Computed tomography-based radiomics machine learning models for differentiating enchondroma and atypical cartilaginous tumor in long bones

Computertomographie-basierte Radiomics Machine Learning Modelle zur Unterscheidung von Enchondromen und atypischen Knorpeltumoren in Röhrenknochen

Authors
Rui Hong¹, Qian Li¹, Jielin Ma², Chunmiao Lu¹, Zhiwei Zhong¹

Affiliations
1 Radiology, The Third Hospital of Hebei Medical University, Shijiazhuang, China
2 Oncology, The Third Hospital of Hebei Medical University, Shijiazhuang, China

Keywords
Radiomics, Machine learning, Enchondroma, Atypical cartilaginous tumor, Long bone, Computed tomography

ABSTRACT

Purpose To explore the value of CT-based radiomics machine learning models for differentiating enchondroma from atypical cartilaginous tumor (ACT) in long bones and methods to improve model performance.

Materials and Methods 59 enchondromas and 53 ACTs in long bones confirmed by pathology were collected retrospectively. The features were extracted from preoperative CT images of these patients, and least absolute shrinkage and selection operator (LASSO) regression was used for feature selection and dimensionality reduction. The selected features were used to construct classification models by thirteen machine learning algorithms. The data set was randomly divided into a training set and a test set at a proportion of 7:3 by ten-fold cross-validation to evaluate the performance of these models.

Results A total of 1199 features were extracted, 9 features were selected, and 13 radiomics machine learning models were constructed. The area under the curve (AUC) of 11 models was more than 0.8, and that of 3 models was more than 0.9. The Extremely Randomized Trees model achieved the best performance (AUC = 0.9375 ± 0.0884), followed by the Adaptive Boosting model (AUC = 0.9188 ± 0.1010) and the Linear Discriminant Analysis model (AUC = 0.9062 ± 0.1459).

Conclusion CT-based radiomics machine learning models had great ability to distinguish enchondroma and ACT in long bones. By using filters to deeply mine high-order features in the original image and selecting appropriate machine learning algorithms, the performance of the model can be improved.

Key points
- CT-based radiomics machine learning models can distinguish enchondroma and ACT in long bones.
- Using filters and selecting advanced machine learning algorithms can improve model performance.
- Clinical features have limited utility in distinguishing enchondroma and ACT in long bones.

Citation Format
Introduction

The most common intermediate (locally aggressive) chondrogenic tumor is the atypical cartilaginous tumor (ACT), which has been specifically found to occur in the appendicular skeletons (long and short tubular bones) according to the 2020 World Health Organization (WHO) classification of bone tumors [1,2]. Enchondroma is the most common benign chondrogenic tumor [2]. Most of them are accidentally discovered without obvious symptoms, and with the widespread use of MRI, the incidence of them being accidentally found in long bones was higher than in short bones [3]. The typical imaging features are round osteolysis with popcorn-like calcifications in the medullary cavity of the metaphysis of long bones [2]. However, most enchondroma patients can choose regular surveillance over surgery, while the main treatments of ACT are surgical intralesional curettage and filling of the tumor cavity. The probability of local recurrence is about 7.5–11 % with only few metastases [1]. Therefore, it is important to improve the accuracy of identifying enchondroma and ACT.

Radiomics is a relatively objective method that can identify tumor heterogeneity and reflect potential structural and functional information by extracting quantitative features with high throughput from standard images and utilizing machine learning algorithms for mathematical operations [4]. MRI-based radiomics models have achieved great results regarding the differentiation of chondrogenic tumors in long bones [5]. When it comes to demonstrating osteolysis and calcification of chondrogenic tumors, CT is more advantageous than MRI. Nevertheless, there aren’t many studies using CT-based radiomics models to identify chondrogenic tumors in long bones. The CT-based radiomics machine learning model developed by Gitto et al. [6] performed admirably with regard to distinguishing ACT from high-grade chondrosarcoma in long bones, but the validation set only included CT images from PET-CT examination. Deng et al. [7] developed the CT-based texture analysis model to classify enchondroma and low-grade chondrosarcoma in long bones due to the limited number of included patients and extracted features, resulting in low accuracy in the model.

The aim of this study is to explore the value of CT-based radiomics machine learning models for distinguishing enchondroma from ACT in long bones and methods to improve model performance.

