inadequate treatment and subsequent need for further surgery with increased morbidity. Imaging plays a crucial role by integrating clinical data and biopsy before surgery is performed, and MRI is the method of choice [24]. On unenhanced MRI, bone expansion, periosteal reaction, soft-tissue mass and tumor length have been demonstrated to yield a diagnostic accuracy greater than 90% [25]. Bone marrow edema, cortical thickening or destruction and soft-tissue edema are also useful signs in the grading of chondrosarcoma [25]. These findings are consistent with our qualitative image analysis as the radiologist yielded an accuracy of 81.3%, which was a little lower than previously reported probably due to the small size of our test cohort. Diffusion-weighted MRI has been shown unable to differentiate low-grade lesions from high-grade chondrosarcomas [26] and was then not included in our analysis. On the other hand, contrast-enhanced MRI and particularly dynamic MRI aid in the diagnosis, as high-grade tumor areas enhance fast because of richly vascularized intralesional septations [27–29]. However, contrast-enhanced sequences were not evaluated in our series because they were not available in all patients.

Current imaging techniques may potentially be further equipped to better grade and safely diagnose cartilaginous bone lesions preoperatively [24]. In this regard, radiomics may prove a valuable aid by providing quantitative data that integrate qualitative image information already available [30,31]. To our knowledge, two studies to date have focused on radiomics and cartilaginous bone tumors [32,33]. Lisson et al. [32] evaluated 22 patients with enchondroma or low-grade cartilaginous tumor and observed that MRI-based volumetric texture analysis could discriminate these two entities by some individual texture

features. They included kurtosis and skewness on contrast-enhanced T1-weighted images, and entropy and uniformity of distribution of positive pixels on non-contrast T1-weighted images. Surprisingly, no texture feature showed a significant difference between benign and low-grade tumors on T2-weighted images [32]. This finding is consistent with our results, as exclusively T1-weighted image-derived features were selected during data dimensionality reduction. More recently, Fritz et al. [33] assessed the diagnostic accuracy of morphologic MRI and MRI-based bidimensional texture analysis for tumor grading in a series of 53 chondromas and 63 low-to-high grade cartilaginous tumors. Independent morphologic MRI and texture analysis predictors were found, and a combination of both achieved the highest diagnostic accuracy for differentiation of benign from malignant as well as benign from low-grade tumors. However, no statistically significant texture analysis predictors existed for differentiation of low-grade from high-grade lesions [33]. This could be interpreted as a limitation of classical statistical approaches in this field, which could be solved by means of data mining and machine learning [34], as shown in our study. Our study attempted to discriminate low-grade/atypical cartilaginous tumors from higher-grade lesions using radiomic data extracted from T1-weighted and T2-weighted sequences, which are cornerstones of morphologic MRI assessment. We integrated texture analysis with machine learning, a branch of computer science that enables algorithms to learn from data without explicitly being programmed [35–38]. A Random Forest wrapper was used to perform feature selection and provided four features, which were all derived from T1-weighted images. Thereafter, a locally weighted ensemble classifier (AdaboostM1) was trained on the training cohort and then its performance was evaluated on the test cohort. It demonstrated a substantial performance (area under the ROC curve = 0.78 in the test cohort), which was not different compared to a musculoskeletal radiologist.

Our classification model might potentially help radiologists express the probability that a cartilaginous bone tumor is low-to-high grade, integrating histological information and directing clinicians towards a conservative or an aggressive approach. Nonetheless, some limitations of this study need to be addressed. First, chondrosarcoma is a rare tumor and our small population of study did not allow us to analyze separately appendicular and axial high-grade lesions, which were grouped together as they are all treated with surgical resection with free margins [6]. Second, we used a bidimensional approach for segmentation and selected the image with the largest tumor area. This decision was based on previous studies

suggesting that bidimensional texture analysis is not inferior to volumetric texture analysis [39], and it is also easier to implement in clinical practice. Third, the retrospective study design accounts for the exclusion of contrast-enhanced MRI sequences, as they were not available in all patients. Finally, an external patient population for testing of the classification model was not available. Future investigations will require data exchange between different institutions to obtain high-volume image databases, also including contrast-enhanced MRI and allowing for testing of the classifier in an external population.

In conclusion, even though qualitative image assessment still plays a central role in the diagnosis, our machine learning classification model of low-to-high grade cartilaginous bone tumors is promising and may prove a valuable aid in the preoperative tumor characterization. Further studies with a larger sample of patients from multiple institutions are warranted.

References

[1] H.D. Dorfman, B. Czerniak, Bone cancers., Cancer. 75 (1995) 203–10. https://doi.org/10.1002/1097-0142(19950101)75:1+<203::aid-cncr2820751308>3.0.co;2-v.

[2] A. Franchi, Epidemiology and classification of bone tumors., Clin. Cases Miner. Bone Metab. 9 (2012) 92–5.

[3] M.D. Murphey, E.A. Walker, A.J. Wilson, M.J. Kransdorf, H.T. Temple, F.H. Gannon, From the archives of the AFIP: imaging of primary chondrosarcoma: radiologic-pathologic correlation., Radiographics. 23 (2003) 1245–78. https://doi.org/10.1148/rg.235035134.

[4] M. Altay, K. Bayrakci, Y. Yildiz, S. Erekul, Y. Saglik, Secondary chondrosarcoma in cartilage bone tumors: report of 32 patients, J. Orthop. Sci. 12 (2007) 415–423. https://doi.org/10.1007/s00776-007-1152-z.

[5] V.M. van Praag (Veroniek), A.J. Rueten-Budde, V. Ho, P.D.S. Dijkstra, M. Fiocco, M.A.J. van de Sande, I.C. van der Geest, J.A. Bramer, G.R. Schaap, P.C. Jutte, H.B. Schreuder, J.J.W. Ploegmakers, Incidence, outcomes and prognostic factors during 25 years of treatment of chondrosarcomas, Surg. Oncol. 27 (2018) 402–408. https://doi.org/10.1016/j.suronc.2018.05.009.

[6] P.G. Casali, S. Bielack, N. Abecassis, H.T. Aro, S. Bauer, R. Biagini, S. Bonvalot, I. Boukovinas, J.V.M.G. Bovee, B. Brennan, T. Brodowicz, J.M. Broto, L. Brugières, A. Buonadonna, E. De Álava, A.P. Dei Tos, X.G. Del Muro, P. Dileo, C. Dhooge, M. Eriksson, F. Fagioli, A. Fedenko, V. Ferraresi, A. Ferrari, S. Ferrari, A.M. Frezza, N. Gaspar, S. Gasperoni, H. Gelderblom, T. Gil, G. Grignani, A. Gronchi, R.L. Haas, B. Hassan, S. Hecker-Nolting, P. Hohenberger, R. Issels, H. Joensuu, R.L. Jones, I. Judson, P. Jutte, S. Kaal, L. Kager, B. Kasper, K. Kopeckova, D.A. Krákorová, R. Ladenstein, A. Le Cesne, I. Lugowska, O. Merimsky, M. Montemurro, B. Morland, M.A. Pantaleo, R. Piana, P. Picci, S. Piperno-Neumann, A.L. Pousa, P. Reichardt, M.H. Robinson, P. Rutkowski, A.A. Safwat, P. Schöffski, S. Sleijfer, S. Stacchiotti, S.J. Strauss, K. Sundby Hall, M. Unk, F. Van Coevorden, W.T.A. van der Graaf, J. Whelan, E. Wardelmann, O. Zaikova, J.Y. Blay, Bone sarcomas: ESMO–PaedCan–EURACAN Clinical Practice Guidelines for diagnosis, treatment and follow-up, Ann. Oncol. 29 (2018)

iv79–iv95. https://doi.org/10.1093/annonc/mdy310.

[7] L. Cannavò, D. Albano, C. Messina, A. Corazza, S. Rapisarda, G. Pozzi, A. Di Bernardo, A. Parafioriti, G. Scotto, G. Perrucchini, A. Luzzati, L.M. Sconfienza, Accuracy of CT and MRI to assess resection margins in primary malignant bone tumours having histology as the reference standard, Clin. Radiol. 74 (2019) 736.e13-736.e21. https://doi.org/10.1016/j.crad.2019.05.022.

[8] J.L. Bloem, I.I. Reidsma, Bone and soft tissue tumors of hip and pelvis, Eur. J. Radiol. 81 (2012) 3793–3801. https://doi.org/10.1016/j.ejrad.2011.03.101.

[9] K.B. Jones, J.A. Buckwalter, E.F. McCarthy, B.R. DeYoung, G.Y. El-Khoury, L. Dolan, F.H. Gannon, C.Y. Inwards, M.J. Klein, M. Kyriakus, A.E. Rosenberg, G.P. Siegal, K.K. Unni, L. Fayad, M.J. Kransdorf, M.D. Murphey, D.M. Panicek, D.A. Rubin, M. Sundararri, D. Vanel, Reliability of Histopathologic and Radiologic Grading of Cartilaginous Neoplasms in Long Bones, J. Bone Joint Surg. Am. 89 (2007) 2113–2123. https://doi.org/10.2106/JBJS.F.01530.

[10] D. Eefting, Y.M. Schrage, M.J.A. Geirnaerdt, S. Le Cessie, A.H.M. Taminiau, J.V.M.G. Bovée, P.C.W. Hogendoorn, Assessment of Interobserver Variability and Histologic Parameters to Improve Reliability in Classification and Grading of Central Cartilaginous Tumors, Am. J. Surg. Pathol. 33 (2009) 50–57. https://doi.org/10.1097/PAS.0b013e31817eec2b.

[11] B. Ganeshan, K.A. Miles, Quantifying tumour heterogeneity with CT, Cancer Imaging. 13 (2013) 140–149. https://doi.org/10.1102/1470-7330.2013.0015.

[12] F. Davnall, C.S.P. Yip, G. Ljungqvist, M. Selmi, F. Ng, B. Sanghera, B. Ganeshan, K.A. Miles, G.J. Cook, V. Goh, Assessment of tumor heterogeneity: an emerging imaging tool for clinical practice?, Insights Imaging. 3 (2012) 573–589. https://doi.org/10.1007/s13244-012-0196-6.

[13] R.J. Gillies, P.E. Kinahan, H. Hricak, Radiomics: Images Are More than Pictures, They Are Data, Radiology. 278 (2016) 563–577. https://doi.org/10.1148/radiol.2015151169.

[14] G. Choy, O. Khalilzadeh, M. Michalski, S. Do, A.E. Samir, O.S. Pianykh, J.R. Geis, P. V. Pandharipande, J.A. Brink, K.J. Dreyer, Current Applications and Future Impact of Machine Learning in Radiology, Radiology. 288 (2018) 318–328. https://doi.org/10.1148/radiol.2018171820.

[15] M.G. Lubner, A.D. Smith, K. Sandrasegaran, D. V. Sahani, P.J. Pickhardt, CT Texture Analysis: Definitions, Applications, Biologic Correlates, and Challenges, Radiographics. 37 (2017) 1483–1503. https://doi.org/10.1148/rg.2017170056.

[16] P.A. Yushkevich, J. Piven, H.C. Hazlett, R.G. Smith, S. Ho, J.C. Gee, G. Gerig, User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability, Neuroimage. 31 (2006) 1116–1128. https://doi.org/10.1016/j.neuroimage.2006.01.015.

[17] J.J.M. van Griethuysen, A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, V. Narayan, R.G.H. Beets-Tan, J.-C. Fillion-Robin, S. Pieper, H.J.W.L. Aerts, Computational Radiomics System to Decode the Radiographic Phenotype, Cancer Res. 77 (2017) e104–e107. https://doi.org/10.1158/0008-5472.CAN-17-0339.

[18] E. Frank, M. Hall, L. Trigg, G. Holmes, I.H. Witten, Data mining in bioinformatics using Weka, Bioinformatics. 20 (2004) 2479–2481. https://doi.org/10.1093/bioinformatics/bth261.

[19] Y. Choi, K.-J. Ahn, Y. Nam, J. Jang, N.-Y. Shin, H.S. Choi, S.-L. Jung, B. Kim, Analysis of heterogeneity of peritumoral T2 hyperintensity in patients with pretreatment glioblastoma: Prognostic value of MRI-based radiomics., Eur. J. Radiol. 120 (2019) 108642. https://doi.org/10.1016/j.ejrad.2019.108642.

[20] R. Ferrari, C. Mancini-Terracciano, C. Voena, M. Rengo, M. Zerunian, A. Ciardiello, S. Grasso, V. Mare', R. Paramatti, A. Russomando, R. Santacesaria, A. Satta, E. Solfaroli Camillocci, R. Faccini, A. Laghi, MR-based artificial intelligence model to assess response to therapy in locally advanced rectal cancer, Eur. J. Radiol. 118 (2019) 1–9. https://doi.org/10.1016/j.ejrad.2019.06.013.

[21] R. Kohavi, G.H. John, Wrappers for feature subset selection, Artif. Intell. 97 (1997) 273–324. https://doi.org/10.1016/S0004-3702(97)00043-X.

[22] M. Wehenkel, A. Sutera, C. Bastin, P. Geurts, C. Phillips, Random Forests Based Group Importance Scores and Their Statistical Interpretation: Application for Alzheimer's Disease, Front. Neurosci. 12 (2018) 411. https://doi.org/10.3389/fnins.2018.00411.

[23] S. Hodel, C. Laux, J. Farei-Campagna, T. Götschi, B. Bode-Lesniewska, D.A. Müller, The impact of biopsy sampling errors and the quality of surgical margins on local recurrence and survival in chondrosarcoma, Cancer Manag. Res. 10 (2018)

3765–3771. https://doi.org/10.2147/CMAR.S178768.

[24] M.A.J. van de Sande, R.J.P. van der Wal, A. Navas Cañete, C.S.P. van Rijswijk, H.M. Kroon, P.D.S. Dijkstra, J.L. (Hans) Bloem, Radiologic differentiation of enchondromas, atypical cartilaginous tumors, and high-grade chondrosarcomas—Improving tumor-specific treatment: A paradigm in transit?, Cancer. 125 (2019) 3288–3291. https://doi.org/10.1002/cncr.32404.

[25] H. Douis, L. Singh, A. Saifuddin, MRI differentiation of low-grade from high-grade appendicular chondrosarcoma, Eur. Radiol. 24 (2014) 232–240. https://doi.org/10.1007/s00330-013-3003-y.

[26] H. Douis, L. Jeys, R. Grimer, S. Vaiyapuri, A.M. Davies, Is there a role for diffusion-weighted MRI (DWI) in the diagnosis of central cartilage tumors?, Skeletal Radiol. 44 (2015) 963–969. https://doi.org/10.1007/s00256-015-2123-7.

[27] M.J.A. Geirnaerdt, P.C.W. Hogendoorn, J.L. Bloem, A.H.M. Taminiau, H.-J. van der Woude, Cartilaginous Tumors: Fast Contrast-enhanced MR Imaging, Radiology. 214 (2000) 539–546. https://doi.org/10.1148/radiology.214.2.r00fe12539.

[28] T. De Coninck, L. Jans, G. Sys, W. Huysse, T. Verstraeten, R. Forsyth, B. Poffyn, K. Verstraete, Dynamic contrast-enhanced MR imaging for differentiation between enchondroma and chondrosarcoma, Eur. Radiol. 23 (2013) 3140–3152. https://doi.org/10.1007/s00330-013-2913-z.

[29] H.J. Yoo, S.H. Hong, J.-Y. Choi, K.C. Moon, H.-S. Kim, J.-A. Choi, H.S. Kang, Differentiating high-grade from low-grade chondrosarcoma with MR imaging, Eur. Radiol. 19 (2009) 3008–3014. https://doi.org/10.1007/s00330-009-1493-4.

[30] S. Gyftopoulos, D. Lin, F. Knoll, A.M. Doshi, T.C. Rodrigues, M.P. Recht, Artificial Intelligence in Musculoskeletal Imaging: Current Status and Future Directions, AJR Am. J. Roentgenol. 213 (2019) 506–513. https://doi.org/10.2214/AJR.19.21117.

[31] A. Hirschmann, J. Cyriac, B. Stieltjes, T. Kober, J. Richiardi, P. Omoumi, Artificial Intelligence in Musculoskeletal Imaging: Review of Current Literature, Challenges, and Trends, Semin. Musculoskelet. Radiol. 23 (2019) 304–311. https://doi.org/10.1055/s-0039-1684024.

[32] C.S. Lisson, C.G. Lisson, K. Flosdorf, R. Mayer-Steinacker, M. Schultheiss, A. von

Baer, T.F.E. Barth, A.J. Beer, M. Baumhauer, R. Meier, M. Beer, S.A. Schmidt, Diagnostic value of MRI-based 3D texture analysis for tissue characterisation and discrimination of low-grade chondrosarcoma from enchondroma: a pilot study, Eur. Radiol. 28 (2018) 468–477. https://doi.org/10.1007/s00330-017-5014-6.

[33] B. Fritz, D.A. Müller, R. Sutter, M.C. Wurnig, M.W. Wagner, C.W.A. Pfirrmann, M.A. Fischer, Magnetic Resonance Imaging–Based Grading of Cartilaginous Bone Tumors, Invest. Radiol. 53 (2018) 663–672. https://doi.org/10.1097/RLI.0000000000000486.

[34] L. Ugga, R. Cuocolo, D. Solari, E. Guadagno, A. D'Amico, T. Somma, P. Cappabianca, M.L. del Basso de Caro, L.M. Cavallo, A. Brunetti, Prediction of high proliferative index in pituitary macroadenomas using MRI-based radiomics and machine learning, Neuroradiology. 61 (2019) 1365–1373. https://doi.org/10.1007/s00234-019-02266-1.

[35] M. Codari, L. Melazzini, S.P. Morozov, C.C. van Kuijk, L.M. Sconfienza, F. Sardanelli, Impact of artificial intelligence on radiology: a EuroAIM survey among members of the European Society of Radiology, Insights Imaging. 10 (2019) 105. https://doi.org/10.1186/s13244-019-0798-3.

[36] B.J. Erickson, P. Korfiatis, Z. Akkus, T.L. Kline, Machine Learning for Medical Imaging, Radiographics. 37 (2017) 505–515. https://doi.org/10.1148/rg.2017160130.

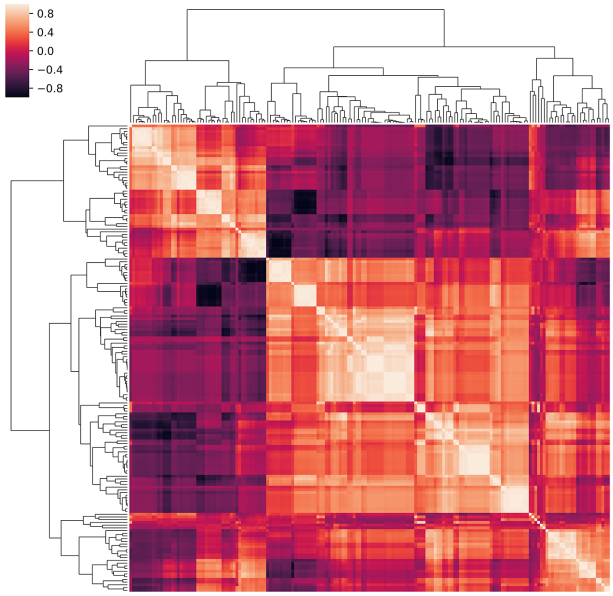
[37] F. Pesapane, C. Volonté, M. Codari, F. Sardanelli, Artificial intelligence as a medical device in radiology: ethical and regulatory issues in Europe and the United States, Insights Imaging. 9 (2018) 745–753. https://doi.org/10.1007/s13244-018-0645-y.

[38] F. Pesapane, M. Codari, F. Sardanelli, Artificial intelligence in medical imaging: threat or opportunity? Radiologists again at the forefront of innovation in medicine, Eur. Radiol. Exp. 2 (2018) 35. https://doi.org/10.1186/s41747-018-0061-6.

[39] M.G. Lubner, N. Stabo, S.J. Lubner, A.M. del Rio, C. Song, R.B. Halberg, P.J. Pickhardt, CT textural analysis of hepatic metastatic colorectal cancer: pre-treatment tumor heterogeneity correlates with pathology and clinical outcomes, Abdom. Imaging. 40 (2015) 2331–2337. https://doi.org/10.1007/s00261-015-0438-4.

Supplementary material

Supplementary file 1

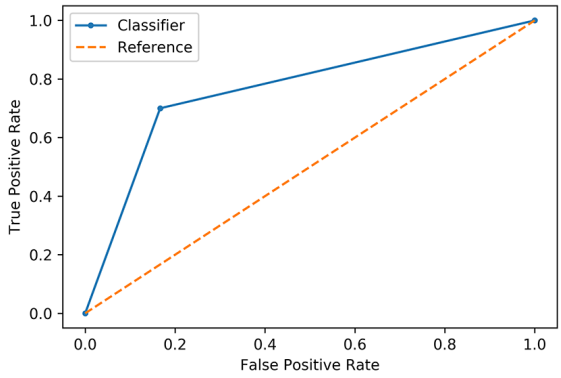


Pairwise feature correlation matrix represented as a hierarchically clustered heatmap.

Supplementary file 2

Machine learning algorithm code, as presented in the Weka data mining software: weka.classifiers.lazy.LWL -U 0 -K 3 -A "weka.core.neighboursearch.LinearNNSearch -A \"weka.core.EuclideanDistance -R first-last\"" -W weka.classifiers.meta.AdaBoostM1 -- -P 100 -S 1 -I

Supplementary file 3



ROC curve showing the diagnostic performance of the classifier in the test cohort.

Chapter 4

Effects of interobserver variability on 2D and 3D CT- and MRI-based texture feature reproducibility of cartilaginous bone tumors

Gitto S, Cuocolo R, Emili I, Tofanelli L, Chianca V, Albano D, Messina C, Imbriaco M, Sconfienza LM

J Digit Imaging 2021; 34:820-832
DOI: 10.1007/s10278-021-00498-3

List of abbreviations (Chapter 4)

2D, bidimensional

3D, volumetric

CT, computed tomography

ET, Extra Trees

GLCM, gray-level cooccurrence matrix

GLDM, gray-level dependence matrix

GLRLM, gray-level run length matrix

GLSZM, gray-level size zone matrix

ICC, intraclass correlation coefficient

LoG, Laplacian of Gaussian

MRI, magnetic resonance imaging

RFE, recursive feature elimination

ROI, region of interest

Abstract

This study aims to investigate the influence of interobserver manual segmentation variability on the reproducibility of 2D and 3D unenhanced computed tomography (CT)- and magnetic resonance imaging (MRI)-based texture analysis. Thirty patients with cartilaginous bone tumors (10 enchondromas, 10 atypical cartilaginous tumors, 10 chondrosarcomas) were retrospectively included. Three radiologists independently performed manual contour-focused segmentation on unenhanced CT, T1-weighted and T2-weighted MRI by drawing both a 2D region of interest (ROI) on the slice showing the largest tumor area and a 3D ROI including the whole tumor volume. Additionally, a marginal erosion was applied to both 2D and 3D segmentations to evaluate the influence of segmentation margins. A total of 783 and 1132 features were extracted from original and filtered 2D and 3D images, respectively. Intraclass correlation coefficient ≥0.75 defined feature stability. In 2D vs. 3D contour-focused segmentation, the rates of stable features were 74.71% vs. 86.57% (p<0.001), 77.14% vs. 80.04% (p=0.142) and 95.66% vs. 94.97% (p=0.554) for CT, T1-weighted and T2-weighted images, respectively. Margin shrinkage did not improve 2D (p=0.343) and performed worse than 3D (p<0.001) contour-focused segmentation in terms of feature stability. In 2D vs. 3D contour-focused segmentation, matching stable features derived from CT and MRI were 65.8% vs. 68.7% (p=0.191), and those derived from T1-weighted and T2-weighted images were 76.0% vs. 78.2% (p=0.285). 2D and 3D radiomic features of cartilaginous bone tumors extracted from unenhanced CT and MRI are reproducible, although some degree of interobserver segmentation variability highlights the need for reliability analysis in future studies.

4.1 Introduction

Cartilaginous tumors of the bone include a broad spectrum of lesions that range from benign to malignant entities [1, 2]. Reliable identification and grading are crucial, as clinical management varies widely. Specifically, asymptomatic benign enchondromas do not require any treatment, appendicular atypical cartilaginous tumors are managed with intralesional curettage or even watchful waiting, appendicular higher-grade lesions and axial skeleton chondrosarcomas are resected with free margins [3]. The diagnosis relies on a combination of clinical presentation, imaging and biopsy [3, 4]. Imaging, and particularly magnetic resonance imaging (MRI), has good accuracy in discriminating atypical cartilaginous tumors from higher-grade lesions [5] but is less reliable in differentiating the former from enchondromas [6]. Biopsy is considered the reference standard but has the disadvantages of sampling errors [7] and discrepancies even among specialized bone pathologists due to overlapping histological findings [8]. Additionally, the risk of biopsy-tract contamination remains a concern. Thus, the need for cutting-edge imaging-based tools, such as radiomics, is advocated to safely diagnose and grade cartilaginous bone tumors non-invasively [9].

Texture analysis is a post-processing method for quantification of tumor heterogeneity, which reflects adverse tumor biology but cannot be captured using conventional imaging modalities or sampling biopsies [10]. It belongs to the growing field of radiomics, which includes extraction, analysis and interpretation of large amounts of quantitative parameters from medical images [11, 12]. To date, texture analysis has been used to discriminate tumor grades and types before treatment, monitor response to therapy and predict outcome [13]. The resulting quantitative parameters, known as texture or radiomic features, may suffer however from interobserver variability, particularly with regard to tumor delineation while performing manual segmentation [14–16]. The influence of segmentation margins is also critical because of textural details of the peritumoral area, which may affect the reproducibility of texture features and therefore their diagnostic performance [17]. In literature, the Intraclass Correlation Coefficient (ICC) is commonly employed to assess radiomic feature reproducibility [17–21].

The aim of this study is to investigate the influence of interobserver manual segmentation variability on the reproducibility of bidimensional (2D) and volumetric (3D)

unenhanced computed tomography (CT)- and MRI-based texture analysis in cartilaginous bone tumors.

4.2 Materials and methods

4.2.1 Design and population

The local Institutional Review Board approved this retrospective study and waived the need for informed consent. According to the ICC guidelines by Koo et al. [22], we designed our study to meet the numerical requirements of a reliability analysis in terms of both patients and observers involved, namely 30 lesions and 3 different readers [22]. A search of the radiology information system was performed and 30 patients with cartilaginous bone tumors were recruited (median age 52 [range, 28-72] years), including 10 benign enchondromas, 10 atypical cartilaginous tumors and 10 malignant chondrosarcomas. Inclusion criteria were: (i) enchondromas proven either by histology or minimum follow-up of 6 years without alteration in shape or size and typical imaging findings of lobulated morphology and T2-weighted hyperintensity on MRI; (ii) histology-proven atypical cartilaginous tumors; (iii) histology-proven primary conventional grade II-III or dedifferentiated chondrosarcomas; (iv) 1.5-T MRI including turbo spin echo T1-weighted and T2-weighted sequences and 64-slice CT performed within one month before biopsy, intralesional curettage or surgical resection for tumors diagnosed by histology. Exclusion criteria were the presence of pathological fracture and ambiguous histology report.

