Radiomic Analysis of Soft Tissues Sarcomas Can Distinguish Intermediate From High-Grade Lesions

Valentina D.A. Corino, PhD,[1]* Eros Montin, PhD,[1] Antonella Messina, MD,[2] Paolo G. Casali, MD,[2,3] Alessandro Gronchi, MD,[2] Alfonso Marchianò, MD,[2] and Luca T. Mainardi, PhD[1]

Purpose: To assess the feasibility of grading soft tissue sarcomas (STSs) using MRI features (radiomics).
Materials and Methods: MRI (echo planar SE, 1.5T) from 19 patients with STSs and a known histological grading, were retrospectively analyzed. The apparent diffusion coefficient (ADC) maps, obtained by diffusion-weighted imaging acquisitions, were analyzed through 65 radiomic features, intensity-based (first order statistics, FOS) and texture (gray level co-occurrence matrix, GLCM; and gray level run length matrix, GLRLM) features. Feature selection (sequential forward floating search) and classification (k-nearest neighbor classifier) were performed to distinguish intermediate- from high-grade STSs. Classification was performed using the three different sub-groups of features separately as well as all the features together. The entire dataset was divided in three subsets: the training, validation and test set, containing, respectively, 60, 30, and 10% of the data.
Results: Intermediate-grade lesions had a higher and less disperse ADC values compared with high-grade ones: most of FOS related to intensity are higher for the intermediate-grade STSs, while FOS related to signal variability were higher in the high grade (e.g., the feature variance is $2.6*10^5 \pm 0.9*10^5$ versus $3.3*10^5 \pm 1.6*10^5$, $P = 0.3$). The GLCM features related to entropy and dissimilarity were higher in the high-grade. When performing classification, the best accuracy is obtained with a maximum of three features for each subgroup, FOS features being those leading to the best classification (validation set: FOS accuracy 0.90 ± 0.11, area under the curve [AUC] 0.85 ± 0.16; test set: FOS accuracy 0.88 ± 0.25, AUC 0.87 ± 0.34).
Conclusion: Good accuracy and AUC could be obtained using only few Radiomic features, belonging to the FOS class.
Level of Evidence: 4
Technical Efficacy: Stage 2

Introduction

Soft tissue sarcomas (STSs) are a rare and heterogeneous group of tumors representing less than 1% of all malignant tumors, with only several tens of-thousands of new diagnoses annually in the United States.[1,2] They pose significant diagnostic and therapeutic challenges.[3]

The pathologic classification of sarcomas is histogenetic.[4] Then, STSs are assigned to a low, intermediate, or high malignancy grade, based on characteristics such as mitotic count, differentiation, and necrosis.[5] According to the histological type and grade of STSs, pre- or postsurgery addition of chemotherapy and/or radiation therapy might be useful.[6,7]

STSs are highly heterogeneous from the spatial point of view.[8] This heterogeneity could provide useful information on tumor aggressiveness and/or its response to treatment. However, it also may suggest that, for example, core needle biopsies can underestimate the malignancy grade. In addition, grading has been a controversial topic in STSs even because the whole group of STSs is considered as a

Drs. Corino and Montin are equal contributors in this study.

*Address reprint requests to: V.D.A.C., Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Via Golgi 39, 20133, Milano, Italy.
E-mail: valentina.corino@polimi.it

From the [1]Department of Electronic, Information, and Bioengineering, Politecnico di Milano, Milan, Italy; [2]Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy; and [3]Oncology and Haematology/Oncology Department, University of Milan, Italy

This is the peer reviewed version of the following article: "Radiomic analysis of soft tissues sarcomas can distinguish intermediate from high-grade lesions", JOURNAL OF MAGNETIC RESONANCE IMAGING, (2018) 47:3, 829-840 which has been published in final form at http://dx.doi.org/10.1002%2Fjmri.25791. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions

TABLE 1. Characteristics of Patients and Tumors

	Patients characteristics		
	All	**Intermediate grade**	**High grade**
No. of patients	19	5	14
Age (years)	56 ± 18 (22-77)	57 ± 25 (22-77)	55 ± 16 (28-75)
Gender (male/female)	6/13	3/2	3/11
Tumor characteristics			
	All	Intermediate grade	High grade
Size (cm)	11.3 ± 3.4	9.4 ± 3.7	11.9 ± 3.1
Location of tumors			
Limb	14	2	12
Torso	5	3	2
Histology			
Leiomyosarcoma	2	1	1
Pleomorphic sarcoma	11	3	8
Synovial sarcoma	2	-	2
Myxofibrosarcoma	2	-	2
MPNST	1	-	1
Chondrosarcoma	1	1	-

single entity, thus underestimating the interplay between grade and histological type.[5] Clearly, a major distinction is made between low-grade STS and the others, but, say, adjuvant chemotherapy is likely to be especially active when malignancy grade is high rather than intermediate, with special regard to some histological types.

Diffusion-weighted imaging (DWI) MRI can capture changes at the cellular level thanks to differences in movement of water protons in the different tissue regions. The apparent diffusion coefficient (ADC) map, derived from different diffusion-weighted MRIs, has been shown to be predictive of treatment response.[9,10] Moreover, advantages of the ADC maps are that they have been shown to be powerful biomarker for assessing tumor cellularity[11–13] and to correlate with malignancy grading of STSs,[14] even when different vendor scanners are used.

Radiomics has applied to oncology recently. Radiomics extracts a large number of image characteristics, or features, in a noninvasive way.[15] The assumption is that image features quantify crucial information regarding the entire tumor phenotype and thus they can highlight intra-tumor heterogeneity.[15] Many studies reported that this heterogeneity could have profound implications on tumor prognosis.[16,17]

The aim of this work was to assess the capability of Radiomic features to characterize and/or differentiate STSs of different malignancy grades, paying attention not just to the distinction between low- and high-grade STSs, but also to the intermediate-grade subset.

