Radiomics-based machine learning classification of bone chondrosarcoma

Gitto, S.

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.

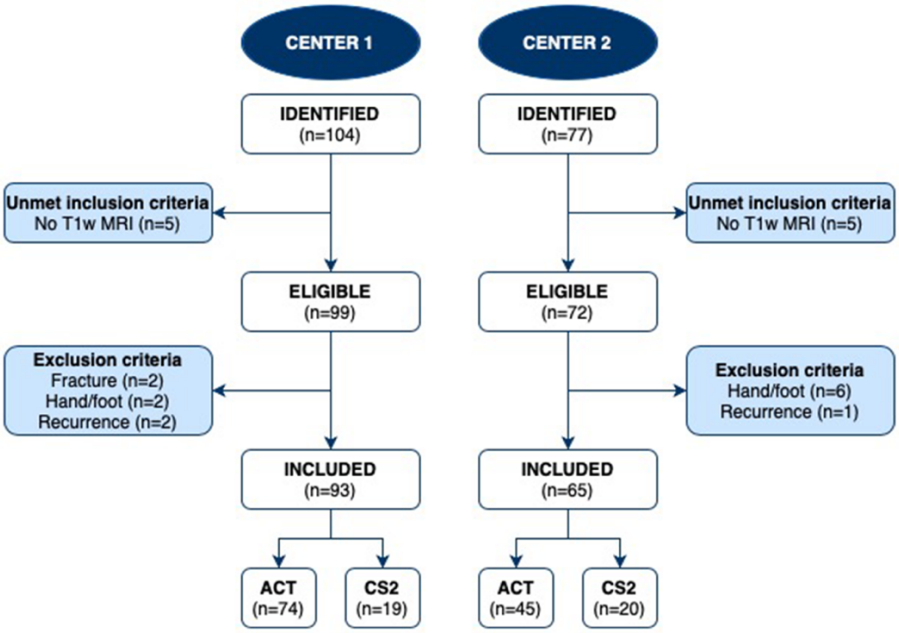

Fig. 1 Flowchart of patient selection.

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].

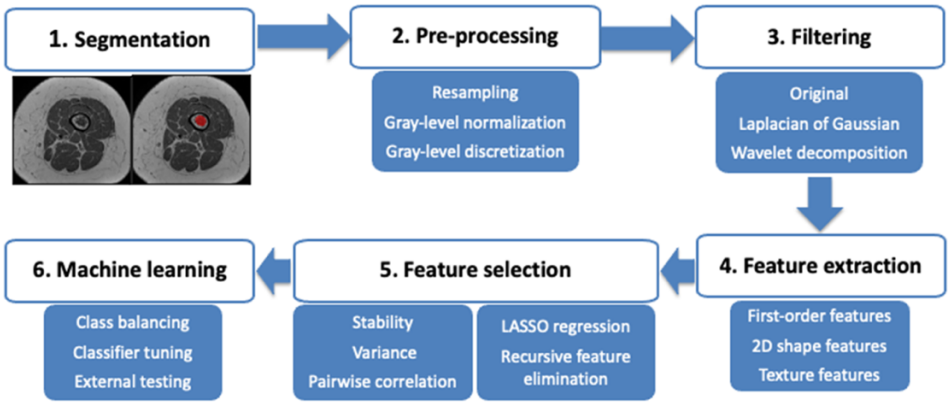

Fig. 2 Radiomics-based machine learning workflow pipeline. This workflow is similar to one recent study from our group [13], with differences mainly related to feature selection process and machine learning classification.

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

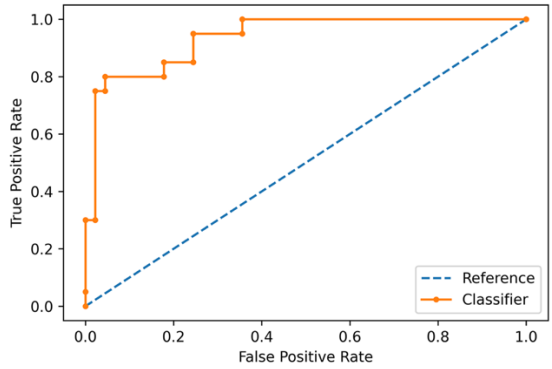

Fig. 3 ROC curve showing the classifier performance in the external test cohort.