Enchondromas were located in the femur (n = 5), fibula (n = 2), foot phalanx (n = 1), humerus (n = 1) and radius (n = 1), atypical cartilaginous tumors in the femur (n = 2), fibula (n = 2) and humerus (n = 6), chondrosarcomas in the calcaneus (n = 1), femur (n = 2), humerus (n = 1), pelvis (n = 2), spine (n = 3) and tibia (n = 1).

4.2.2 Image segmentation

A musculoskeletal radiologist (S.G.) and two last-year radiology residents trained in musculoskeletal and oncologic imaging (I.E. and L.T.) independently performed manual image segmentation using the open-source software ITK-SNAP (v3.6) [23]. The readers knew the study would deal with cartilaginous bone tumors, but they were blinded to any other information regarding histological grade, disease course and additional imaging studies. All tumors were segmented on axial CT scans and on axial MRI sequences as first

choice and coronal or sagittal sequences as second choice. Manual contour-focused segmentation was performed on unenhanced bone-window CT, T1-weighted and T2-weighted MRI by drawing both a 2D region of interest (ROI) on the slice showing the largest tumor area and a 3D ROI including the whole tumor volume. The "polygon mode" ITK-SNAP tool was used for all segmentations. While segmenting the tumors on CT, the readers used the MRI sequences to aid contour identification of each tumor. Thereafter, margin shrinkage segmentation was computed by applying a marginal erosion to both 2D and 3D segmentations in order to evaluate the influence of segmentation margins on feature reproducibility (Figure 1). In detail, ROI shrinkage was performed using the fslmaths erosion function of the FMRIB Software Library [24]. The default 2D and 3D kernels, which are 3x3x1 and 3x3x3 boxes centered at the target voxel, were employed as appropriate. During the erosion process, each voxel in the ROI is targeted sequentially, and its value is changed to 0 (i.e. removed from the ROI) if a zero-value voxel is found within the kernel. Therefore, the shrinkage was usually more extensive for 3D ROIs compared to 2D ones.

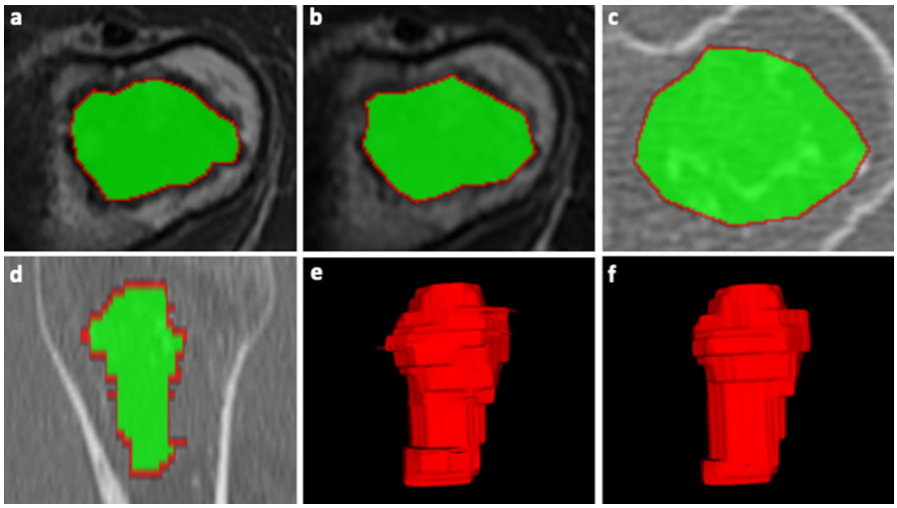


Fig. 1 Contour-focused and margin shrinkage segmentation of an atypical cartilaginous tumor of the humerus in a 45-year-old woman. **a-c** 2D contour-focused segmentation was performed on axial T1-weighted MRI (**a**), T2-weighted MRI (**b**) and bone-window CT (**c**) on the slice showing the largest tumor extension. **d** 3D contour-focused segmentation was performed slice by slice in the axial plane to include the whole tumor volume, as shown in the sagittal CT image. Contour-focused segmentation provided the ROI including both green and red areas. Margin shrinkage segmentation provided the ROI including only the green area by computing a marginal erosion, which is shown in red. **e-f** Segmented tumor volumes obtained with 3D contour-focused (**e**) and margin shrinkage (**f**) segmentation are shown, where the latter has smoother margins as a result of marginal erosion.

4.2.3 Texture analysis

Image pre-processing consisted in resampling to a 2x2 isotropic pixel or 2x2x2 isotropic voxel, whole image intensity normalization (mean value of 300 and standard deviation of 100) and discretization with a fixed bin width of 5. Original CT and MRI and 2D and 3D ROIs were used for feature extraction on PyRadiomics (v2.2.0) [25], an open-source Python software. The extracted features were grouped according to PyRadiomics official documentation (https://pyradiomics.readthedocs.io/en/latest/features.html), as follows:

- 18 first-order features, which describe the distribution of pixel or voxel gray-level values;
- 9 shape-based 2D and 14 shape-based 3D features, which respectively describe the 2D and 3D size and shape of the ROI;
- 22 gray-level cooccurrence matrix (GLCM) features, which quantify how often pairs of pixels or voxels with certain values occur in a specified spatial range;
- 16 gray-level size zone matrix (GLSZM) features, which quantify gray-level zones, i.e. the number of connected pixels or voxels sharing the same gray-level value;
- 16 gray-level run length matrix (GLRLM) features, which quantify gray-level runs, i.e. the length in number of consecutive pixels or voxels having the same gray-level value;
- 14 gray-level dependence matrix (GLDM) features, which quantify gray-level dependencies, i.e. the number of connected pixels or voxels within a set distance that are dependent on the center pixel and voxel.

In addition to the original CT and MRI, Laplacian of Gaussian (LoG)-filtered (sigma = 2, 3, 4, 5) and wavelet-transformed 2D and 3D images (all possible low and high pass filter combinations) were obtained for extraction of first-order and matrix features. Shape-based features are independent from gray-level value distribution and therefore were only computed on the original images. A total of 783 and 1132 features were extracted from original, LoG-filtered and wavelet-transformed 2D and 3D images, respectively.

4.2.4 Statistical analysis

Texture feature interobserver reliability was assessed using a two-way, random-effects, single-rater, absolute agreement ICC. Features were considered stable when

achieving good ($0.75 \leq ICC < 0.9$) to excellent ($ICC \geq 0.9$) interobserver reliability [22]. Differences among variables were evaluated using Chi-square test. A 2-sided p-value < 0.05 indicated statistical significance [26]. Data analysis was performed using the pandas and numpy Python software and the "irr" R package [27, 28].

4.2.5 Machine learning analysis

To assess the potential value of CT and MRI texture features extracted from 2D and 3D annotations, an exploratory data analysis was performed with an Extra Trees (ET) ensemble model. The same pipeline was employed on all available datasets, consisting of feature selection through cross-validated recursive feature elimination (RFE) and random search hyperparameter tuning nested within a leave-one-out cross-validation on the entire dataset. RFE was conducted using 10-fold cross-validation and an ET estimator with default hyperparameters. Then, in the training folds of the leave-one-out cross-validation, the synthetic oversampling technique was applied to balance the 3 classes (i.e., creating a synthetic instance to substitute the lesion in the test fold), followed by 100 iterations of ET hyperparameter random search. Given the presence of 3 classes with balanced cases, accuracy was used as the reference score for both RFE and ET tuning. The hyperparameter search space was as follows:

1. Number of trees = 100-1000
2. Criterion = entropy or Gini
3. Max depth = 1-10
4. Bootstrap = True or False
5. Max samples = 0-100%

4.3 Results

In 2D contour-focused vs. margin shrinkage segmentation, the stable feature rates were 74.71% (n = 585) vs. 71.65% (n = 561), 77.14% (n = 604) vs. 76.12% (n = 596) and 95.66% (n = 749) vs. 96.42% (n = 755) for CT, T1-weighted and T2-weighted images, respectively. The number of stable features derived from 2D contour-focused segmentation showed no difference in comparison with 2D margin shrinkage segmentation ($p = 0.343$). Table 1 details the number and percentage of stable features that were obtained with 2D contour-focused segmentation, grouped according to feature class and image type.

In 3D contour-focused vs. margin shrinkage segmentation, the stable feature rates were 86.57% (n = 980) vs. 83.66% (n = 947), 80.04% (n = 906) vs. 71.47% (n = 809) and 94.97% (n = 1075) vs. 65.72% (n = 744) for CT, T1-weighted and T2-weighted images, respectively. The number of stable features derived from 3D contour-focused segmentation was higher compared to 3D margin shrinkage segmentation ($p < 0.001$). Table 2 details the number and percentage of stable features that were obtained with 3D contour-focused segmentation, grouped according to feature class and image type.

The rate of stable features derived from CT was higher for 3D compared to 2D contour-focused segmentation ($p < 0.001$), while no difference was found for features derived from T1-weighted and T2-weighted MRI between 3D and 2D contour-focused segmentation ($p = 0.142$ and 0.554, respectively). In Figure 2, box and whisker plots show the interobserver reliability of feature classes derived from 3D and 2D contour-focused segmentation, grouped according to image type.

In 2D vs. 3D contour-focused segmentation, matching stable features derived from CT and MRI were 65.77% (n = 515) vs. 68.73% (n = 778), and those derived from T1-weighted and T2-weighted images were 75.99% (n = 595) vs. 78.18% (n = 885), respectively ($p = 0.191$ and 0.285). Tables 3 and 4 respectively detail the number and percentage of matching stable features obtained with 2D and 3D contour-focused segmentation, as well as overall interobserver reliability across different imaging modalities and MRI sequences, grouped according to feature class and image type. In Figure 3, box and whisker plots show the overall interobserver reliability of matching feature classes derived 3D and 2D contour-focused segmentation of CT and MRI, as well as MRI including T1-weighted and T2-weighted sequences, grouped according to image type. Most shape-based 2D and 3D features were stable even across different imaging modalities and MRI sequences.

Regarding the machine learning pipeline, the number of selected features ranged from 1 (from 2D annotations on T2-weighted images) to 236 (2D annotations on CT images). The accuracy of the ET models was fair to good, ranging between 77% (2D annotations on CT images) and 90% (3D annotations on T2-weighted images). Table 5 reports the results of each annotation and image type combination.

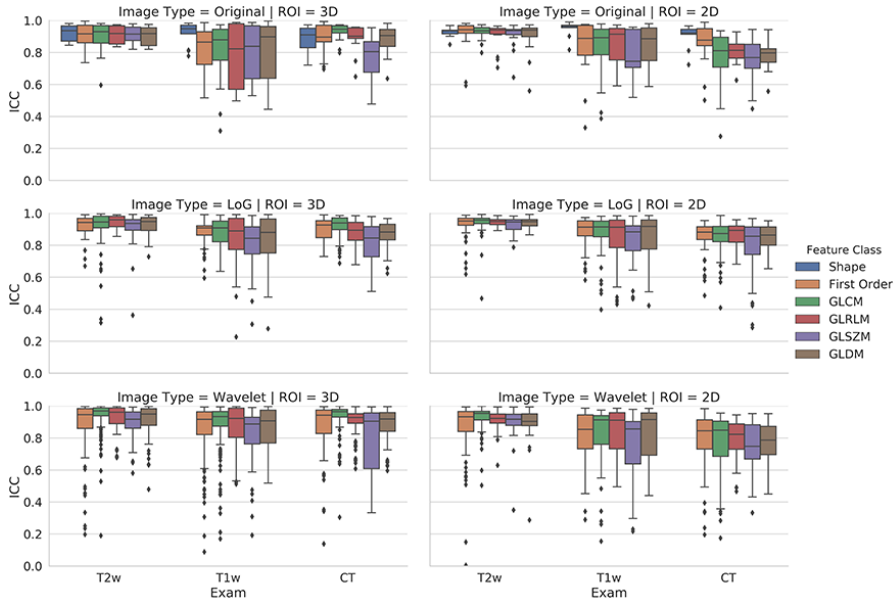


Fig. 2 3D and 2D contour-focused segmentation. Box and whisker plots show the interobserver reliability of feature classes grouped according to image type.

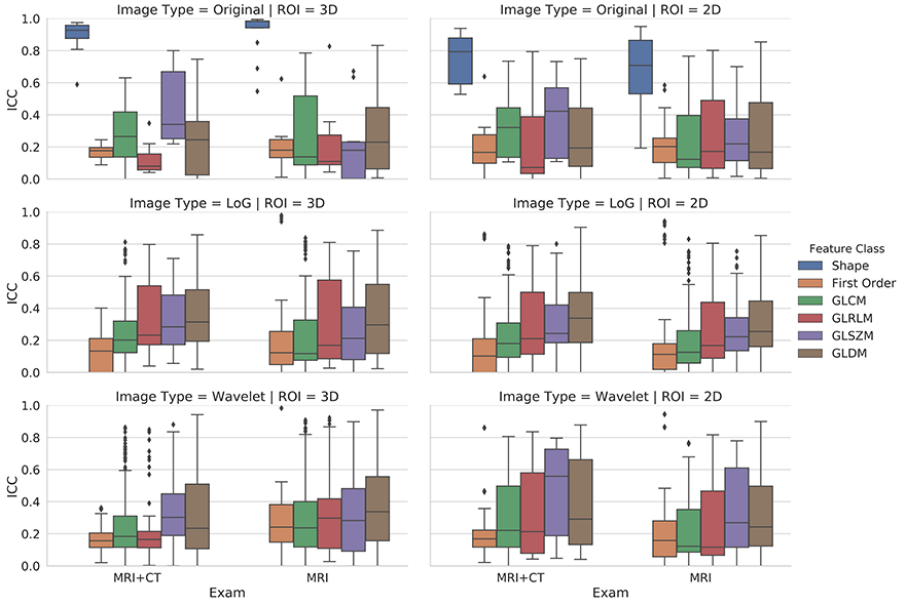


Fig. 3 3D and 2D contour-focused segmentation. Box and whisker plots show the overall interobserver reliability of matching feature classes derived from CT and MRI, as well as T1-weighted and T2-weighted MRI sequences, grouped according to image type.

Table 1 2D contour-focused segmentation. Number and percentage of stable features with good (0.75≤ICC<0.9) and excellent (ICC≥0.9) interobserver reliability grouped according to feature class and image type. GLCM, gray-level cooccurrence matrix; GLDM, gray-level dependence matrix; GLRLM, gray-level run length matrix; GLSZM, gray-level size zone matrix; ICC, intraclass correlation coefficient; LoG, Laplacian of Gaussian.

2D	Feature Class	Image Type	Total Features (n)	ICC ≥ 0.75 (n)	ICC ≥ 0.75 (%)	ICC ≥ 0.90 (n)	ICC ≥ 0.90 (%)
CT	First Order	LoG	72	63	87.50	29	40.28
		Original	18	16	88.89	8	44.44
		Wavelet	72	53	73.61	20	27.78
	GLCM	LoG	88	78	88.64	31	35.23
		Original	22	13	59.09	5	22.73
		Wavelet	88	60	68.18	27	30.68
	GLDM	LoG	56	49	87.50	18	32.14
		Original	14	10	71.43	2	14.29
		Wavelet	56	34	60.71	10	17.86
	GLRLM	LoG	64	58	90.63	27	42.19
		Original	16	13	81.25	2	12.50
		Wavelet	64	43	67.19	12	18.75
	GLSZM	LoG	64	46	71.88	20	31.25
		Original	16	9	56.25	3	18.75
		Wavelet	64	32	50.00	14	21.88
	Shape	Original	9	8	88.89	7	77.78
	OVERALL		783	585	74.71	235	30.01
T1w	First Order	LoG	72	65	90.28	42	58.33
		Original	18	15	83.33	8	44.44
		Wavelet	72	52	72.22	27	37.50
	GLCM	LoG	88	81	92.05	48	54.55
		Original	22	17	77.27	10	45.45
		Wavelet	88	67	76.14	50	56.82
	GLDM	LoG	56	43	76.79	29	51.79
		Original	14	10	71.43	7	50.00
		Wavelet	56	38	67.86	30	53.57
	GLRLM	LoG	64	51	79.69	34	53.13
		Original	16	12	75.00	9	56.25
		Wavelet	64	46	71.88	35	54.69
	GLSZM	LoG	64	50	78.13	26	40.63
		Original	16	8	50.00	6	37.50
		Wavelet	64	40	62.50	19	29.69
	Shape	Original	9	9	100.00	8	88.89
	OVERALL		783	604	77.14	388	49.55
T2w	First Order	LoG	72	68	94.44	61	84.72
		Original	18	16	88.89	15	83.33
		Wavelet	72	60	83.33	48	66.67
	GLCM	LoG	88	86	97.73	79	89.77
		Original	22	22	100.00	18	81.82
		Wavelet	88	84	95.45	71	80.68
	GLDM	LoG	56	56	100.00	48	85.71
		Original	14	12	85.71	10	71.43
		Wavelet	56	53	94.64	31	55.36
	GLRLM	LoG	64	64	100.00	60	93.75
		Original	16	15	93.75	13	81.25
		Wavelet	64	63	98.44	45	70.31
	GLSZM	LoG	64	64	100.00	47	73.44
		Original	16	15	93.75	12	75.00
		Wavelet	64	62	96.88	41	64.06
	Shape	Original	9	9	100.00	8	88.89
	OVERALL		783	749	95.66	607	77.52

Table 2 3D contour-focused segmentation. Number and percentage of stable features with good (0.75≤ICC<0.9) and excellent (ICC≥0.9) interobserver reliability grouped according to feature class and image type. GLCM, gray-level cooccurrence matrix; GLDM, gray-level dependence matrix; GLRLM, gray-level run length matrix; GLSZM, gray-level size zone matrix; ICC, intraclass correlation coefficient; LoG, Laplacian of Gaussian.

3D	Feature Class	Image Type	Total Features (n)	ICC ≥ 0.75 (n)	ICC ≥ 0.75 (%)	ICC ≥ 0.90 (n)	ICC ≥ 0.90 (%)
CT	First Order	LoG	72	64	88.89	44	61.11
		Original	18	14	77.78	9	50.00
		Wavelet	144	114	79.17	93	64.58
	GLCM	LoG	88	86	97.73	65	73.86
		Original	22	22	100.00	19	86.36
		Wavelet	176	169	96.02	153	86.93
	GLDM	LoG	56	50	89.29	24	42.86
		Original	14	13	92.86	8	57.14
		Wavelet	112	98	87.50	71	63.39
	GLRLM	LoG	64	62	96.88	30	46.88
		Original	16	14	87.50	9	56.25
		Wavelet	128	112	87.50	86	67.19
	GLSZM	LoG	64	46	71.88	19	29.69
		Original	16	11	68.75	2	12.50
		Wavelet	128	93	72.66	67	52.34
	Shape	Original	14	12	85.71	7	50.00
	OVERALL		1132	980	86.57	706	62.37
T1w	First Order	LoG	72	67	93.06	43	59.72
		Original	18	12	66.67	7	38.89
		Wavelet	144	121	84.03	89	61.81
	GLCM	LoG	88	77	87.50	47	53.41
		Original	22	16	72.73	10	45.45
		Wavelet	176	151	85.80	125	71.02
	GLDM	LoG	56	42	75.00	24	42.86
		Original	14	9	64.29	7	50.00
		Wavelet	112	85	75.89	60	53.57
	GLRLM	LoG	64	50	78.13	31	48.44
		Original	16	9	56.25	6	37.50
		Wavelet	128	99	77.34	77	60.16
	GLSZM	LoG	64	47	73.44	21	32.81
		Original	16	10	62.50	5	31.25
		Wavelet	128	97	75.78	55	42.97
	Shape	Original	14	14	100.00	11	78.57
	OVERALL		1132	906	80.04	618	54.59
T2w	First Order	LoG	72	70	97.22	53	73.61
		Original	18	17	94.44	11	61.11
		Wavelet	144	126	87.50	94	65.28
	GLCM	LoG	88	81	92.05	69	78.41
		Original	22	21	95.45	15	68.18
		Wavelet	176	169	96.02	145	82.39
	GLDM	LoG	56	55	98.21	41	73.21
		Original	14	14	100.00	9	64.29
		Wavelet	112	106	94.64	73	65.18
	GLRLM	LoG	64	64	100.00	53	82.81
		Original	16	16	100.00	11	68.75
		Wavelet	128	122	95.31	92	71.88
	GLSZM	LoG	64	62	96.88	46	71.88
		Original	16	16	100.00	11	68.75
		Wavelet	128	122	95.31	70	54.69
	Shape	Original	14	14	100.00	9	64.29
	OVERALL		1132	1075	94.97	802	70.85

Table 3 2D matching features. Number and percentage of matching stable features obtained with 2D contour-focused segmentation, as well as number and percentage of matching stable features with good (ICC≥0.75) overall interobserver reliability across different imaging modalities and MRI sequences, grouped according to feature class and image type. GLCM, gray-level cooccurrence matrix; GLDM, gray-level dependence matrix; GLRLM, gray-level run length matrix; GLSZM, gray-level size zone matrix; ICC, intraclass correlation coefficient; LoG, Laplacian of Gaussian.

2D	Feature Class	Image Type	Total features (n)	Matching features (n)	Matching features (%)	ICC ≥ 0.75 (n)	ICC ≥ 0.75 (%)
CT + MRI (T1w + T2w)	First Order	LoG	72	61	84.72	4	6.56
		Original	18	15	83.33	0	0
		Wavelet	72	45	62.50	3	6.67
	GLCM	LoG	88	74	84.09	2	2.70
		Original	22	11	50.00	0	0
		Wavelet	88	55	62.50	2	3.64
	GLDM	LoG	56	41	73.21	4	9.76
		Original	14	7	50.00	0	0
		Wavelet	56	29	51.79	6	20.69
	GLRLM	LoG	64	48	75.00	1	2.08
		Original	16	10	62.50	1	10.00
		Wavelet	64	36	56.25	7	19.44
	GLSZM	LoG	64	40	62.50	1	2.50
		Original	16	7	43.75	0	0
		Wavelet	64	28	43.75	3	10.71
	Shape	Original	9	8	88.89	4	50.00
	OVERALL		783	515	65.77	38	7.38
MRI (T1w + T2w)	First Order	LoG	72	63	87.50	8	12.70
		Original	18	15	83.33	0	0
		Wavelet	72	50	69.44	6	12.00
	GLCM	LoG	88	80	90.91	2	2.50
		Original	22	17	77.27	1	5.88
		Wavelet	88	65	73.86	2	3.08
	GLDM	LoG	56	43	76.79	2	4.65
		Original	14	9	64.29	1	11.11
		Wavelet	56	37	66.07	6	16.22
	GLRLM	LoG	64	51	79.69	1	1.96
		Original	16	12	75.00	2	16.67
		Wavelet	64	46	71.88	2	4.35
	GLSZM	LoG	64	50	78.13	1	2.00
		Original	16	8	50.00	0	0
		Wavelet	64	40	62.50	2	5.00
	Shape	Original	9	9	100.00	4	44.44
	OVERALL		783	595	75.99	40	6.72

Table 4 3D matching features. Number and percentage of matching stable features obtained with 3D contour-focused segmentation, as well as number and percentage of matching stable features with good (ICC≥0.75) overall interobserver reliability across different imaging modalities and MRI sequences, grouped according to feature class and image type. GLCM, gray-level cooccurrence matrix; GLDM, gray-level dependence matrix; GLRLM, gray-level run length matrix; GLSZM, gray-level size zone matrix; ICC, intraclass correlation coefficient; LoG, Laplacian of Gaussian.

3D	Feature Class	Image Type	Total features (n)	Matching features (n)	Matching features (%)	ICC ≥ 0.75 (n)	ICC ≥ 0.75 (%)
CT + MRI (T1w + T2w)	First Order	LoG	72	57	79.17	0	0
		Original	18	10	55.56	0	0
		Wavelet	144	97	67.36	0	0
	GLCM	LoG	88	75	85.23	4	5.33
		Original	22	16	72.73	0	0
		Wavelet	176	147	83.52	6	4.08
	GLDM	LoG	56	37	66.07	5	13.51
		Original	14	8	57.14	0	0
		Wavelet	112	72	64.29	6	8.33
	GLRLM	LoG	64	48	75.00	1	2.08
		Original	16	7	43.75	0	0
		Wavelet	128	81	63.28	3	3.70
	GLSZM	LoG	64	34	53.13	0	0
		Original	16	5	31.25	1	20.00
		Wavelet	128	72	56.25	6	8.33
	Shape	Original	14	12	85.71	11	91.67
	OVERALL		1132	778	68.73	43	5.53
MRI (T1w + T2w)	First Order	LoG	72	65	90.28	8	12.31
		Original	18	12	66.67	0	0
		Wavelet	144	116	80.56	14	12.07
	GLCM	LoG	88	75	85.23	10	13.33
		Original	22	16	72.73	2	12.50
		Wavelet	176	149	84.66	16	10.74
	GLDM	LoG	56	42	75.00	6	14.29
		Original	14	9	64.29	1	11.11
		Wavelet	112	83	74.11	10	12.05
	GLRLM	LoG	64	50	78.13	3	6.00
		Original	16	9	56.25	1	11.11
		Wavelet	128	96	75.00	10	10.42
	GLSZM	LoG	64	47	73.44	2	4.26
		Original	16	10	62.50	0	0
		Wavelet	128	92	71.88	6	6.52
	Shape	Original	14	14	100.00	12	85.71
	OVERALL		1132	885	78.18	101	11.41

Table 5 Feature selection process and exploratory machine learning pipeline in the reproducible feature datasets. The results of each annotation and image type combination are reported.

Annotation type	Imaging modality	Selected features (n)	Accuracy (%)
2D	T1w	5	83
	T2w	1	83
	CT	236	77
3D	T1w	67	87
	T2w	14	90
	CT	108	80

4.4 Discussion

The main finding of our study is that the rates of stable radiomic features extracted from unenhanced CT and MRI were 75% or higher for 2D and 80% or higher for 3D contour-focused segmentation. 3D CT-based texture analysis provided more stable features than 2D approach, while no difference in feature stability rates was found between 2D and 3D MRI-based texture analysis. Overall, a certain degree of segmentation variability highlighted the need to include a reliability analysis in future studies.

