DOI: 10.4274/imj.galenos.2021.87528

Machine Learning-Based Computed Tomography Texture Analysis of Lytic Bone Lesions Needing Biopsy: A Preliminary Study

Biyopsi Gerektiren Litik Kemik Lezyonlarının Makine Öğrenme-Tabanlı Bilgisayarlı Tomografi Yapısal Analizi: Bir Ön Çalışma

İlhan Nahit Mutlu[1], Burak Koçak[1], Ece Ateş Kuş[2], Melis Baykara Ulusan[2], Özgür Kılıçkesmez[1]

[1]University of Health Sciences Turkey, Başakşehir Çam ve Sakura City Hospital, Clinic of Radiology, İstanbul, Turkey
[2]University of Health Sciences Turkey, İstanbul Training and Research Hospital, Clinic of Radiology, İstanbul, Turkey

ABSTRACT

Introduction: Currently, medical imaging has a limited capacity to achieve a final histopathological diagnosis of bone lesions. This study aimed to evaluate the use of machine learning (ML)-based computed tomography (CT) texture analysis to determine benign and malignant behaviors of lytic bone lesions needing a biopsy.

Methods: This retrospective study included 58 patients with lytic bone lesions. Lesion segmentation was independently performed by two observers. After evaluating unenhanced CT images, a total of 744 texture features were obtained. Reproducibility analysis and feature selection were used for dimension reduction. A training data set with a nested cross-validation approach was used for feature selection, optimization, and validation. Testing was executed on the remaining unseen data set. Classifications were done using five base ML classifiers and three voting strategies.

Results: The best predictive performance was achieved using the k-nearest neighbors algorithm with six features. The area under the curve, accuracy, sensitivity, and specificity of the best algorithm were, respectively, 0.774%, 78.1%, 78%, and 78.1% for the validation data set; and 0.861, 82.4%, 82.4%, and 81.5% for the unseen test data set.

Conclusion: The ML-based CT texture analysis may be a promising non-invasive technique for determining benign and malignant behaviors of lytic bone lesions that need a biopsy.

Keywords: Bone, texture analysis, radiomics, machine learning, artificial intelligence

ÖZ

Amaç: Günümüzde, sadece tıbbi görüntüleme ile kemik lezyonlarının kesin histopatolojik tanısını koymak mümkün olmamaktadır. Bu çalışmada, biyopsi gerektiren litik kemik lezyonlarının benign veya malign olduklarını belirleyebilmek için makine öğrenme (MÖ) tabanlı bilgisayarlı tomografi (BT) yapısal analizinin değerini ölçmeyi amaçladık.

Yöntemler: Bu retrospektif çalışmaya litik kemik lezyonu olan 58 hasta dahil edilmiştir. Lezyon segmentasyonu bağımsız iki gözlemci tarafından gerçekleştirilmiştir. Toplamda, kontrastsız BT görüntülerinden 744 yapısal özellik çıkartılmıştır. Boyut küçültme, tekrarlanabilirlik analizi ve özellik seçimi ile yapılmıştır. Özellik seçimi, optimizasyon ve doğrulama, iç içe geçmiş çapraz doğrulama yaklaşımına sahip bir eğitim veri kümesi kullanılarak yapılmıştır. Geriye kalan görünmeyen veri seti üzerinde test yapılmıştır. Sınıflandırmalar, beş temel MÖ sınıflandırıcısı ve üç farklı oylama stratejisi kullanılarak yapılmıştır.

Bulgular: En iyi tahmin performansı, altı özelliğe sahip k-nearest neighbors algoritması ile elde edilmiştir. En iyi algoritma değerinin eğri altındaki alan, doğruluk, duyarlılık ve özgüllük değerleri doğrulama veri seti için sırasıyla; %0,774, %78,1, %78 ve %78,1; görünmeyen test veri seti için ise sırasıyla; %0,861, %82,4, %82,4 ve %81,5 idi.

Sonuç: MÖ tabanlı BT yapısal analizi, biyopsi gerektiren litik kemik lezyonlarının benign ve malign davranışlarını tahmin etmek için ümit verici, invazif olmayan bir teknik olabilir.

Anahtar Kelimeler: Kemik, yapısal analiz, radyomik, makine öğrenme, yapay zeka

Address for Correspondence/Yazışma Adresi: İlhan Nahit Mutlu MD, University of Health Sciences Turkey, Başakşehir Çam ve Sakura City Hospital, Clinic of Radiology, İstanbul, Turkey
Phone: +90 533 635 54 52 **E-mail:** ilhannahit@gmail.com **ORCID ID:** orcid.org/0000-0002-9326-5432

Received/Geliş Tarihi: 24.03.2021
Accepted/Kabul Tarihi: 16.06.2021

Cite this article as/Atıf: Mutlu İN, Koçak B, Ateş Kuş E, Baykara Ulusan M, Kılıçkesmez Ö. Machine Learning-Based Computed Tomography Texture Analysis of Lytic Bone Lesions Needing Biopsy: A Preliminary Study. İstanbul Med J 2021; 22(3): 223-31.

Introduction

Texture analysis, which is a vital part of radiomics, is used to change standard medical images into high-dimensional quantitative data by calculating distribution and patterns of voxels or pixels (1,2). The literature has widely suggested that texture analysis might have a potential value in predicting certain underlying pathology or outcomes in different organs or systems (2). Contrary to standard qualitative evaluation, texture analysis may provide an objective, non-invasive assessment of the medical images, possibly leading to better decision-making in patient management (2,3). Moreover, artificial intelligence offers robust and reliable tools that learn data patterns and then make predictions on unseen instances for better decision support to manage such high-dimensional quantitative data that the texture analysis supplies (2,4).