Materials and Methods

Patient selection

Approval from the Institutional Review Board was obtained, and in keeping with the policies for a retrospective review, informed consent was not required. Inclusion criteria: 1) enchondroma and ACT in long bones confirmed by pathology; 2) CT performed within 1 month before pathology. Exclusion criteria: 1) complicated with pathological fracture; 2) no first CT scans of recurrent ACT; 3) radiotherapy or chemotherapy before CT scans; 4) secondary ACT (Fig. 1).

Image acquisition

All enrolled patients underwent a multidetector row CT examination (Philips IQon Spectral CT; Siemens Somatom Definition AS 128; Siemens Somatom Sensatim 64). The CT scan parameters were: voltage: 120 kV; variable tube current; slice thickness: 0.65 to 1 mm; matrix: 512×512; field of view: from 120×120 mm to 400×400 mm.

Image segmentation

Two musculoskeletal radiologists, observer 1 with three years of experience and observer 2 with fifteen years of experience, used the Research Oncology Suite of IntelliSpace Discovery (ISD, Version 3.0, Philips Healthcare, The Netherlands) to perform semi-automatic 3D volume of interest (VOI) segmentation based on threshold intensity (Fig. 2). The Interclass Correlation Coeffi-
cient (ICC) was used to evaluate the repeatability of VOI segmentation between observers. The CT images of 30 patients (enchondroma = 16, ACT = 14) were randomly selected from all patients with chondrogenic tumors by observer 2 for repeatability verification. When the ICC value of the feature was $\geq 0.90$, it was considered to be a stable feature with great repeatability, and the follow-up steps were continued.

**Image Preprocessing and Feature extraction**

The PyRadiomics plugin of ISD was used for image preprocessing and feature extraction. CT images were resampled at a spatial resolution of voxels with a size of $1 \text{ mm} \times 1 \text{ mm} \times 3 \text{ mm}$ and discretized with a fixed bin width of 25HU [8]. All radiomic features were extracted with high throughput from the original and the filtered image, including first-order statistics, shape-based (3D), gray level co-occurrence matrix (GLCM), gray level size zone matrix (GLSZM), and gray level run length matrix (GLRLM). Filters included logarithm, exponential, square, square root, wavelet, and Laplacian of Gaussian (LoG).

**Feature selection and model development**

R software (Version 2023.12.0 + 369) was used for feature selection and dimensionality reduction, and the IntelliSpace Medicina Scientia Research Platform (ISMS, Version 3.0, Philips Healthcare, The Netherlands) was used for model development. After standardizing the features, ICC, t-tests, and least absolute shrinkage and selection operator (LASSO) regression were used for feature selection and dimensionality reduction in R software. The selected features were imported into ISMS, utilizing 13 machine learning algorithms to develop the radiomics model. To evaluate the performance of these models using ten-fold cross-validation, the data set was randomly divided into a training set (n = 78) and a test set (n = 34) at a 7:3 proportion in each cross-validation.

**Statistical analysis**

All statistical analyses were performed with SPSS (Version 27.0), R software, and ISMS. The statistically significant level was set to a two-sided P-value $< 0.05$. The T-test was used to analyze continuous variables, and the chi-square test was used to compare categorical variables. Indicators for evaluating the performance of the model included the area under the curve (AUC), accuracy (ACC), recall, precision, F1 score, Kappa, and Matthews correlation coefficient (MCC). The receiver operating characteristic curve (ROC), precision-recall curve (P-R), confusion matrix, and feature importance plot were drawn.

**Results**

Finally, 112 patients with enchondroma (n = 59) or ACT (n = 53) in long bones met the inclusion and exclusion criteria. The age of patients with enchondroma (42.19 ± 18.30) was less than ACT (50.42 ± 11.94) ($p = 0.005$), and there was no significant difference between sex ($p = 0.480$) and tumor location ($p = 0.909$) (▶Table 1).

A total of 1199 radiomic features were extracted, including 1172 stable features with an ICC $\geq 0.90$, then 388 features with statistical significance ($P < 0.05$) by t-test, and finally nine most valuable features by LASSO regression for selection and dimensionality reduction (▶Fig. 3 and ▶Table 2).

Among the 13 models constructed, eleven models had AUC values above 0.8 and three models above 0.9 (▶Table 3). The Extremely Randomized Trees (ERT) model had the best performance (AUC = 0.9375 ± 0.0884, ACC = 0.8500 ± 0.1225), followed by the Adaptive Boosting (ADA) model (AUC = 0.9188 ± 0.1010, ACC = 0.8732 ± 0.0970), and the Linear Discriminant Analysis (IDA) model (AUC = 0.9062 ± 0.1459, ACC = 0.8500 ± 0.1346). The ROC curve, P-R curve, confusion matrix, and feature importance plot were drawn according to the analysis results of the ERT model, ADA model, and IDA model shown in ▶Fig. 4, ▶Fig. 5 and ▶Fig. 6.