Material and Methods

Study Population

Nineteen arbitrarily selected patients with STSs were retrospectively analyzed. The entire dataset was divided in three groups: (i) the training set containing the data used to train the models (60% of the data); (ii) the validation set containing the data used to validate the model and to choose the best one (30% of the data); and (iii) the test set containing the data used to test the model and examine its behavior with never-seen data (10% of the data), see the Statistical Analysis section. They had a histological diagnosis of STS of intermediate (5 patients) or high (14 patients) malignancy grade according to the FNCLCC (French Fédération Nationale des Centres de Lutte Contre le Cancer) system.[18] The FNCLCC system is based on tumor differentiation, mitotic rate, and amount of tumor necrosis. A score is attributed independently to each parameter, and the grade is obtained by adding the three attributed scores (Grading of Soft Tissue Sarcomas). Patient and tumor characteristics are shown in Table 1; age and gender were not statistically different in the two groups. All patients underwent to a DWI MRI acquisition before starting the treatment. The study was approved by the ethical committee of Fondazione IRCCS (Istituto Nazionale dei Tumori of Milan, Italy). At the time of the acquisition, all patients filled out a generic consent to the use of data, including

TABLE 2. MRI Sequence Parameters by MRI Scanner

Sequence parameter	Siemens Avanto MRI (n = 13)	Philips Achieva (n = 6)
Sequence	Echo planar SE	Single-shot echo planar SE
Matrix (pixels)	192 × 192	255 × 255
Resolution (voxel/mm)	1.98 × 1.98	1.37 × 1.37
Field of View (mm)	380 × 380	350 × 350
TR (msec)	5400	7410
TE (msec)	78	63
Slice thikness (mm)	4 (no gap)	5 (no gap)
NEX	4	3

NEX = number of excitations.

the acquired images, and biological material for research. All patients' data were anonymized before the analysis.

Image Acquisition

DWI MRI images were acquired using Achieva 1.5T system (Philips Medical system Achieva, Nederlands) (6 patients) or a Magnetom Avanto 1.5T system (Siemens Medical Solutions, Erlangen, Germany) (13 patients), both with a body-matrix coil and spine array coil for signal reception. The data were acquired axially by means of echo planar imaging, the sequences' parameters (for both equipment) are reported in Table 2. DWI images were acquired using four b-values (namely, 50, 400, 800, and 1000 s/mm^2)

Preprocessing of MRI Images

ADC maps creation: For each acquisition, the ADC was computed as the slope of the linear regression of the logarithm of the DWI exponential signal decay on the four b-values.[8] The calculation was performed pixel-wise using ITK 4.8.[9] An expert radiologist manually segmented the lesion (region of interest, ROI). The segmentation was performed using three-dimensional (3D) slicer.[19] Contouring of the ROI was performed on images acquired with the b-value 50 s/mm^2, on which the anatomical details are maintained and the heterogeneity of the tumor is more visible.[20]

Radiomic Features Extraction

We assessed 65 radiomic features, pertaining to two main classes: (i) intensity-based features and (ii) texture features. The list of features is reported in Table 3. Features belonging to the intensity-based (first order statistics, FOS) group were computed on the ROI volume and/or on the intensity histogram, evaluated between 0 and the maximum of the image (mm^2/s) using 32 bins. Texture features were based on the gray level co-occurrence matrix (GLCM)[21] and the gray level run length matrix (GLRLM).[22] Before matrix computation, the discretization of gray levels was reduced to 32 to avoid sparseness of matrix. For a given direction α, the GLCM is a $N \times N$ matrix (where N is the number of bins used to discretize the gray values ($N = 32$ in this study)), whose *(i, j)* element is the counting of pixels of gray intensity level *i* which

TABLE 3. Features Used in the Analysis

FOS features	Energy, Kurtosis, Mad, Max, Mean, Median, Min, Range, RMS, Skewness, SD, Variance, Quantile 0.01, Quantile 0.1, Quantile 0.2, Quantile 0.3, Quantile 0.4, Quantile 0.5, Quantile 0.6, Quantile 0.7, Quantile 0.8, Quantile 0.9, Quantile 0.99, Histogram Entropy, Histogram Kurtosis, Histogram Mad, Histogram Max, Histogram Mean, Histogram Median, Histogram Min, Histogram Range, Histogram RMS, Histogram Skewness, Histogram SD, Histogram Variance, Histogram Uniformity, Histogram Total Frequency,
Texture features – GLCM	Autocorrelation, Cluster Prominence, Cluster Shade, Cluster Tendency, Contrast, Correlation, Difference Entropy, Dissimilarity, Energy, Entropy, Homogeneity, Homogeneity2, IMOC1, IMOC2, Inverse Difference moment, Inverse Difference moment2, Inverse Variance, Max Probability, Sum Average, Sum Entropy, Inertia
Texture features – GLRLM	Short Run Emphasis, Long Run Emphasis, Gray Level Non Uniformity, Run Length Non Uniformity, Run Percentage, Low Gray Level Run Emphasis, high Gray Level Run Emphasis, Short Run Low Gray Level Emphasis, Short Run High Gray Level Emphasis, Long Run Low Gray Level Emphasis, Long Run High Gray Level Emphasis

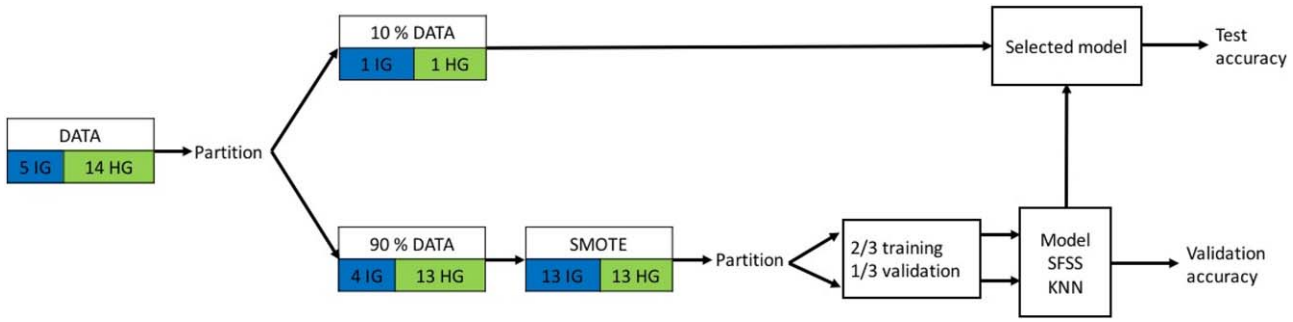

FIGURE 1: Schematic representation of the overall method. As first step, 10% of the data (one patient of intermediate and one high grade group) are removed to be used as test group. The remaining 90% of the data enter the oversampling (obtained using SMOTE) and then the feature selection and classification algorithms, using 2/3 of the data as training and 1/3 as validation group. The classification algorithm gives as output the optimal model to be used on the test group.