Imaging of bone lesions has heavily relied on radiographs for a long time (5). Nowadays, it comprises technologically more advanced armamentarium, including positron emission tomography (PET), ultrasound, magnetic resonance imaging (MRI), and computed tomography (CT) (5,6). From a diagnostic point of view, all of these imaging methods can contribute to tissue characterization by narrowing the range of differential diagnoses and then indicating the most appropriate course of action afterward (7). In other words, these approaches have a limited capacity to achieve a final histopathological diagnosis (6). In particular, CT can help determine the calcification pattern in the bone lesion matrix, identify occult destruction, or even localize the nidus of an osteoid osteoma (5,6). However, if we consider all possible lesion types that can be encountered in clinical practice, it has a limited capability in tissue characterization. Interestingly, unenhanced CT texture analysis is being used to evaluate many different pathologies located in various organs or systems with a promising predictive performance (8-11).

In our work, we assessed the future value of unenhanced CT texture analysis for foreseeing benign and malignant behaviors of lytic bone lesions that need biopsy in clinical practice using various state-of-the-art machine learning (ML) algorithms and strategies.

Methods

Ethics

All study procedures, including waiver of informed consent for medical records review, were approved by our institutional review board. The approval form the University of Health Sciences Turkey, İstanbul Training and Research Hospital Local Ethics Committee was obtained (approval number: 1965, date: 29.08.2019).

Patients

Biopsy-proven bone lesions examined between January 2016 and May 2019 were obtained from our picture archiving and communication system. The exclusion criteria of patients were as follows: 1) mixed or sclerotic bone lesions, 2) unavailability of unenhanced CT in our archive, 3) quality problems in CT study, and 4) indefinite border or small (≤5 mm) lesions. No other criterion regarding the malignancy status of the patients was applied. A simplified flowchart for patient selection is presented in Figure 1.

Computed Tomography Protocol

CT scans were performed using different scanners as follows: a 128-slice multidetector CT (Ingenuity, Philips Healthcare, Cleveland, OH, USA), a 64-slice multidetector CT (Aquilion, Canon Medical Systems, Otawara, Japan), and a 2-slice helical CT (HiSpeed, General Electric Company, Fairfield, CT, USA). Overall, the CT parameters were as follows: 1) tube voltage of 100-140 kV, 2) tube current of 97-500 mAs, 3) slice thickness of 0.5-5 mm, 4) pixel size of 0.162-0.976 mm, and 5) no contrast medium administration.

Technical Workflow

To provide a basic understanding and a larger view to the reviewers, we summarized our technical workflow in a flowchart in Figure 2.

Preprocessing

To reduce inter-scanner differences, all CT images were normalized using the ±3 sigma technique that centers voxel gray-level values at the mean with the standard deviation (SD) (12). Normalization was established on the following expression:

$$f(x)=\frac{s[x-\mu(x)]}{\sigma(x)}$$

where $f\ (x)$ is the normalized image gray-level value, x is the original gray-level value, $\mu\ (x)$ is the mean gray-level value, $\sigma\ (x)$ is the SD of the image gray-level value, and s is the scaling factor, which was 100 in this study.

Resampling and rescaling of pixel spaces to an in-plane resolution of 1x1 mm^2 were performed because comparing texture characteristics necessitates identical spatial resolution (13).

We used a fixed bin-width to obtain an ideal bin count between 16 and 128 for gray-level discretization (14,15). To obtain this, a preliminary extraction of first-order parameters was performed in all patients included in this work to calculate the gray-level range and optimal bin

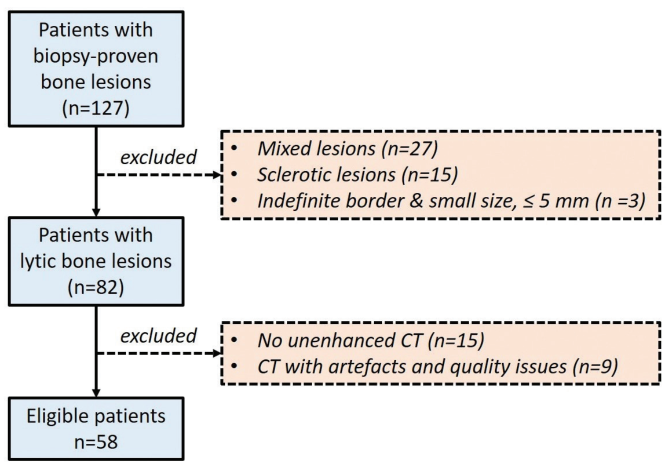
Patients with biopsy-proven bone lesions (n=127)
excluded
• Mixed lesions (n=27)
• Sclerotic lesions (n=15)
• Indefinite border & small size, ≤ 5 mm (n =3)
Patients with lytic bone lesions (n=82)
excluded
• No unenhanced CT (n=15)
• CT with artefacts and quality issues (n=9)
Eligible patients n=58

Figure 1. Patient selection flowchart

CT: Computed tomography

width. The discretization was established on the following **mathematical** expression:

$$\mathrm{Xb,i} = \left\lfloor \frac{\mathrm{Xgl,i}}{\mathrm{W}} \right\rfloor - \left\lfloor \frac{\min(\mathrm{Xgl,i})}{\mathrm{W}} \right\rfloor + 1$$

where $X_{b,i}$ is the gray-level value after discretization, $X_{gl,i}$ is the gray-level value before discretization, and W is the bin-width value, which was 2 in this study.

Padding was applied with a distance value of 5 using original gray-level intensity.