Despite its great potential as a non-invasive biomarker to quantify several tumor characteristics, radiomics still faces challenges to clinical implementation, both standalone and paired to machine learning [13, 29]. A great variability in radiomic features has emerged as a major issue across studies, and segmentation is the most critical step [12]. Image segmentation represents the basis of radiomic image analysis pipelines and can be time-consuming if performed manually. Therefore, methodological analyses are advisable prior to conducting radiomic studies in order to assess the robustness of different segmentation approaches and avoid biases due to non-reproducible, noisy features. These analyses have been previously performed in kidney [30, 31], lung and head and neck [15] lesions. With regard to cartilaginous bone tumors, radiomic studies to date have focused on discriminating among benign, atypical and malignant lesions [32–35], differentiating chondrosarcoma from other entities such as skull chordoma [36], or predicting recurrence of chondrosarcoma [37]. To our knowledge, our work is the first comprehensively addressing the influence of interobserver manual segmentation variability on the reproducibility of 2D and 3D CT- and MRI-based texture analysis in cartilaginous bone tumors. Nonetheless, Fritz et al. [33] and Gitto et al. [34] performed an interobserver reliability assessment as a feature-reduction method in their radiomic analysis, which provided a model for prediction of tumor grade. Particularly, Fritz et al. found that most 2D features derived from unenhanced (15 out of 19) and contrast-enhanced (18 out of 19) T1-weighted MRI had at least good agreement between two observers, using an ICC cutoff of 0.6 [33]. In this study, however, the number of extracted features was only 19 per sequence, the impact of different feature classes was not analyzed, and filtered and transformed images were not used. Despite these issues, a common conclusion that can be drawn from this and our studies is that most MRI radiomic features of cartilaginous bone tumors have good reproducibility, even though a certain degree of segmentation variability exists. In a more recent study by Gitto et al., stability was assessed

as a feature-reduction method and CT radiomic features were considered stable if ICC 95% confidence interval lower bound was 0.75 or higher. This resulted in a lower feature stability rate (30%) [34] compared to our current study.

In our study, all imaging modalities demonstrated good reproducibility both employing 2D and 3D annotations, with a robust feature percentage ranging from 75 to 96% for the former and 80 to 95% for the latter. Stable features also proved quite informative for predictive modeling at our preliminary analysis, with accuracies of 77-90%. Given the limited sample size and presence of 3 class labels, this result is promising and supports the use of radiomic data in this research domain. These findings are encouraging for future radiomic analyses, even though they confirm the need for a preliminary assessment of feature stability, and in line with recent literature emphasizing the importance of reproducibility in artificial intelligence and radiology [38]. The higher spatial resolution of CT did not seem to influence feature reproducibility and was probably offset by the better contrast resolution of T1-weighted and T2-weighted images. Furthermore, margin shrinkage did not lead to improvements in terms of feature reproducibility, contrary to a previous investigation on renal cell carcinoma CT images [17]. It should be noted that in this investigation, however, the authors reported that margin shrinkage produced less informative features even with improved reproducibility [17].

We found higher rates of stable features derived from CT for 3D compared to 2D segmentation, but no difference in the rates of 2D and 3D MRI-derived stable features. This finding is in favor of a 2D approach in future radiomic studies dealing with MRI-based texture analysis of cartilaginous bone tumors, as this is less time-consuming and easier to be employed in clinical practice, particularly in large atypical cartilaginous tumors and chondrosarcomas. Furthermore, most 2D (66-76%) and 3D (69-78%) stable features matched between CT and MRI, as well as T1-weghted and T2-weighted images. Finally, shape-based features were stable even across different imaging modalities and MRI sequences, and were thus reproducible and independent descriptors of tumor size and shape. On the other hand, overall interobserver reliability of other feature classes was unsurprisingly low across different imaging modalities and MRI sequences, indicating that their quantitative values depend on the specific image used.

Some limitations of our study should be acknowledged. First, it has a retrospective design as a prospective analysis is not strictly necessary for radiomic studies [13]. The

retrospective design accounts for the exclusion of contrast-enhanced images, as they were not performed for all enchondromas. Contrast-enhanced and dynamic contrast-enhanced MRI improve the accuracy of cartilaginous bone tumor assessment [39–41] and future radiomic studies focusing on these sequences are warranted. Finally, due to its scope, this was a single institution study and generalizability of our findings need to be confirmed on more varied datasets.

4.5 Conclusions

In conclusion, radiomic features of cartilaginous bone tumors extracted from 2D and 3D segmentations on CT and MRI examinations are reproducible, although some degree of segmentation variability highlights the need to perform a preliminary reliability analysis in radiomic studies. 3D and 2D MRI-based texture analysis provide similar rates of stable features. Thus, a 2D approach can be favored in future studies, as this is easier to implement in clinical practice.

References

1. Murphey MD, Walker EA, Wilson AJ, Kransdorf MJ, Temple HT, Gannon FH: From the archives of the AFIP: imaging of primary chondrosarcoma: radiologic-pathologic correlation. Radiographics 23:1245–1278, 2003

2. Albano D, Messina C, Gitto S, Papakonstantinou O, Sconfienza L: Differential Diagnosis of Spine Tumors: My Favorite Mistake. Semin Musculoskelet Radiol 23:26–35, 2019

3. Casali PG, Bielack S, Abecassis N, Aro HT, Bauer S, Biagini R, Bonvalot S, Boukovinas I, Bovee JVMG, Brennan B, Brodowicz T, Broto JM, Brugières L, Buonadonna A, De Álava E, Dei Tos AP, Del Muro XG, Dileo P, Dhooge C, Eriksson M, Fagioli F, Fedenko A, Ferraresi V, Ferrari A, Ferrari S, Frezza AM, Gaspar N, Gasperoni S, Gelderblom H, Gil T, Grignani G, Gronchi A, Haas RL, Hassan B, Hecker-Nolting S, Hohenberger P, Issels R, Joensuu H, Jones RL, Judson I, Jutte P, Kaal S, Kager L, Kasper B, Kopeckova K, Krákorová DA, Ladenstein R, Le Cesne A, Lugowska I, Merimsky O, Montemurro M, Morland B, Pantaleo MA, Piana R, Picci P, Piperno-Neumann S, Pousa AL, Reichardt P, Robinson MH, Rutkowski P, Safwat AA, Schöffski P, Sleijfer S, Stacchiotti S, Strauss SJ, Sundby Hall K, Unk M, Van Coevorden F, van der Graaf WTA, Whelan J, Wardelmann E, Zaikova O, Blay JY: Bone sarcomas: ESMO–PaedCan–EURACAN Clinical Practice Guidelines for diagnosis, treatment and follow-up. Ann Oncol 29:iv79–iv95, 2018

4. Cannavò L, Albano D, Messina C, Corazza A, Rapisarda S, Pozzi G, Di Bernardo A, Parafioriti A, Scotto G, Perrucchini G, Luzzati A, Sconfienza LM: Accuracy of CT and MRI to assess resection margins in primary malignant bone tumours having histology as the reference standard. Clin Radiol 74:736.e13-736.e21, 2019

5. Douis H, Singh L, Saifuddin A: MRI differentiation of low-grade from high-grade appendicular chondrosarcoma. Eur Radiol 24:232–240, 2014

6. Crim J, Schmidt R, Layfield L, Hanrahan C, Manaster BJ: Can imaging criteria distinguish enchondroma from grade 1 chondrosarcoma? Eur J Radiol 84:2222–2230, 2015

7. Hodel S, Laux C, Farei-Campagna J, Götschi T, Bode-Lesniewska B, Müller DA: The impact of biopsy sampling errors and the quality of surgical margins on local recurrence and survival in chondrosarcoma. Cancer Manag Res 10:3765–3771, 2018

8. Eefting D, Schrage YM, Geirnaerdt MJA, Le Cessie S, Taminiau AHM, Bovée JVMG,

Hogendoorn PCW: Assessment of Interobserver Variability and Histologic Parameters to Improve Reliability in Classification and Grading of Central Cartilaginous Tumors. Am J Surg Pathol 33:50–57, 2009
9. van de Sande MAJ, van der Wal RJP, Navas Cañete A, van Rijswijk CSP, Kroon HM, Dijkstra PDS, Bloem JL: Radiologic differentiation of enchondromas, atypical cartilaginous tumors, and high-grade chondrosarcomas—Improving tumor-specific treatment: A paradigm in transit? Cancer 125:3288–3291, 2019
10. Davnall F, Yip CSP, Ljungqvist G, Selmi M, Ng F, Sanghera B, Ganeshan B, Miles KA, Cook GJ, Goh V: Assessment of tumor heterogeneity: an emerging imaging tool for clinical practice? Insights Imaging 3:573–589, 2012
11. Codari M, Melazzini L, Morozov SP, van Kuijk CC, Sconfienza LM, Sardanelli F: Impact of artificial intelligence on radiology: a EuroAIM survey among members of the European Society of Radiology. Insights Imaging 10:105, 2019
12. Gillies RJ, Kinahan PE, Hricak H: Radiomics: Images Are More than Pictures, They Are Data. Radiology 278:563–577, 2016
13. Lubner MG, Smith AD, Sandrasegaran K, Sahani D V., Pickhardt PJ: CT Texture Analysis: Definitions, Applications, Biologic Correlates, and Challenges. Radiographics 37:1483–1503, 2017
14. Berenguer R, Pastor-Juan M, Canales-Vázquez J, Castro-García M, Villas MV, Mansilla Legorburo F, Sabater S: Radiomics of CT Features May Be Nonreproducible and Redundant: Influence of CT Acquisition Parameters. Radiology 288:407–415, 2018
15. Pavic M, Bogowicz M, Würms X, Glatz S, Finazzi T, Riesterer O, Roesch J, Rudofsky L, Friess M, Veit-Haibach P, Huellner M, Opitz I, Weder W, Frauenfelder T, Guckenberger M, Tanadini-Lang S: Influence of inter-observer delineation variability on radiomics stability in different tumor sites. Acta Oncol 57:1070–1074, 2018
16. Bologna M, Corino VDA, Montin E, Messina A, Calareso G, Greco FG, Sdao S, Mainardi LT: Assessment of Stability and Discrimination Capacity of Radiomic Features on Apparent Diffusion Coefficient Images. J Digit Imaging 31:879–894, 2018
17. Kocak B, Ates E, Durmaz ES, Ulusan MB, Kilickesmez O: Influence of segmentation margin on machine learning–based high-dimensional quantitative CT texture analysis: a reproducibility study on renal clear cell carcinomas. Eur Radiol 29:4765–4775, 2019
18. Kocak B, Durmaz ES, Kadioglu P, Polat Korkmaz O, Comunoglu N, Tanriover N:

Predicting response to somatostatin analogues in acromegaly: machine learning-based high-dimensional quantitative texture analysis on T2-weighted MRI. Eur Radiol 29:2731–2739, 2019
19. Schwier M, van Griethuysen J, Vangel MG, Pieper S, Peled S, Tempany C, Aerts HJWL, Kikinis R, Fennessy FM, Fedorov A: Repeatability of Multiparametric Prostate MRI Radiomics Features. Sci Rep 9:9441, 2019
20. Ugga L, Cuocolo R, Solari D, Guadagno E, D'Amico A, Somma T, Cappabianca P, del Basso de Caro ML, Cavallo LM, Brunetti A: Prediction of high proliferative index in pituitary macroadenomas using MRI-based radiomics and machine learning. Neuroradiology 61:1365–1373, 2019
21. Zwanenburg A, Vallières M, Abdalah MA, Aerts HJWL, Andrearczyk V, Apte A, Ashrafinia S, Bakas S, Beukinga RJ, Boellaard R, Bogowicz M, Boldrini L, Buvat I, Cook GJR, Davatzikos C, Depeursinge A, Desseroit M, Dinapoli N, Dinh CV, Echegaray S, El Naqa I, Fedorov AY, Gatta R, Gillies RJ, Goh V, Götz M, Guckenberger M, Ha SM, Hatt M, Isensee F, Lambin P, Leger S, Leijenaar RTH, Lenkowicz J, Lippert F, Losnegård A, Maier-Hein KH, Morin O, Müller H, Napel S, Nioche C, Orlhac F, Pati S, Pfaehler EAG, Rahmim A, Rao AUK, Scherer J, Siddique MM, Sijtsema NM, Socarras Fernandez J, Spezi E, Steenbakkers RJHM, Tanadini-Lang S, Thorwarth D, Troost EGC, Upadhaya T, Valentini V, van Dijk LV, van Griethuysen J, van Velden FHP, Whybra P, Richter C, Löck S: The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping. Radiology 295:328–338, 2020
22. Koo TK, Li MY: A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. J Chiropr Med 15:155–163, 2016
23. Yushkevich PA, Piven J, Hazlett HC, Smith RG, Ho S, Gee JC, Gerig G: User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability. Neuroimage 31:1116–1128, 2006
24. Jenkinson M, Beckmann CF, Behrens TEJ, Woolrich MW, Smith SM: FSL. Neuroimage 62:782–790, 2012
25. van Griethuysen JJM, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V, Beets-Tan RGH, Fillion-Robin J, Pieper S, Aerts HJWL: Computational Radiomics System to Decode the Radiographic Phenotype. Cancer Res 77:e104–e107, 2017
26. Di Leo G, Sardanelli F: Statistical significance: p value, 0.05 threshold, and

applications to radiomics—reasons for a conservative approach. Eur Radiol Exp 4:18, 2020
27. R Core Team: R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, 2020
28. van der Walt S, Colbert SC, Varoquaux G: The NumPy Array: A Structure for Efficient Numerical Computation. Comput Sci Eng 13:22–30, 2011
29. Cuocolo R, Caruso M, Perillo T, Ugga L, Petretta M: Machine Learning in oncology: A clinical appraisal. Cancer Lett 481:55–62, 2020
30. Kocak B, Durmaz ES, Kaya OK, Ates E, Kilickesmez O: Reliability of Single-Slice–Based 2D CT Texture Analysis of Renal Masses: Influence of Intra- and Interobserver Manual Segmentation Variability on Radiomic Feature Reproducibility. AJR Am J Roentgenol 213:377–383, 2019
31. Kocak B, Durmaz ES, Erdim C, Ates E, Kaya OK, Kilickesmez O: Radiomics of Renal Masses: Systematic Review of Reproducibility and Validation Strategies. AJR Am J Roentgenol 214:129–136, 2020
32. Gitto S, Cuocolo R, Albano D, Chianca V, Messina C, Gambino A, Ugga L, Cortese MC, Lazzara A, Ricci D, Spairani R, Zanchetta E, Luzzati A, Brunetti A, Parafioriti A, Sconfienza LM: MRI radiomics-based machine-learning classification of bone chondrosarcoma. Eur J Radiol 128:109043, 2020
33. Fritz B, Müller DA, Sutter R, Wurnig MC, Wagner MW, Pfirrmann CWA, Fischer MA: Magnetic Resonance Imaging–Based Grading of Cartilaginous Bone Tumors. Invest Radiol 53:663–672, 2018
34. Gitto S, Cuocolo R, Annovazzi A, Anelli V, Acquasanta M, Cincotta A, Albano D, Chianca V, Ferraresi V, Messina C, Zoccali C, Armiraglio E, Parafioriti A, Sciuto R, Luzzati A, Biagini R, Imbriaco M, Sconfienza LM: CT radiomics-based machine learning classification of atypical cartilaginous tumours and appendicular chondrosarcomas. EBioMedicine 68:103407, 2021
35. Lisson CS, Lisson CG, Flosdorf K, Mayer-Steinacker R, Schultheiss M, von Baer A, Barth TFE, Beer AJ, Baumhauer M, Meier R, Beer M, Schmidt SA: Diagnostic value of MRI-based 3D texture analysis for tissue characterisation and discrimination of low-grade chondrosarcoma from enchondroma: a pilot study. Eur Radiol 28:468–477, 2018
36. Li L, Wang K, Ma X, Liu Z, Wang S, Du J, Tian K, Zhou X, Wei W, Sun K, Lin Y, Wu Z, Tian J: Radiomic analysis of multiparametric magnetic resonance imaging for

differentiating skull base chordoma and chondrosarcoma. Eur J Radiol 118:81–87, 2019
37. Yin P, Mao N, Liu X, Sun C, Wang S, Chen L, Hong N: Can clinical radiomics nomogram based on 3D multiparametric MRI features and clinical characteristics estimate early recurrence of pelvic chondrosarcoma? J Magn Reson Imaging 51:435–445, 2020
38. Mongan J, Moy L, Kahn CE: Checklist for Artificial Intelligence in Medical Imaging (CLAIM): A Guide for Authors and Reviewers. Radiol Artif Intell 2:e200029, 2020
39. De Coninck T, Jans L, Sys G, Huysse W, Verstraeten T, Forsyth R, Poffyn B, Verstraete K: Dynamic contrast-enhanced MR imaging for differentiation between enchondroma and chondrosarcoma. Eur Radiol 23:3140–3152, 2013
40. Geirnaerdt MJA, Hogendoorn PCW, Bloem JL, Taminiau AHM, van der Woude H-J: Cartilaginous Tumors: Fast Contrast-enhanced MR Imaging. Radiology 214:539–546, 2000
41. Yoo HJ, Hong SH, Choi J, Moon KC, Kim H, Choi J, Kang HS: Differentiating high-grade from low-grade chondrosarcoma with MR imaging. Eur Radiol 19:3008–3014, 2009

Chapter 5

CT radiomics-based machine learning classification of atypical cartilaginous tumours and appendicular chondrosarcomas

Gitto S, Cuocolo R, Annovazzi A, Anelli V, Acquasanta M, Cincotta A, Albano D, Chianca V, Ferraresi V, Messina C, Zoccali C, Armiraglio E, Parafioriti A, Sciuto R, Luzzati A, Biagini R, Imbriaco M, Sconfienza LM

EBioMedicine 2021; 68:103407
DOI: 10.1016/j.ebiom.2021.103407

List of abbreviations (Chapter 5)

CT, computed tomography

ICC, intraclass correlation coefficient

MRI, magnetic resonance imaging

PET-CT, positron emission tomography-computed tomography

SMOTE, synthetic minority oversampling technique

Abstract

Background. Clinical management ranges from surveillance or curettage to wide resection for atypical to higher-grade cartilaginous tumours, respectively. Our aim was to investigate the performance of computed tomography (CT) radiomics-based machine learning for classification of atypical cartilaginous tumours and higher-grade chondrosarcomas of long bones.

Methods. One-hundred-twenty patients with histology-proven lesions were retrospectively included. The training cohort consisted of 84 CT scans from centre 1 (n=55 G1 or atypical cartilaginous tumours; n=29 G2-G4 chondrosarcomas). The external test cohort consisted of the CT component of 36 positron emission tomography-CT scans from centre 2 (n=16 G1 or atypical cartilaginous tumours; n=20 G2-G4 chondrosarcomas). Bidimensional segmentation was performed on preoperative CT. Radiomic features were extracted. After dimensionality reduction and class balancing in centre 1, the performance of a machine-learning classifier (LogitBoost) was assessed on the training cohort using 10-fold cross-validation and on the external test cohort. In centre 2, its performance was compared with preoperative biopsy and an experienced radiologist using McNemar's test.

Findings. The classifier had 81% (AUC=0.89) and 75% (AUC=0.78) accuracy in identifying the lesions in the training and external test cohorts, respectively. Specifically, its accuracy in classifying atypical cartilaginous tumours and higher-grade chondrosarcomas was 84% and 78% in the training cohort, and 81% and 70% in the external test cohort, respectively. Preoperative biopsy had 64% (AUC=0.66) accuracy (p=0.29). The radiologist had 81% accuracy (p=0.75).

Interpretation. Machine learning showed good accuracy in classifying atypical and higher-grade cartilaginous tumours of long bones based on preoperative CT radiomic features.

Funding. ESSR Young Researchers Grant.

Research in context

Evidence before this study. To date, radiomic studies have dealt with MRI of cartilaginous bone lesions with the aim of discriminating among benign enchondroma, atypical cartilaginous tumour and malignant chondrosarcoma, predicting local recurrence of chondrosarcoma and differentiating chondrosarcoma from other entities such as skull chordoma. Machine learning was used in combination with radiomics to address some of these issues. Particularly, an adaptive boosting classifier (AdaBoostM1) was a good predictor of tumour grade based on MRI radiomic features derived from unenhanced T1-weighted sequences, showing 75% accuracy in the test cohort for classification of atypical cartilaginous tumours and chondrosarcomas. This previous study included 58 patients from the same institution and the machine-learning classifier was internally tested using a hold-out set as a test cohort. To our knowledge, no published study has focused on machine learning and CT radiomics of cartilaginous bone lesions, as done in our study.

Added value of this study. We also attempted to differentiate atypical cartilaginous tumours from chondrosarcomas of long bones, as this is the most relevant clinical question and orientates towards a conservative approach or aggressive surgery. Our CT radiomics-based machine-learning classifier (boosted [LogitBoost] linear logistic regression classifier) achieved 75% accuracy overall, 81% accuracy in identifying atypical cartilaginous tumours and 70% accuracy in identifying higher-grade chondrosarcomas in the external test cohort, respectively, with no difference in comparison with an experienced radiologist (p=0.75). These results agree with those previously reported for tumour grading based on MRI radiomics. Furthermore, our findings were obtained in a more than twice larger population and validated in an independent test cohort from a second institution, thus ensuring their generalizability in clinical practice. Finally, although statistical significance was not reached (p=0.29), the machine-learning classifier's accuracy was slightly superior compared to preoperative biopsy. We may speculate that this difference could become significant in a larger population.

Implications of all the available evidence. Radiomics-based machine learning may potentially aid in preoperative tumour characterization by integrating the multidisciplinary approach currently based on clinical, conventional imaging and histological assessment.

5.1 Introduction

Chondrosarcoma accounts for 20 to 30% of primary malignant bone lesions (1). Clinical management primarily depends on tumour grading. Particularly, low-grade (G1) chondrosarcomas of long bones, recently downgraded from malignant to locally aggressive lesions and renamed “atypical cartilaginous tumours” (2), are managed with intralesional curettage or even watchful waiting. Appendicular higher-grade and axial skeleton chondrosarcomas require wide resection with free margins (3). The 10-year overall survival decreases from 88% for atypical cartilaginous tumour/G1 to 62% and 26% for G2 and G3 chondrosarcoma, respectively (4). Both imaging and biopsy integrate clinical information before any treatment is started (3). Magnetic resonance imaging (MRI) is the best imaging modality for local staging (5). Computed tomography (CT) is used for biopsy guidance (6) and provides additional information, such as matrix mineralization and cortex changes (3). CT and positron emission tomography-CT (PET-CT) can be both employed for general staging (3). Biopsy is considered the reference standard for preoperative assessment but suffers from the disadvantages of sampling errors (7) and overlapping histological findings leading to discrepancies even among expert bone pathologists (8). Thus, the need for new imaging-based tools like radiomics is advocated to better characterize cartilaginous bone lesions preoperatively (9).

Radiomics includes extraction and analysis of large numbers of quantitative characteristics, known as radiomic features, from imaging studies (10). This research field has gained much attention in oncologic imaging as a potential tool for quantification of tumour heterogeneity, which is hard to capture with conventional imaging assessment or sampling biopsies (11). Most radiomic studies to date have focused on discriminating tumour grades and types before treatment, monitoring response to therapy and predicting outcome (11). Due to its high-dimensional nature consisting of numerous radiomic features, radiomics benefits from powerful analytic tools and artificial intelligence with machine learning perfectly addresses this issue (12). Machine learning algorithms can be trained using subsets of radiomic features creating classification models for the diagnosis of interest (13–15).

Machine learning has recently shown good accuracy in discriminating between atypical cartilaginous tumours and higher-grade bone chondrosarcomas based on unenhanced MRI radiomic features (16). The aim of this study is to investigate the diagnostic

performance of CT radiomics-based machine learning for classification of atypical cartilaginous tumours and higher-grade chondrosarcomas of long bones.

5.2 Methods

5.2.1 Ethics

Our Institutional Review Board approved this retrospective study and waived the need for informed consent (Protocol: "AI tumori MSK"). Our database was anonymized according to the General Data Protection Regulation for Research Hospitals.

5.2.2 Study design and inclusion criteria

Information was retrieved through electronic records from the orthopaedic surgery and pathology departments. Consecutive patients with an atypical cartilaginous tumour or appendicular chondrosarcoma and CT or PET-CT performed over the last 10 years at one of two tertiary bone tumour centres (centre 1, IRCCS Orthopaedic Institute Galeazzi in Milan, Italy; centre 2, IRCCS Regina Elena National Cancer Institute in Rome, Italy) were considered for inclusion. Inclusion criteria were: (i) atypical cartilaginous tumour or conventional G2-G3-G4 (dedifferentiated) chondrosarcoma of long bones that was surgically treated with intralesional curettage or resection; (ii) definitive histological diagnosis defined on the basis of the surgical specimen assessment; (iii) CT (centre 1) or PET-CT (centre 2) scan performed before biopsy and within 1 month before surgery; and (iv) in centre 2, preoperative biopsy performed within 1 month before surgery. Patients with pathological fractures, secondary tumours arising from pre-existing cartilaginous lesions, recurrent tumours or metal devices resulting in beam hardening artifacts were excluded. A flowchart of patient selection is shown in Fig. 1.

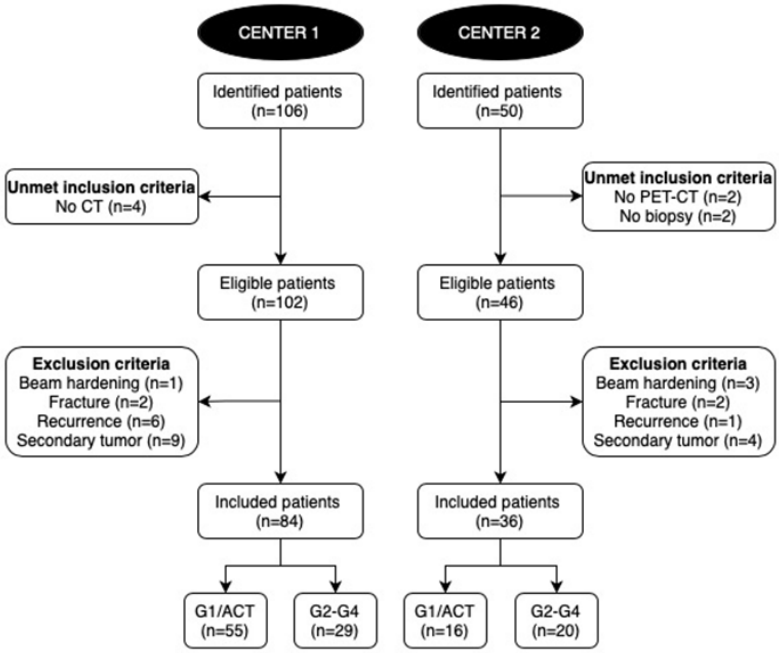


Fig. 1 Flowchart of patient selection. ACT, atypical cartilaginous tumours.