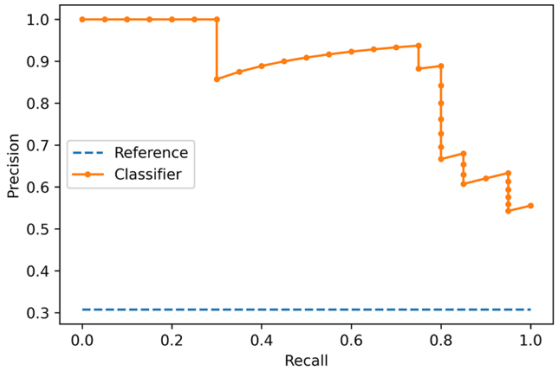

Fig. 4 Precision-recall curve illustrating the classifier performance in the external test cohort.

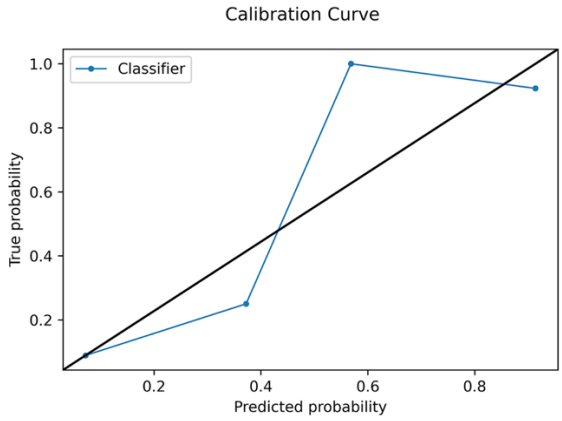

Fig. 5 Calibration curve in the external test cohort.

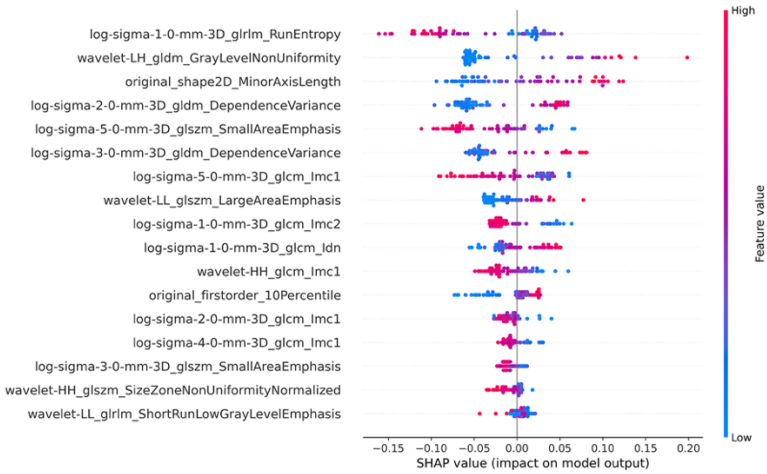

Fig. 6 Beeswarm plot of feature Shapley values in the final model.

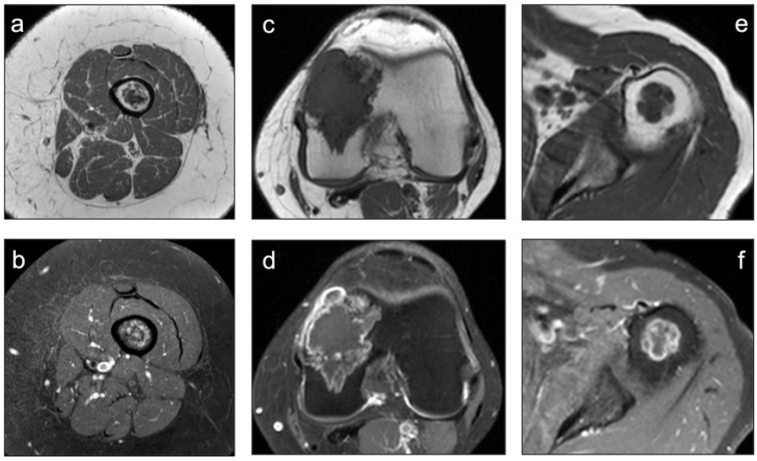

Fig. 7 Native and fat-saturated post-contrast T1-weighted sequences show three different cases of cartilaginous bone tumors, including ACT of the femur (a-b), CS2 of the femur (c-d) and CS2 of the humerus (e-f). Cortical breakthrough and soft-tissue extension are highly suspicious of high-grade lesion 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