Texture Analysis

By depicting a polygonal region of interest (ROI), the lytic bone lesions were segmented using the largest representative axial image slice of unenhanced CT. To minimize the partial volume effect from visually healthy structures, the ROI was carefully depicted, considering the clear lesion margin. The segmentation style is presented in Figure 3. Two observers segmented the lesions for a feature reproducibility analysis. The possible influence of the slice selection bias was considered since this might be a major concern of texture analysis, which is based on a single slice (16,17). Therefore, each observer was blind to the selected slices by the other observer.

PyRadiomics software program (PyRadiomics 2.0.1; Python 2.7.13; Numpy 1.13.1; SimpleITK 1.1.0; PyWavelet 0.5.2) was used for extracting texture features (18). Original, filtered, and wavelet-transformed images were used for extracting the features. Laplacian of Gaussian (LoG) filter was used for image filtrations with values of 2, 4, and 6 mm presenting with fine, medium, and coarse patterns, respectively. The following are the extracted texture features: 1) 18 first-order features, 2) 14 gray-level dependence matrix features, 3) 24 gray-level co-occurrence matrix features, 4) 16 gray-level run-length matrix features. 5) 16 gray-level size zone matrix features, and 6) 5 neighboring gray-tone difference matrix features. These six groups of features were derived from one original, three LoG-filtered, and four wavelet-transformed images, which add up to 744 features per lesion in total. Detailed feature names and classes are presented in the Online Supplement. Comprehensive descriptions and mathematical expressions for these features can be found on the software program website and in the references (19-23).

Data Handling

The radiomic data underwent randomization, standardization, and stratified sampling.

The standardization of radiomic feature values was performed by centering and scaling the values by the mean and SD, respectively. Then, before data sampling, the order of data sets related to patient identifiers was randomized.

The whole data set was sampled to create training and unseen test data set splits, with proportions of 70% and 30%, respectively. The data split was created to prevent possible information leakage. Class balance in the training and unseen test data sets was ensured using stratified sampling, which sets the same class balance in the whole data set to the training and testing data sets. Training data set was used in classifier-specific feature selection, model development, optimization, and validation. The remaining data set, which is unseen by the feature selection and ML algorithms, was used for the unbiased testing of the model.

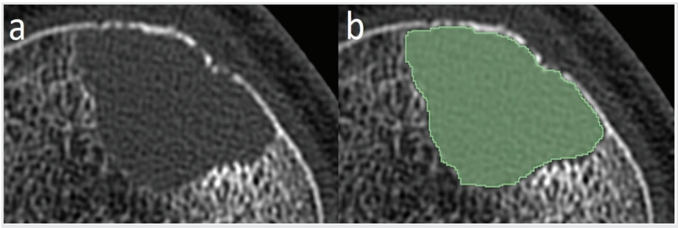

Figure 3. Lesion segmentation. (a) Unenhanced computed tomography image of a 51-year-old man with a giant cell tumor in the proximal tibia. (b) Lesion is segmented using the largest representative axial image slice with a particular focus on the contour

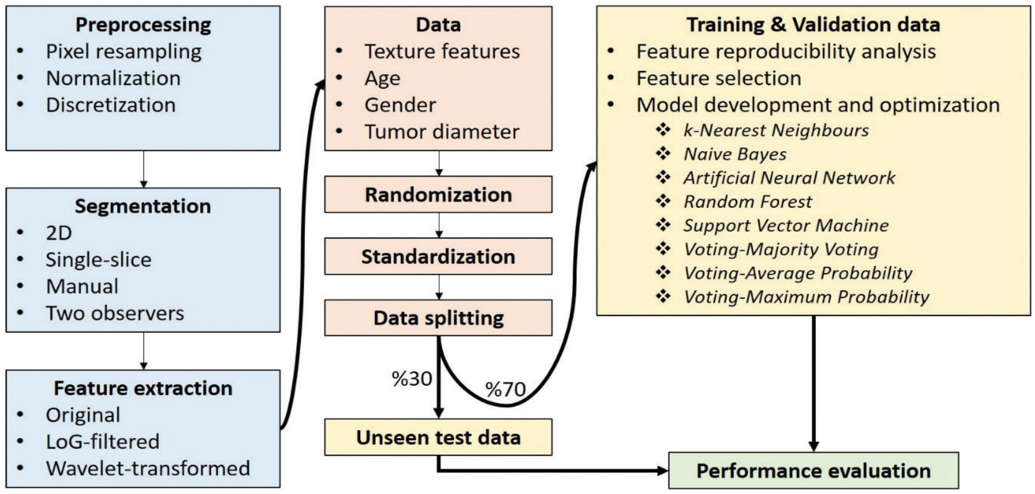

Figure 2. Technical study pipeline

2D: Two-dimensional, LoG: Laplacian of Gaussian

Dimension Reduction

A feature reproducibility analysis was initially used, followed by a classifier-specific feature selection algorithm to reduce the dimension of the training data.

For the reproducibility analysis, the intra-class correlation coefficient (ICC) was calculated for each texture feature and a clinical variable (maximum tumor diameter) using a two-way model, single rating, and absolute agreement. In the following dimension reduction step, which is also called feature selection, only features with ICC ≥0.75 that indicated good and excellent inter-observer reproducibility were included.

Waikato Environment for Knowledge Analysis (WEKA) toolkit version 3.8.2 was used for the feature selection (24). In the feature selection and model validation process, a nested cross-validation method with 5-fold inner and 10-fold outer loops was used. An incremental wrapper-based subset search method along with a wrapper attribute evaluator was used (25,26). The features were graded by their probabilistic significance, which was computed as a two-way function in the search method. The attributes undergoing more than one inner cross-validations were classified in the outer loop. Of note, age, gender, and maximal lesion diameter were also included in the feature selection as clinical variables besides the texture features.