According to the feature importance plot, the most important feature in the ERT model was the zone entropy (ZE) of the GLSZM ▶Fig. 2.
feature from the logarithm-filtered image. By t-test again, the mean of the ZE in ACT (7.4503 ± 0.1799) was higher than in enchondroma (7.2035 ± 0.2375) (P < 0.05). In the ADA model, the most important feature was the surface area to volume ratio (SA/V) of the shape-based (3D) feature from the original image. The mean of the SA/V in ACT (0.2869 ± 0.0855) was discovered to be lower by t-test than in enchondroma (0.4782 ± 0.1421) (P < 0.05). The small area low gray level emphasis (SALGLE) of the GLSZM feature from the logarithm-filtered image was the most significant feature in the IDA model. The mean of the SALGLE in ACT (0.0014 ± 0.0010) was higher than in enchondroma (0.006 ± 0.003) by t-test once again (P < 0.05).

**Discussion**

This study used a variety of filters to extract high-order radiomic features from CT images and constructed radiomics models utilizing 13 machine learning algorithms to identify the enchondroma and ACT in long bones. The results revealed that the AUC value of
eleven models was more than 0.8 and that of three models was more than 0.9.

In the clinical data, the age of patients with enchondroma is lower than in ACT. Although some researchers have pointed out that most chondrosarcoma patients are older than enchondroma patients [9], in this study the difference in mean age between enchondroma and ACT patients is just eight years with a sizable standard deviation. Consequently, we propose that age has limited utility in the differential diagnosis of ACT and enchondroma in long bones. There were no significant differences in terms of sex or tumor location between ACT and enchondroma in long bones in this study. Pan et al. [10] found that the location of chondrogenic tumors was the most significant clinical risk factor. In their study, there were notably more enchondroma patients in short bones than chondrosarcoma, and the number of chondrosarcoma patients in the pelvis was much higher than enchondroma, while there was no significant difference in the long bones. According to the 2020 WHO classification of bone tumors [1], ACT has been specifically referred to as an intermediate chondrogenic tumor that occurs in the long and short tubular bones. Therefore, we think that the long bone is the most valuable location for differential diagnosis between enchondroma and ACT.

Among the nine most valuable features selected, three shape-based (3D) features were extracted from the original image, and

<table>
<thead>
<tr>
<th>Feature name</th>
<th>Feature class</th>
<th>Source image</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maximum 2D diameter (slice)</td>
<td>Shape (3D)</td>
<td>Original</td>
</tr>
<tr>
<td>Compactness 1</td>
<td>Shape (3D)</td>
<td>Original</td>
</tr>
<tr>
<td>Surface area to volume ratio</td>
<td>Shape (3D)</td>
<td>Original</td>
</tr>
<tr>
<td>Small area low gray level emphasis</td>
<td>GLSZM</td>
<td>Logarithm</td>
</tr>
<tr>
<td>Zone entropy</td>
<td>GLSZM</td>
<td>Logarithm</td>
</tr>
<tr>
<td>Size-zone non-uniformity normalized</td>
<td>GLSZM</td>
<td>Exponential</td>
</tr>
<tr>
<td>Root mean squared</td>
<td>First order</td>
<td>Wavelet (low-high-low pass filter)</td>
</tr>
<tr>
<td>Sum variance</td>
<td>GLCM</td>
<td>Wavelet (low-high-high pass filter)</td>
</tr>
<tr>
<td>Informational measure of correlation 1</td>
<td>GLCM</td>
<td>Wavelet (high-high-high pass filter)</td>
</tr>
</tbody>
</table>

Note: GLSZM: gray level size zone matrix; GLCM, gray level co-occurrence matrix
six high-order features were extracted from the filtered image. Because shape features are independent of the gray value of the voxel, they can only be extracted from the original image, while other features are calculated based on the gray value of the voxel and can be extracted from both the original image and the filtered image [4]. By using a variety of filters, plenty of high-order features were extracted, and more valuable features were used to construct models than in previous research [7]. Compared with the logistic regression (LR) model in their research, it had a higher AUC value and accuracy in this study, and there were other models that performed better than the LR model. The performance of the radiomics model can be improved by increasing the sample size, using filters to deeply mine high-order features from the original image, and selecting appropriate machine learning algorithms.