5.2.3 Study cohorts

One-hundred-twenty patients were retrospectively included. The training cohort consisted of 84 CT scans by as many patients from centre 1 (n=55 G1 or atypical cartilaginous tumours; n=29 G2-G4 chondrosarcomas). The external test cohort was constituted by the CT component of 36 PET-CT scans by as many patients from centre 2 (n=16 G1 or atypical cartilaginous tumours; n=20 G2-G4 chondrosarcomas). Patients' demographics and data regarding lesion location, grading and surgical treatment are detailed in Table 1. In centre 1, all examinations were performed using a 64-slice CT unit (Siemens SOMATOM Emotion, Erlangen, Germany). CT specifications were: matrix, 512 x 512; field of view (range), 138-380 mm; slice thickness, 1 mm. In centre 2, all examinations were performed using a 16-slice PET-CT unit (Siemens Biograph, Erlangen, Germany). PET-CT specifications were: matrix, 512 x 512; field of view, 500 mm; slice thickness, 4 mm. All DICOM images were exported and converted to the NiFTI format prior to the analysis (17).

Table 1 Demographics and clinical data. Age is presented as median and interquartile (1^{st}-3^{rd}) range.

	Centre 1	Centre 2
Age	52 (45-65) years	57 (46-69) years
Sex	Men: n=30 Women: n=54	Men: n=13 Women: n=23
Lesion location	Femur: n=40 Fibula: n=9 Humerus: n=30 Radius: n=1 Tibia: n=4	Femur: n=21 Fibula: n=6 Humerus: n=5 Tibia: n=4
Lesion grading	G1: n=55 G2: n=13 G3: n=9 G4 (dedifferentiated): n=7	G1: n=16 G2: n=12 G3: n=3 G4 (dedifferentiated): n=5
Surgery	G1/Atypical cartilaginous tumours Curettage: n=47 Wide resection: n=8*	G1/Atypical cartilaginous tumours Wide resection: n=16*
	G2-G4 chondrosarcomas Curettage + wide resection: n=5** Wide resection: n=24	G2-G4 chondrosarcomas Wide resection: n=20

*Wide resection was performed in n=8 G1/atypical cartilaginous tumours from centre 1 in case of specific anatomic location (like fibular head) or to prevent the risk of biopsy sampling errors. It was performed in all cases from centre 2 to prevent the risk of biopsy sampling errors, as per routine procedure.

**Curettage was initially performed in n=5 G2 chondrosarcomas from centre 2, as preoperative biopsy downgraded the lesions as G1. A second surgery consisting of wide resection was thus required.

5.2.4 Segmentation

A recently-boarded musculoskeletal radiologist (S.G.) manually performed contour-focused segmentation using a freely available, open-source software, ITK-SNAP (v3.6) (18). In detail, bidimensional regions of interest were annotated on the axial slice showing the maximum lesion extension. Unenhanced CT scan or CT scan performed as part of PET-CT protocol was used. According to the intraclass correlation coefficient (ICC) guidelines by Koo et al. (19), in a subgroup of 30 patients randomly selected from centre 1, segmentations were additionally performed independently by two radiology residents experienced in musculoskeletal and oncologic imaging (M.A. and A.C.) to meet the requirements of a reliability analysis in terms of patients and readers involved. All the readers knew the study would deal with cartilaginous bone lesions, but they were unaware of tumour grading and disease course, as well as the slice other readers used for segmentation.

5.2.5 Feature extraction

Image preprocessing and feature extraction were performed using PyRadiomics (v3.0.0) (20). Regarding preprocessing, image resampling (to an 1x1 mm in-plane resolution) was performed to ensure the correct calculation of texture features, following current guidelines (21). Grey level normalization and discretization followed. For the first, after z-score normalization, grey level values were scaled by a factor of 100. The resulting arrays were shifted by a value of 300 to avoid negative-valued pixels that could cause issues with texture analysis. After this process, the final image grey level range is expected to fall between 0 and 600, excluding outliers. To select the correct bin width for discretization, an exploratory extraction of first order parameters (i.e., grey level range) was performed exclusively on the training set, to avoid any information leak from the external test cohort. In this step, bin widths 2, 3, 4 and 5 were used to analyse grey level range of the normalized images. In addition to the original images, features were also extracted from filtered ones, i.e. after Laplacian of Gaussian (sigma=1, 2, 3, 4, 5) filtering and wavelet decomposition (all combinations of high and low-pass filtering on the x and y axes). All available first-order (histogram analysis), 2D shape-based and texture features were extracted, described in detail in the PyRadiomics official documentation (https://pyradiomics.readthedocs.io/en/latest/features.html).

5.2.6 Machine learning analysis

Radiomic data processing and machine learning analysis were performed using the Weka data mining platform (v3.8.4), R and scikit-learn Python package (22–24). A normalization (min-max range=0-1) scaler was fitted on the training data and applied to both training and external test cohorts prior to the analysis. Feature selection was performed exclusively using the training cohort data and included stability assessment as well as variance and intercorrelation analyses. The first was performed by obtaining feature ICC with a two-way random effect, single rater, absolute agreement model. Features were considered stable if the ICC 95% confidence interval lower bound was ≥0.75. Next, low variance (0.15 threshold) or highly inter-correlated (Pearson correlation coefficient threshold 0.80) features were removed. Finally, features with an information gain ratio >0.35 were selected.

Given the unbalanced nature of the training dataset, the synthetic minority oversampling technique (SMOTE) was used to balance this data by creating new instances from the minority class in centre 1, thus increasing the number of G2-G4 chondrosarcomas to 55 (25). The test set underwent no oversampling as it was not employed to build the classification model but only to assess its performance. Thereafter, a boosted (LogitBoost) linear logistic regression machine-learning classifier was trained and validated on the training cohort using 10-fold cross validation and tested on the external cohort. The Brier score was obtained, together with calibration curves, for the external test set in order to evaluate prediction and calibration loss. Our radiomics-based machine-learning workflow pipeline is shown in Fig. 2.

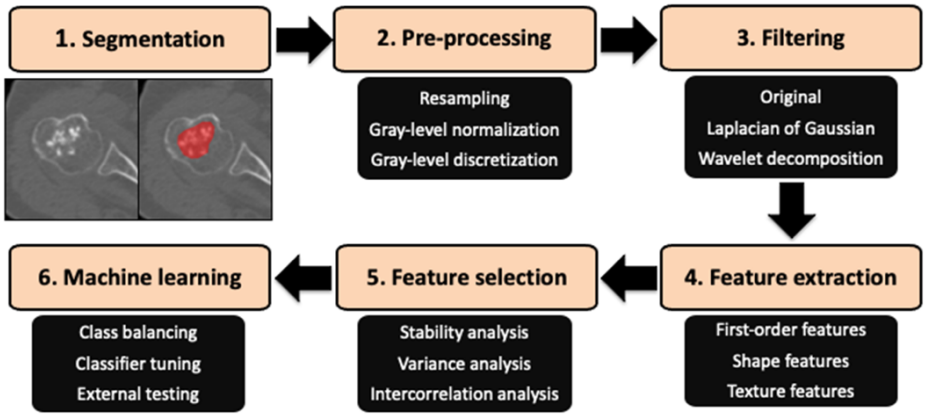


Fig. 2 Radiomics-based machine learning workflow pipeline.

5.2.7 Qualitative imaging assessment

A musculoskeletal radiologist with 12 years of experience in bone sarcomas (V.A.) read all CT studies from centre 2 blinded to any information regarding tumour grading, disease course and radiomics-based machine learning analysis. G2-G4 chondrosarcomas were differentiated from atypical cartilaginous tumours based on the presence of at least one of the following parameters: medullary cavity expansion with thinner cortex, cortical breakthrough, aggressive periosteal reaction, soft-tissue mass (5,26). Additionally, maximum lesion diameter was measured.

5.2.8 Statistical analysis

Continuous data are presented as median and interquartile (1^{st}-3^{rd}) range. Categorical data are presented as value counts and proportions. Data management was performed using the pandas Python software package. The "irr" and "stats" R packages were used for ICC assessment and remaining statistical tests, respectively. In the external test cohort, the classifier's performance was compared with preoperative biopsy and the radiologist's performance using McNemar's test. Mann-Whitney and Fisher's tests were used to assess age and sex differences between the two cohorts. A 2-sided p-value <0.05 indicated statistical significance.

Accuracy measures of the machine-learning classifier performance included, among others:

- F-score, i.e. the harmonic average of the precision (i.e. positive predictive value) and recall (i.e. sensitivity), ranging from 0 to 1 (perfect accuracy)
- Area under the precision-recall curve, i.e. an alternative to the area under the ROC curve, which is more informative for imbalanced classes.

A radiologist with experience in radiomics and artificial intelligence (R.C.) assessed Radiomics Quality Score in the attempt to estimate the methodological rigor of our study, as suggested by Lambin et al. (27).

5.2.9 Role of funding source

This research was partially funded by the Young Researchers Grant awarded by the European Society of Musculoskeletal Radiology (S.G.). The funding source provided financial support without any influence on the collection, analysis, and interpretation of data; on the writing of the report; and on the decision to submit the paper for publication.

5.3 Results

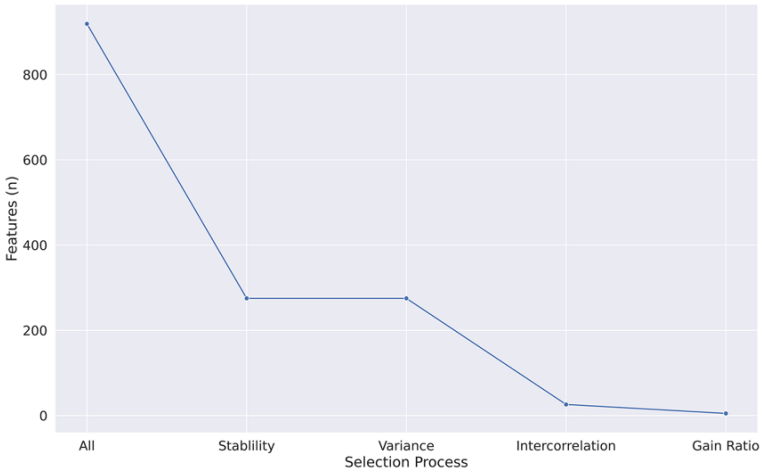
No difference in patients' age (p=0.25 [Mann-Whitney test]) and sex (p>0.99 [Fisher's test]) was found between the training and the external test cohorts. In our population, a bin width value of 3 presented the best results for feature extraction, with an average of 59 bins (± 30) in the training set. A total of 919 radiomic features were extracted from each segmentation. The rate of stable features was 30% (n=275), none of which had low variance. Removing all inter-correlated features yielded a dataset of 26 non-colinear features. Of these, the five with the highest gain ratio were selected and included: Major Axis

Length (2D shape-based) derived from the original images; Contrast (Neighbouring Gray Tone Difference Matrix) derived from wavelet-transformed images (Low-High pass filter); Short Run Low Gray Level Emphasis (Gray Level Run Length Matrix) from LoG-filtered images (sigma=5); Difference Entropy (Gray Level Co-occurrence Matrix) from the original images; Inverse Difference Moment (Gray Level Co-occurrence Matrix) derived from LoG-filtered images (sigma=2). Feature dimensionality reduction is shown in Fig. 3.

The machine learning classifier had 81% (89/110) and 75% (27/36) accuracy in identifying the cartilaginous bone lesions in the training and external test cohorts, respectively. Area under the ROC curve was, respectively, 0.89 and 0.78. In detail, its accuracy in classifying atypical cartilaginous tumours and higher-grade chondrosarcoma was 84% (46/55) and 78% (43/55) in the training cohort, and 81% (13/16) and 70% (14/20) in the external test cohort, respectively. Other evaluation metrics are derived from confusion matrix in Table 2 and reported in Table 3. Fig. 4 shows the ROC curve illustrating the classifier performance in the external test cohort. Fig. 5 shows the precision-recall curve illustrating the classifier performance for G2-G4 chondrosarcoma identification in the external test cohort. The final model had a Brier score of 0.25, while Fig. 6 depicts its calibration curve in the external test cohort. Our Radiomics Quality Score was 47% (Supplementary material).

In patients from centre 2, preoperative biopsy had 64% (23/36 correct tumour grade provided) accuracy in grading the cartilaginous bone lesions. Area under the ROC curve was 0.66. Preoperative biopsy provided an inconclusive result (n=5) or downgraded the lesion (n=8) in the remaining patients. Biopsy accuracy was slightly lower in comparison with the machine-learning classifier's accuracy, although this difference was not statistically significant (p=0.29 [McNemar's test]). The experienced radiologist had 81% (29/36 correct diagnosis provided) accuracy in identifying the cartilaginous bone lesions with no statistical difference compared to the classifier (p=0.75 [McNemar's test]). The radiologist's accuracy was 75% (4/16) and 85% (17/20) in classifying atypical cartilaginous tumours and higher-grade chondrosarcomas, respectively, as detailed in Table 4.

(2A)



(2B)

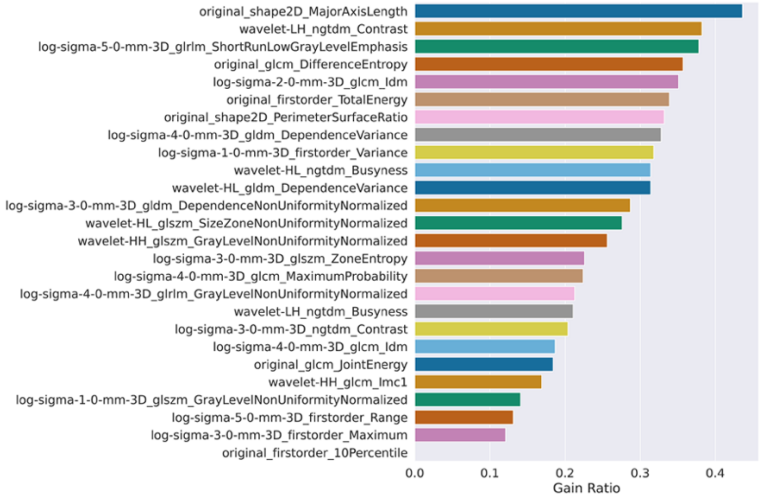


Fig. 3 Feature dimensionality reduction. **A** Feature selection process was performed exclusively using the training cohort data and included stability assessment as well as variance and intercorrelation analyses. The rate of stable features was 30% (n=275), none of which had low variance. Removing all inter-correlated features yielded a dataset of 26 non-colinear features. **B** The five features with the highest gain ratio were selected and included.

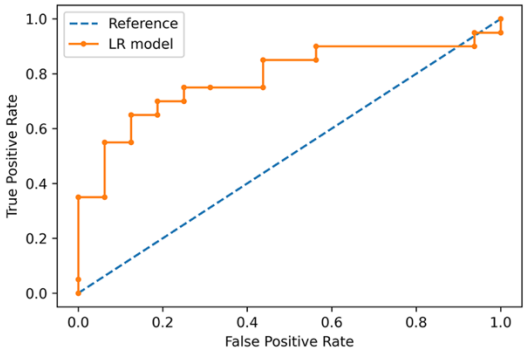


Fig. 4 ROC curve showing the classifier performance in the external test cohort.

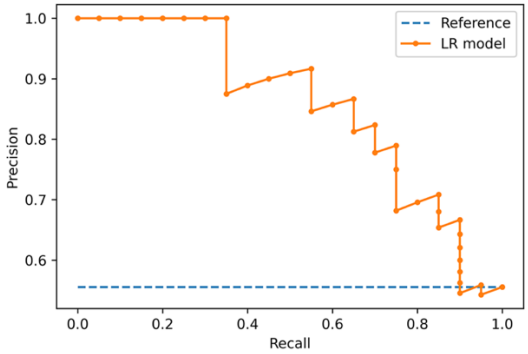


Fig. 5 Precision-recall curve illustrating the classifier performance for G2-G4 chondrosarcoma identification in the external test cohort.

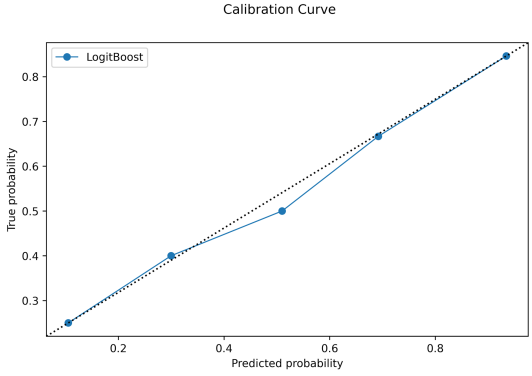


Fig. 6 Calibration curve in the external test cohort. The data is divided into bins, with the *y*-axis representing the distribution of positive cases in each bin while the *x*-axis the probability as predicted by the classifier. The closer the resulting calibration curve is to the reference line. the better the model's predictions reflect the actual class distribution in the dataset.

Table 2 Confusion matrix for the training and external test cohorts. ACT, atypical cartilaginous tumour; CS, higher-grade chondrosarcoma.

			Predicted class	
			ACT	CS
Actual class	Training	ACT	46	9
		CS	12	43
	External test	ACT	13	3
		CS	6	14

Table 3 Classifier accuracy metrics weighted average and by class in both the training and external test cohorts. ACT, atypical cartilaginous tumour; CS, higher-grade chondrosarcoma; FP, false positive; PRC, precision-recall curve; ROC, receiver operator curve; TP, true positive; WA, weighted average.

Cohort	Class	TP rate	FP rate	Precision	Recall	F-score	ROC	PRC
Training	ACT	0.836	0.218	0.793	0.836	0.814	0.891	0.876
	CS	0.782	0.164	0.827	0.782	0.804	0.891	0.915
	WA	0.809	0.191	0.810	0.809	0.809	0.891	0.895
External test	ACT	0.813	0.300	0.684	0.813	0.743	0.784	0.661
	CS	0.700	0.188	0.824	0.700	0.757	0.784	0.857
	WA	0.750	0.238	0.762	0.750	0.751	0.784	0.770

Table 4 Qualitative imaging assessment performed by the experienced radiologist. Lesion diameter is presented as median and interquartile (1^{st}-3^{rd}) range. Other variables are presented as proportions. ACT, atypical cartilaginous tumour; CS, higher-grade chondrosarcoma.

Class	Bone expansion	Cortical breakthrough	Aggressive periostitis	Soft-tissue mass	Maximum diameter	Correct diagnosis
ACT	1/16	3/16	1/16	0/16	45 (31-54) mm	12/16
CS	13/20	16/20	14/20	13/20	91 (59-124) mm	17/20
Overall	14/36	19/36	15/36	13/36	60 (42-100) mm	29/36

5.4 Discussion

The main finding of this study is that we developed a machine-learning classifier for discrimination between atypical cartilaginous tumours and higher-grade chondrosarcomas of long bones based on preoperative CT radiomic features, which achieved good accuracy in an independent test cohort from an external institution. Its performance did not differ in comparison with both an experienced bone tumour radiologist and preoperative biopsy.

Atypical cartilaginous tumours are locally aggressive lesions of the extremities, relatively indolent as compared with higher-grade tumours, and have a very low metastatic rate (2). Curettage is the standard of care (3), but its effectiveness in preventing transformation into higher-grade chondrosarcoma has not been demonstrated. Hence, given the similarity to benign enchondroma on both imaging (28) and histology (8), watchful

waiting has been proposed as an alternative strategy to prevent overtreatment and morbidity associated with surgery (29–31). An accurate differentiation from higher-grade chondrosarcomas requiring wide resection is thus necessary for treatment planning, and currently based on a multidisciplinary approach combining clinical presentation with imaging and biopsy (3). On imaging, MRI is the method of choice for local staging, while CT and PET-CT are employed for general staging (3). Both MRI (5) and PET-CT based on standard uptake values (32) are accurate in discriminating between atypical cartilaginous tumours and chondrosarcomas. On the other hand, biopsy may erroneously lead to tumour down-grading in large heterogenous lesions, as only small areas are sampled (7). Additionally, low reliability in tumour grading has been reported even among specialized bone pathologists (8) and the risk of biopsy-tract contamination also remains a concern. Thus, current imaging techniques may be further equipped to safely grade cartilaginous bone lesions non-invasively, and radiomics looks promising in this regard (9).

To date, radiomic studies have dealt with MRI of cartilaginous bone lesions with the aim of discriminating among benign enchondroma, atypical cartilaginous tumour and malignant chondrosarcoma (16,33,34), predicting local recurrence of chondrosarcoma (35) and differentiating chondrosarcoma from other entities such as skull chordoma (36). Machine learning was used in combination with radiomics to address some of these issues (16,35,36). Particularly, machine learning was a good predictor of tumour grade based on MRI radiomic features derived from unenhanced T1-weighted sequences, showing 75% accuracy in the test cohort for classification of atypical cartilaginous tumours and chondrosarcomas (16). This previous study included 58 patients from the same institution and the machine-learning classifier was internally tested using a hold-out set as a test cohort (16). To our knowledge, no published study has focused on machine learning and CT radiomics of cartilaginous bone lesions, as done in this study. We also attempted to differentiate atypical cartilaginous tumours from chondrosarcomas of long bones, as this is the most relevant clinical question and orientates towards a conservative approach or aggressive surgery. Our machine-learning classifier achieved 75% accuracy overall, 81% accuracy in identifying atypical cartilaginous tumours and 70% accuracy in identifying higher-grade chondrosarcomas in the external test cohort, respectively, with no difference compared to a dedicated radiologist with 12 years of experience in bone sarcomas (p=0.75 [McNemar's test]). These results agree with those previously reported for tumour grading based on MRI radiomics (16). Furthermore, our

findings were obtained in a more than twice larger population and validated in an independent test cohort from a second institution, thus ensuring their generalizability in clinical practice. Finally, although statistical significance was not reached (p=0.29 [McNemar's test]), the machine-learning classifier's accuracy was slightly superior compared to preoperative biopsy. We may speculate that this difference could become significant in a larger population.

Some limitations of our study need to be taken into account. First, our study is retrospective, as this design allowed including relatively large numbers of patients with an uncommon disease, such as chondrosarcoma, and imaging data already available. Additionally, a prospective analysis is not strictly needed for radiomic studies [13]. Second, we performed bidimensional segmentation and chose the image showing the maximum lesion extension. This decision was taken according to a recent study emphasizing that bidimensional segmentation yields better performance than volumetric approach (37), which would also be time-consuming in clinical practice. Third, feature stability was assessed by 3 readers only in a subgroup of 30 patients randomly selected from the training cohort, as 3 observers and 30 samples are the minimum numerical requirements for a reliability analysis according to the ICC guidelines by Koo et al. (19). Fourth, atypical cartilaginous tumours were twice more numerous than higher-grade chondrosarcomas in the training cohort. However, an imbalance of 2/3 is acceptable in machine-learning studies (38) and SMOTE was used to artificially oversample the minority class in the training cohort (25). Fifth, the training and external test cohorts respectively included CT scans and the CT portion of combined PET-CT scans with different acquisition parameters. Nonetheless, this is a further point in favour of the reliability of our findings, as the classifier performed well in both cohorts of patients. Sixth, only non-contrast CT was used in this study. However, contrast-enhanced CT was not available in patients from centre 2, as PET-CT was used. It was available only for a limited number of patients from centre 1, where preoperative assessment routinely included both CT and contrast-enhanced MRI; contrast was also administered before CT according to need, mainly to assess tumour-vessel relationships in case of high-grade chondrosarcoma. Our findings open the possibility for future studies to shed light on the value of contrast-enhanced CT radiomics and machine-learning assessment of cartilaginous bone tumours.

In conclusion, our machine-learning classifier showed good accuracy in differentiating atypical cartilaginous tumours from higher-grade chondrosarcomas of long bones based on radiomic features derived from preoperative CT scans. Our large population of study relative to such an uncommon disease, along with the good performance achieved in an independent cohort of patients from an external institution, supports the generalizability of our findings and their transferability into clinical practice. Our method may potentially aid in preoperative tumour characterization by integrating the multidisciplinary approach currently based on clinical, conventional imaging and histological assessment. Future investigations with prospective design are warranted to further validate our findings.

Acknowledgements

This research was partially funded by the Young Researchers Grant awarded by the European Society of Musculoskeletal Radiology (S. Gitto). The funding source provided financial support without any influence on the collection, analysis, and interpretation of data; on the writing of the report; and on the decision to submit the paper for publication.

Data sharing

The trained model and Weka result buffer, containing model weights as well as the entire training, validation and test run, are available in a publicly accessible repository (https://github.com/rcuocolo/csa_ct).

References

1. Murphey MD, Walker EA, Wilson AJ, Kransdorf MJ, Temple HT, Gannon FH. From the archives of the AFIP: imaging of primary chondrosarcoma: radiologic-pathologic correlation. Radiographics 2003;23:1245-1278. doi:10.1148/rg.235035134
2. Fletcher CDM, Bridge JA, Hogendoorn PCW, Mertens F. World Health Organization Classification of Tumours of Soft Tissue and Bone. Lyon, France: International Agency for Research on Cancer Press, 2013
3. Casali PG, Bielack S, Abecassis N, et al. Bone sarcomas: ESMO–PaedCan–EURACAN Clinical Practice Guidelines for diagnosis, treatment and follow-up. Ann Oncol 2018;29:iv79-iv95. doi:10.1093/annonc/mdy310
4. van Praag (Veroniek) VM, Rueten-Budde AJ, Ho V, et al. Incidence, outcomes and prognostic factors during 25 years of treatment of chondrosarcomas. Surg Oncol 2018;27:402-408. doi:10.1016/j.suronc.2018.05.009
5. Douis H, Singh L, Saifuddin A. MRI differentiation of low-grade from high-grade appendicular chondrosarcoma. Eur Radiol 2014;24:232-240. doi:10.1007/s00330-013-3003-y
6. Cannavò L, Albano D, Messina C, et al. Accuracy of CT and MRI to assess resection margins in primary malignant bone tumours having histology as the reference standard. Clin Radiol 2019;74:736.e13-736.e21. doi:10.1016/j.crad.2019.05.022
7. Hodel S, Laux C, Farei-Campagna J, Götschi T, Bode-Lesniewska B, Müller DA. The impact of biopsy sampling errors and the quality of surgical margins on local recurrence and survival in chondrosarcoma. Cancer Manag Res 2018;10:3765-3771. doi:10.2147/CMAR.S178768
8. Eefting D, Schrage YM, Geirnaerdt MJA, et al. Assessment of Interobserver Variability and Histologic Parameters to Improve Reliability in Classification and Grading of Central Cartilaginous Tumours. Am J Surg Pathol 2009;33:50-57. doi:10.1097/PAS.0b013e31817eec2b
9. van de Sande MAJ, van der Wal RJP, Navas Cañete A, et al. Radiologic differentiation of enchondromas, atypical cartilaginous tumours, and high-grade chondrosarcomas—Improving tumour-specific treatment: A paradigm in transit?