Machine Learning-based Classifications

WEKA toolkit was used for ML-based classifications (24). The five base ML algorithms used were as follows: k-nearest neighbors, naive Bayes, random forest, support vector machine, and artificial neural network (27). In addition, these algorithms were also used in an ensemble learning technique called voting with three strategies (majority voting, average probability, and maximum probability) (28).

Considering the potential bias of the internal validation techniques, the performance evaluation was performed both in the training data as validation and in the unseen data as testing (29). The performance of the algorithms was mainly evaluated using the area under the receiver operating characteristic curve (AUC). Moreover, the accuracy, sensitivity, specificity, precision, F-measure, and Matthews correlation coefficient were also determined for further evaluation. For sensitivity and specificity, the weighted averages were also calculated.

Reference Standards for Classifications

The reference standards for the classifications were based on the official histopathological reports. Primary malignancies, secondary malignancies, and systemic malignancies were grouped as malignant lesions (30). Other lesions were grouped as benign lesions (30).

Conventional Statistical Analysis of Baseline Characteristics

Based on the value distribution, the parametric or non-parametric statistical tests were used to compare age and maximum lesion diameter between training and testing data sets. The chi-square test was used to compare the proportions of the gender. A p-value of less than 0.05 indicated statistical significance.

Results

Baseline Characteristics

A total of 58 patients met our eligibility criteria. Of them, 41 and 17 patients were randomly assigned to the training and test datasets, respectively. The following class distributions were in almost perfect balance using stratified sampling: 28 benign vs 30 malignant for the whole data, 20 benign vs 21 malignant for the training data, and 8 benign vs 9 malignant for the test data. There was no statistically significant difference in age (p=0.191), gender (p=0.488), and maximum lesion diameter (p=0.447) between the training and unseen test data sets. The baseline characteristics of the patients and their lesions are presented in Table 1.

Dimension Reduction

Inter-observer reproducibility was good or excellent in 464 of 744 texture features (ICC: ≥0.75). Additionally, the inter-observer agreement for the maximum lesion diameter was excellent (ICC: 0.905). All the reproducible texture features (ICC: ≥0.75) and clinical variables (age, gender, and maximum lesion diameter) were contained in the following feature selection process based on an algorithm.

In total, 15 texture features and 1 clinical variable were selected (Figure 4). Using the classifier-specific feature selection algorithm for each ML classifier to perform optimization, selected feature subsets were substantially different across the models created. Selected feature subsets are presented in Table 2. The selected feature numbers for each classifier ranged from 2 to 6. Considering all selected features, there was a predominance of the features extracted from the wavelet-transformed images. Based on feature classes, first-order and gray-level co-occurrence matrix features outnumbered the others. The only clinical variable selected by the algorithms was the maximum lesion diameter, which was included in the feature subset of k-nearest neighbors (Figure 5).

Machine Learning-based Training and Nested Cross-validation

Considering the five base ML classifiers, the AUC and accuracy metrics ranged from 0.724 to 0.774 and from 73.2% to 78.1%, respectively. The best performance was achieved by the k-nearest neighbors, with a weighted average sensitivity and specificity of 78% and 78.1%, respectively. Regarding the three voting strategies, the AUC and accuracy metrics ranged from 0.712 to 0.757 and from 75.6% to 82.9%, respectively. The majority voting achieved the best performance, with a weighted average sensitivity and specificity of 75.6% and 75.8%, respectively. Nested cross-validation performance metrics of the ML algorithms on the training data set are presented in Table 3.

Machine Learning-based Unseen Testing

Considering the five base ML classifiers, the AUC and accuracy metrics ranged from 0.715 to 0.861 and from 70.6% to 82.4%, respectively. The k-nearest neighbors achieved the best performance, with a weighted average sensitivity and specificity of 82.4% and 81.5%, respectively. Regarding the three voting strategies, the AUC and accuracy ranges were

Table 1. Baseline characteristics of 58 study patients

Characteristics	Whole data (n=58)	Training and validation data (n=41)	Unseen test data (n=17)
Age (years)	56.6	54.9	60.8
Gender, n (%)			
Female	21 (36.2)	16 (39)	5 (29.4)
Male	37 (63.8)	25 (61)	12 (70.6)
Mean lesion diameter (mm)*	31.1	28.9	36.6
Benign lesions, n (%)	28 (48.3)	20 (48.8)	8 (47.1)
Neoplastic	3 (5.2)	3 (7.3)	0 (0.0)
Non-neoplastic	25 (43.1)	17 (41.5)	8 (47.1)
Malign lesions, n (%)	30 (51.7)	21 (51.2)	9 (52.9)
Metastasis	17 (29.3)	12 (23.3)	5 (29.4)
Plasma cell malignancy	8 (13.8)	5 (12.2)	3 (17.6)
Other malignant lesions	5 (8.6)	4 (9.7)	1 (5.9)
Lesion location, n (%)			
Acetabulum	2 (3.4)	2 (4.9)	0 (0.0)
Femur	6 (10.3)	3 (7.3)	3 (17.6)
Fibula	1 (1.7)	1 (2.4)	0 (0.0)
Humerus	4 (6.9)	3 (7.3)	1 (5.9)
Iliac	8 (13.8)	6 (14.6)	2 (11.8)
Ischium	2 (3.4)	1 (2.4)	1 (5.9)
Calcaneus	1 (1.7)	1 (2.4)	0 (0.0)
Rib	2 (3.4)	1 (2.4)	1 (5.9)
Mandibula	1 (1.7)	1 (2.4)	0 (0.0)
Pubis	1 (1.7)	0 (0.0)	1 (5.9)
Radius	1 (1.7)	1 (2.4)	0 (0.0)
Sacrum	7 (12.1)	5 (12.2)	2 (11.8)
Scapula	2 (3.4)	2 (4.9)	0 (0.0)
Sternum	1 (1.7)	1 (2.4)	0 (0.0)
Tibia	4 (6.9)	1 (2.4)	3 (17.6)
Vertebra	15 (25.9)	12 (29.3)	3 (17.6)

