European Journal of Radiology 131 (2020) 109266

Contents lists available at ScienceDirect

European Journal of Radiology

journal homepage: www.elsevier.com/locate/ejrad

Differential diagnosis and mutation stratification of desmoid-type fibromatosis on MRI using radiomics

Milea J.M. Timbergen [a,b,*,1], Martijn P.A. Starmans [c,d,1], Guillaume A. Padmos [d], Dirk J. Grünhagen [a], Geert J.L.H. van Leenders [e], D.F. Hanff [d], Cornelis Verhoef [a], Wiro J. Niessen [c,d,f], Stefan Sleijfer [b], Stefan Klein [c,d], Jacob J. Visser [d]

[a] Department of Surgical Oncology, Erasmus MC Cancer Institute Rotterdam, the Netherlands
[b] Department of Medical Oncology, Erasmus MC Cancer Institute Rotterdam, the Netherlands
[c] Department of Medical Informatics, Erasmus MC, Rotterdam, the Netherlands
[d] Department of Radiology and Nuclear Medicine, Erasmus MC, Rotterdam, the Netherlands
[e] Department of Pathology, Erasmus MC, Rotterdam, the Netherlands
[f] Faculty of Applied Sciences, Delft University of Technology, Delft, the Netherlands

ARTICLE INFO

Keywords:
Fibromatosis
Aggressive
Magnetic resonance imaging
Beta catenin
Machine learning
Radiomics

ABSTRACT

Purpose: Diagnosing desmoid-type fibromatosis (DTF) requires an invasive tissue biopsy with β-catenin staining and *CTNNB1* mutational analysis, and is challenging due to its rarity. The aim of this study was to evaluate radiomics for distinguishing DTF from soft tissue sarcomas (STS), and in DTF, for predicting the *CTNNB1* mutation types.

Methods: Patients with histologically confirmed extremity STS (non-DTF) or DTF and at least a pretreatment T1-weighted (T1w) MRI scan were retrospectively included. Tumors were semi-automatically annotated on the T1w scans, from which 411 features were extracted. Prediction models were created using a combination of various machine learning approaches. Evaluation was performed through a 100x random-split cross-validation. The model for DTF vs. non-DTF was compared to classification by two radiologists on a location matched subset.

Results: The data included 203 patients (72 DTF, 131 STS). The T1w radiomics model showed a mean AUC of 0.79 on the full dataset. Addition of T2w or T1w post-contrast scans did not improve the performance. On the location matched cohort, the T1w model had a mean AUC of 0.88 while the radiologists had an AUC of 0.80 and 0.88, respectively. For the prediction of the *CTNNB1* mutation types (S45 F, T41A and wild-type), the T1w model showed an AUC of 0.61, 0.56, and 0.74.

Conclusions: Our radiomics model was able to distinguish DTF from STS with high accuracy similar to two radiologists, but was not able to predict the *CTNNB1* mutation status.

1. Introduction

Sporadic desmoid-type fibromatosis (DTF) is a rare borderline, soft tissue tumor arising in musculoaponeurotic structures [1]. Worldwide epidemiological data is lacking, but population studies in Scandinavia and the Netherlands show a low incidence of 2.4–5.4 cases per million per year [2,3]. Early recognition and diagnosis of DTF is therefore challenging.

On MRI, DTF can display a wide variety of enhancement patterns [4]. DTF has imaging characteristics that are often associated with soft tissue sarcomas (STS), such as crossing fascial boundaries, an invasive growth pattern, little central necrosis, mildly hyperintense on T1-weighted (T1w) MRI, and hyperintense and heterogeneous on T2-weighted (T2w) MRI with hypointense bands [5]. Hence, the distinction between DTF and STS, i.e. non-DTF, can be difficult. An invasive tissue biopsy, with additional immunohistochemical staining for β-catenin and