are adjacent (within a distance $\rho = 1$, in our case) to pixels of the gray intensity level *j*. We computed GLCM for the 26 directions in the three dimensions, obtaining globally 26 matrices. The GLRLM is a NxN matrix whose *(i, j)* element counts the number of runs of pixels of gray level *i* (run step 1) and run length *j* in a given direction. As before, we computed GLRLM for the 26 directions in the three dimensions, obtaining globally 26 matrices. On each matrix (GLCMs or GLRLM), the texture features of Table 3 were computed and the results averaged on all angles, thus obtaining two sets of features, one for the GLCM and one form the GLRLM. All the algorithms were implemented in Insight Segmentation and Registration Toolkit (ITK 4.8).[9,23]

Statistical Analysis

Based on the computed radiomic features, we aimed at distinguishing intermediate grade STS from high grade STS. First, we observed that our dataset is imbalanced as the classes (intermediate and high-grade STS) are not equally represented, this imbalance may produce classifiers with poor predictive accuracy for the minority class, tending to classify most new samples in the majority class. As the classification accuracy would be influenced by the imbalanced classes, a way to overcome this problem is to re-sample the original dataset, by oversampling the minority class.[24] To this purpose, the Synthetic Minority Over-sampling Technique (SMOTE) is used. In SMOTE, the minority class is over-sampled by taking each minority class sample and introducing synthetic examples along the line segments joining any/all of the Q nearest neighbors in the minority class.[25] Briefly, for each sample of the minority class, the Q nearest neighbors of the same class are found identified and one of them is chosen randomly. The new synthetic sample lies on a random point of the line joining the two original samples. In this study, Q was chosen equal to 3. After SMOTE application, both classes have the same number of instances to be classified.

Figure 1 shows a schematic representation of the method used for classification. As first step, a test set was created, containing 10% of the data (one patient of intermediate and one high grade group), used to test the classifier and examine its behavior with never-seen data. The remaining 90% of the data underwent the oversampling by SMOTE, thus both classes contain 13 patients each. Then the feature selection and classification algorithms were run, using 2/3 of the data as training and 1/3 as validation group. The classification algorithm gave as output the optimal model that was used to classify elements of the test group.

In particular, for feature selection, a sequential forward floating search (SFFS) algorithm[26] was used to identify the best subset of features differentiating the two STS grades. Briefly, starting from the empty set of features, the feature x_i that maximizes the objective function (Y_k+X_i) when combined with the features Y_k that have already been selected, is added. After this forward step, SFFS performs backward steps as long as the objective function increases. A backward step consists in removing from Y_k the feature that makes the objective function increase. A schematic representation of the algorithm is shown in Figure 2.

After the feature selection step, the classification was performed by using the k-nearest neighbor classifier with k equal to 3.[27] In the training phase 2/3 of the 90% of the data were used to build the model, whereas in the validation phase the remaining 1/3 of the data were classified according to the model generated in the training phase.[28] Leave-p-out cross-validation was performed with 100 bootstrap repetitions, i.e., all the above steps are repeated 100 times, randomizing images, allowing images from the two scanners to be part of the groups. Performance metrics were averaged over the 100 repetitions.

In this study, we tested the different classes of features alone and in combination, i.e., we performed the features selection and classification using only intensity-based features, GLCM, GLRLM, and their combinations.

The difference in feature values between intermediate- and high-grade STSs was assessed by the Wilcoxson test. Spearman correlation coefficient was computed between the feature values and the tumor grade. Accuracy of the 100 repetitions for the model using N features was compared with the accuracy of the 100 repetitions for the model using N+1 features, using an unpaired t-test.

A P-value < 0.05 was considered statistically significant.

Results

Features Results

Figure 3a shows two ADC maps of an intermediate- (top panel) and high- (bottom panel) grade STSs. It is apparent that the intermediate-grade lesions have a higher and less