Cancer 2019;125:3288-3291. doi:10.1002/cncr.32404
10. Gillies RJ, Kinahan PE, Hricak H. Radiomics: Images Are More than Pictures, They Are Data. Radiology 2016;278:563-577. doi:10.1148/radiol.2015151169
11. Lubner MG, Smith AD, Sandrasegaran K, Sahani D V., Pickhardt PJ. CT Texture Analysis: Definitions, Applications, Biologic Correlates, and Challenges. Radiographics 2017;37:1483-1503. doi:10.1148/rg.2017170056
12. Kocak B, Durmaz ES, Ates E, Kilickesmez O. Radiomics with artificial intelligence: a practical guide for beginners. Diagn Interv Radiol 2019;25:485-495. doi:10.5152/dir.2019.19321
13. Chianca V, Cuocolo R, Gitto S, et al. Radiomic Machine Learning Classifiers in Spine Bone Tumors: A Multi-Software, Multi-Scanner Study. Eur J Radiol 2021;137:109586. doi: 10.1016/j.ejrad.2021.109586
14. Choy G, Khalilzadeh O, Michalski M, et al. Current Applications and Future Impact of Machine Learning in Radiology. Radiology 2018;288:318-328. doi:10.1148/radiol.2018171820
15. Cuocolo R, Caruso M, Perillo T, Ugga L, Petretta M. Machine Learning in oncology: A clinical appraisal. Cancer Lett 2020;481:55-62. doi:10.1016/j.canlet.2020.03.032
16. Gitto S, Cuocolo R, Albano D, et al. MRI radiomics-based machine-learning classification of bone chondrosarcoma. Eur J Radiol 2020;128:109043. doi:10.1016/j.ejrad.2020.109043
17. Li X, Morgan PS, Ashburner J, Smith J, Rorden C. The first step for neuroimaging data analysis: DICOM to NIfTI conversion. J Neurosci Methods 2016;264:47-56. doi:10.1016/j.jneumeth.2016.03.001
18. Yushkevich PA, Piven J, Hazlett HC, et al. User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability. Neuroimage 2006;31:1116-1128. doi:10.1016/j.neuroimage.2006.01.015
19. Koo TK, Li MY. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. J Chiropr Med 2016;15:155-163. doi:10.1016/j.jcm.2016.02.012
20. van Griethuysen JJM, Fedorov A, Parmar C, et al. Computational Radiomics

System to Decode the Radiographic Phenotype. Cancer Res 2017;77:e104-e107. doi:10.1158/0008-5472.CAN-17-0339

21. Zwanenburg A, Vallières M, Abdalah MA, et al. The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping. Radiology 2020;295:328–38. doi: 10.1148/radiol.2020191145.
22. Frank E, Hall M, Trigg L, Holmes G, Witten IH. Data mining in bioinformatics using Weka. Bioinformatics 2004;20:2479-2481. doi:10.1093/bioinformatics/bth261
23. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res 2011:12; 825-2830.
24. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2020.
25. Chawla N V., Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. J Artif Intell Res 2002;16:321-357. doi:10.1613/jair.953
26. Parlier-Cuau C, Bousson V, Ogilvie CM, Lackman RD, Laredo J-D. When should we biopsy a solitary central cartilaginous tumour of long bones? Literature review and management proposal. Eur J Radiol 2011;77:6-12. doi:10.1016/j.ejrad.2010.06.051
27. Lambin P, Leijenaar RTH, Deist TM, et al. Radiomics: the bridge between medical imaging and personalized medicine. Nat Rev Clin Oncol 2017;14:749–62. doi: 10.1038/nrclinonc.2017.141
28. Crim J, Schmidt R, Layfield L, Hanrahan C, Manaster BJ. Can imaging criteria distinguish enchondroma from grade 1 chondrosarcoma? Eur J Radiol 2015;84:2222-2230. doi:10.1016/j.ejrad.2015.06.033
29. Zoccali C, Baldi J, Attala D, et al. Intralesional vs. extralesional procedures for low-grade central chondrosarcoma: a systematic review of the literature. Arch Orthop Trauma Surg 2018;138:929-937. doi:10.1007/s00402-018-2930-0
30. Deckers C, Schreuder BHW, Hannink G, de Rooy JWJ, van der Geest ICM. Radiologic follow-up of untreated enchondroma and atypical cartilaginous tumours in the long bones. J Surg Oncol 2016;114:987-991.

doi:10.1002/jso.24465

31. Omlor GW, Lohnherr V, Lange J, et al. Outcome of conservative and surgical treatment of enchondromas and atypical cartilaginous tumours of the long bones: retrospective analysis of 228 patients. BMC Musculoskelet Disord 2019;20:134. doi:10.1186/s12891-019-2502-7
32. Annovazzi A, Anelli V, Zoccali C, et al. 18F-FDG PET/CT in the evaluation of cartilaginous bone neoplasms: the added value of tumour grading. Ann Nucl Med 2019;33:813-821. doi:10.1007/s12149-019-01392-3
33. Fritz B, Müller DA, Sutter R, et al. Magnetic Resonance Imaging–Based Grading of Cartilaginous Bone Tumours. Invest Radiol 2018;53:663-672. doi:10.1097/RLI.0000000000000486
34. Lisson CS, Lisson CG, Flosdorf K, et al. Diagnostic value of MRI-based 3D texture analysis for tissue characterisation and discrimination of low-grade chondrosarcoma from enchondroma: a pilot study. Eur Radiol 2018;28:468-477. doi:10.1007/s00330-017-5014-6
35. Yin P, Mao N, Liu X, et al. Can clinical radiomics nomogram based on 3D multiparametric MRI features and clinical characteristics estimate early recurrence of pelvic chondrosarcoma? J Magn Reson Imaging 2020;51:435-445. doi:10.1002/jmri.26834
36. Li L, Wang K, Ma X, et al. Radiomic analysis of multiparametric magnetic resonance imaging for differentiating skull base chordoma and chondrosarcoma. Eur J Radiol 2019;118:81-87. doi:10.1016/j.ejrad.2019.07.006
37. Ren J, Yuan Y, Qi M, Tao X. Machine learning–based CT texture analysis to predict HPV status in oropharyngeal squamous cell carcinoma: comparison of 2D and 3D segmentation. Eur Radiol 2020;30:6858-6866. doi:10.1007/s00330-020-07011-4
38. Sollini M, Antunovic L, Chiti A, Kirienko M. Towards clinical application of image mining: a systematic review on artificial intelligence and radiomics. Eur J Nucl Med Mol Imaging 2019;46:2656-2672. doi:10.1007/s00259-019-04372-x

Supplementary material

	Radiomics Quality Score
Item 1	1
Item 2	1
Item 3	0
Item 4	0
Item 5	3
Item 6	0
Item 7	1
Item 8	0
Item 9	2
Item 10	1
Item 11	0
Item 12	3
Item 13	2
Item 14	2
Item 15	0
Item 16	1
Total	17
Total (%)	47,22

Reference: Lambin 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

Chapter 6

MRI radiomics-based machine learning classification of atypical cartilaginous tumour and grade II chondrosarcoma of long bones

Gitto S, Cuocolo R, van Langevelde K, van de Sande MAJ, Parafioriti A, Luzzati A, Imbriaco M, Sconfienza LM, Bloem JL

EBioMedicine 2022; 75:103757
DOI: 10.1016/j.ebiom.2021.103757

List of abbreviations (Chapter 6)

ACT, atypical cartilaginous tumour

CT, computed tomography

CS, chondrosarcoma

CS2, grade II chondrosarcoma

ICC, intraclass correlation coefficient

LASSO, least absolute shrinkage and selection operator

MRI, magnetic resonance imaging

RFE, recursive feature elimination

SMOTE, synthetic minority oversampling technique

WHO, World Health Organization

Abstract

Background. Atypical cartilaginous tumour (ACT) and grade II chondrosarcoma (CS2) of long bones are respectively managed with watchful waiting or curettage and wide resection. Preoperatively, imaging diagnosis can be challenging due to interobserver variability and biopsy suffers from sample errors. The aim of this study is to determine diagnostic performance of MRI radiomics-based machine learning in differentiating ACT from CS2 of long bones.

Methods. One-hundred-fifty-eight patients with surgically treated and histology-proven cartilaginous bone tumours were retrospectively included at two tertiary bone tumour centres. The training cohort consisted of 93 MRI scans from centre 1 (n=74 ACT; n=19 CS2). The external test cohort consisted of 65 MRI scans from centre 2 (n=45 ACT; n=20 CS2). Bidimensional segmentation was performed on T1-weighted MRI. Radiomic features were extracted. After dimensionality reduction and class balancing in centre 1, a machine-learning classifier (Extra Trees Classifier) was tuned on the training cohort using 10-fold cross-validation and tested on the external test cohort. In centre 2, its performance was compared with an experienced musculoskeletal oncology radiologist using McNemar's test.

Findings. After tuning on the training cohort (AUC=0.88), the machine-learning classifier had 92% accuracy (60/65, AUC=0.94) in identifying the lesions in the external test cohort. Its accuracies in correctly classifying ACT and CS2 were 98% (44/45) and 80% (16/20), respectively. The radiologist had 98% accuracy (64/65) with no difference compared to the classifier (p=0.134).

Interpretation. Machine learning showed high accuracy in classifying ACT and CS2 of long bones based on MRI radiomic features.

Funding. ESSR Young Researchers Grant.

Research in context

Evidence before this study. Radiomic studies to date have focused on the classification of bone chondrosarcoma, including atypical cartilaginous tumour and high-grade chondrosarcoma, using radiomics alone or combined with machine learning. In long bones, therapeutic strategies for those lesions are entirely different and mainly based on imaging. In a recent study, we focused on CT radiomics-based machine learning and the distinction between atypical cartilaginous tumour and high-grade (II and higher) chondrosarcoma of long bones, including 120 patients from two institutions. Machine learning had 75% accuracy with no difference compared to an experienced radiologist. Previously, we used machine learning in combination with MRI radiomics to discriminate atypical cartilaginous tumour from high-grade chondrosarcoma. Only 58 patients from the same centre were included and the machine learning classifier was internally tested using a hold-out set as a test cohort, achieving 75% accuracy.

Added value of this study. In the current study, we attempted to differentiate atypical cartilaginous tumours from grade II chondrosarcoma of long bones using MRI radiomics-based machine learning. Higher-grade chondrosarcomas are more easily identified on MRI and were thus not included. The population of our current study was larger than previous publications, including 158 patients from two specialized institutions, which allowed for model validation on independent data from the external test cohort. Our classifier had 92% accuracy based on T1-weighted MRI radiomics, overlapping a dedicated bone tumour radiologist with 35-year experience who read all available MRI sequences. Thus, compared to previous studies, our method showed better performance to solve the most relevant clinical problem of atypical cartilaginous tumour/grade II chondrosarcoma differentiation.

Implications of all the available evidence. Radiomics-based machine learning is an objective method that may be used in clinical decision making by accurately differentiating atypical cartilaginous tumour from chondrosarcoma of long bones.

6.1 Introduction

Chondrosarcoma (CS) accounts for 20-30% of primary bone tumours in adulthood [1]. Based upon pathology, conventional CS was graded into three categories, where grade I, also called atypical cartilaginous tumour (ACT), has an indolent biologic behaviour, whereas grades II-III are aggressive malignant tumours with metastatic potential and high recurrence rates after surgery [2]. In the 2020 edition of the World Health Organization (WHO) classification, the term ACT is reserved for formerly named ACT/grade I CS only when located in long bones [3]. Cartilaginous tumours with the same histology, but located in the axial skeleton, are classified as grade I CS [3]. ACTs of long bones are indolent as compared to axial grade I CS and appendicular or axial grade II-III CS. Also, the increase of prevalence of ACT secondary to increased use of MRI over the past decades, relative to the lack of increase of grade II-III CS in the long bones, does not support the previous opinion that there is a risk of higher-grade CS developing in ACT [4]. Thus, this new classification better connects to therapeutic options that are different between ACT and CS grades I-III. Intralesional curettage, or even watchful waiting has been proposed for ACT, whereas for CS grades I-III, wide resection remains the therapy of choice [5–8].

As a consequence of these therapeutic options, clinical management currently depends on our ability to differentiate between ACT and grade II CS (CS2) of long bones [8]. Biopsy suffers from sample errors and is no longer standard of care in many tertiary centres [9]. MRI is the method of choice for diagnosis and differentiating between ACT and CS2 in long bones [10]. There is, however, discussion on accuracy of the various subjective MRI parameters, and there is the inherent interobserver variability [11,12]. New imaging-based tools like radiomics have recently been proposed to characterize cartilaginous bone tumours more objectively [13,14]. Radiomics includes the analysis of quantitative features extracted from imaging studies, known as radiomic features, which can be combined with machine learning algorithms to create classification models for the diagnosis of interest [15–17].

Machine learning has already shown good accuracy in discriminating ACT from all-grade CS based on computed tomography (CT) [13] and MRI [14] radiomics. However, no validated study to date has addressed the more relevant and specific distinction between ACT and CS2. Thus, the aim of this study is to determine diagnostic performance of MRI radiomics-based machine learning for classification of ACT and CS2 of long bones.

6.2 Methods

6.2.1 Ethics

Institutional Review Board from each involved centre approved this retrospective study and waived the need for informed consent (Protocols: "RETRORAD" in centre 1 and "G19.047" in centre 2). Patients included in this study granted written permission for anonymized data use for research purposes at the time of the MRI. After matching imaging, pathological, and surgical data, our database was completely anonymized to delete any connections between data and patients' identity according to the General Data Protection Regulation for Research Hospitals.

6.2.2 Study design and inclusion/exclusion criteria

Consecutive patients with ACT or CS2 of long bones and MRI available at one of two tertiary bone tumour centres (centre 1, IRCCS Orthopaedic Institute Galeazzi, Milan, Italy; centre 2, Leiden University Medical Centre, Leiden, The Netherlands) were considered for inclusion. Information was retrieved through medical records from the orthopaedic surgery and pathology departments. Inclusion criteria were: (i) ACT or primary central CS2 of long bones that was surgically treated with curettage or resection; (ii) definitive pathological diagnosis based on the surgical specimen assessment; (iii) MRI scan with at least T1-weighted and fluid-sensitive sequences in two directions performed within 3 months before surgery. Exclusion criteria were: (i) metacarpal, metatarsal, and phalangeal lesions; (ii) recurrent lesions; (iii) presence of pathological fracture. A flowchart of the patient selection process is shown in Fig. 1.

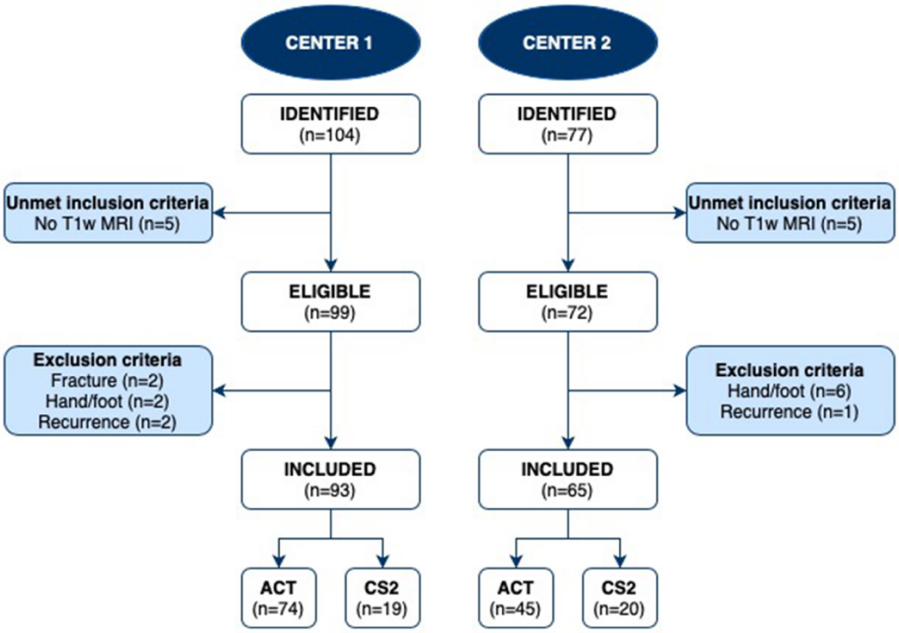


Fig. 1 Flowchart of patient selection.

6.2.3 Study cohorts

One-hundred-fifty-eight patients were retrospectively included. The training cohort consisted of 93 MRI scans from Centre 1 (n=74 ACT; n=19 CS2). The external test cohort consisted of 65 MRI scans from Centre 2 (n=45 ACT; n=20 CS2). Patients' demographics and data regarding lesion location are detailed in Table 1. In Centre 1, examinations were performed on one of two 1.5-T MRI systems (Magnetom Avanto, Siemens Healthineers, Erlangen, Germany; or Magnetom Espree, Siemens Healthineers, Erlangen, Germany). In Centre 2, examinations were performed on a 3-T (Ingenia or Intera, Philips Medical System, The Netherlands) or 1.5-T (Ingenia, Philips Medical System, The Netherlands) MRI system. Also, externally obtained MRI scans of patients referred to centre 2 were included in this study as long as the minimal MRI protocol was available. MRI specifications for Centre 1 and Centre 2 are summarized in Supplementary Table 1. All DICOM images were extracted and converted to the NiFTI format prior to the analysis using the dcm2niix software [18].

Table 1 Demographics and clinical data. Age is presented as median and interquartile (1st-3rd) range.

	Center 1	Center 2
Age	53 (45-62) years	62 (49-72) years
Sex	Men: n=29 Women: n=64	Men: n=31 Women: n=34
Lesion location	Femur: n=41 Fibula: n=9 Humerus: n=37 Radius: n=1 Tibia: n=5	Femur: n=46 Humerus: n=10 Tibia: n=9

6.2.4 Segmentation

A 2-year-experienced musculoskeletal radiologist (S.G.) performed contour-focused segmentation on preoperative T1-weighted MRI using the freely available, open-source software ITK-SNAP (v3.8) [19]. The axial, as first choice, or coronal or sagittal sequence was used based on availability and lesion location. In detail, bidimensional regions of interest were manually annotated on the slice showing the maximum lesion diameter. Radiomic analysis was not performed on fluid-sensitive sequences based on previous findings that, when extracting both T2- and T1-weighted MRI features, only the latter passed feature selection during dimensionality reduction [14]. Contrast-enhanced MRI was not available in all our cases, particularly ACT in centre 1, and was also not used.

In order to meet the numerical requirements of a reliability analysis according to the intraclass correlation coefficient (ICC) guidelines by Koo et al. [20], namely 3 observers and 30 observations, segmentations were additionally performed by other two radiologists in a subgroup of 30 patients randomly extracted from the training cohort. The additional segmentations performed by the second and third readers on this subset of 30 patients were exclusively used to assess feature reproducibility. The segmentations employed to build and test the classification model were all performed by the first reader. Each radiologist was independent and unaware of the slice other readers selected for segmentation, as well as blinded regarding lesion grading and disease course.

6.2.5 Feature extraction

Image pre-processing and feature extraction were performed using PyRadiomics (v3.0.1) [21]. The suggested pre-processing steps were employed [22]: image resampling, grey level normalization and discretization. In particular, pixels were resampled to a 1×1 mm in-plane resolution, z-score normalized to a 0-600 grey level value range and discretized with a

fixed bin width. In order to determine the ideal bin width value, a preliminary extraction exclusively of the first order range parameter was performed on training data alone. The parameter file for the radiomic data extraction is available in a freely accessible online repository (https://github.com/rcuocolo/mri_act_cs2).

Radiomic features were obtained from original and filtered images, including Laplacian of Gaussian filtering and wavelet decomposition. All available radiomic features for bidimensional masks were extracted (https://pyradiomics.readthedocs.io/en/latest/features.html), subdivided into the following classes: first-order (histogram analysis), 2D shape-based, Gray Level Co-occurrence Matrix, Gray Level Size Zone Matrix, Gray Level Run Length Matrix, Neighbouring Gray Tone Difference Matrix and Gray Level Dependence Matrix.

6.2.6 Machine learning analysis

Radiomic data processing and machine learning analysis were performed using the "irr" R package [23], "pandas" and "scikit-learn" Python packages [24]. Radiomic feature selection was performed using the training cohort data alone and consisted of stability, variance and pairwise correlation analyses as well as cross-validation based least absolute shrinkage and selection operator (LASSO) regression and recursive feature elimination (RFE). Feature stability was assessed by obtaining feature ICC using a two-way random effect, single rater, absolute agreement model. Features were considered stable if the ICC 95% confidence interval lower bound was ≥ 0.75. Then, low variance (threshold = 0.01) and highly intercorrelated (Pearson correlation coefficient threshold ≥ 0.80) were removed. LASSO regression coefficient analysis followed by RFE were finally used to determine the feature set to employ for model training. RFE used an Extra Trees model with default hyperparameters as its estimator and area under the ROC curve as the reference score. Both LASSO and RFE employed 10-fold stratified cross-validation.

Given the unbalanced nature of the training cohort, the synthetic minority oversampling technique (SMOTE) was used to balance the dataset by creating new instances from the minority class in Centre 1, thus increasing the number of CS2 to n=74 [25]. No oversampling was performed in the external test cohort. Thus, a machine-learning classifier (Extra Trees Classifier) was tuned via 10-fold stratified cross-validation using a random hyperparameter search on the training cohort. Decision tree forests are a commonly

employed ensemble machine learning architecture. As decision trees alone have a tendency to overfit the training data, the use of random resampling through bootstrapping and a subsample of the available features reduces model variance by introducing a degree of randomness. Compared to Random Forests, Extra Trees also perform random selection of feature thresholds within each tree node. This leads to further reduce the variance of the final ensemble (https://scikit-learn.org/stable/modules/ensemble.html#forest). The random search hyperparameter space was defined as follows:

1. Number of trees = 100-1000
2. Criterion = entropy, Gini
3. Maximum tree depth = 1-10
4. Maximum number of features per tree = 1-All
5. Bootstrap = true, false
6. Maximum number of samples per tree = 0-100%

The training process also included sigmoid model calibration via 5-fold stratified cross-validation nested within each loop of the 10-fold stratified cross-validation. The final model consisted of the best performing pipeline which was then fitted on the entire training dataset and tested on the external test cohort. Our radiomics-based machine learning workflow is illustrated in Fig. 2. This workflow is similar to one recent study from our group [13], with differences mainly related to feature selection process and machine learning classification. To offer some insights on the model's predictions, Shapley values were obtained for each feature using the "SHAP" Python package [26]. These provide a game-theory based assessment of the contribution of each parameter to the final output of the classifier.

6.2.7 Qualitative imaging assessment

An expert bone tumour radiologist with 35 years of work experience in a tertiary sarcoma centre (J.L.B.) read all MRI studies from the external test cohort blinded to any information about lesion grading, disease course and radiomics-based machine learning analysis. All available MRI sequences were used for qualitative assessment. The following parameters were assessed to differentiate CS2 from ACT and give the final impression: peritumoral bone marrow oedema, expansion of the medullary canal with thinner cortex, cortical breakthrough, periosteal reaction and cortical remodelling, reactive soft-tissue oedema and soft-tissue extension [10].

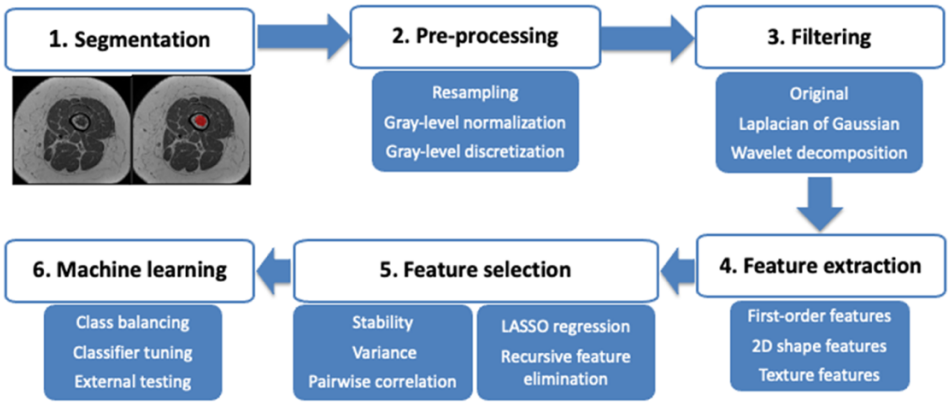


Fig. 2 Radiomics-based machine learning workflow pipeline. This workflow is similar to one recent study from our group [13], with differences mainly related to feature selection process and machine learning classification.

6.2.8 Statistical analysis

Continuous data are presented as median and interquartile (1^{st}-3^{rd}) range. Categorical data are presented as value counts and proportions. The R "stats" package was used for the following statistical analyses. Chi-square test and Mann-Whitney tests were used to evaluate sex and age differences between the training and external test cohorts, respectively. In the external test cohort, McNemar's test was used to compare the classifier performance with the radiologist's one. A two-sided p-value <0.05 indicated statistical significance.

Accuracy measures of the classifier performance included, among others: F-score, which is the harmonic average of precision (also known as positive predictive value) and recall (also known as sensitivity) and ranges from 0 to 1 (perfect accuracy); area under the precision-recall curve, which is an alternative to the area under the ROC curve and more informative for imbalanced classes.

6.2.9 Role of funding source

This research was partially funded by the Young Researchers Grant awarded by the European Society of Musculoskeletal Radiology (S.G.). The funding source provided financial support without any influence on the study design; on the collection, analysis, and

interpretation of data; and on the writing of the report. The first author had the final responsibility for the decision to submit the paper for publication.

6.3 Results

No statistical difference in sex (p=0.053 [Chi-square test]) was present between the training (64 women and 29 men) and external test (34 women and 31 men) cohorts. Age was younger (p=0.001 [Mann-Whitney test]) in patients from the training cohort (53 [45-62] years) compared to the external test cohort (62 [49-72] years). A bin width value of 3 presented the best results for feature extraction, with a median of 34 (22-55) bins in the training cohort. A total of 919 radiomic features were extracted from each lesion. The rate of stable features was 78% (n = 720). Removing low variance (n = 2) and highly inter-correlated (n=633) features yielded a dataset of 87 features. Next, features with LASSO coefficients shrinking to zero (n=67) were removed. Of the remaining features, an optimal number of 17 features was identified with RFE, as summarized in Table 2.

After tuning on the training cohort (AUC=0.88), the machine-learning classifier had 92% accuracy (60/65) in identifying the cartilaginous bone lesions in the external test cohort. Specifically, its accuracy in classifying ACT and CS2 was 98% (44/45) and 80% (16/20), respectively. Areas under the ROC (Fig. 3) and precision-recall (Fig. 4) curves were 0.94 and 0.90, respectively. Other evaluation metrics are derived from confusion matrix in Table 3 and detailed in Table 4. Fig. 5 depicts the calibration curve of the classifier in the external test cohort. The Brier score was 0.09, with lower values suggestive for better calibration. Shapley values for the model are presented in Fig. 6. The model, its implementation instructions, all required files for data extraction and processing are available in the online study repository (https://github.com/rcuocolo/mri_act_cs2).