*Based on the three-dimensional maximum lesion diameter. Unless otherwise stated, data represent number of patients or lesions

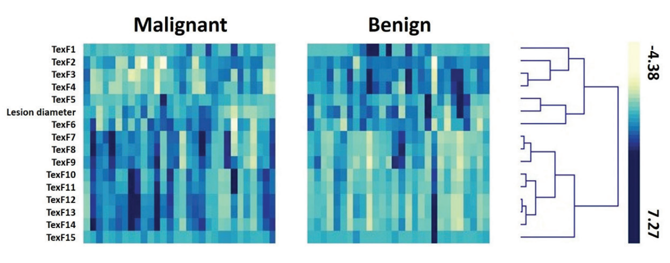

Figure 4. Heatmap and unsupervised clustering of the texture features selected for all machine learning algorithms

TexF1: Strength in wavelet image (high-low), TexF2: informational measure of correlation-2 in wavelet image (low-low), TexF3: entropy in wavelet image (low-low), TexF4: difference entropy in wavelet image (low-low), TexF5: size zone nonuniformity in original image, TexF6: inverse difference normalized in wavelet image (low-low), TexF7: dependence variance in original image, TexF8: large dependence emphasis in original image, TexF9: inverse variance in wavelet image (low-high), TexF10: maximum probability in wavelet image (low-low), TexF11: joint energy in wavelet image (low-low), TexF12: uniformity in wavelet image (low-low), TexF13: gray-level nonuniformity normalized in wavelet image (low-low), TexF14: gray-level nonuniformity normalized in original image, TexF15: large dependence low gray-level emphasis in Laplacian of Gaussian-filtered image with 4 mm lesion diameter, maximum lesion diameter

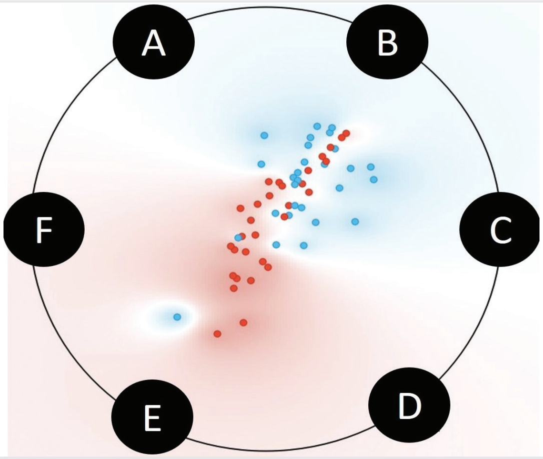

Figure 5. Two-dimensional projection of the selected features for the best machine learning algorithm (k-nearest neighbors)

A: Strength in wavelet image (high-low), B: informational measure of correlation-2 in wavelet image (low-low), C: joint energy in wavelet image (low-low), D: maximum lesion diameter, E: gray-level nonuniformity normalized in wavelet image (low-low), F: uniformity in wavelet image (low-low)

Table 2. Selected feature subsets for each machine learning algorithm

Algorithm	Selected features (feature class and image type)	ICC
k-Nearest neighbors	Uniformity (first-order, wavelet-LL)	0.835
	Gray-level nonuniformity normalized (GLRLM, wavelet-LL)	0.819
	Joint energy (GLCM, wavelet-LL)	0.897
	Maximum tumor diameter	0.905
	Informational measure of correlation 2 (GLCM, wavelet-LL)	0.808
	Strength (NGTDM, wavelet-HL)	0.908
Naive Bayes	Uniformity (first-order, wavelet-LL)	0.835
	Large dependence low gray-level emphasis (GLDM, LoG-4 mm)	0.887
Support vector machine	Uniformity (first-order, wavelet-LL)	0.835
	Maximum probability (GLCM, wavelet-LL)	0.849
Random forest	Uniformity (first-order, wavelet-LL)	0.835
	Difference entropy (GLCM, wavelet-LL)	0.849
	Joint energy (GLCM, wavelet-LL)	0.897
	Inverse difference normalized (GLCM, wavelet-LL)	0.760
Artificial neural network	Uniformity (first-order, wavelet-LL)	0.835
	Gray-level **non**-uniformity normalized (GLRLM, wavelet-LL)	0.819
	Entropy (first-order, wavelet-LL)	0.796
	Dependence variance (GLDM, original)	0.819
	Large dependence emphasis (GLDM, original)	0.826
	Inverse variance (GLCM, wavelet-LH)	0.772
Voting-majority voting	Uniformity (first-order, wavelet-LL)	0.835
	Gray-level **non**-uniformity normalized (GLRLM, wavelet-LL)	0.819
	Size zone **non**-uniformity (GLSZM, original)	0.908
Voting-average probability	Uniformity (first-order, wavelet-LL)	0.835
	Joint energy (GLCM, wavelet-LL)	0.897
Voting-maximum probability	Uniformity (first-order, wavelet-LL)	0.835
	Gray-level **non**-uniformity normalized (GLRLM, original)	0.861

ICC: Intra-class correlation coefficient, GLRLM: gray-level run-length matrix, GLCM: gray-level co-occurrence matrix, NGTDM: neighboring gray-tone difference matrix, GLDM: gray-level dependence matrix, GLSZM: gray-level size zone matrix, LoG: Laplacian of Gaussian, LL: low-low, HL: high-low, LH: low-high

0.708 to 0.806 and 70.6% to 76.5%, respectively. Voting strategy based on maximum probability achieved the best performance, with a weighted average sensitivity and specificity of 75.6% and 75.8%, respectively. Performance metrics of the ML algorithms on the testing data set are presented in Table 4.