* Corresponding author at: Department of Surgical Oncology and Department of Medical Oncology, Dr. Molewaterplein 40, 3015 GD, Rotterdam, the Netherlands.
E-mail addresses: m.timbergen@erasmusmc.nl (M.J.M. Timbergen), m.starmans@erasmusmc.nl (M.P.A. Starmans), guill5@hotmail.com (G.A. Padmos), d.grunhagen@erasmusmc.nl (D.J. Grünhagen), g.vanleenders@erasmusmc.nl (G.J.L.H. van Leenders), d.hanff@erasmusmc.nl (D.F. Hanff), c.verhoef@erasmusmc.nl (C. Verhoef), w.niessen@erasmusmc.nl (W.J. Niessen), s.sleijfer@erasmusmc.nl (S. Sleijfer), s.klein@erasmusmc.nl (S. Klein), j.j.visser@erasmusmc.nl (J.J. Visser).
[1] Equal contributions.

https://doi.org/10.1016/j.ejrad.2020.109266
Received 30 March 2020; Received in revised form 18 July 2020; Accepted 31 August 2020
Available online 8 September 2020

mutation analysis of the *CTNNB1* (β-catenin) gene, is therefore currently required to differentiate DTF from non-DTF [6].

As DTF is a borderline tumor who is unable to metastasize, and requires a different treatment regimen than malignant STS, this distinction is highly relevant. Differentiation between DTF and STS based on imaging would be beneficial because of the rarity of DTF, making clinical and pathological recognition challenging. Furthermore, DTF exhibits an aggressive growth pattern and growth might be stimulated after (surgical) trauma, including biopsies [7]. Avoiding (multiple) harmful biopsies which potentially cause tumor growth is therefore of great importance.

Several studies have addressed the prognostic role of the *CTNNB1* mutation in DTF [8–10], as serine 45 (S45 F) tumors appear to have a higher risk of recurrence after surgery compared to threonine 41 (T41A) and wild type (WT) (i.e. no *CTNNB1* mutation [11]) tumors [12]. Obtaining the *CTNNB1* mutation status is for diagnostic purposes and to guide the clinical work-up, but, for now, the *CTNNB1* mutation status has no therapeutic consequences [13]. The majority of DTF harbors a *CTNNB1* mutation at either T41A or S45 F [8]. Assessment of the mutation status is currently done by Sanger Sequencing or Next Generation Sequencing, which are time consuming and expensive.

In radiomics, large amounts of quantitative imaging features are related to clinical outcome [14]. Radiomics may serve as a non-invasive surrogate to contribute to diagnosis, prognosis and treatment planning [15,16]. Based on the results of previous studies in cancer [17], we hypothesized that radiomics may also be useful in DTF.

This study investigated whether a radiomics model based on MRI is able to 1) distinguish DTF from non-DTF in the extremities, and 2) to predict the *CTNNB1* mutation status of DTF. Additionally, in the DTF vs. non-DTF distinction, we evaluated which of the included MRI sequences has the highest predictive value.

2. Material and methods

2.1. Data collection

Approval by the Erasmus Medical Center (MC) institutional review board (MEC-2016-339) was obtained. Patients diagnosed or referred to the Erasmus MC between 1990-2018 with a histologically proven primary or recurrent DTF were included. This resulted in a multicenter imaging dataset as patients referred to our sarcoma expert institute often received imaging at their referring hospital. The most frequently used imaging modality prior to treatment was T1w-MRI, and its availability was used as an inclusion criterion [18]. When available, other sequences such as T2w, T1w post-contrast, dynamical contrast enhanced (DCE), proton density (PD) and diffusion weighted imaging (DWI) MRI were collected.