In the feature importance plot of the ERT model, the most important feature was the ZE:

\[
\text{Zone Entropy} = - \sum_{i=1}^{N_i} \sum_{j=1}^{N_j} P(i,j) \log_2(P(i,j) + \epsilon)
\]

ZE measures the uncertainty or randomness in the distribution of zone sizes and gray levels, and the higher the value, the higher the heterogeneity in the texture patterns [4]. By t-test again, the mean of the ZE in ACT was higher than in enchondroma, indicating that the heterogeneity of ACT in long bones is higher than that of enchondroma. The most important feature in the ADA model was the SA/V:

\[
\text{Surface Area to Volume ratio} = \frac{A}{V}
\]

This feature is dependent on the volume of the segmented VOI, and a lower value indicates that the shape of the segmented VOI is closer to a sphere [8]. The mean of the SA/V in ACT was discovered to be lower by t-test than in enchondroma, meaning that ACT has a more spherical-shaped volume in long bones than enchondroma, which may be related to the former’s more locally aggressive growth tendency [11]. The SALGLE was the most significant feature in the IDA model:

\[
\text{Small Area Low Gray Level Emphasis} = \sum_{i=1}^{N_i} \sum_{j=1}^{N_j} \frac{P(i,j)}{P(i,j)^2}
\]
**Fig. 5** The ability of the Adaptive Boosting model to correctly classify an enchondroma (label = 0, in figure) and atypical cartilaginous tumor (label = 1, in figure) in long bones in the test set. **a** Receiver operating characteristic curve; **b** Precision-recall curve; **c** Confusion matrix; **d** Feature importance plot.

**Fig. 6** The ability of the Linear Discriminant Analysis model to correctly classify an enchondroma (label = 0, in figure) and atypical cartilaginous tumor (label = 1, in figure) in long bones in the test set. **a** Receiver operating characteristic curve; **b** Precision-recall curve; **c** Confusion matrix; **d** Feature importance plot.
SALGLE measures the proportion of joint distribution of smaller size zones with lower gray values in the image. The larger its value, the greater the prevalence of these areas in the image and the more uneven the distribution of these areas [8]. The mean of the SALGLE in ACT was higher than in enchondroma by t-test once again, which indicates that ACT in long bones may be more prone to small zone necrosis, leading to an uneven density of the image compared to enchondroma.

The ERT model, ADA model, IDA model, random forest (RF) model, and gradient boosting classifier (GBC) model are the first few models that performed well in this study. They all utilize ensemble learning algorithms. Ensemble learning algorithms integrate many different machine learning algorithms in order to construct multiple models to enhance prediction accuracy and reduce generalization errors [12]. Ishaq et al. [13] used nine machine learning algorithms to construct machine learning models for predicting the survival of 299 patients with heart failure. The results showed that the ERT model achieved the best accuracy (ACC = 0.9262), followed by the RF model (ACC = 0.9188), the ADA model (ACC = 0.8852), and the GBC model (ACC = 0.8852), which performed better than the decision tree model (ACC = 0.8778) and the LR model (ACC = 0.8442). Erdem et al. [14] asked two researchers to extract features from MRI images and utilize seven machine learning algorithms to construct radiomics models to classify enchondroma (n = 57) and chondrosarcoma (n = 31). The results showed that when using all features to construct models, both researchers found the best model to be the neural network (NN) model (AUC = 0.979, AUC = 0.984, respectively). When using selected features to construct models, the best model for the two researchers was the GBC (AUC = 0.990) model and the NN model (AUC = 0.979). The NN algorithm is the fundamental algorithm of deep learning that can automatically identify specific structures when combined with machine learning, but it generally requires a large sample size [15]. Combined with the findings of these studies, it is found that utilizing advanced machine learning algorithms such as ensemble learning algorithms and deep learning algorithms can improve the prediction performance of models.

There are some limitations: First, the number of patients is relatively small due to the low incidence rate of enchondroma and ACT, the absence of treatment for most enchondroma patients, and the limiting location and grading of chondrogenic tumors. Second, this study did not employ MRI images of chondrogenic tumor patients to construct models concurrently, making it impossible to directly assess the performance differences between CT-based and MRI-based radiomics machine learning models.

**Conclusion**

This study found that CT-based radiomics machine learning models have great ability to distinguish enchondroma from ACT in long bones. The prediction performance of the model can be improved by using filters to deeply mine high-order features from the original image and selecting advanced machine learning algorithms.

**Conflict of Interest**

The authors declare that they have no conflict of interest.

**References**