Discussion

Study Overview

We assessed the future predictive value of the ML-based CT texture analysis to distinguish benign and malignant behaviors of lytic bone lesions that need a biopsy procedure in clinical practice. We created models using five base ML classifiers and three different voting strategies. The predictive performance of the models was evaluated using two approaches: 1) training along with a nested cross-validation approach and 2) testing on an unseen data set (or a random holdout). For each base classifier and voting strategy, a different feature subset was selected. The k-nearest neighbors achieved the best predictive performance. Using this base ML algorithm, more than 80% of the patients were sorted rightly. The voting strategy yielded no improvement in the predictive performance.

Practical Implications

Differential diagnosis of benign lytic bone lesions usually includes fibrous dysplasia, eosinophilic granuloma, enchondroma, giant cell tumor, non-ossifying fibroma, osteoblastoma, aneurysmal bone cyst, solitary bone cyst, chondroblastoma, brown tumor, and infection-related pathologies (30). On the other hand, the differential diagnosis for the malignant lytic lesion category is a little shorter and mainly includes metastasis, myeloma, rare osteosarcoma, and chondrosarcoma (30). Also, Ewing's sarcoma and leukemia should be considered in the pediatric age group. Conventionally, differentiation of these lesions is made by age, lesion localization, and qualitative imaging features such as periosteal reaction, cortical destruction, lesion margins, matrix pattern, and transition zone (30). However, these features may overlap between benign and malignant lesions, leading to diagnostic confusion

Table 3. Nested cross-validation in training data

Algorithm	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	F-Measure	MCC	AUC	Confusion matrix*		
								B	M	R
k-Nearest neighbors	78.1	80.0	76.2	76.2	0.780	0.562	0.774	16	4	Benign
		76.2	80.0	80.0	0.780			5	16	Malign
Naive Bayes	75.6	90.0	61.9	69.2	0.783	0.539	0.757	18	2	Benign
		61.9	90.0	86.7	0.722			8	13	Malign
Support vector machine	75.6	95.0	57.1	67.9	0.792	0.560	0.761	19	1	Benign
		57.1	95.0	92.3	0.706			9	12	Malign
Random forest	75.6	75.0	76.2	75.0	0.750	0.512	0.742	15	5	Benign
		76.2	75.0	76.2	0.762			5	16	Malign
Artificial neural network	73.2	80.0	66.7	69.6	0.744	0.470	0.724	16	4	Benign
		66.7	80.0	77.8	0.718			7	14	Malign
Voting-majority voting	75.6	80.0	71.4	72.7	0.762	0.516	0.757	16	4	Benign
		71.4	80.0	78.9	0.750			6	15	Malign
Voting-average probability	82.9	95.0	71.4	76.0	0.844	0.681	0.719	19	1	Benign
		71.4	95.0	93.8	0.811			6	15	Malign
Voting-maximum probability	78.0	90.0	66.7	72.0	0.800	0.581	0.712	18	2	Benign
		66.7	90.0	87.5	0.757			7	14	Malign

*B and M indicate classification results. Benign and malign indicate reference standards.
MCC: Matthews correlation coefficient, AUC: area under the curve, B: benign, M: malign

Table 4. Testing on the remaining unseen data

Algorithm	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	F-Measure	MCC	AUC	Confusion matrix*		
								B	M	R
k-Nearest neighbors	82.4	75.0	88.9	85.7	0.800	0.648	0.861	6	2	Benign
		88.9	75.0	80.0	0.842			1	8	Malign
Naive Bayes	82.4	87.5	77.8	77.8	0.824	0.653	0.806	7	1	Benign
		77.8	87.5	87.5	0.824			2	7	Malign
Support vector machine	70.6	87.5	55.6	63.6	0.737	0.450	0.715	7	1	Benign
		55.6	87.5	83.3	0.667			4	5	Malign
Random forest	76.5	75.0	77.8	75.0	0.750	0.528	0.792	6	2	Benign
		77.8	75.0	77.8	0.778			2	7	Malign
Artificial neural network	82.4	87.5	77.8	77.8	0.824	0.653	0.792	7	1	Benign
		77.8	87.5	87.5	0.824			2	7	Malign
Voting-majority voting	70.6	75.0	66.7	66.7	0.706	0.417	0.708	6	2	Benign
		66.7	75.0	75.0	0.706			3	6	Malign
Voting-average probability	70.6	87.5	55.6	63.6	0.737	0.450	0.778	7	1	Benign
		55.6	87.5	83.3	0.667			4	5	Malign
Voting-maximum probability	76.5	87.5	66.7	70.0	0.778	0.549	0.806	7	1	Benign
		66.7	87.5	85.7	0.750			3	6	Malign

*B and M indicate classification results. Benign and Malign indicate reference standard.
MCC: Matthews correlation coefficient, AUC: area under the curve, B: benign, M: malign

(30). Some conditions called tumor mimickers may even make the diagnosis much more challenging (31). While evaluating such conditions, unnecessary diagnostic work-up, including invasive procedures, might be related to patient morbidity, discomfort, and high economic cost. Therefore, differentiating benign lesions from malignant ones by non-invasive methods is necessary.

It may seem that the technique proposed might have low clinical applicability since the biopsy would be performed on all such lesions. In other words, even if the clinicians were about 80% sure that a lesion was benign; one would still likely perform a biopsy. However, this investigation should be considered a preliminary work, with its many obvious limitations. We think there is more room for improvement of this technique to make it more useful in clinical practice.