For the differential diagnosis (DTF vs. non-DTF), histologically confirmed malignant extremity STS were included. Benign STS were excluded, because this distinction is clinically less relevant. Non-extremity STS were excluded because of the infrequent use of MRI. Although DTF tumors commonly occur in the abdominal wall, their differential diagnosis is broad and includes pseudo-tumors such as myositis, nodular fasciitis and hematomas, and tumors such as lipomas, STS, endometriosis, carcinomas, lymphomas and metastasis [19]. Hence, we decided to focus on the distinction between DTF and STS, and included patients with a histologically proven primary fibromyxosarcoma, myxoid liposarcoma or leiomyosarcoma of the extremities. Similar to the DTF, patients with at least a pre-treatment T1w-MRI were retrospectively included.

Sex, age at diagnosis, and tumor location were collected. For the DTF, in case of a missing *CTNNB1* mutation status, Sanger Sequencing was performed after review of formalin-fixed paraffin-embedded tumor sections by a pathologist. Cases with a known *CTNNB1* mutation did not undergo additional review by a pathologist. Poor scan quality (e.g. artifacts), poor DTF DNA quality with failure of sequencing, and *CTNNB1* mutation other than S45F, T41A or WT led to exclusion.

2.2. Radiomics feature extraction

The tumors were all manually segmented once on the T1w-MRI by one of two clinicians under supervision of a musculoskeletal radiologist (4 years of experience). A subset of 30 DTF was segmented by both clinicians, in which intra-observer variability was evaluated through the pairwise Dice Similarity Coefficient (DSC), with DSC > 0.70 indicating good agreement [20]. To transfer the segmentations to the other sequences, all sequences were automatically aligned to the T1w-MRI using image registration with the Elastix software [21]. For each lesion, per MRI sequence, 411 features quantifying intensity, shape and texture were extracted. Details can be found in Appendix A and Table A.2.

2.3. Decision model creation

To create a decision model from the features, the WORC toolbox was used, see Fig. 1 [22–24]. In WORC, the decision model creation consists of several steps, e.g. feature selection, resampling, and machine learning. WORC performs an automated search amongst a variety of algorithms for each step and determines which combination of algorithms maximizes the prediction performance on the training set. More details can be found in Appendix B.

For the differential diagnosis cohort, a binary classification model was created using a variety of machine learning models. For the DTF cohort (predicting the *CTNNB1* mutation), a multiclass classification model was created using random forests.

2.4. Evaluation

Evaluation of all models was done through a 100x random-split cross-validation. In each iteration, the data was randomly split in 80 % for training and 20 % for testing in a stratified manner, to make sure the distribution of the classes in all sets was similar to the original (Fig. A.1). Within the training set, model optimization was performed using an internal cross-validation (5x). Hence, all optimization was done on the training set to eliminate any risk of overfitting on the test set.

Performance was evaluated using the Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) curve, balanced classification accuracy (BCA), sensitivity, specificity, negative predictive value (NPV), and positive predictive value (PPV). For the multiclass models, we reported the multiclass AUC [25] and overall BCA [26]. The positive classes included: DTF in the differential diagnosis, and the presence of the mutation in the mutation analysis. The 95 % confidence intervals were constructed using the corrected resampled t-test, thereby taking into account that the samples in the cross-validation splits are not statistically independent [27]. Both the mean and the confidence intervals are reported. ROC confidence bands were constructed using fixed-width bands [28].

To assess the predictive value of the various features, models were trained based on: 1) volume; 2) age and sex; 3) T1w-MRI imaging; 4) T1w-MRI imaging, age and sex. Model 1 was created to verify that the imaging models were not solely based on volume. Model 2 was created to evaluate potential age and gender biases. In model 4, the imaging and clinical characteristics are combined by using both the imaging features and age and sex as features for a total of 413 features. This allows WORC to combine the imaging and clinical characteristics in the most optimal way. Additionally, a model was made for each combination of T1w-MRI and one of the other included MRI sequences (e.g. based on T1w-MRI and T2w-MRI) to evaluate the added value of these other sequences. When a sequence was missing for a patient, feature imputation was used to estimate the missing values.

The code for the feature extraction, model creation and evaluation has been published open-source [29].