The experienced radiologist had 98% accuracy (64/65 correct diagnosis provided) in classifying the lesions with no statistical difference compared to the classifier (p=0.134 [McNemar's test]). The radiologist's accuracy was 100% (45/45) and 95% (19/20) in classifying ACT and CS2, respectively. The radiologist and the classifier agreed on the final diagnosis in 94% (61/65) of cases, as one case was misdiagnosed by both. Fig. 7 shows cartilaginous lesions of long bones in three different patients from the external test cohort.

Table 2 List of selected features by feature class and source image, including original, Laplacian of Gaussian-filtered (LoG) and wavelet-transformed images.

Feature name	Feature class	Source image
10^{th} percentile	First Order	Original
Minor Axis Length	2D shape	Original
Informational Measure of Correlation 2	GLCM	LoG (sigma = 1)
Inverse Difference Normalized	GLCM	LoG (sigma = 1)
Run Entropy	GLRLM	LoG (sigma = 1)
Informational Measure of Correlation 1	GLCM	LoG (sigma = 2)
Dependence Variance	GLDM	LoG (sigma = 2)
Small Area Emphasis	GLSZM	LoG (sigma = 3)
Dependence Variance	GLDM	LoG (sigma = 3)
Informational Measure of Correlation 1	GLCM	LoG (sigma = 4)
Informational Measure of Correlation 1	GLCM	LoG (sigma = 5)
Small Area Emphasis	GLSZM	LoG (sigma = 5)
Gray Level Non-Uniformity	GLDM	Wavelet (low-high pass filter)
Informational Measure of Correlation 1	GLCM	Wavelet (high-high pass filter)
Size-Zone Non-Uniformity Normalized	GLSZM	Wavelet (high-high pass filter)
Short Run Low Gray Level Emphasis	GLRLM	Wavelet (low-low pass filter)
Large Area Emphasis	GLSZM	Wavelet (low-low pass filter)

Abbreviations. GLCM, Gray Level Co-occurrence Matrix; GLDM, Gray Level Dependence Matrix; GLRLM, Gray Level Run Length Matrix; GLSZM, Gray Level Size Zone Matrix.

Table 3 Confusion matrix for the external test cohort.

		Predicted class	
		ACT	**CS2**
Actual class	**ACT**	44	1
	CS2	4	16

Table 4 Classifier accuracy metrics weighted average and by class in the external test cohort.

Class	Precision	Recall	F-score
ACT	0.92	0.98	0.95
CS2	0.94	0.80	0.86
Weighted average	0.92	0.92	0.92

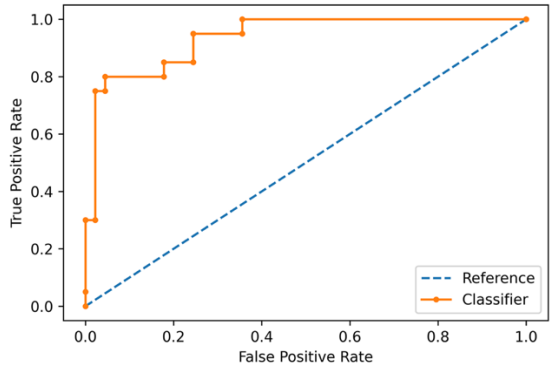


Fig. 3 ROC curve showing the classifier performance in the external test cohort.

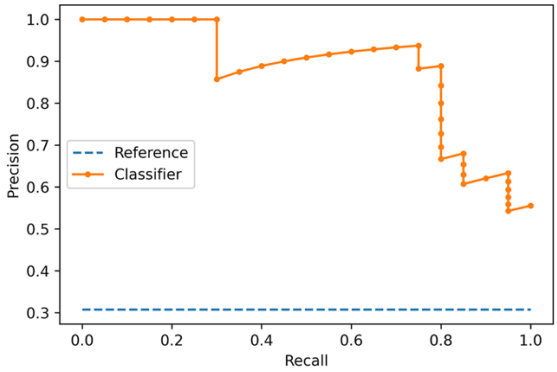


Fig. 4 Precision-recall curve illustrating the classifier performance in the external test cohort.

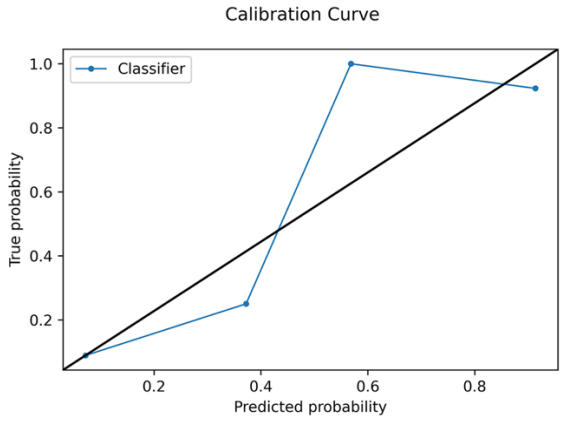


Fig. 5 Calibration curve in the external test cohort.

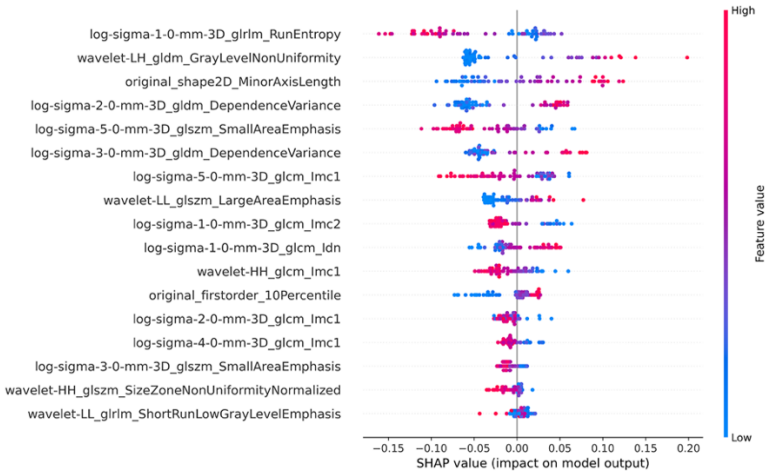


Fig. 6 Beeswarm plot of feature Shapley values in the final model.

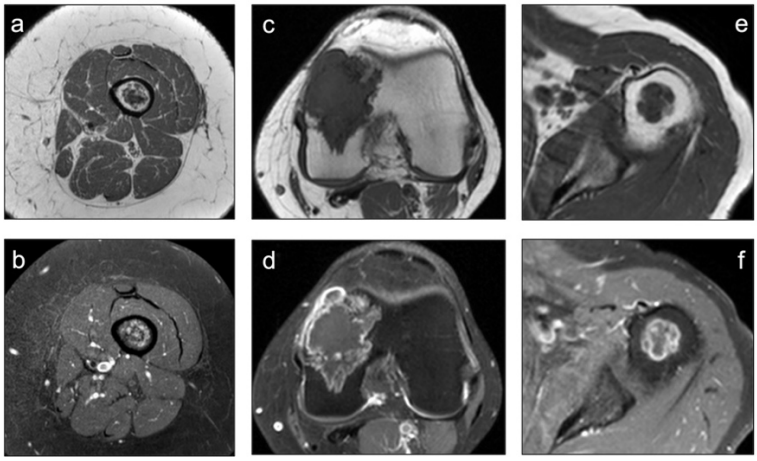


Fig. 7 Native and fat-saturated post-contrast T1-weighted sequences show three different cases of cartilaginous bone tumors, including ACT of the femur (a-b), CS2 of the femur (c-d) and CS2 of the humerus (e-f). Cortical breakthrough and soft-tissue extension are highly suspicious of high-grade lesion in the femur (c-d), whereas no suspicious feature is qualitatively seen in the humerus (e-f). Post-contrast images were qualitatively assessed by the radiologists, but they were not included in the radiomics-based machine learning analysis.

6.4 Discussion

The main finding of our study was that our machine learning method was 92% accurate in differentiating ACT from CS2 of long bones based on T1-weighted MRI radiomic features. This result was achieved in an independent cohort of patients from a second institution (external test cohort) and did not differ compared to a dedicated bone tumour radiologist with 35-year experience.

Our findings have clinical relevance as therapeutic strategies for ACT and CS2 in long bones are entirely different and mainly based on MRI. The difference in treatment strategies between ACT and enchondroma is disappearing, as watchful waiting in ACT has become an increasingly favoured option over intralesional curettage [6–8]. Thus, radiological focus has shifted from differentiating enchondroma from ACT towards identifying high grade CS. The exact, conservative, options for managing enchondroma and ACT are currently under evaluation, but there is consensus that CS2 needs wide resection [8]. Additionally, clinical outcome strongly depends on tumor grading, as reported 5- and 10-year overall survival rates are 87-99% and 88-95% for ACT/grade I CS, while they are 74-99% and 58-86% for CS2, respectively [3,4].

Radiomic studies to date have focused on the classification of cartilaginous bone tumours, such as enchondroma, ACT and high-grade CS, using radiomics alone [27–29] or combined with machine learning [13,14]. Particularly, in a recent study we focused on CT radiomics-based machine learning and the distinction between ACT and high-grade CS of long bones, including CS2, grade III and dedifferentiated CS in the latter group [13]. One-hundred-twenty patients were included from two institutions (IRCCS Orthopaedic Institute Galeazzi in Milan and IRCCS Regina Elena National Cancer Institute in Rome, Italy) and split into training and external test cohorts, as done in our current study. Machine learning had 75% accuracy in identifying the lesions in the external test cohort with no difference compared to an experienced radiologist [13]. Previously, we used machine learning in combination with non-contrast MRI radiomics to discriminate ACT from high-grade CS [14]. Only 58 patients from the same centre were included and the machine learning classifier was internally tested using a hold-out set as a test cohort, achieving 75% accuracy. In this work, radiomic features were extracted from both T1-weighted and T2-weighted sequences, but only T1-weighted MRI features were selected during dimensionality reduction (i.e. feature selection) process [14]. Based on this preliminary finding, in the current study we intentionally focused on T1-weighted MRI radiomics. Our current study addressed the most relevant clinical issue of differentiating between ACT and CS2 of long bones [8], thus excluding higher-grade CS, that more easily identified on MRI. The population of our study was larger than previous publications, including 158 patients from two specialized institutions (IRCCS Orthopaedic Institute Galeazzi in Milan, Italy and Leiden University Medical Centre in The Netherlands), which allowed for model validation on independent data from the external test

cohort. In the present study, the workflow was similar to the above discussed CT-based study from our group [13], although some differences mainly related to feature selection process and machine learning classification existed. Particularly, the pipeline was improved by employing a random search hyperparameter tuning process and classifier calibration through nested cross-validation. Our classifier (Extra Trees classifier) had 92% accuracy overall, 98% in identifying ACT and 80% in identifying CS2 in the external test cohort based on T1-weighted MRI radiomics, respectively, overlapping a dedicated bone tumour radiologist with 35-year experience who read all available MRI sequences. Thus, although the different outcome cannot be distinctly attributed to larger population, differences in workflow or input image (MRI, rather than CT as in [13]), our current method showed better performance than previous studies [13,14] to solve the clinical problem of ACT/CS2 differentiation.

Some limitations of our study need to be addressed. First, the design of our study is retrospective that, however, allowed including a large number of patients with a relatively uncommon disease. Also, a prospective analysis is not strictly necessary for radiomic studies [30]. Second, we performed bidimensional segmentation on the MRI slice showing the largest lesion diameter. This decision was taken following our recent finding that no difference in reproducible feature rates exists between bidimensional and volumetric MRI-based texture analysis [31], and the latter would also be less easily performed in clinical practice. Third, ACT was over-represented compared to CS2 in our population of study. However, this accurately reflects the incidence of ACT and CS2 in clinical practice [4], and class balancing was performed to artificially oversample the minority class in the training cohort [25]. Fourth, contrast-enhanced MRI was not used for radiomics-based machine learning analysis. On one hand, our intention was to keep our model as simple as possible by focusing on a single sequence and non-contrast T1-weighted images are almost always part of MRI protocols in these patients. On the other hand, we favoured having a large population of study over including contrast-enhanced MRI, which was not available in all our cases. Our findings open the possibility for future studies to investigate the added value of machine learning and contrast-enhanced MRI radiomics for classification of cartilaginous bone tumours. Finally, while a clear correlation of specific radiomic features with lesion phenotypical characteristics remains complex to identify, the Shapley value plot offers a degree of explainability and insight on the inner workings of our model.

In conclusion, our machine learning method was highly accurate in discriminating ACT from CS2 of long bones based on radiomic features obtained from T1-weighted MRI. Our large population of study and the excellent performance achieved using independent data from different institutions ensure the generalizability of our findings. Thus, radiomics-based machine learning is an objective MRI method that may be used in clinical decision making by accurately differentiating between ACT and CS2. Future studies are warranted to verify the transferability of our findings into clinical practice, particularly involving inexperienced radiologists, who may mostly benefit in using this tool. Additionally, our findings from the present and previous works may be compared with other studies from different groups, using meta-analysis, in order to deeper investigate the theoretical aspects of radiomics and machine learning regarding cartilaginous bone tumours.

Acknowledgements

This research was partially funded by the Young Researchers Grant awarded by the European Society of Musculoskeletal Radiology (S. Gitto). The funding source provided financial support without any influence on the study design; on the collection, analysis, and interpretation of data; and on the writing of the report. The first author had the final responsibility for the decision to submit the paper for publication.

Data sharing

The model, its implementation instructions, all required files for data extraction and processing are available in the online study repository (https://github.com/rcuocolo/mri_act_cs2).

References

1. Murphey MD, Walker EA, Wilson AJ, Kransdorf MJ, Temple HT, Gannon FH. From the archives of the AFIP: imaging of primary chondrosarcoma: radiologic-pathologic correlation. *Radiographics* 2003;**23**:1245–1278
2. Fletcher CDM, Bridge JA, Hogendoorn PCW, Mertens F. World Health Organization Classification of Tumours of Soft Tissue and Bone. Lyon, France: International Agency for Research on Cancer Press, 2013
3. WHO Classification of Tumours Editorial Board. WHO Classification of Tumours: Soft Tissue and Bone Tumours. Lyon, France: International Agency for Research on Cancer Press, 2020
4. van Praag (Veroniek) VM, Rueten-Budde AJ, Ho V, et al. Incidence, outcomes and prognostic factors during 25 years of treatment of chondrosarcomas. *Surg Oncol* 2018;**27**:402–408
5. Casali PG, Bielack S, Abecassis N, et al. Bone sarcomas: ESMO–PaedCan–EURACAN Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Ann Oncol* 2018;**29**:iv79–iv95
6. Deckers C, Schreuder BHW, Hannink G, de Rooy JWJ, van der Geest ICM. Radiologic follow-up of untreated enchondroma and atypical cartilaginous tumors in the long bones. *J Surg Oncol* 2016;**114**:987–991
7. Omlor GW, Lohnherr V, Lange J, et al. Outcome of conservative and surgical treatment of enchondromas and atypical cartilaginous tumors of the long bones: retrospective analysis of 228 patients. *BMC Musculoskelet Disord* 2019;**20**:134
8. van de Sande MAJ, van der Wal RJP, Navas Cañete A, et al. Radiologic differentiation of enchondromas, atypical cartilaginous tumors, and high-grade chondrosarcomas—Improving tumor-specific treatment: A paradigm in transit? *Cancer* 2019;**125**:3288–3291
9. Hodel S, Laux C, Farei-Campagna J, Götschi T, Bode-Lesniewska B, Müller DA. The impact of biopsy sampling errors and the quality of surgical margins on local recurrence and survival in chondrosarcoma. *Cancer Manag Res* 2018;**10**:3765–3771
10. Douis H, Singh L, Saifuddin A. MRI differentiation of low-grade from high-grade appendicular chondrosarcoma. *Eur Radiol* 2014;**24**:232–240

11. Jones KB, Buckwalter JA, McCarthy EF, et al. Reliability of Histopathologic and Radiologic Grading of Cartilaginous Neoplasms in Long Bones. *J Bone Joint Surg Am* 2007;**89**:2113–2123
12. Zamora T, Urrutia J, Schweitzer D, Amenabar PP, Botello E. Do Orthopaedic Oncologists Agree on the Diagnosis and Treatment of Cartilage Tumors of the Appendicular Skeleton? *Clin Orthop Relat Res* 2017;**475**:2176–2186
13. Gitto S, Cuocolo R, Annovazzi A, et al. CT radiomics-based machine learning classification of atypical cartilaginous tumours and appendicular chondrosarcomas. *EBioMedicine* 2021;**68**:103407
14. Gitto S, Cuocolo R, Albano D, et al. MRI radiomics-based machine-learning classification of bone chondrosarcoma. *Eur J Radiol* 2020;**128**:109043
15. Cuocolo R, Caruso M, Perillo T, Ugga L, Petretta M. Machine Learning in oncology: A clinical appraisal. *Cancer Lett* 2020;**481**:55–62
16. Gillies RJ, Kinahan PE, Hricak H. Radiomics: Images Are More than Pictures, They Are Data. *Radiology* 2016;**278**:563–577
17. Gitto S, Cuocolo R, Albano D, et al. CT and MRI radiomics of bone and soft-tissue sarcomas: a systematic review of reproducibility and validation strategies. *Insights Imaging* 2021;**12**:68
18. Li X, Morgan PS, Ashburner J, Smith J, Rorden C. The first step for neuroimaging data analysis: DICOM to NIfTI conversion. *J Neurosci Methods* 2016;**264**:47–56
19. Yushkevich PA, Piven J, Hazlett HC, et al. User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability. *Neuroimage* 2006;**31**:1116–1128
20. Koo TK, Li MY. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *J Chiropr Med* 2016;**15**:155–163
21. van Griethuysen JJM, Fedorov A, Parmar C, et al. Computational Radiomics System to Decode the Radiographic Phenotype. *Cancer Res* 2017;**77**:e104–e107
22. Zwanenburg A, Vallières M, Abdalah MA, et al. The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping. *Radiology* 2020;**295**:328–338
23. R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing, 2020

24. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python. *J Mach Learn Res* 2011;**12**:2825–2830
25. Chawla N V., Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. *J Artif Intell Res* 2002;**16**:321–357
26. Lundberg S, Lee S-I. A Unified Approach to Interpreting Model Predictions. *Adv Neural Inf Process Syst* 2017;**2017**:4766–4775
27. Fritz B, Müller DA, Sutter R, et al. Magnetic Resonance Imaging–Based Grading of Cartilaginous Bone Tumors. *Invest Radiol* 2018;**53**:663–672
28. Lisson CS, Lisson CG, Flosdorf K, et al. Diagnostic value of MRI-based 3D texture analysis for tissue characterisation and discrimination of low-grade chondrosarcoma from enchondroma: a pilot study. *Eur Radiol* 2018;**28**:468–477
29. Pan J, Zhang K, Le H, Jiang Y, Li W, Geng Y, et al. Radiomics Nomograms Based on Non-enhanced MRI and Clinical Risk Factors for the Differentiation of Chondrosarcoma from Enchondroma. *J Magn Reson Imaging* 2021;**54**:1314–23.
30. Lubner MG, Smith AD, Sandrasegaran K, Sahani DV, Pickhardt PJ. CT Texture Analysis: Definitions, Applications, Biologic Correlates, and Challenges. *Radiographics* 2017;**37**:1483–1503
31. Gitto S, Cuocolo R, Emili I, et al. Effects of Interobserver Variability on 2D and 3D CT- and MRI-Based Texture Feature Reproducibility of Cartilaginous Bone Tumors. *J Digit Imaging* 2021;**34**:820–32.

Supplementary material

Supplementary Table 1 MRI specifications for turbo spin echo T1-weighted axial sequence in both center 1 and center 2, expressed in millimeters. FOV, field of view.

	Center 1		**Center 2**	
	1.5T	1.5T	3T	1.5T
Humerus	FOV: 200 Thickness: 4.5 Pixel: 0.8x0.6	FOV: 160 Thickness: 3 Pixel: 0.8x0.6	FOV: 200 Thickness: 6 Pixel: 0.65x0.79	FOV: 200 Thickness: 6 Pixel: 0.55x0.69
Radius	FOV: 160 Thickness: 3 Pixel: 0.7x0.5	//	//	//
Proximal femur	FOV: 370 Thickness: 3 Pixel: 1x0.8	//	FOV: 300 Thickness: 8 Pixel: 0.96x0.96	FOV: 300 Thickness: 8 Pixel: 0.85x0.86
Distal femur	FOV: 180 Thickness: 3 Pixel: 0.7x0.5	FOV: 180 Thickness: 3 Pixel: 0.7x0.5	FOV: 300 Thickness: 8 Pixel: 0.96x0.96	FOV: 300 Thickness: 8 Pixel: 0.85x0.86
Fibula **Tibia**	FOV: 180 Thickness: 3 Pixel: 0.7x0.5	FOV: 180 Thickness: 3 Pixel: 0.7x0.5	FOV: 150 Thickness: 7 Pixel: 0.6x0.71	FOV: 150 Thickness: 7 Pixel: 0.6x0.7

Chapter 7

Summary and general discussion

List of abbreviations (Chapter 7)

2D, bidimensional

3D, volumetric

ACT, atypical cartilaginous tumor

CT, computed tomography

CS, chondrosarcoma

ICC, intraclass correlation coefficient

MRI, magnetic resonance imaging

PET-CT, positron emission tomography-computed tomography

ROI, region of interest

7.1 Summary

Chapter 1 provided a general introduction to the doctoral thesis. The aim of this thesis was to determine diagnostic performance of machine learning in differentiating between atypical cartilaginous tumor (ACT) and high-grade chondrosarcoma (CS) based on radiomic features derived from cross-sectional imaging, such as magnetic resonance imaging (MRI) and computed tomography (CT), in comparison with experienced musculoskeletal oncology radiologists.

In chapter 2 we introduced the concept of CT and MRI radiomics of bone and soft-tissue sarcomas by reviewing the issue of radiomic feature reproducibility and predictive model validation strategies. The ultimate goal of this systematic review was to promote achieving a consensus on these aspects in radiomic workflows and facilitate clinical transferability. Out of 278 identified papers, forty-nine papers published between 2008 and 2020 were included. They dealt with radiomics of bone (N=12) or soft-tissue (N=37) tumors. Eighteen (37%) studies included a feature reproducibility analysis. Inter/intra-reader segmentation variability was the theme of reproducibility analysis in 16 (33%) investigations, outnumbering the analyses focused on image acquisition or post-processing (N=2,4%). The intraclass correlation coefficient (ICC) was the most commonly used statistical method to assess reproducibility, which ranged from 0.6 and 0.9. At least one machine learning validation technique was used for model development in 25 (51%) papers and K-fold cross validation was the most commonly employed. A clinical validation of the model was reported in 19 (39%) papers. It was performed using a separate dataset from the primary institution (i.e., internal validation) in 14 (29%) studies and an independent dataset related to different scanners or from another institution (i.e., independent validation) in 5 (10%) studies. In conclusion, the issues of radiomic feature reproducibility and model validation varied largely among the studies dealing with musculoskeletal sarcomas. This should be addressed in future investigations to bring the field of radiomics from a preclinical research area to the clinical stage.

In chapter 3, the diagnostic performance of MRI radiomics-based machine learning in discriminating ACT from high-grade CS was evaluated in a preliminary single-center study. We retrospectively included 58 patients with histology-proven ACT (N=26) or high-grade CS (N=32, including 16 appendicular and 16 axial CS). They were randomly divided into training (N=42) and test (N=16) groups for model tuning and testing, respectively. All

tumors were manually segmented on T1-weighted and T2-weighted MRI by drawing bidimensional (2D) regions of interest (ROIs), which were used for radiomic feature extraction. After feature selection, an ensemble classifier (AdaBoostM1) was tuned on the training set using 10-fold cross-validation and then tested on the previously unseen test set. Thereafter, an experienced musculoskeletal radiologist blinded to histology and radiomic data qualitatively evaluated the lesions in the test group. The dataset was reduced to 4 T1-weighted MRI radiomic features after feature selection. The classifier correctly identified 85.7% (AUC=0.85) and 75% (AUC=0.78) of the lesions in the training and test groups, respectively. The radiologist correctly identified 81.3% of the lesions, with no difference compared to the classifier (p=0.453). In conclusion, our machine learning approach showed good diagnostic performance for classification of ACT and high-grade CS.

In chapter 4, the influence of interobserver manual segmentation variability on the reproducibility of 2D and volumetric (3D) CT- and MRI-based texture analysis was investigated. Thirty patients with cartilaginous bone tumors (N=10 enchondroma; N=10 ACT; N=10 high-grade CS) were retrospectively included. Three radiologists independently performed manual contour-focused segmentation on unenhanced CT, T1-weighted and T2-weighted MRI by drawing both a 2D ROI on the slice showing the largest tumor area and a 3D ROI including the whole tumor volume. Additionally, a marginal erosion was applied to both 2D and 3D segmentations to evaluate the influence of segmentation margins. A total of 783 and 1132 features were extracted from original and filtered 2D and 3D images, respectively. ICC≥0.75 defined feature stability. In 2D vs. 3D contour-focused segmentation, the rates of stable features for respectively CT, T1- and T2-weighted MRI were 74.71% vs. 86.57% (p<0.001), 77.14% vs. 80.04% (p=0.142) and 95.66% vs. 94.97% (p=0.554). Margin shrinkage did not improve 2D (p=0.343) and performed worse than 3D (p<0.001) contour-focused segmentation in terms of feature stability. In 2D vs. 3D contour-focused segmentation, matching stable features derived from CT and MRI were 65.8% vs. 68.7% (p=0.191), and those derived from T1-weighted and T2-weighted images were 76.0% vs. 78.2% (p=0.285). In conclusion, 2D and 3D CT and MRI radiomic features of cartilaginous bone tumors were reproducible, although some degree of interobserver segmentation variability highlighted the need for reliability analysis in radiomic studies.

Chapter 5 described a multicenter study that investigated the performance of CT radiomics-based machine learning in discriminating ACT from high-grade CS of long bones.

One-hundred-twenty patients with histology-proven lesions were retrospectively included. The training cohort consisted of 84 CT scans from center 1 (N=55 ACT; N=29 CS grade II-IV). The external test cohort consisted of the CT component of 36 positron emission tomography-CT (PET-CT) scans from center 2 (N=16 ACT; N=20 CS grade II-IV). 2D segmentation was performed on preoperative CT. Radiomic features were extracted. After dimensionality reduction and class balancing in center 1, the performance of a machine-learning classifier (LogitBoost) was assessed on the training cohort using 10-fold cross-validation and on the external test cohort. In center 2, its performance was compared with preoperative biopsy and with the classification of an experienced radiologist using McNemar's test. The classifier had 81% (AUC=0.89) and 75% (AUC=0.78) accuracy in identifying the lesions in the training and external test cohorts, respectively. Specifically, its accuracy in classifying ACT and high-grade CS was 84% and 78% in the training cohort, and 81% and 70% in the external test cohort, respectively. Preoperative biopsy had 64% (AUC=0.66) accuracy (p=0.29). The radiologist had 81% accuracy (p=0.75). In conclusion, machine learning showed good accuracy in classifying ACT and high-grade CS of long bones based on preoperative CT radiomic features.