Previous Works

Most of the radiomic studies about lytic bone lesions evaluated the performance of computer-aided or automatic lesion detection systems, with a particular focus on spinal bone lesions (32-36). Meanwhile, other papers worked on lesion classification problems using texture analysis. Larhmam et al. (37) worked on spinal metastasis classification using conventional MRI images and achieved an accuracy of 90.1%. Reischauer et al. (38) published a prospective study in patients with prostate cancer along with bone metastasis by investigating the potential value of texture features extracted from apparent diffusion coefficient maps in treatment response assessment (38). Acar et al. (39) worked on ML-based CT texture analysis to distinguish metastatic and completely responded sclerotic bone lesions in patients with prostate cancer (39). In our study, rather than lesion detection, we focused on the lesion classification with a different perspective, i.e., distinguishing benign and malignant lesions that need a biopsy in clinical practice. Furthermore, we included spinal lesions and ones from different locations.

Study Limitations

There are a few limitations to the generalizability of our results. First, the main limitations were the retrospective nature of the study and the relatively small number of patients. Second, lesions from different locations were included in this work. It would be worth looking at specific locations such as the spine or pelvic bones. Nonetheless, the major constraint for this was the limited number of patients considering major locations. Third, we only used unenhanced CT with a rather heterogeneous protocol in this preliminary work. The imaging protocol heterogeneity was due to the usage of different scanners, which may represent the clinical practice and improve the generalizability of the findings. The patients' other imaging studies, such as PET-CT and MRI, were also heterogeneous, and some had been performed in different centers. Because of these, we were unable to use other techniques for texture analysis. On the other hand, other imaging methods including MRI and PET-CT/MRI should be evaluated in future works. Fourth, we used a single-slice two-dimensional manual segmentation. Although a few slice-based or three-dimensional volumetric segmentation would be much more illustrative for the lesion texture, it is too difficult to use in clinical practice unless it is performed with automated methods. The major problem of the two-dimensional texture analysis in such large lesions is the slice selection bias, which was considered in this work (16,17). Fifth, although we performed separate testing on the unseen data set or a holdout data set, there was no external data set. On the other hand, we plan to perform independent external validation when appropriate data are available. Sixth, we only conducted quantitative analysis on the lesions that we performed a biopsy procedure. Because some of the patients were referred to our hospital for the biopsy procedure, not all the imaging data were available to conduct a proper qualitative analysis for comparison.

Conclusion

This preliminary work suggests that the ML-based CT texture analysis may be a promising non-invasive technique to distinguish benign and malignant behaviors of lytic bone lesions that need a biopsy. By improving the above mentioned limitations of this work, future research may have the potential to increase the predictive performance of this method. We hope this comprehensive work will provide a base for future research.

Ethics Committee Approval: The approval form the University of Health Sciences Turkey, İstanbul Training and Research Hospital Local Ethics Committee was obtained (approval number: 1965, date: 29.08.2019).

Informed Consent: All study procedures, including waiver of informed consent for medical records review, were approved by our institutional review board.

Peer-review: Externally and internally peer-reviewed.

Authorship Contributions: Surgical and Medical Practices- İ.N.M., Ö.K.; Concept - İ.N.M., B.K., Ö.K.; Design - B.K., E.A.K., M.B.U.; Data Collection or Processing - İ.N.M., B.K., E.A.K., M.B.U.; Analysis or Interpretation - B.K., Ö.K.; Literature Search - İ.N.M., B.K., E.A.K., M.B.U.; Writing - İ.N.M., B.K.

Conflict of Interest: No conflict of interest was declared by the authors.

Financial Disclosure: The authors declared that this study received no financial support.

References

1. Varghese BA, Cen SY, Hwang DH, Duddalwar VA. Texture Analysis of Imaging: What Radiologists Need to Know. AJR Am J Roentgenol 2019; 212: 520-8.
2. Koçak B, Durmaz EŞ, Ateş E, Kılıçkesmez Ö. Radiomics with artificial intelligence: a practical guide for beginners. Diagn Interv Radiol 2019; 25: 485-95.
3. Gillies RJ, Kinahan PE, Hricak H. Radiomics: Images Are More than Pictures, They Are Data. Radiology 2016; 278: 563-77.
4. Bi WL, Hosny A, Schabath MB, Giger ML, Birkbak NJ, Mehrtash A, et al. Artificial intelligence in cancer imaging: Clinical challenges and applications. CA Cancer J Clin 2019; 69: 127-57.
5. Mintz DN, Hwang S. (2014) Bone tumor imaging, then and now: review article. HSS J 2014; 10: 230-9.
6. Miwa S, Otsuka T. Practical use of imaging technique for management of bone and soft tissue tumors. J Orthop Sci 2017; 22: 391-400.
7. Greenspan A, Jundt G, Remagen W, Technologies O. Differential diagnosis of orthopaedic oncology. Lippincott Williams & Wilkins, Philadelphia; 2007.
8. Kocak B, Durmaz ES, Ates E, Kaya OK, Kilickesmez O. Unenhanced CT Texture Analysis of Clear Cell Renal Cell Carcinomas: A Machine Learning-Based Study for Predicting Histopathologic Nuclear Grade. AJR Am J Roentgenol 2019; W1-8.
9. Yi X, Guan X, Chen C, Zhang Y, Zhang Z, Li M, et al. Adrenal incidentaloma: machine learning-based quantitative texture analysis of unenhanced CT can effectively differentiate sPHEO from lipid-poor adrenal adenoma. J Cancer 2018; 9: 3577-82.