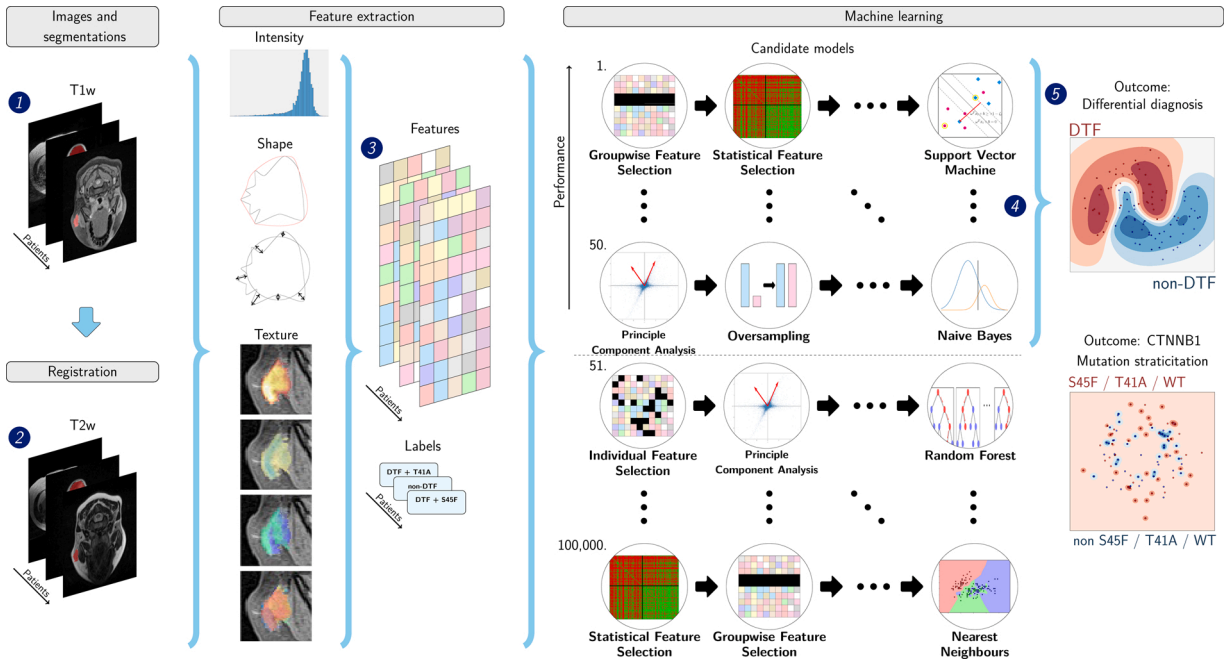

Fig. 1. Schematic overview of the radiomics approach: adapted from [24]. Processing steps include segmentation of the tumor on the T1-weighted (T1w) MRI (1), registration of the T1w to the T2-weighted (T2w) MRI to transform this segmentation to the T2w-MRI (2), feature extraction from both the T1w-MRI and the T2w-MRI (3) and the creation of machine learning decision models (5), using an ensemble of the best 50 workflows from 100,000 candidates (4), where the workflows are different combinations of the processing and analysis steps.
DTF, desmoid-type fibromatosis.

2.5. Model insight

To explore the predictive value of individual features, the Mann-Whitney U univariate statistical test was used. P-values were corrected for multiple testing using the Bonferroni correction, and were considered statistically significant at a p-value <0.05. Feature robustness to variations in the segmentations was assessed on the subset of 30 DTF segmented by two observers using the intra-class correlation coefficient (ICC), were an ICC > 0.75 indicated good reliability [30]. To evaluate model reliability, a separate model was trained using only these features with a good reliability. To gain insight into the models, the patients were ranked based on the consistency of the model predictions. Typical examples for each class consisted of the patients that were correctly classified in all cross-validation iterations; atypical vice versa.