Chapter 6 described a multicenter study that determined diagnostic performance of MRI radiomics-based machine learning in differentiating ACT from grade II CS of long bones. One-hundred-fifty-eight patients with surgically treated and histology-proven cartilaginous bone tumors were retrospectively included at two tertiary bone tumor centers. The training cohort consisted of 93 MRI scans from center 1 (N=74 ACT; N=19 CS grade II). The external test cohort consisted of 65 MRI scans from center 2 (N=45 ACT; N=20 CS grade II). 2D segmentation was manually performed on T1-weighted MRI sequences. First-order, shape-based and texture features were extracted. Dimensionality reduction consisting of stability, variance and inter-correlation analyses and recursive feature elimination, after class balancing in center 1 (CS grade II oversampled to N=74), was performed. Thus, a machine learning classifier (Extra Trees Classifier) was automatically tuned on the training cohort using 10-fold cross-validation and tested on the external test cohort. In center 2, its performance was compared with an experienced musculoskeletal oncology radiologist using McNemar's test. Nine-hundred-nineteen radiomic features were extracted and then reduced to 17 through dimensionality reduction. After tuning on the training cohort (AUC=0.88), the machine learning classifier had 92% accuracy (60/65, AUC=0.94) in identifying the lesions

in the external test cohort. Specifically, its accuracies in correctly classifying ACT and grade II CS were 98% (44/45) and 80% (16/20), respectively. The experienced radiologist had 98% accuracy (64/65) with no significant difference compared to the classifier (p=0.134). In conclusion, machine learning showed high accuracy in classifying ACT and grade II CS of long bones based on MRI radiomic features.

7.2 General discussion, limitations and future perspectives

This thesis focuses on the concept of ACT, which has been defined in 2013 and 2020 World Health Organization classifications [1,2]. This relatively new definition reflects the indolent biological behavior of ACT, which is now considered an intermediate cartilaginous tumor of long bones rather than a malignancy [2]. This concept connects better to therapeutic options that are entirely different for ACT compared to high-grade (II or higher) appendicular CS and all-grade axial CS [3]. Particularly, wide resection with negative margins is the therapy of choice for the latter group. The treatment of ACT has remarkably changed over the last three decades. Wide resection was performed until the nineties, then therapy changed towards intralesional curettage and nowadays watchful waiting with imaging follow-up is an increasingly favored option [4].

Because of the risk of sample errors that makes biopsy no longer standard of care in many tertiary centers [5] and given the increasing incidence of ACT due to an increase in incidental findings on MRI [6], there is a need for clear imaging criteria to differentiate between ACT and high-grade CS. Imaging assessment suffers, however, from interobserver variability [7,8]. Thus, in this thesis, radiomics has been proposed as an imaging method to differentiate ACT from high-grade CS more objectively, in combination with machine learning algorithms. Other radiomic studies to date have addressed the issue of discriminating enchondroma from ACT/high-grade CS [9–11]. However, the enchondroma/ACT differentiation has progressively become less relevant, due to the above mentioned, new concept of ACT and the ensuing new treatment options in which watchful waiting is an alternative to curettage [4,12,13]. Resection of ACT is no longer indicated. Because enchondroma and high-grade CS are easily differentiated on conventional imaging, diagnostic focus has moved from differentiating enchondroma from ACT to differentiating ACT from grade II CS [4].

In chapters 3, 5 and 6 of this thesis, machine learning algorithms were used to create classification models based on radiomic features extracted from MRI and CT. Particularly, chapters 5 and 6 described large multicenter studies including a clinical validation of the model on independent data (external test cohort) from a different institution. They focused on CT and MRI radiomics, respectively, and achieved good-to-high accuracy rates in correctly classifying ACT and high-grade CS of long bones, with no difference compared to experienced musculoskeletal oncology radiologists. Undoubtedly, the most relevant finding was achieved in chapter 6, where T1-weighted MRI radiomics-based machine learning had 92% accuracy in differentiating ACT from grade II CS. In addition, there was no statistical difference ($p=0.134$) in results between our approach and those of a dedicated bone tumor radiologist with 35 years of experience. In chapter 4 we methodologically analyzed the influence of interobserver segmentation variability on the reproducibility of 2D and 3D CT- and MRI radiomic features of cartilaginous bone tumors. All imaging modalities demonstrated good reproducibility both employing 2D and 3D annotations, although a certain degree of variability highlighted the need for a preliminary assessment of feature stability. This was also highlighted in the systematic review on feature reproducibility and validation strategies in chapter 2 and performed in our subsequent studies (chapters 5 and 6).

Some limitations of this thesis need to be addressed. First, all performed studies were retrospective, as this design allowed including a larger number of patients (with a relatively uncommon disease). Additionally, a prospective analysis was not strictly needed in radiomic studies [14]. Second, the segmentation method may have an impact on results. In studies described in chapters 3, 5 and 6, we performed 2D segmentation and selected the image slice showing the largest tumor dimension. A 2D approach was preferred as it would be easier to implement in clinical practice. In addition, recent literature suggested that it could yield better performance than 3D segmentation [15]. Furthermore, our findings in chapter 4 showed no difference in terms of feature reproducibility between 2D and 3D MRI-based texture analysis. Third, ACT was over-represented compared to high-grade CS in chapter 5, particularly in the training cohort from center 1, and in chapter 6. However, this accurately reflected the incidence of ACT and high-grade CS [6]. Furthermore, class balancing was performed in both studies to artificially oversample the minority class in the training cohort [16]. Fourth, contrast-enhanced CT and MRI were not used for radiomics-based machine learning analysis, as they were not available in all patients. Still, our good results are

encouraging and open the possibility for future research to shed light on the value of machine learning and contrast-enhanced CT or MRI radiomics for ACT/high-grade CS classification.

In conclusion, CT and MRI radiomics-based machine learning demonstrated good-to-high accuracy in differentiating ACT from high-grade CS and looks promising as an objective imaging method that may be used in clinical decision making. This has potential, especially in general practice without presence of specialized expertise, in identifying the commonly encountered ACT. Our large population of study and the very good performance achieved using independent data from different institutions, as presented in chapters 5 and 6, ensure the generalizability of our results. Future studies will have to verify the applicability of our findings in clinical practice.

References

[1] Fletcher CDM, Bridge JA, Hogendoorn PCW, Mertens F. World Health Organization Classification of Tumours of Soft Tissue and Bone. Lyon, France: International Agency for Research on Cancer Press; 2013.

[2] WHO Classification of Tumours Editorial Board. WHO Classification of Tumours: Soft Tissue and Bone Tumours. Lyon, France: International Agency for Research on Cancer Press; 2020.

[3] Casali PG, Bielack S, Abecassis N, Aro HT, Bauer S, Biagini R, et al. Bone sarcomas: ESMO–PaedCan–EURACAN Clinical Practice Guidelines for diagnosis, treatment and follow-up. Ann Oncol 2018;29:iv79–95.

[4] van de Sande MAJ, van der Wal RJP, Navas Cañete A, van Rijswijk CSP, Kroon HM, Dijkstra PDS, et al. Radiologic differentiation of enchondromas, atypical cartilaginous tumors, and high-grade chondrosarcomas—Improving tumor-specific treatment: A paradigm in transit? Cancer 2019;125:3288–91.

[5] Hodel S, Laux C, Farei-Campagna J, Götschi T, Bode-Lesniewska B, Müller DA. The impact of biopsy sampling errors and the quality of surgical margins on local recurrence and survival in chondrosarcoma. Cancer Manag Res 2018;10:3765–71.

[6] van Praag (Veroniek) VM, Rueten-Budde AJ, Ho V, Dijkstra PDS, Fiocco M, van de Sande MAJ, et al. Incidence, outcomes and prognostic factors during 25 years of treatment of chondrosarcomas. Surg Oncol 2018;27:402–8.

[7] Jones KB, Buckwalter JA, McCarthy EF, DeYoung BR, El-Khoury GY, Dolan L, et al. Reliability of Histopathologic and Radiologic Grading of Cartilaginous Neoplasms in Long Bones. J Bone Joint Surg Am 2007;89:2113–23.

[8] Zamora T, Urrutia J, Schweitzer D, Amenabar PP, Botello E. Do Orthopaedic Oncologists Agree on the Diagnosis and Treatment of Cartilage Tumors of the Appendicular Skeleton? Clin Orthop Relat Res 2017;475:2176–86.

[9] Fritz B, Müller DA, Sutter R, Wurnig MC, Wagner MW, Pfirrmann CWA, et al. Magnetic Resonance Imaging–Based Grading of Cartilaginous Bone Tumors. Invest Radiol 2018;53:663–72.

[10] Lisson CS, Lisson CG, Flosdorf K, Mayer-Steinacker R, Schultheiss M, von Baer A, et al. Diagnostic value of MRI-based 3D texture analysis for tissue characterisation and discrimination of low-grade chondrosarcoma from

enchondroma: a pilot study. Eur Radiol 2018;28:468–77.

[11] Pan J, Zhang K, Le H, Jiang Y, Li W, Geng Y, et al. Radiomics Nomograms Based on Non-enhanced MRI and Clinical Risk Factors for the Differentiation of Chondrosarcoma from Enchondroma. J Magn Reson Imaging 2021;54:1314–23.

[12] Deckers C, Schreuder BHW, Hannink G, de Rooy JWJ, van der Geest ICM. Radiologic follow-up of untreated enchondroma and atypical cartilaginous tumors in the long bones. J Surg Oncol 2016;114:987–91.

[13] Omlor GW, Lohnherr V, Lange J, Gantz S, Mechtersheimer G, Merle C, et al. Outcome of conservative and surgical treatment of enchondromas and atypical cartilaginous tumors of the long bones: retrospective analysis of 228 patients. BMC Musculoskelet Disord 2019;20:134.

[14] Lubner MG, Smith AD, Sandrasegaran K, Sahani D V., Pickhardt PJ. CT Texture Analysis: Definitions, Applications, Biologic Correlates, and Challenges. Radiographics 2017;37:1483–503.

[15] Ren J, Yuan Y, Qi M, Tao X. Machine learning–based CT texture analysis to predict HPV status in oropharyngeal squamous cell carcinoma: comparison of 2D and 3D segmentation. Eur Radiol 2020;30:6858–66.

[16] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. J Artif Intell Res 2002;16:321–57.

Samenvatting en algemene discussie

Samenvatting

Hoofdstuk 1 is een algemene inleiding tot het doctoraal onderzoek. Het doel van dit proefschrift was het bepalen van de diagnostische waarde van kunstmatige intelligentie (in deze Nederlandstalige samenvatting wordt de meer algemene term kunstmatige intelligentie gebruikt om de specifiekere term machine learning te vertalen) bij het differentiëren tussen atypische cartilagineuze tumor (ACT) en hooggradig chondrosarcoom (CS), gebaseerd op radiomics kenmerken die zijn afgeleid van dwarsdoorsnedebeeldvorming, zoals magnetische resonantie beeldvorming (MRI) en computertomografie (CT), in vergelijking met ervaren radiologen gespecialiseerd in musculoskeletale oncologie.

In hoofdstuk 2 introduceerden we het concept van CT- en MRI-radiomics van bot- en weke delen sarcomen door middel van een literatuurstudie van de reproduceerbaarheid van radiomics karakteristieken en validatiestrategieën voor voorspellende modellen. Het uiteindelijke doel van deze systematische evaluatie was om een consensus te bereiken over deze aspecten die van belang zijn voor de translatie en toepasbaarheid van radiomics in de klinische praktijk. Van de 278 geïdentificeerde artikelen werden negenenveertig artikelen opgenomen die tussen 2008 en 2020 werden gepubliceerd. Ze behandelden radiomics van bottumoren (n=12) of tumoren van de weke delen (n=37). Achttien (37%) onderzoeken bevatten een analyse van de reproduceerbaarheid van de berekende radiomics karakteristieken. De variabiliteit van de beeldsegmentatie gedaan door dezelfde persoon (intra-observervariabiliteit) en tussen verschillende personen (interobserver variabiliteit) was het thema van de reproduceerbaarheidsanalyse in 16 (33%) onderzoeken, hetgeen meer was dan het aantal analyses die gericht waren op beeldacquisitie of beeldbewerking (n=2, 4%). De intraclass correlatiecoëfficiënt (ICC) was de meest gebruikte statistische methode om de reproduceerbaarheid te beoordelen. Deze varieerde van 0.6 tot 0.9. In 25 (51%) artikelen werd ten minste één validatietechniek voor kunstmatige intelligentie gebruikt ten behoeve van modelontwikkeling. K-voudige kruisvalidatie werd het meest gebruikt. Een klinische validatie van het model werd gerapporteerd in 19 (39%) artikelen. Dit werd uitgevoerd met behulp van een separate dataset afkomstig van de primaire instelling (d.w.z. interne validatie)

in 14 (29%) studies en met een onafhankelijke dataset waarbij data werden gegenereerd met verschillende scanners of in een andere instelling (d.w.z externe validatie) in 5 (10%) studies. Concluderend bleek dat de reproduceerbaarheid van radiomics kenmerken en modelvalidatie tussen de studies over musculoskeletale sarcomen in hoge mate varieerden. Dit zou in toekomstige onderzoeken moeten worden aangepakt om het gebied van radiomics van een onderzoeksgebied naar het klinische domein te brengen.

In hoofdstuk 3 werd de diagnostische waarde van op MRI-radiomics gebaseerde modellering met kunstmatige intelligentie (specifiek machine learning) gevalueerd in het onderscheiden van ACT en hooggradig CS in een eerste studie in één centrum. We hebben retrospectief 58 patiënten bestudeerd met een histologisch bewezen ACT (n=26) of een hooggradig CS (n=32, inclusief 16 gelokaliseerd in het appendiculaire en 16 in het axiale skelet). Ze werden willekeurig verdeeld in trainings- (n=42) en test- (n=16) groepen voor respectievelijk de ontwikkeling en validatie van het model. Alle tumoren werden handmatig gesegmenteerd op T1- en T2-gewogen MRI door 2D interessegebieden (Regions Of Interest: ROIs) te tekenen die werden gebruikt voor extractie van radiomics kenmerken. Na het selecteren van een functie werd een ensemble-classificeerder (AdaBoostM1) op de training set afgestemd met behulp van 10-voudige kruisvalidatie en vervolgens werd getest op een niet eerder gebruikte testset. Daarna heeft een ervaren musculoskeletale oncologische radioloog geblindeerd voor histologie en radiologische gegevens de laesies in de testgroep kwalitatief geëvalueerd. De gegevensset werd na functieselectie teruggebracht tot 4 T1-gewogen MRI-radiomics karakteristieken. Het model identificeerde respectievelijk 85.7% (AUC=0.85) en 75% (AUC=0.78) van de laesies in de trainings- en testgroepen. De radioloog identificeerde 81.3% van de laesies correct, zodat er geen verschil was met het model (p=0.453). Concluderend toonde dat ons ontwikkelde model gebaseerd op kunstmatige intelligentie accuraat het onderscheid tussen ACT en hooggradig CS kon maken.

In hoofdstuk 4 werd de invloed onderzocht van de handmatige segmentatievariabiliteit tussen observatoren op de reproduceerbaarheid van 2D en 3D CT- en MRI-gebaseerde textuuranalyse. Dertig patiënten met kraakbeentumoren (n=10 enchondroom; N=10 ACT; N=10 hooggradig CS) werden retrospectief opgenomen. Drie radiologen voerden onafhankelijk van elkaar handmatige beeldsegmentatie uit op natieve CT, T1-gewogen en T2-gewogen MRI door zowel een 2D ROI op de coupe te tekenen die het grootste tumoroppervlak toonde als een 3D ROI te tekenen die het gehele tumorvolume

omlijnde. Bovendien werd een marginale reductie van de ROI toegepast op zowel 2D- als 3D-segmentaties om de invloed van segmentatiemarges te evalueren. In totaal werden 783 en 1132 radiomics kenmerken geëxtraheerd uit respectievelijk originele en gefilterde 2D- en 3D-beelden. Functiestabiliteit werd gedefinieerd als een ICC≥0.75. In 2D versus 3D beeldsegmentatie waren de waarden van stabiele eigenschappen voor respectievelijk CT-, T1-gewogen en T2-gewogen beelden 74.71% versus 86.57% ($p<0.001$), 77.14% versus 80.04% ($p=0.142$) en 95.66% versus 94.97% ($p=0.554$). De reductie van de marge verbeterde 2D segmentatie ($p=0.343$) niet en presteerde slechter dan 3D ($p<0.001$) beeldsegmentatie wat betreft stabiliteit van de radiomics kenmerken. Bij 2D- versus 3D-beeldsegmentatie bedroegen de overeenkomende stabiele kenmerken die waren afgeleid van CT en MRI 65.8% versus 68.7% ($p=0.191$), en die afgeleid van T1-gewogen en T2-gewogen beelden 76.0% versus 78.2% ($p=0.285$). Concluderend waren 2D en 3D CT en MRI radiomics kenmerken van kraakbeentumoren reproduceerbaar, hoewel een zekere mate van segmentatievariabiliteit tussen observatoren de noodzaak van betrouwbaarheidsanalyse in radiomics studies aantoonde.

In hoofdstuk 5 werd een multicentrische studie beschreven die de prestaties van CT op radiomics-gebaseerde kunstmatige intelligentie modellen onderzocht in het onderscheiden tussen ACT en hooggradig CS van lange pijpbeenderen. Honderdtwintig patiënten met histologie-bewezen laesies werden retrospectief opgenomen. Het trainingscohort bestond uit 84 CT-scans uit centrum 1 (n=55 ACT; N=29 CS graad II-IV). Het externe testcohort bestond uit de CT-beelden afkomstig van 36 PET-CT-scans uit centrum 2 (n=16 ACT; N=20 CS graad II-IV). 2D-segmentatie werd uitgevoerd op preoperatieve CT-scans. Radiomics kenmerken werden geëxtraheerd. Na reductie van datadimensionaliteit en uitvoer van een klassenbalans in centrum 1 werd een kunstmatige intelligentie gebaseerde classificeerder (LogitBoost) gevalideerd van een intern testcohort (tot stand gekomen met behulp van 10-voudige kruisvalidatie) en op het externe testcohort. In centrum 2 werden, m.b.v. de McNemar test, de resultaten vergeleken met die van een preoperatieve biopsie en met de beoordeling van een ervaren radioloog. Het model had een nauwkeurigheid van 81% (AUC=0.89) en 75% (AUC=0.78) bij het identificeren van de laesies in respectievelijk de training en de externe testcohorten. In het bijzonder was de nauwkeurigheid bij het classificeren van ACT en hooggradig CS respectievelijk 84% en 78% in het trainingscohort, en 81% en 70% in het externe testcohort. Preoperatieve biopsie had

een nauwkeurigheid van 64% (AUC=0.66) (p=0.29). De radioloog bereikte een nauwkeurigheid van 81% (p=0.75). Conclusie was dat het model een goede nauwkeurigheid bereikte bij het classificeren van ACT en hooggradige CS van lange pijpbeenderen op basis van radiomics kenmerken afkomstig van preoperatieve CT-scans.

In hoofdstuk 6 werd een multicentrische studie beschreven die de prestaties van MR op radiomics-gebaseerde kunstmatige intelligentie onderzocht in het onderscheiden tussen ACT en hooggradig CS van lange pijpbeenderen. Honderdachtenvijftig patiënten uit twee tertiaire centra voor bottumoren met chirurgisch behandelde en histologisch bewezen kraakbeentumoren werden retrospectief geïncludeerd. Het trainingscohort bestond uit 93 MRI-scans uit centrum 1 (n=74 ACT; N=19 CS graad II). Het externe testcohort bestond uit 65 MRI-scans uit centrum 2 (n=45 ACT; N=20 CS graad II). 2D-segmentatie werd handmatig uitgevoerd op T1-gewogen MRI-sequenties. Eerste orde, morfologische- en textuureigenschappen werden geëxtraheerd. Reductie van datadimensionaliteit werd uitgevoerd op basis van stabiliteit, variatie en inter-correlatie analyses en recursieve kenmerk eliminatie op de data afkomstig van centrum 1 na balancering van klassen (CS graad II oversampling naar N=74). Zo werd een model gebaseerd op kunstmatige intelligentie (Extra Trees Classifier) getraind op een cohort tot stand gekomen door 10-voudige kruis-validatie en getest op een extern test cohort. In centrum 2 werden de prestaties, m.b.v. de McNemar's test, vergeleken met die van een radioloog ervaren op het gebied van musculoskeletale oncologie. Negenhonderdnegentien radiomics kenmerken werden geëxtraheerd en vervolgens teruggebracht tot 17 door middel van datadimensionaliteitsvermindering. Na training had het model (AUC=0.88), een nauwkeurigheid van 92% (60/65, AUC=0.94) voor het identificeren van de laesies in het externe testcohort. In het bijzonder waren de nauwkeurigheden voor het correct classificeren van ACT en graad II CS respectievelijk 98% (44/45) en 80% (16/20). De ervaren radioloog had een nauwkeurigheid van 98% (64/65) en er was geen significant verschil met het getrainde model (p>0.99). Concluderend bereikte het model een hoge nauwkeurigheid voor het classificeren van ACT en graad II CS van lange pijpbeenderen op basis van MRI-radiomics kenmerken.

Algemene discussie, beperkingen en toekomstperspectieven

Dit proefschrift richt zich op het concept ACT, dat is gedefinieerd volgens de 2013 en 2020 classificatie van de Wereldgezondheidsorganisatie [1,2]. Deze relatief nieuwe

definitie weerspiegelt het indolente biologische gedrag van ACT, dat nu wordt beschouwd als een intermediaire kraakbeentumor van lange pijpbeenderen in plaats van een maligne tumor [2]. Dit concept sluit beter aan bij therapeutische opties die voor ACT volledig verschillen met die van hooggradig (II of hoger) appendiculair CS en axiaal CS van elke graad [3]. Wijde resectie met tumorvrije grenzen is de therapie van keuze voor de laatste groep. De behandeling van ACT is de afgelopen drie decennia opmerkelijk veranderd. Tot in de jaren negentig werd er een wijde resectie uitgevoerd, waarna de therapie veranderde in intra-lesionale curettage en tegenwoordig is in toenemende mate waakzaam afwachten met beeldvorming, maar zonder chirurgisch ingrijpen een optie [4].

Vanwege het risico van het niet verkrijgen van representatief biopsiemateriaal afkomstig uit het meest kwaadaardige deel van de tumor, wordt een biopsie in veel tertiaire centra niet langer gebruikt [5]. Gezien de toenemende incidentie van ACT als gevolg van een toename van incidentele bevindingen op MRI [6] is er behoefte aan duidelijke beeldvormingscriteria om onderscheid te maken tussen ACT enerzijds en hooggradige CS anderzijds. Beoordeling van beeldvorming lijdt echter aan interobserver-variabiliteit [7,8]. In dit proefschrift worden modellen op basis van radiomics kenmerken voorgesteld om op basis van beeldvorming het onderscheid tussen ACT en hooggradig CS objectiever te maken. Andere radiomics studies tot nu toe hebben zich gericht op het differentiëren tussen enchondroom enerzijds en ACT/hooggradig CS anderzijds [9–11]. De differentiatie tussen enchondroom en ACT is echter geleidelijk minder relevant geworden, vanwege de hierboven genoemde herwaardering van ACT en de daarbij horende nieuwe inzichten omtrent behandeling waarbij het controleren van de laesie met behulp van beeldvorming een alternatief is voor het curetteren van de laesie [4,12,13]. Resectie van ACT is, in tegenstelling tot behandeling van CS graad II, niet meer geïndiceerd. Aangezien enchondroom en hooggradig CS eenvoudig te onderscheiden zijn met behulp van röntgenfoto's, heeft het diagnostisch dilemma zich verplaatst van onderscheid tussen enchondroom en ACT naar onderscheid tussen ACT en CS graad II [4].

In de hoofdstukken 3, 5 en 6 van dit proefschrift werden kunstmatige intelligentie gebaseerde technieken gebruikt om classificatiemodellen te maken op basis van radiomics kenmerken die zijn afgeleid van MRI en CT. In de hoofdstukken 5 en 6 werden met name grote multicentrische studies beschreven, inclusief klinische validatie van het model op onafhankelijke data (externe testcohort) van een andere instelling. Deze studies

concentreerden zich op respectievelijk CT- en MRI-radiomics en behaalden een hoge nauwkeurigheid bij het correct classificeren van ACT en hooggradig CS van lange pijpbeenderen, zonder verschil met ervaren radiologen op het gebied van musculoskeletale oncologie. Ongetwijfeld werd de meest relevante bevinding gedaan in hoofdstuk 6, waarbij een getraind model op basis van radiomics kenmerken afkomstig uit T1-gewogen MRI een nauwkeurigheid van 92% bereikte bij het onderscheiden tussen ACT en CS graad II. Bovendien was er geen statistisch significant verschil ($p>0.99$) tussen resultaten van het model en die van een radioloog, gespecialiseerd in bottumoren met 35 jaar ervaring. In hoofdstuk 4 hebben we de invloed van de segmentatievariabiliteit tussen observatoren op de reproduceerbaarheid van 2D en 3D CT- en MRI-radiomics kenmerken van kraakbeentumoren methodologisch geanalyseerd. Er was een goede reproduceerbaarheid van radiomics kenmerken op alle beeldvormende modaliteiten, zowel van kenmerken op basis van 2D- als 3D-segmentaties, hoewel voorafgaande kennis m.b.t. stabiliteit van radiomics kenmerken een belangrijke voorwaarde is om nauwkeurige classificatie te bewerkstelligen. Dit werd ook benadrukt in de systematische review van hoofdstuk 2 m.b.t. reproduceerbaarheid van kenmerken en validatie strategieën, en dit werd ook gedaan in onze daaropvolgende studies (hoofdstukken 5 en 6).