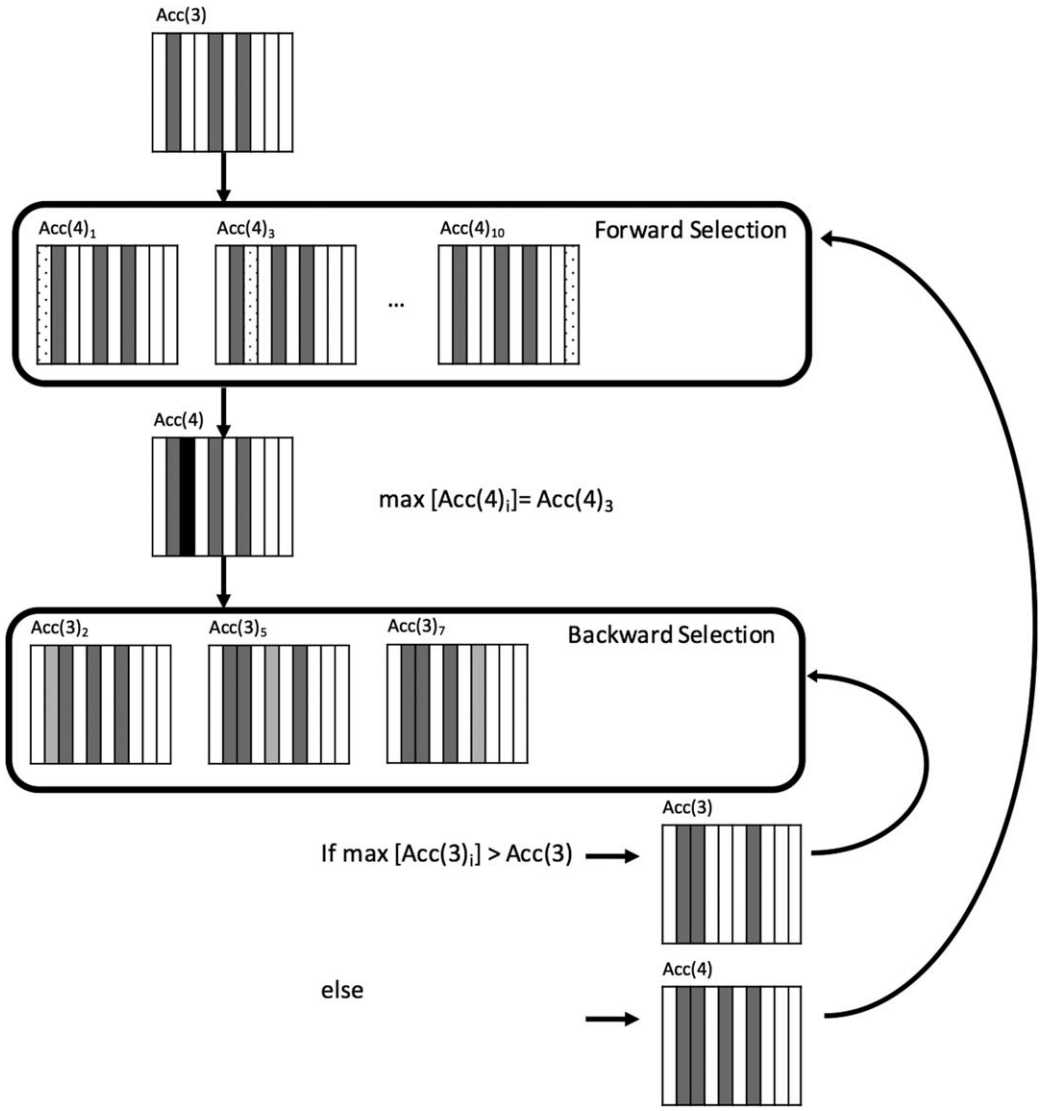

FIGURE 2: Schematic representation of the SFFS algorithm, as an example starting with three features already selected. Each rectangle represents the whole set of features, and each vertical line a feature. The top rectangle represents the current set of chosen features (three gray lines) along with all the others (white lines), these three features make an accuracy equal to Acc(3). The first block is the Forward Selection: each not yet selected feature is added (dotted line) and the corresponding accuracy computed ($Acc(4)_i$). The feature producing the maximum accuracy is finally added (black line). The following block is the Backward Selection: each of the selected features (light gray line), but the last one, is removed from the selected features set and the corresponding accuracy computed. If the maximum accuracy $Acc(3)_i$ is bigger than the previous Acc(3), the feature is removed from the selected features set. If a feature is removed, the Backward Selection is repeated, otherwise the following step is the Forward Selection.

disperse ADC values in comparison to high-grade ones, as also shown by the corresponding histograms of Figure 3b.

Figure 4 shows radar plots of the features in the two groups of patients. The Radar plot assumes only two values (0 and 1) to emphasize the differences between intermediate- and high-grade tumors (for each feature, the radar plot is equal to 1 for the group having the higher value). Most of the intensity-based features are higher for the intermediate-grade STSs (Fig. 4a), while the features related to the signal variability (like SD and Variance) are higher in the high-grade (Fig. 3a). Almost all the histogram-based features (Fig. 4b) are higher for the intermediate grade STSs. The GLCM features (Fig. 4c) related to entropy and dissimilarity were higher in the high grade. Regarding the GLRLM texture features (Fig. 4d), the ones related to the high gray run were higher in the intermediate grade STSs, whereas those related to the low gray run were higher in the high grade STSs.

All differences, analyzed one by one, were not statistically significant between intermediate and high-grade STSs.

Classification Results

The distinction between the two types of STSs was performed using the three different sub-groups of features separately (FOS, GLCM, and GLRLM) as well as all the features together. Figure 5 shows the accuracy obtained

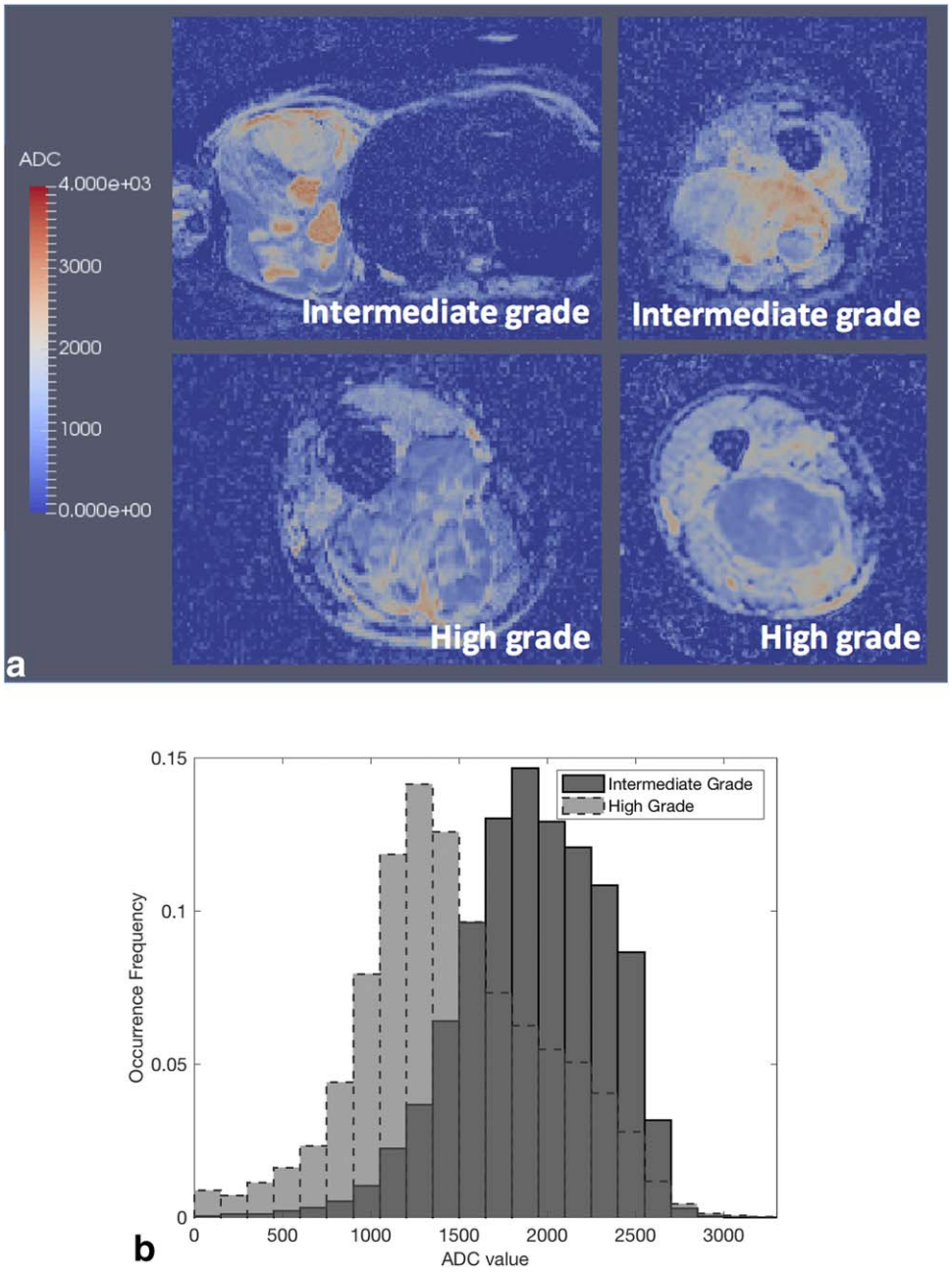

FIGURE 3: Four ADC maps for two intermediate- and two high-grade STSs (a), and two intensity histograms corresponding to an intermediate- (gray) and high-grade (light gray) STSs (b).

using an increasing number of features, as selected by the SFFS algorithm, for each sub-group of features for the validation and test sets. In Figure 5, the *n*-th dot represents the mean accuracy (over the 100 repetitions) obtained using *n* features. In each sub-group the mean accuracy increases when the number of features in increased from one to two. Adding more features, the mean accuracy further increases for the FOS and GLCM groups only. Then, when new parameters are added, it slightly decreases.