2.6. Classification by radiologists

To compare the models with clinical practice, the tumors were classified by two musculoskeletal radiologists (5 and 4 years of experience), which had access to all available MRI sequences, age, and sex. They were specifically instructed to distinguish between STS and DTF. Classification was made on a ten-point scale to indicate the radiologists' certainty. As only extremity STS were selected for the non-DTF group, a location-matched database was used. This included all extremity DTF and the same number of non-DTF. Agreement between the radiologists was evaluated using Cohen's kappa. The radiomics models were evaluated as well in this cohort. In each cross-validation iteration, these models were trained on 80 % of the full dataset, but tested only on patients from the location-matched cohort in the other 20 % of the dataset. The DeLong test was used to compare the AUCs [31].

3. Results

3.1. Study selection and population

The dataset included 203 patients; see Table 1 for the clinical characteristics. The differential diagnosis cohort consisted of 64 fibromyxosarcomas, 31 leiomyosarcomas, 36 myxoid liposarcomas, and 72 DTFs (65 primary, 7 recurrent), of which 61 were suitable for the mutation analysis.

The dataset originated from 68 scanners, resulting in a large heterogeneity in the acquisition protocols, see Table 2. From the 72 patients in the DTF cohort, there were 30 T1w post-contrast (42 %), 49 T1w post-contrast FatSat (68 %), 34 T2w (47 %), 33 T2w FatSat (46 %), 3 proton density (PD) (4%), 18 DCE (25 %) and 3 DWI (4%) MRI scans. Due to the limited availability of the PD, DCE, and DWI sequences, besides the T1w-MRI, only the T1w post-contrast and T2w (with/without FatSat) sequences were analyzed.

On the subset of 30 DTF that was segmented by both observers, the mean DSC was 0.77 (standard deviation of 0.20), indicating good agreement. An example of the image registration results is depicted in Fig. 2.

3.2. Differential diagnosis

The performance of models 1–6 for the differential diagnosis is shown in Table 3. Model 1, based on volume, showed little predictive value (mean AUC of 0.69). Model 2, based on age and sex, performed better (mean AUC of 0.86). Model 3, based on T1w-MRI, had a mean AUC of 0.79, thus performing worse than age and sex, but better than volume alone. Model 4, combining the T1w-MRI, age, and sex, showed little improvement in terms of mean AUC (0.88) over model 2. Addition of a T2w-MRI, i.e. model 5, or T1 post-contrast MRI, i.e. model 6, both with or without FatSat, both yielded a minor overall improvement over

Table 1
Clinical characteristics of both cohorts.

	Differential diagnosis cohort				Mutation analysis cohort
	DTF n = 72	Fibro-myxosarcoma n = 64	Leiomyosarcoma n = 31	Myxoid liposarcoma n = 36	DTF n = 61
Sex					
Male	16 (22 %)	41 (64 %)	19 (61 %)	22 (61 %)	15 (25 %)
Female	56 (78 %)	23 (36 %)	12 (39 %)	14 (39 %)	46 (75 %)
Age median (IQR)	36 (23–47)	67 (54–77)	66 (55–73)	42 (35–56)	36 (22–47)
Tumor location					
Head/neck	12 (17 %)	–	–	–	11 (18 %)
Chest aperture	4 (6 %)	–	–	–	3 (5 %)
Abdominal wall	24 (33 %)	–	–	–	16 (26 %)
Back	11 (15 %)	–	–	–	10 (16 %)
Intra-abdominal	1 (1 %)	–	–	–	1 (2 %)
Upper extremity	5 (7 %)	6 (9 %)	7 (23 %)	1 (3 %)	5 (8 %)
Lower extremity	15 (21 %)	58 (91 %)	24 (77 %)	35 (97 %)	15 (25 %)
Tumor size in cm[a] median (IQR)	6.3 (4.1–9.8)	7.0 (4.9–12.9)	8.3 (5.2–9.4)	12.8 (8.5–15.3)	6.3 (4.1–9.5)
Volume in cl median (IQR)	2.0 (0.5–9.8)	5.6 (1.1–34.1)	8.2 (1.7–11.4)	16.8 (5.2–37.4)	2.2 (0.7–9.6)
Mutation type					
T41A	NA	NA	NA	NA	24 (39 %)
S45F	NA	NA	NA	NA	16 (26 %)
Wild-type	NA	NA	NA	NA	21 (34 %)
MRI sequences					
T2w FS	33 (46 %)	37 (58 %)	15 (48 %)	16 (44 %)	26 (43 %)
T2w non-FS	32 (70 %)	37 (64 %)	19 (39 %)	19 (43 %)	26 (61 %)
T1w PC FS	49 (70 %)	32 (50 %)	19 (48 %)	22 (51 %)	43 (70 %)
T1w PC non-FS	30 (43 %)	24 (48 %)	11 (23 %)	17 (33 %)	25 (35 %)