Dit proefschrift kent enkele beperkingen. In de eerste plaats waren alle uitgevoerde onderzoeken retrospectief, omdat hiermee een groter aantal patiënten (met een relatief zeldzame aandoening) kon worden geïncludeerd. Bovendien is een prospectieve analyse niet strikt nodig in radiomics studies [14]. In de tweede plaats kan de segmentatietechniek de resultaten beïnvloeden. In onderzoeken die in hoofdstukken 3, 5 en 6 worden beschreven hebben we een 2D-segmentatie uitgevoerd op de coupe geselecteerd met de grootste tumordimensie. Een 2D-benadering had de voorkeur, omdat het potentieel gemakkelijker is om dit in de klinische praktijk te implementeren, bovendien werd in recente literatuur gemeld dat 2D-segmentatie betere resultaten zou opleveren dan 3D-segmentatie [15]. Onze bevindingen in hoofdstuk 4 toonden geen verschil tussen 2D en 3D MRI-gebaseerde textuuranalyse. Ten derde was ACT oververtegenwoordigd in vergelijking met hooggradig CS in hoofdstuk 5, vooral in het trainingscohort uit centrum 1, en in hoofdstuk 6. Dit weerspiegelde echter nauwkeurig de incidentie van ACT en hooggradig CS [6], bovendien werd balansering uitgevoerd in beide studies om de minderheidsklasse kunstmatig te oversampelen in de trainingscohorten [16]. Ten vierde werden geen contrastmiddelen

gebruikt bij CT en MRI voor de ontwikkelde modellen, omdat deze niet bij alle patiënten beschikbaar waren. Tot slot is het gebruik van de lage dosis CT-beelden van PET-CT in het externe testcohort zoals beschreven in hoofdstuk 5 suboptimaal. De goede resultaten stemmen ons echter positief en openen de mogelijkheid voor toekomstig onderzoek om licht te werpen op de waarde van het gebruik van kunstmatige intelligentie modellen die gebruik maken van radiomics kenmerken afkomstig van CT- of MRI.

Concluderend toonden CT- en MRI-modellen die gebruik maken van radiomics kenmerken een hoge nauwkeurigheid aan bij het onderscheiden van ACT en hooggradig CS en ziet het er veelbelovend uit als een objectieve beeldvormingsmethode die kan worden gebruikt bij de klinische besluitvorming. Dit kan met name belangrijk zijn in de algemene praktijk waar gespecialiseerde expertise niet voorhanden is, bij het correct identificeren van de veel voorkomende ACT. Onze grote onderzoekspopulatie en de zeer goede prestaties die zijn bereikt met onafhankelijke gegevens van verschillende instellingen, zoals gepresenteerd in de hoofdstukken 5 en 6, garanderen de generaliseerbaarheid van onze resultaten. Toekomstige studies zullen de toepasbaarheid van onze bevindingen in de klinische praktijk moeten verifiëren.

Referenties

[1] Fletcher CDM, Bridge JA, Hogendoorn PCW, Mertens F. World Health Organization Classification of Tumours of Soft Tissue and Bone. Lyon, France: International Agency for Research on Cancer Press; 2013.

[2] WHO Classification of Tumours Editorial Board. WHO Classification of Tumours: Soft Tissue and Bone Tumours. Lyon, France: International Agency for Research on Cancer Press; 2020.

[3] Casali PG, Bielack S, Abecassis N, Aro HT, Bauer S, Biagini R, et al. Bone sarcomas: ESMO–PaedCan–EURACAN Clinical Practice Guidelines for diagnosis, treatment and follow-up. Ann Oncol 2018;29:iv79–95.

[4] van de Sande MAJ, van der Wal RJP, Navas Cañete A, van Rijswijk CSP, Kroon HM, Dijkstra PDS, et al. Radiologic differentiation of enchondromas, atypical cartilaginous tumors, and high-grade chondrosarcomas—Improving tumor-specific treatment: A paradigm in transit? Cancer 2019;125:3288–91.

[5] Hodel S, Laux C, Farei-Campagna J, Götschi T, Bode-Lesniewska B, Müller DA. The impact of biopsy sampling errors and the quality of surgical margins on local recurrence and survival in chondrosarcoma. Cancer Manag Res 2018;10:3765–71.

[6] van Praag (Veroniek) VM, Rueten-Budde AJ, Ho V, Dijkstra PDS, Fiocco M, van de Sande MAJ, et al. Incidence, outcomes and prognostic factors during 25 years of treatment of chondrosarcomas. Surg Oncol 2018;27:402–8.

[7] Jones KB, Buckwalter JA, McCarthy EF, DeYoung BR, El-Khoury GY, Dolan L, et al. Reliability of Histopathologic and Radiologic Grading of Cartilaginous Neoplasms in Long Bones. J Bone Joint Surg Am 2007;89:2113–23.

[8] Zamora T, Urrutia J, Schweitzer D, Amenabar PP, Botello E. Do Orthopaedic Oncologists Agree on the Diagnosis and Treatment of Cartilage Tumors of the Appendicular Skeleton? Clin Orthop Relat Res 2017;475:2176–86.

[9] Fritz B, Müller DA, Sutter R, Wurnig MC, Wagner MW, Pfirrmann CWA, et al. Magnetic Resonance Imaging–Based Grading of Cartilaginous Bone Tumors. Invest Radiol 2018;53:663–72.

[10] Lisson CS, Lisson CG, Flosdorf K, Mayer-Steinacker R, Schultheiss M, von Baer A, et al. Diagnostic value of MRI-based 3D texture analysis for tissue characterisation and discrimination of low-grade chondrosarcoma from

enchondroma: a pilot study. Eur Radiol 2018;28:468–77.

[11] Pan J, Zhang K, Le H, Jiang Y, Li W, Geng Y, et al. Radiomics Nomograms Based on Non-enhanced MRI and Clinical Risk Factors for the Differentiation of Chondrosarcoma from Enchondroma. J Magn Reson Imaging 2021; 54:1314–23.

[12] Deckers C, Schreuder BHW, Hannink G, de Rooy JWJ, van der Geest ICM. Radiologic follow-up of untreated enchondroma and atypical cartilaginous tumors in the long bones. J Surg Oncol 2016;114:987–91.

[13] Omlor GW, Lohnherr V, Lange J, Gantz S, Mechtersheimer G, Merle C, et al. Outcome of conservative and surgical treatment of enchondromas and atypical cartilaginous tumors of the long bones: retrospective analysis of 228 patients. BMC Musculoskelet Disord 2019;20:134.

[14] Lubner MG, Smith AD, Sandrasegaran K, Sahani D V., Pickhardt PJ. CT Texture Analysis: Definitions, Applications, Biologic Correlates, and Challenges. Radiographics 2017;37:1483–503.

[15] Ren J, Yuan Y, Qi M, Tao X. Machine learning–based CT texture analysis to predict HPV status in oropharyngeal squamous cell carcinoma: comparison of 2D and 3D segmentation. Eur Radiol 2020;30:6858–66.

[16] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. J Artif Intell Res 2002;16:321–57.

List of publications

18th December 2021

Original research papers

1) Gitto S, Cuocolo R, van Langevelde K, van de Sande MAJ, Parafioriti A, Luzzati A, Imbriaco M, Sconfienza LM, Bloem JL
MRI radiomics-based machine learning classification of atypical cartilaginous tumour and grade II chondrosarcoma of long bones
EBioMedicine 2022; 75:103757

2) Gitto S, Cuocolo R, Emili I, Tofanelli L, Chianca V, Albano D, Messina C, Imbriaco M, Sconfienza LM
Effects of Interobserver Variability on 2D and 3D CT- and MRI-Based Texture Feature Reproducibility of Cartilaginous Bone Tumors
J Digit Imaging 2021; 34:820-832

3) Schiaffino S, Codari M, Cozzi A, Albano D, Alì M, Arioli R, Avola E, Bnà C, Cariati M, Carriero S, Cressoni M, Danna PSC, Della Pepa G, Di Leo G, Dolci F, Falaschi Z, Flor N, Foà RA, Gitto S, Leati G, Magni V, Malavazos AE, Mauri G, Messina C, Monfardini L, Paschè A, Pesapane F, Sconfienza LM, Secchi F, Segalini E, Spinazzola A, Tombini V, Tresoldi S, Vanzulli A, Vicentin I, Zagaria D, Fleischmann D, Sardanelli F
Machine Learning to Predict In-Hospital Mortality in COVID-19 Patients Using Computed Tomography-Derived Pulmonary and Vascular Features
J Pers Med 2021; 11:501

4) Gitto S, Cuocolo R, Annovazzi A, Anelli V, Acquasanta M, Cincotta A, Albano D, Chianca V, Ferraresi V, Messina C, Zoccali C, Armiraglio E, Parafioriti A, Sciuto R, Luzzati A, Biagini R, Imbriaco M, Sconfienza LM
CT radiomics-based machine learning classification of atypical cartilaginous tumours and appendicular chondrosarcomas
EBioMedicine 2021; 68:103407

5) Schiaffino S, Albano D, Cozzi A, Messina C, Arioli R, Bnà C, Bruno A, Carbonaro LA, Carriero A, Carriero S, Danna PSC, D'Ascoli E, De Berardinis C, Della Pepa G, Falaschi Z, Gitto S, Malavazos AE, Mauri G, Monfardini L, Paschè A, Rizzati R, Secchi F, Vanzulli A, Tombini V, Vicentin I, Zagaria D, Sardanelli F, Sconfienza LM
CT-derived Chest Muscle Metrics for Outcome Prediction in Patients with COVID-19
Radiology 2021; 300:E328-E336

6) Chianca V, Cuocolo R, Gitto S, Albano D, Merli I, Badalyan J, Cortese MC, Messina C, Luzzati A, Parafioriti A, Galbusera F, Brunetti A, Sconfienza LM
Radiomic Machine Learning Classifiers in Spine Bone Tumors: A Multi-Software, Multi-Scanner Study
Eur J Radiol 2021; 137:109586

7) Tagliafico AS, Albano D, Torri L, Messina C, Gitto S, Bruno F, Barile A, Giovagnoni A, Miele V, Grassi R, Sconfienza LM
Impact of coronavirus disease 2019 (COVID-19) outbreak on radiology research: An Italian survey
Clin Imaging 2021; 76:144-148

8) Casale S, Bortolotto C, Stella GM, Filippi AR, Gitto S, Bottinelli OM, Carnevale S, Morbini P, Preda L
Recent advancement on PD-L1 expression quantification: the radiologist perspective on CT-guided FNAC
Diagn Interv Radiol 2021; 27:214-218

9) Albano D, Messina C, Zagra L, Andreata M, De Vecchi E, Gitto S, Sconfienza LM
Failed Total Hip Arthroplasty: Diagnostic Performance of Conventional MRI Features and Locoregional Lymphadenopathy to Identify Infected Implants
J Magn Res Imaging 2021; 53:201-210

10) Messina C, Buzzoni AC, Gitto S, Almolla J, Albano D, Sconfienza LM
Disruption of bone densitometry practice in a Northern Italy Orthopedic Hospital during the COVID-19 pandemic
Osteoporos Int 2020; 10:1-5

11) Mauri G, Gitto S, Pescatori LC, Albano D, Messina C, Sconfienza LM
Technical Feasibility of Electromagnetic US/CT Fusion Imaging and Virtual Navigation in the Guidance of Spine Biopsies
Ultraschall Med 2020; doi: 10.1055/a-1194-4225

12) Albano D, Messina C, Gambino A, Gurgitano M, Sciabica C, Oliveira Pavan GR, Gitto S, Sconfienza LM.
Segmented lordotic angles to assess lumbosacral transitional vertebra on EOS
Eur Spine J 2020; 29:2470-2476

13) Sconfienza LM, Albano D, Messina C, Gitto S, Guarrella V, Perfetti C, Taverna E, Arrigoni P, Randelli PS
Ultrasound-Guided Percutaneous Tenotomy of the Long Head of Biceps Tendon in Patients with Symptomatic Complete Rotator Cuff Tear: In Vivo Non-controlled Prospective Study
J Clin Med 2020; 9:2114

14) Gitto S, Cuocolo R, Albano D, Chianca V, Messina C, Gambino A, Ugga L, Cortese MC, Lazzara A, Ricci D, Spairani R, Zanchetta E, Luzzati A, Brunetti A, Parafioriti A, Sconfienza LM
MRI radiomics-based machine-learning classification of bone chondrosarcoma
Eur J Radiol 2020; 128:109043

15) Cabitza F, Campagner A, Albano D, Aliprandi A, Bruno A, Chianca V, Corazza A, Di Pietto F, Gambino A, Gitto S, Messina C, Orlandi D, Pedone L, Zappia M, Sconfienza LM
The Elephant in the Machine: Proposing a New Metric of Data Reliability and its Application to a Medical Case to Assess Classification Reliability
Appl Sci 2020; 10:4014

16) Albano D, Gambino A, Messina C, Chianca V, Gitto S, Faenza S, Galia M, Sconfienza LM
Ultrasound-Guided Percutaneous Irrigation of Rotator Cuff Calcific Tendinopathy (US-PICT): Patient Experience
Biomed Res Int 2020; 2020:3086395

17) Messina C, Vitale JA, Pedone L, Chianca V, Vicentin I, Albano D, Gitto S, Sconfienza LM
Critical appraisal of papers reporting recommendation on sarcopenia using the AGREE II tool: a EuroAIM initiative
Eur J Clin Nutr 2020; 74:1164-1172

18) Albano D, Cortese MC, Duarte A, Messina C, Gitto S, Vicentin I, Coppola A, Galia M, Sconfienza LM
Predictive role of ankle MRI for tendon graft choice and surgical reconstruction
Radiol Med 2020; 125:763-769

19) Mauri G, Gitto S, Cantisani V, Vallone G, Schiavone C, Papini E, Sconfienza LM
Use of the Thyroid Imaging Reporting and Data System (TIRADS) in clinical practice: an Italian survey
Endocrine 2020; 68:329-335

20) Chianca V, Albano D, Cuocolo R, Messina C, Gitto S, Brunetti A, Sconfienza LM
T2 mapping of the trapeziometacarpal joint and triangular fibrocartilage complex: a feasibility and reproducibility study at 1.5 T
Radiol Med 2020; 125:306-312

21) Messina C, Usuelli FG, Maccario C, Di Silvestri CA, Gitto S, Cortese MC, Albano D, Sconfienza LM
Precision of bone mineral density measurements around total ankle replacement using dual energy X-ray absorptiometry
J Clin Densitom 2020; 23:656-663

22) Messina C, Albano D, Orlandi D, Chianca V, Corazza A, Ferrari F, Gitto S, Sconfienza LM
Potential use of a diluted high-relaxivity gadolinium-based intra-articular contrast agent for magnetic resonance arthrography: an in-vitro study
BMC Med Imaging 2019; 19:83

23) Gitto S, Lee SC, Miller TT
Ultrasound-guided percutaneous treatment of volar radiocarpal ganglion cysts: safety and efficacy
J Clin Ultrasound 2019; 47:339-44

24) Gitto S, Bisdas S, Emili I, Nicosia L, Pescatori LC, Bhatia K, Lingam RK, Sardanelli F, Sconfienza LM, Mauri G
Clinical practice guidelines on ultrasound-guided fine needle aspiration biopsy of thyroid nodules: a critical appraisal using AGREE II
Endocrine 2019; 65:371-8

25) Messina C, Buonomenna C, Menon G, Magnani S, Albano D, Gitto S, Ulivieri FM, Sconfienza LM
Fat mass does not increase the precision error of trabecular bone score measurements
J Clin Densitom 2019; 22:359-66

26) Gitto S, Grassi G, De Angelis C, Monaco CG, Sdao S, Sardanelli F, Sconfienza LM, Mauri G
A computer-aided diagnosis system for the assessment and characterization of low-to-high suspicion thyroid nodules on ultrasound
Radiol Med 2019; 124:118-25

27) Galbusera F, Bassani T, Casaroli G, Gitto S, Zanchetta E, Costa F, Sconfienza LM
Generative models: an upcoming innovation in musculoskeletal radiology? A preliminary test in spine imaging
Eur Radiol Exp 2018; 2:29

28) Draghi F, Torresi M, Urciuoli L, Gitto S
MR signal abnormalities within the pericruciate fat pad: a possible secondary sign for acute anterior cruciate ligament tears
Can Assoc Radiol J 2017; 68:438-44

29) Orlandi D, Gitto S, Perugin Bernardi S, Corazza A, De Flaviis L, Silvestri E, Cimmino MA, Sconfienza LM
Advanced power Doppler technique increases synovial vascularity detection in patients with rheumatoid arthritis
Ultrasound Med Biol 2017; 43:1880-87

30) Felisaz PF, Maugeri G, Busi V, Vitale R, Balducci F, Gitto S, Leporati P, Pichiecchio A, Baldi M, Calliada F, Chiovato L, Bastianello S
MR micro-neurography and a segmentation protocol applied to diabetic neuropathy
Radiol Res Pract 2017; 2017:2761818

31) Aryal M, Fischer K, Gentile C, Gitto S, Zhang YZ, McDannold N
Effects on P-Glycoprotein expression after blood-brain barrier disruption using focused ultrasound and microbubbles
Plos One 2017; 12:e0166061

32) Draghi F, Bortolotto C, Coscia DR, Canepari M, Gitto S
Magnetic resonance imaging of degenerative changes of the posterior cruciate ligament
Acta Radiol 2017; 58:338-343

33) Zappia M, Aliprandi A, Pozza S, Doniselli F, Gitto S, Sconfienza LM
How Is Shoulder Ultrasound Done in Italy? A Survey of Clinical Practice
Skeletal Radiol 2016; 45:1629-34

34) Felisaz PF, Balducci F, Gitto S, Carne I, Montagna S, De Icco R, Pichiecchio A, Baldi M, Calliada F, Bastianello S
Nerve fascicles and epineurium volume segmentation of peripheral nerve using magnetic resonance micro-neurography
Acad Radiol 2016; 23:1000-7

Narrative and systematic review articles

35) Gitto S, Cuocolo R, Albano D, Morelli F, Pescatori LC, Messina C, Imbriaco M, Sconfienza LM
CT and MRI radiomics of bone and soft-tissue sarcomas: a systematic review of reproducibility and validation strategies
Insights Imaging 2021; 12:68

36) Chianca V, Albano D, Messina C, Gitto S, Ruffo G, Guarino S, Del Grande F, Sconfienza LM
Sarcopenia: imaging assessment and clinical application
Abdom Radiol (NY) 2021; doi: 10.1007/s00261-021-03294-3

37) Tortora S, Messina C, Albano D, Serpi F, Corazza A, Carrafiello G, Sconfienza LM, Gitto S
Ultrasound-guided musculoskeletal interventional procedures around the elbow, hand and wrist excluding carpal tunnel procedures
J Ultrason 2021; 21:e169-e176

38) Tortora S, Messina C, Gitto S, Chianca V, Serpi F, Gambino A, Pedone L, Carrafiello G, Sconfienza LM, Albano D
Ultrasound-guided musculoskeletal interventional procedures around the shoulder
J Ultrason 2021; 21:e162-e168

39) Rosskopf AB, Taljanovic MS, Sconfienza LM, Gitto S, Martinoli C, Picasso R, Klauser A
Pulley, Flexor, and Extensor Tendon Injuries of the Hand
Semin Musculoskelet Radiol 2021;25:203-215

40) Albano D, Coppola A, Gitto S, Rapisarda S, Messina C, Sconfienza LM
Imaging of calcific tendinopathy around the shoulder: usual and unusual presentations and common pitfalls
Radiol Med 2021; 126:608-619

41) Gitto S, Messina C, Chianca V, Tuscano B, Lazzara A, Pedone L, Albano D, Sconfienza LM
Superb microvascular imaging (SMI) in the evaluation of musculoskeletal disorders: a systematic review
Radiol Med 2020; 125:481-490
Awarded with 2021 best published musculoskeletal ultrasound paper prize ("Premio Giovani Generazioni Maurizio Pinto") from the Italian Society for Ultrasound in Medicine and Biology (SIUMB)

42) Gitto S, Messina C, Vitale N, Albano D, Sconfienza LM
Quantitative musculoskeletal ultrasound
Semin Musculoskelet Radiol 2020; 24:367-374

43) Messina C, Albano D, Gitto S, Tofanelli L, Bazzocchi A, Ulivieri FM, Guglielmi G, Sconfienza LM
Body composition with dual energy X-ray absorptiometry: from basics to new tools
Quant Imaging Med Surg 2020; 10:1687-1698

44) Jimenez F, Gitto S, Sconfienza LM, Draghi F
Ultrasound of iliotibial band syndrome
J Ultrasound 2020; 23:379-385

45) Bianchi S, Gitto S, Draghi F
Ultrasound features of trigger finger: review of the literature
J Ultrasound Med 2019; 38:3141-3154

46) Albano D, Messina C, Gitto S, Papakonstantinou O, Sconfienza LM
Differential diagnosis of spine tumors: my favorite mistake
Semin Musculoskelet Radiol 2019; 23:26-35

47) Moraux A, Gitto S, Bianchi S
Ultrasound features of the normal and pathologic periosteum
J Ultrasound Med 2019; 38:775-84

48) Draghi F, Bortolotto C, Draghi AG, Gitto S
Intrasheath instability of the peroneal tendons: dynamic ultrasound imaging
J Ultrasound Med 2018; 37:2753-8

49) Draghi F, Gitto S, Bianchi S
Injuries to the collateral ligaments of the metacarpophalangeal and interphalangeal joints: sonographic appearance
J Ultrasound Med 2018; 37:2117-33

50) Chianca V, Albano D, Messina C, Midiri F, Mauri G, Aliprandi A, Catapano M, Pescatori LC, Monaco CG, Gitto S, Pisani Mainini A, Corazza A, Rapisarda S, Pozzi G, Barile A, Masciocchi C, Sconfienza LM
Rotator cuff calcific tendinopathy: from diagnosis to treatment
Acta Biomed 2018; 89:186-96

51) Gitto S, Draghi AG, Draghi F
Sonography of non-neoplastic disorders of the hand and wrist tendons
J Ultrasound Med 2018; 37:51-68

52) Gitto S, Messina C, Mauri G, Aliprandi A, Sardanelli F, Sconfienza LM
Dynamic high-resolution ultrasound of intrinsic and extrinsic ligaments of the wrist: how to make it simple
Eur J Radiol 2017; 87:20-35

53) Draghi F, Gitto S, Bortolotto C, Draghi AG, Ori Belometti G
Imaging of plantar fascia disorders: findings on plain radiography, ultrasound and magnetic resonance imaging
Insights Imaging 2017; 8:69-78

54) Gitto S, Draghi AG, Bortolotto C, Draghi F
Sonography of the Achilles tendon after complete rupture repair: what the radiologist should know
J Ultrasound Med 2016; 35:2529-36

55) Gitto S, Draghi F
Normal sonographic anatomy of the wrist with emphasis on assessment of tendons, nerves, and ligaments
J Ultrasound Med 2016; 35:1081-94
Awarded with 2016 best published paper prize from the European Federation of Societies for Ultrasound in Medicine and Biology (EFSUMB)

Guidelines

56-62) Ultrasound and Interventional Subcommittees of the ESSR (coauthor)
Clinical indications for image guided interventional procedures in the musculoskeletal system: a Delphi-based consensus paper from the European Society of Musculoskeletal Radiology (ESSR)
part I, shoulder. Eur Radiol 2020; 30:903-913
part II, elbow and wrist. Eur Radiol 2020; 30:2220-2230
part III, nerves of the upper limb. Eur Radiol 2020; 30:1498-1506
part IV, hip. Eur Radiol 2022; 32:551-560
part V, knee. Eur Radiol; doi: 10.1007/s00330-021-08258-1
part VI, foot and ankle. Eur Radiol; doi: 10.1007/s00330-021-08125-z
part VII, nerves of the lower limb. Eur Radiol; doi: 10.1007/s00330-021-08283-0

Case reports

63) Gitto S, Doeleman T, van de Sande MAJ, van Langevelde K
Intraosseous hibernoma of the appendicular skeleton
Skeletal Radiol 2021; doi: 10.1007/s00256-021-03956-9

64) Draghi F, Draghi AG, Gitto S
Myotendinous strains of the vastus lateralis as a result of sport-related trauma
J Sports Med Phys Fitness 2018; 58:947-949

65) Gitto S, Vaiani M, Cascella T, Lanocita R
Penile metastases from renal cell carcinoma: pre- and postcontrast sonographic findings
Ultrasound Q 2018; 34:285-7

66) Gitto S, Draghi F
Spontaneous distal rupture of the plantar fascia
J Clin Ultrasound 2018; 46:419-20

67) Draghi F, Gitto S
Flexor digitorum superficialis tear in a wakeboarder: an unusual clinical case
Clin J Sport Med 2017; 27:e9-e10

68) Gitto S, Bloem JL
Insufficiency vertical fracture of the proximal femur after bisphosphonate treatment
EURORAD 2019; https://www.eurorad.org/case/16588

69) Gitto S, Pescatori LC, Aliprandi A, Sconfienza LM
Subperiosteal hematoma of the ilium: an unusual complication of acetabular fracture
EURORAD 2016; https://www.eurorad.org/case/14182

70) Gitto S, Pescatori LC, D'Elicio DG, Sconfienza LM
An unusual case of latero-cervical swelling
EURORAD 2015; https://www.eurorad.org/case/13218

71) Gitto S, Draghi F
Bone marrow oedema in acute Osgood-Schlatter disease: a possible cause of knee pain
EURORAD 2015; https://www.eurorad.org/case/12999

Curriculum Vitae

Salvatore Gitto was born in 1990 in Messina, Italy. He finished his medical training at the University of Pavia in 2014 and registered as radiologist in 2019 at the University of Milan, in Italy, receiving *cum laude* distinction on both occasions. He currently practices as musculoskeletal radiologist at the Orthopedic Institute Galeazzi in Milan, and he is also PhD candidate in Radiology at the University of Milan in Italy and Leiden University in The Netherlands. His main research interest is musculoskeletal radiology with a focus on artificial intelligence, radiomics and bone and soft-tissue tumors. He was trained in radiology and was actively involved in research activities at Harvard Medical School/Brigham and Women's Hospital in Boston in 2014, St. Mary's Hospital in London in 2015, Hospital for Special Surgery in New York in 2018 and Leiden University Medical Center in The Netherlands in 2019-2021. He is an active member of professional organizations such as Radiological Society of North America (RSNA), European Society of Radiology (ESR), European Society of Musculoskeletal Radiology (ESSR), European Federation of Societies for Ultrasound in Medicine and Biology (EFSUMB) and Italian Society of Medical and Interventional Radiology (SIRM). He was awarded for several scientific papers and presentations at national and international meetings since 2016 and received international grants for his research activity from ESSR in 2020 (Young Researchers Grant) and International Skeletal Society in 2021 (Early Career Grant). To date he published more than 65 scientific papers in peer-reviewed journals and is co-author of 2 books.

Acknowledgments

I would like to express my sincere gratitude to my supervisors, Prof. dr. Johan L. Bloem, MD PhD and Prof. dr. Luca Maria Sconfienza, MD PhD, for their insightful comments, constant inspiration and enthusiastic guidance during my doctoral studies.

I would also like to thank all co-authors from my Italian and Dutch research teams, particularly Dr. Renato Cuocolo, MD PhD from Naples University and Dr. Kirsten van Langevelde, MD PhD from Leiden University Medical Center, for their valuable help and collaborative efforts.