Table 4 shows the best selected features of each sub-group using up to six parameters with the corresponding mean accuracy and mean AUC. The best accuracy is obtained with a maximum of three features for each subgroup, representing the best classification models for our problem. The model obtained using the FOS feature group is the one leading to the best classification. Using two or three parameters of the FOS features, namely STD, Histogram Uniformity, Histogram Quantile 0.3, leads to the best compromise balancing the number of features and the accuracy values.

Finally, Figure 6 shows the accuracy obtained using all the features for the validation and test sets. Beyond a slight improvement in accuracy by adding features, the accuracy is almost constant and when using all the features the mean accuracy is much lower than that using a few of them.

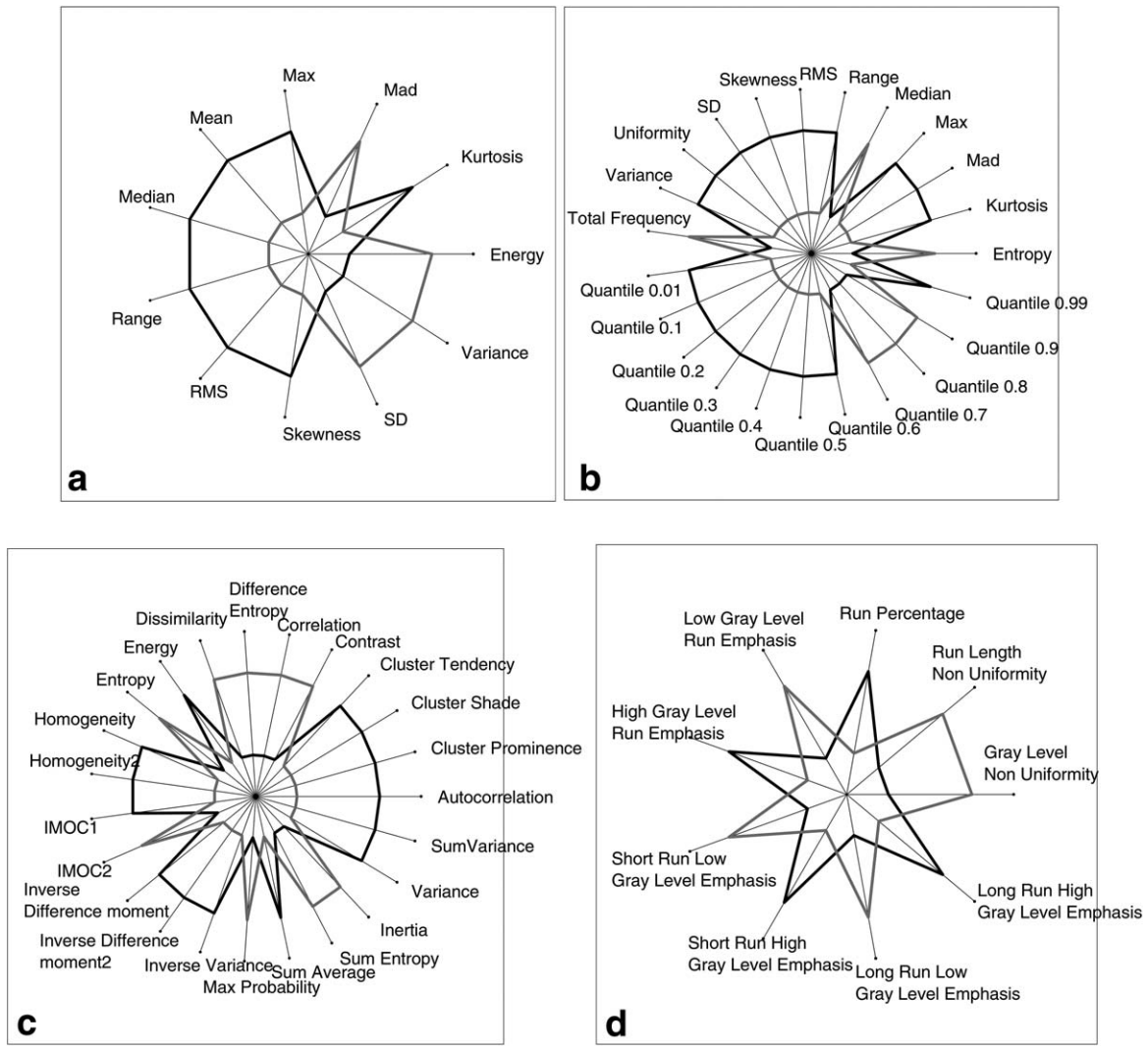

FIGURE 4: Radar charts for signal intensity-based features (a), histogram-based features (b), GLCM features (c), and GLRLM features (d). Each spoke represents one of the features, the data length of a spoke is proportional to the magnitude of the variable for the data point relative to the maximum magnitude of the variable across all data points. The spokes are normalized so that the difference between intermediate (black line) and high (gr y line) grades STSs is emphasized.