*Abbreviations: DTF: desmoid-type fibromatosis; IQR: interquartile range; cm: centimeter; cl: centiliter; MRI: magnetic resonance imaging; FS: FatSat; PC: post-contrast.
Percentages might not add up to 100 % in total because of rounding.
[a] Maximum diameter automatically measured in three planes.

Table 2
Properties of the acquisition protocols of the 203 T1-weighted MRI sequences in the dataset.

Property		Number		%
Magnetic field strength				
1T		20		10
1.5T		167		82
3T		16		8
Manufacturer				
Siemens		93		46
Philips		79		39
General Electrics		27		13
Toshiba		4		2
Setting (Unit)	Mean	Std.	Min	Max
Slice thickness (mm)	4.66	1.45	1.0	11.0
Repetition time (ms)	619	533	0.0	4620
Echo time (ms)	14	7	2.0	94.0

*Abbreviations: T: tesla; Std: standard deviation; mm: millimeter; ms: milliseconds.

model 3 (mean AUC of 0.84 and 0.84, respectively). These observations were confirmed by the ROC curves in Fig. 3. The models using either only non-FatSat or FatSat scans, both for the T2w and T1w post-contrast MRI, faired similar, see Table A.1.

3.3. Comparison with radiologists

As described in the methods, for the comparison with radiologists, a location-matched cohort consisting of all extremity DTFs and an equal amount of extremity non-DTF was used. To this end, all 20 extremity DTFs and 20 randomly selected extremity non-DTFs were included in the location-matched cohort. The performance of radiomics and the radiologists in this cohort is shown in Table 4: model 1 and 5–6 were omitted from the results for brevity. The AUCs of the radiomics models (model 2: 0.93; model 3: 0.88; model 4: 0.98) were generally higher than both radiologists 1 (0.80) and 2 (0.88). This is confirmed by the ROC curves in Fig. 4. Cohen's kappa between the two radiologists was 0.40, indicating intermediate observer agreement. A DeLong power analysis of the AUCs resulted in a power of only 0.1. Due to the limited power, the p-values of the DeLong test were omitted.