10. Liu S, Zheng H, Pan X, Chen L, Shi M, Guan Y, et al. Texture analysis of CT imaging for assessment of esophageal squamous cancer aggressiveness. J Thorac Dis 2017; 9: 4724-32.

11. Liu S, Pan X, Liu R, Zheng H, Chen L, Guan W, et al. Texture analysis of CT images in predicting malignancy risk of gastrointestinal stromal tumours. Clin Radiol 2018; 73: 266-74.

12. Collewet G, Strzelecki M, Mariette F. Influence of MRI acquisition protocols and image intensity normalization methods on texture classification. Magn Reson Imaging 2004; 22: 81-91.

13. Shafiq-Ul-Hassan M, Latifi K, Zhang G, Ullah G, Gillies R, Moros E. Voxel size and gray level normalization of CT radiomic features in lung cancer. Sci Rep 2018; 8: 10545.

14. Duron L, Balvay D, Vande Perre S, Bouchouicha A, Savatovsky J, Sadik JC, et al. Gray-level discretization impacts reproducible MRI radiomics texture features. PLoS One 2019; 14: e0213459.

15. Leijenaar RT, Nalbantov G, Carvalho S, van Elmpt WJ, Troost EG, Boellaard R, et al. The effect of SUV discretization in quantitative FDG-PET Radiomics: the need for standardized methodology in tumor texture analysis. Sci Rep 2015; 5: 11075.

16. 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 2019; 214: 1-8.

17. 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 2019; 213: 377-83.

18. van Griethuysen JJM, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V, et al. Computational Radiomics System to Decode the Radiographic Phenotype. Cancer Res 2017; 77: e104-7.

19. Haralick RM, Shanmugam K, Dinstein I. Textural Features for Image Classification. IEEE Trans Syst Man Cybern 1973; 3:610-21.

20. Thibault G, Fertil B, Navarro C, Pereira S, Cau P, Lévy N, et al. Shape and texture indexes application to cell nuclei classification. Int J Pattern Recognit Artif Intell 2013; 27: 1357002.

21. Galloway MM. Texture analysis using gray level run lengths. Comput Graph Image Process 1975; 4: 172-9.

22. Amadasun M, King R. Textural features corresponding to textural properties. IEEE Trans Syst Man Cybern 1989; 19: 1264-74.

23. Sun C, Wee WG. Neighboring gray level dependence matrix for texture classification. Comput Vis Graph Image Process 1983; 23: 341-52.

24. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: an update. SIGKDD Explor Newsl 2009; 11: 10-8.

25. Kohavi R, John GH. Wrappers for feature subset selection. Artif Intell 1997; 97: 273-324.

26. Bermejo P, Gamez JA, Puerta JM. Improving incremental wrapper-based subset selection via replacement and early stopping. Int J Pattern Recognit Artif Intell 2011; 25: 605-25.

27. Koçak B, Durmaz EŞ, Ateş E, Kılıçkesmez Ö. () Radiomics with artificial intelligence: a practical guide for beginners. Diagn Interv Radiol 2019; 25: 485-95.

28. Kittler J, Hatef M, Duin RPW, Matas J. On Combining Classifiers. IEEE Trans Pattern Anal Mach Intell 1998; 20: 226-39.

29. Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics 2006; 7: 91.

30. Miller TT. Bone tumors and tumorlike conditions: analysis with conventional radiography. Radiology 2008; 246: 662-74.

31. Gould CF, Ly JQ, Lattin GE Jr, Beall DP, Sutcliffe JB 3rd. Bone tumor mimics: avoiding misdiagnosis. Curr Probl Diagn Radiol 2007; 36: 124-41.

32. Yao J, O'Connor SD, Summers R. Computer aided lytic bone metastasis detection using regular CT images. 2006.

33. O'Connor SD, Yao J, Summers RM. Lytic metastases in thoracolumbar spine: computer-aided detection at CT--preliminary study. Radiology 2007; 242: 811-6.

34. Hammon M, Dankerl P, Tsymbal A, Wels M, Kelm M, May M, et al. Automatic detection of lytic and blastic thoracolumbar spine metastases on computed tomography. Eur Radiol 2013; 23: 1862-70.

35. Wels M, Kelm BM, Tsymbal A, Hammon M, Soza G, Sühling M, et al. Multi-stage osteolytic spinal bone lesion detection from CT data with internal sensitivity control. Proc. SPIE 8315, Medical Imaging 2012: Computer-Aided Diagnosis, 831513 (23 February 2012); https://doi.org/10.1117/12.911169.

36. Roth HR, Yao J, Lu L, Stieger J, Burns JE, Summers RN. Detection of Sclerotic Spine Metastases via Random Aggregation of Deep Convolutional Neural Network Classifications. In: Yao J, Glocker B, Klinder T, Li S, (eds). Recent Advances in Computational Methods and Clinical Applications for Spine Imaging. Springer International Publishing, Cham; 2015; pp. 3-12.

37. Larhmam MA, Mahmoudi S, Drisis S, Benjelloun M. A Texture Analysis Approach for Spine Metastasis Classification in T1 and T2 MRI. In: Rojas I, Ortuño F (eds) Bioinformatics and Biomedical Engineering. Springer International Publishing, Cham; 2018; pp. 198-211.

38. Reischauer C, Patzwahl R, Koh DM, Froehlich JM, Gutzeit A. Texture analysis of apparent diffusion coefficient maps for treatment response assessment in prostate cancer bone metastases-A pilot study. Eur J Radiol 2018; 101: 184-90.

39. Acar E, Leblebici A, Ellidokuz BE, Başbınar Y, Kaya GÇ. Machine learning for differentiating metastatic and completely responded sclerotic bone lesion in prostate cancer: a retrospective radiomics study. Br J Radiol 2019; 92: 20190286.