Discussion and Conclusions

In this study, we investigated whether features derived from ADC maps of patients with STSs could be used to differentiate intermediate- versus high-grade lesions. The main result was that a high accuracy and AUC can be obtained by considering only few features. Although we reported the average accuracy using all the possible features for the sake of completeness, the classifications performed with a high

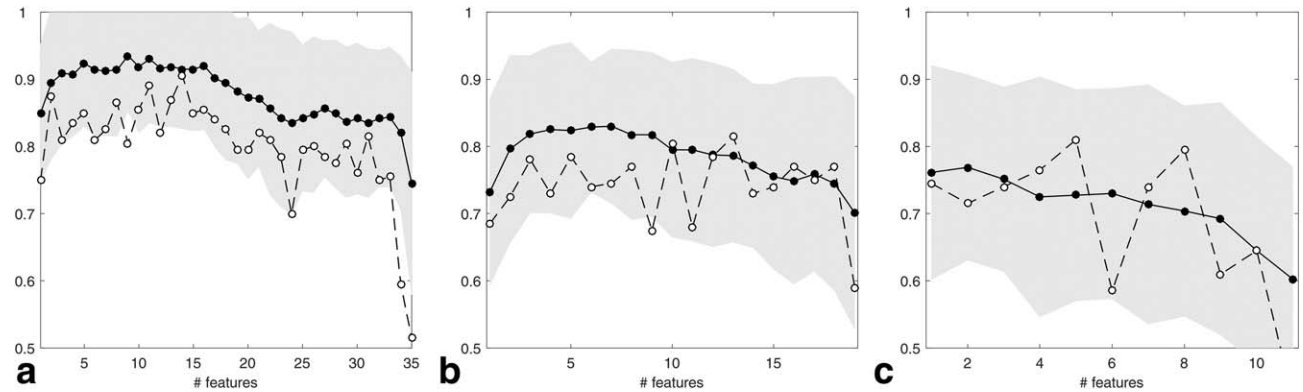

FIGURE 5: Mean ± the standard deviation of the accuracy of the validation set (black line and the grey area, respectively) with superimposed the average accuracy of the test set (dashed line) for FOS features (a), GLCM features (b), and GLRLM features (c).

TABLE 4. Best Selected Features of Each Sub-group

FOS features

Selected features	Accuracy Validation	AUC Validation	Accuracy Test	AUC Test
STD	0.85 ± 0.10	0.76 ± 0.19	0.75 ± 0.31	0.71 ± 0.46
STD, histogram quantile 0.1	0.90 ± 0.11[a]	0.85 ± 0.16[a]	0.88 ± 0.25[a]	0.87 ± 0.34[a]
STD, histogram quantile 0.1, histogram uniformity	0.91 ± 0.10	0.86 ± 0.17	0.81 ± 0.26	0.81 ± 0.39
STD, histogram quantile 0.1, histogram uniformity, histogram Variance	0.91 ± 0.09	0.87 ± 0.14	0.84 ± 0.25	0.85 ± 0.36
STD, histogram quantile 0.1, Mad, histogram STD, histogram skewness	0.92 ± 0.09	0.90 ± 0.14	0.85 ± 0.26	0.87 ± 0.34
STD, histogram quantile 0.1, Mad, histogram STD, histogram skewness, histogram kurtosis	0.91 ± 0.10	0.87 ± 0.14	0.81 ± 0.26	0.81 ± 0.39

GLCM features

Selected features	Accuracy Validation	AUC Validation	Accuracy Test	AUC Test
Entropy	0.73 ± 0.14	0.62 ± 0.17	0.69 ± 0.33	0.60 ± 0.49
Sum entropy, difference entropy	0.80 ± 0.14[a]	0.70 ± 0.17[a]	0.73 ± 0.34	0.67 ± 0.47
Sum entropy, difference entropy, energy	0.82 ± 0.12	0.71 ± 0.17	0.78 ± 0.30	0.78 ± 0.42
Sum entropy, difference entropy, energy, Homogeneity2	0.83 ± 0.12	0.73 ± 0.17	0.73 ± 0.34	0.68 ± 0.47
Sum entropy, difference entropy, energy, Homogeneity, IMOC2	0.82 ± 0.13	0.74 ± 0.19	0.79 ± 0.28	0.75 ± 0.43
Sum entropy, difference entropy, energy, Homogeneity, IMOC2, IMOC1	0.83 ± 0.10	0.73 ± 0.21[a]	0.74 ± 0.33	0.66 ± 0.48

GLRLM features

Selected features	Accuracy Validation	AUC Validation	Accuracy Test	AUC Test
Long run low gray level emphasis	0.76 ± 0.16	0.70 ± 0.21	0.75 ± 0.32	0.74 ± 0.44
Long run low gray level emphasis, short run low gray level emphasis	0.77 ± 0.14	0.72 ± 0.19	0.72 ± 0.32	0.68 ± 0.47
Long run low gray level emphasis, short run low gray level emphasis, low gray level run emphasis	0.75 ± 0.14	0.71 ± 0.18	0.74 ± 0.29	0.70 ± 0.46
Long run low gray level emphasis, short run emphasis, low gray level run emphasis, short run high gray level emphasis	0.73 ± 0.18	0.61 ± 0.21	0.77 ± 0.31	0.67 ± 0.47

TABLE 4: Continued

GLRLM features

Selected features	Accuracy Validation	AUC Validation	Accuracy Test	AUC Test
Long run low gray level emphasis, low gray level run emphasis, short run emphasis, short run high gray level emphasis, run percentage	0.73 ± 0.16	0.61 ± 0.18	0.81 ± 0.27	0.77 ± 0.42
Long run low gray level emphasis, low gray level run emphasis, short run high gray level emphasis, run percentage, long run emphasis 16, short run low gray level emphasis	0.73 ± 0.16	0.60 ± 0.48	0.59 ± 0.33	0.44 ± 0.50

All features

Selected features	Accuracy Validation	AUC Validation	Accuracy Test	AUC Test
STD	0.85 ± 0.10	0.76 ± 0.17	0.78 ± 0.30	0.75 ± 0.43
STD, long run high gray level emphasis	0.90 ± 0.10[a]	0.85 ± 0.17[a]	0.79 ± 0.30	0.65 ± 0.48
STD, long run high gray level emphasis, Histogram Median	0.93 ± 0.09[a]	0.89 ± 0.14	0.86 ± 0.24[a]	0.85 ± 0.36
STD, long run high gray level emphasis, histogram Median, Homogeneity	0.93 ± 0.08	0.90 ± 0.15	0.84 ± 0.27	0.82 ± 0.39
STD, long run high gray level emphasis, histogram median, homogeneity, difference entropy	0.92 ± 0.09	0.88 ± 0.14	0.84 ± 0.26	0.84 ± 0.37
STD, long run high gray level emphasis, histogram median, homogeneity, difference entropy, dissimilarity	0.92 ± 0.09	0.88 ± 0.14	0.88 ± 0.23	0.88 ± 0.33

[a]$p < 0.05$ comparison between a set with n features and the set with $n+1$ features.