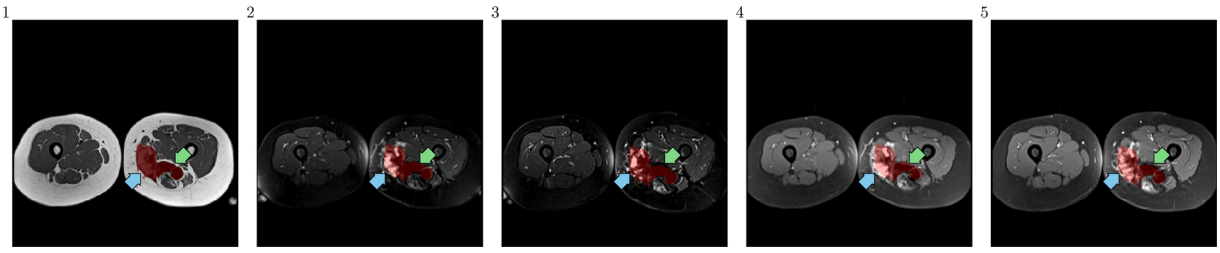

Fig. 2. Segmentations on various MRI sequences before and after applying image registration in a desmoid-type fibromatosis case. The arrows are at the same position in each image and point at two details where the (mis)alignment is evident. (1) Original T1-weighted (T1w) MRI; (2) Original T2w-MRI; (3) Registered T2w-MRI; (4) Original T1w post-contrast MRI ; (5) Registered T1w post-contrast MRI.

Table 3
Performance of the radiomics models for the DTF differential diagnosis based on: model 1: volume only; model 2: age and sex only; model 3: T1w imaging features, including volume; model 4: the combination of T1w imaging features and age and sex; model 5: the combination of T1w and T2w imaging features; and model 6: the combination of T1w and T1w post-contrast imaging features. Outcomes are presented with the 95% confidence interval.

	Model 1 Volume	Model 2 Age + Sex	Model 3 T1w	Model 4 T1w + Age + Sex	Model 5 T1w + T2w	Model 6 T1w + T1w post-contrast
AUC	0.69 [0.61, 0.76]	0.86 [0.79, 0.92]	0.79 [0.73, 0.85]	0.88 [0.82, 0.93]	0.84 [0.78, 0.89]	0.84 [0.78, 0.90]
BCA	0.59 [0.53, 0.65]	0.78 [0.71, 0.86]	0.71 [0.65, 0.77]	0.79 [0.72, 0.86]	0.68 [0.62, 0.75]	0.75 [0.69, 0.81]
Sensitivity	0.80 [0.70, 0.91]	0.78 [0.66, 0.90]	0.61 [0.49, 0.72]	0.70 [0.57, 0.83]	0.43 [0.31, 0.55]	0.62 [0.52, 0.73]
Specificity	0.39 [0.28, 0.49]	0.79 [0.71, 0.87]	0.81 [0.73, 0.89]	0.88 [0.82, 0.94]	0.94 [0.88, 0.99]	0.88 [0.82, 0.95]
NPV	0.50 [0.71, 0.89]	0.88 [0.81, 0.94]	0.80 [0.76, 0.75]	0.85 [0.80, 0.91]	0.76 [0.72, 0.80]	0.81 [0.76, 0.85]
PPV	0.41 [0.36, 0.46]	0.72 [0.57, 0.76]	0.64 [0.53, 0.75]	0.76 [0.67, 0.86]	0.80 [0.66, 0.94]	0.76 [0.65, 0.88]

*Abbreviations: T1w: T1-weighted; T2w: T2-weighted; AUC: area under the receiver operator characteristic curve; BCA: balanced classification accuracy; NPV: negative predictive value; PPV: positive predictive value.