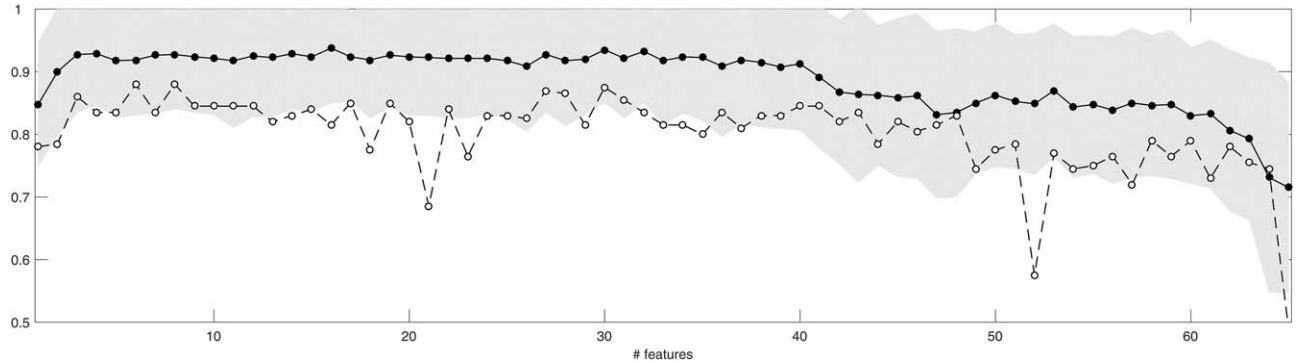

FIGURE 6: Mean ± the standard deviation of the accuracy of the validation set (black line and the gray area, respectively) with superimposed the average accuracy of the test set (dashed line) using all the features.

number of features suffers from the curse of dimensionality, i.e., increasing the number of features results in a decrease of performance.[29] The best classification models (balancing accuracy, AUC, and number of features) for our problem were those obtained using three features. The best average accuracy we obtained was not excellent, but these preliminary results are encouraging as with such a small study population, the accuracy is good, thus we expect that with larger population the accuracy may increase. Features belonging to the FOS class are the best performing in terms of accuracy and AUC. In the FOS class, features describing the histogram distribution, i.e., the gray levels distribution, are those first selected in the classification. Accordingly, the histogram of intermediate-grade tumors is narrower than for high-grade tumors. From a functional point of view, high grade is more heterogeneous and this is what the histogram distribution highlights and thus may be reason for a better performance of the FOS features belonging to the class of histogram descriptors.

Grading of STSs is widely held as their main prognostic factor. The Fédération Nationale des Centres de Lutte Contre le Cancer (FNCLCC) grading system is often used, taking into account the mitotic rate, the degree of necrosis and tumor differentiation.[18] It applies to several histologies, but not to all of them, and indeed some histologies automatically correspond to a given grading, irrespective of those characteristics.[30,31] Of course, the value of grading may be limited when the diagnosis is achieved through core needle biopsies, because the tumor may be heterogeneous and grading can be underestimated.[5] In other words, some STSs whose biopsy points to an intermediate malignancy grade remain uncertain as to their actual potential of aggressiveness, because higher-grade areas may co-exist. Indeed, appropriate grading on biopsy may be crucial for decisions about neoadjuvant treatment, and data were recently provided that neoadjuvant chemotherapy can provide significant benefit to a subset of high-risk (thus also high-grade) STSs patients.[7] Sometimes, the obvious radiologic features of lesions are taken into account. For example, one could factor the degree of macroscopic necrosis which is visible radiologically, even if the histological necrosis is low. Thus, there is room for radiological improvement to assist in assessing the actual malignancy grade when a biopsy points to an intermediate-grade STS. In this respect, radiomic analysis could be crucial: many features describing the tumor are computed on the 3D-volumes obtained from MRI. Radiomic analysis is a noninvasive, fast, low-cost and reproducible way of investigating phenotypic information.

In this study, we computed the radiomic features on ADC maps only for two main reasons. First, the ADC maps have been shown to assess tumor cellularity even when different scanners are used,[11] provided that the same range of b-values and the same field strength are chosen.[12,13] This property of the ADC map suggests that they are useful in multicenter studies, where scanners are usually different. Moreover, it has been shown that repetition time (TR) and echo time (TE) values may be chosen so that the ADC value is not affected. A selection of TR as short as the longest T1 relaxation time of the tissue of interest may result in overestimation of ADC values. However, it has been shown that choosing a TR approximately five-times longer than the tissue T1 relaxation time solves the problem.[32] This recommendation, in our study, is translated into a TR value larger than 5000 ms[33] and both scanners satisfied this condition. Selection of TE has a small effect on ADC maps and minimum TE selection has been recommended for DWI protocol, being TE values lower than 100 ms sufficient for not affecting the ADC computation.[32] The second reason is that the study population was small and computing the features on other N images would have created several features N-time larger than the present number. Several features hundreds of time bigger than the number of patients may create problems in classification.

A limitation of the study is the imbalanced dataset, i.e., the classes are not approximately equally represented and imbalanced data on minority class and high dimensionality problem may cause a misclassification. As the performance of machine learning algorithms for classification is

typically evaluated using predictive accuracy, imbalanced data make the assessment of accuracy not appropriate.[25] We overcame this problem by oversampling the minority class, i.e., the intermediate grade class, using SMOTE. SMOTE has been shown to be useful when using kNN classifiers with feature selection to reduce the number of variables.[34] We performed feature selection and classification without SMOTE, but results were poorer: the average accuracy was significantly lower than when using SMOTE for all the number of selected features. This solution proved to be satisfactory, leading to an average high accuracy using only two FOS features. When running the classification algorithm without SMOTE, the results were less accurate: the maximum average accuracy without performing oversampling was obtained with a higher number of features and was significantly lower.

A second limitation of this exploratory study is that the number of patients was low and the selection process essentially arbitrary. However, we believe that these data are suggestive enough as to encourage larger clinical studies on the value of radiomics to grade a subset of soft tissue sarcoma patients, allowing better decisions as to the indication to neo-adjuvant chemotherapy. Moreover, these preliminary results lay the groundwork for future studies where radiomic features may be used for grading or classification of other clinical characteristics of interest. In the end, future studies should compare radiomics with what standard radiology can allow as of today. Moreover, future studies will be needed to evaluate radiomic biomarkers in independent and prospective validation cohorts with large sample sizes also to assess feature repeatability.

In conclusion, these preliminary results show that good accuracy and AUC could be obtained using only few Radiomic features, when grading STSs using MRI features.

Acknowledgments

Contract grant sponsor: BD2Decide Project of the Horizon 2020 Research and Innovation Program; contract grant number: 689715

The methods used in this study have been developed for the BD2Decide Project of the Horizon 2020 research and innovation program that has been inspiring to this work.

References

1. Sherman KL, Wayne JD, Chung J, et al. Assessment of multimodality therapy use for extremity sarcoma in the United States. J Surg Oncol 2014;109:395–404.
2. Siegel R, Naishadham D, Jemal A. Cancer statistics, 2012. CA Cancer J Clin 2012;62:10–29.
3. Demetri GD, Baker LH, Beech D, et al. Soft tissue sarcoma: clinical practice guidelines in oncology. JNCCN J Natl Compr Cancer Netw 2005;3:158–194.
4. Li N, Yang R, Zhang W, Dorfman H, Rao P, Gorlick R. Genetically transforming human mesenchymal stem cells to sarcomas: changes in cellular phenotype and multilineage differentiation potential. Cancer 2009;115:4795–4806.
5. Coindre J-M. Grading of soft tissue sarcomas: review and update. Arch Pathol Lab Med 2006;130:1448–1453.
6. Wang X, Jacobs MA, Fayad L. Therapeutic response in musculoskeletal soft tissue sarcomas: evaluation by MRI. NMR Biomed 2011;24:750–763.
7. Gronchi A, Ferrari S, Quagliuolo V, et al. Histotype-tailored neoadjuvant chemotherapy versus standard chemotherapy in patients with high-risk soft-tissue sarcomas (ISG-STS 1001): an international, open-label, randomised, controlled, phase 3, multicentre trial. Lancet Oncol 2017;18:812–822.
8. Eary JF, O'Sullivan F, O'Sullivan J, Conrad EU. Spatial heterogeneity in sarcoma 18F-FDG uptake as a predictor of patient outcome. J Nucl Med 2008;49:1973–1979.
9. Sun YS, Zhang XP, Tang L, et al. Locally advanced rectal carcinoma treated with preoperative chemotherapy and radiation therapy: preliminary analysis of diffusion-weighted MR imaging for early detection of tumor histopathologic downstaging. Radiology 2009;254:170–178.
10. Padhani AR, Liu G, Koh DM, et al. Diffusion-weighted magnetic resonance imaging as a cancer biomarker: consensus and recommendations. Neoplasia 2009;11:102–125.
11. Jafar MM, Parsai A, Miquel ME. Diffusion-weighted magnetic resonance imaging in cancer: reported apparent diffusion coefficients, in-vitro and in-vivo reproducibility. World J Radiol 2016;8:21–49.
12. Belli G, Busoni S, Ciccarone A, et al. Quality assurance multicenter comparison of different MR scanners for quantitative diffusion-weighted imaging. J Magn Reson Imaging 2016;43:213–219.
13. Ye XH, Gao JY, Yang ZH, Liu Y. Apparent diffusion coefficient reproducibility of the pancreas measured at different MR scanners using diffusion-weighted imaging. J Magn Reson Imaging 2014;40:1375–1381.
14. Fisher R, Pusztai L, Swanton C. Cancer heterogeneity: implications for targeted therapeutics. Br J Cancer 2013;108:479–485.
15. Parmar C, Leijenaar RT, Grossmann P, et al. Radiomic feature clusters and prognostic signatures specific for lung and head & neck cancer. Sci Rep 2015;5:1–10.
16. Gronchi A, Stacchiotti S, Verderio P, et al. Short, full-dose adjuvant chemotherapy in high-risk adult soft tissue sarcomas: a randomized clinical trial from the Italian Sarcoma Group and the Spanish Sarcoma Group. J Clin Oncol 2012;30:850–856.
17. Fisher R, Pusztai L, Swanton C. Cancer heterogeneity: implications for targeted therapeutics. Br J Cancer 2013;108:479–485.
18. Fisher SM, Joodi R, Madhuranthakam AJ, Öz OK, Sharma R, Chhabra A. Current utilities of imaging in grading musculoskeletal soft tissue sarcomas. Eur J Radiol 2016;85:1336–1344.
19. Fedorov A, Beichel R, Kalpathy-Cramer J, et al. 3D Slicer as an image computing platform for the Quantitative Imaging Network. Magn Reson Imaging 2012;30:1323–1341.
20. Chawla PH, Kim S, Wang S, Poptani H. Diffusion-weighted imaging in head and neck cancers. Future Oncol 2009;5:959–975.
21. Haralick RM. Statistical and structural approaches to texture. Proc IEEE 1979;67:786–804.
22. Tang X. Texture information in run-length matrices. IEEE Trans Image Process 1998;7:1602–1609.
23. Yoo TS. Insight into images: principles and practice for segmentation, registration, and image analysis. Boca Raton, FL: CRC Press; 2004.
24. Kubat M, Matwin S. Addressing the curse of imbalanced training sets: one sided selection. Proc Int Conf Mach Learn 1997;97:179–186.
25. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE. J Artif Intell Res 2002;16:321–357.
26. Pearl J. Heuristics: intelligent search strategies for computer problem solving. Boston, MA: Addison-Wesley Longman Publishing Co. Inc; 1984.

27. Kononenko I. Machine learning for medical diagnosis: history, state of the art and perspective. Artif Intell Med 2001;23:89–109.

28. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. Int Jt Conf Artif Intell 1995;14:1137–1143.

29. Hughes GF. On the mean accuracy of statistical pattern recognizers on the accuracy of statistical pattern recognizers. IEEE Trans Inf Theory 1968;30:55–63.

30. Presant CA, Russell WO, Alexander RW, Fu YS. Soft-tissue and bone sarcoma histopathology peer review: the frequency of disagreement in diagnosis and the need for second pathology opinions. The Southeastern Cancer Study Group experience. J Clin Oncol 1986;4:1658–1661.

31. Harris M, Hartley AL, Blair V, et al. Sarcomas in North-West England. 1. Histopathological peer-review. Br J Cancer 1991;64:315–320.

32. Celik A. Effect of imaging parameters on the accuracy of apparent diffusion coefficient and optimization strategies. Diagn Interv Radiol 2016;22:101–107.

33. Stanisz GJ, Odrobina EE, Pun J, et al. T1, T2 relaxation and magnetization transfer in tissue at 3T. Magn Reson Med 2005;54:507–512.

34. Blagus R, Lusa L. SMOTE for high-dimensional class-imbalanced data. BMC Bioinformatics 2013;14:106.