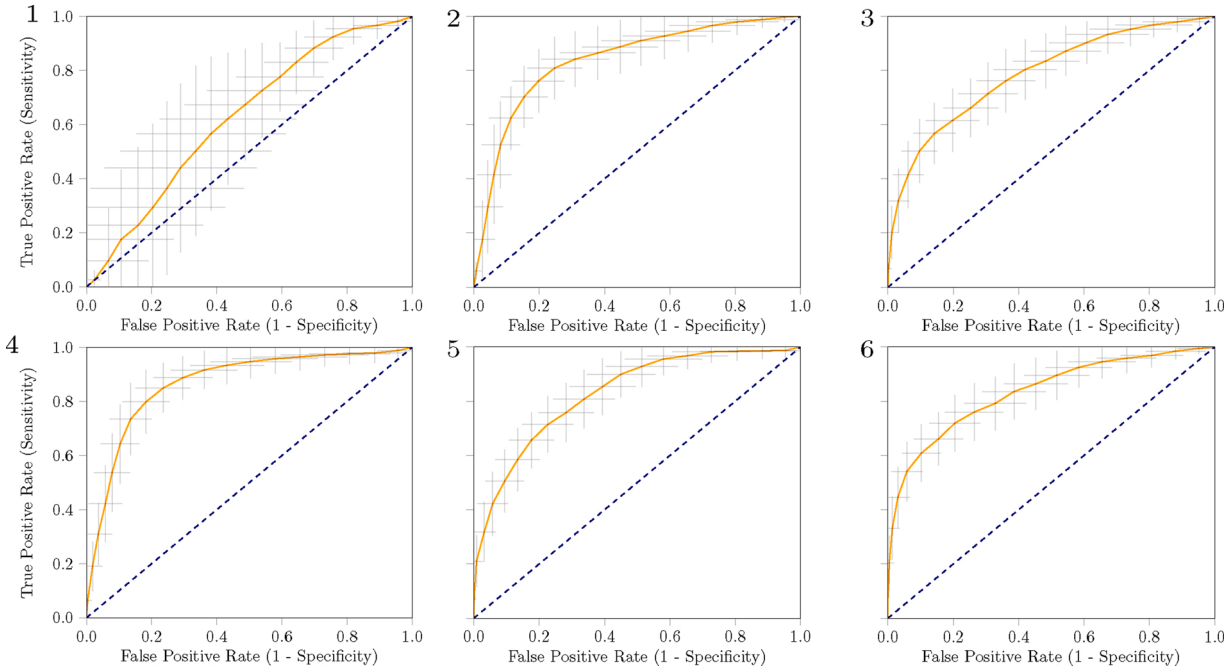

Fig. 3. Receiver operating characteristic curves of the radiomics models based on volume (**1**); age and sex (**2**); T1-weighted (T1w) features (**3**); T1w features, age, and sex (**4**); T1w + T2weighted imaging features (**5**); and T1w + T1w post-contrast imaging features (**6**). The grey crosses identify the 95 % confidence intervals of the 100x random-split cross-validation; the orange curve depicts the mean.

3.4. CTNNB1 mutation status stratification

Table 5 depicts the performance of the radiomics models for the *CTNNB1* mutation stratification. Model 4, using T1w-MRI, age, and sex, had a high specificity (S45 F: 0.83, T41A: 0.59 and WT: 0.72), but a sensitivity similar to guessing (S45 F: 0.15, T41A: 0.49 and WT: 0.56). This indicates a strong bias in the models towards the negative classes, i.e. not-S45 F, not-T41A and not-WT. As model 4 did not perform well, models 1, 2, and 3 were omitted from the results, as these contain a subset of these features. Adding the T2w or T1w post-contrast imaging, i.e. models 5 and 6, did not improve the performance. Hence, the models using either only non-FatSat or FatSat scans were omitted, as these

Table 4
Performance of the two radiologists and the radiomics models in differentiating between DTF (n = 20) and non-DTF (n = 20) in the location-matched cohort. Outcomes are presented with the 95% confidence interval.

	Model 2 Age + Sex	Model 3 T1w	Model 4 T1w + Age + Sex	Rad 1	Rad 2
AUC	0.93 [0.84, >1]	0.87 [0.73, >1]	0.98 [0.92, >1]	0.80	0.88
BCA	0.85 [0.71, 1.00]	0.71 [0.56, 0.87]	0.88 [0.77, 0.99]	0.75	0.90
Sensitivity	0.79 [0.57, >1]	0.49 [0.21, 0.77]	0.78 [0.57, 1.00]	0.65	0.90
Specificity	0.90 [0.71, >1]	0.93 [0.78, >1]	0.98 [0.91, >1]	0.85	0.89
NPV	0.82 [0.61, >1]	0.65 [0.43, 0.76]	0.82 [0.64, >1]	0.71	0.89
PPV	0.91 [0.72, >1]	0.81 [0.47, >1]	0.98 [0.91, >1]	0.81	0.90

*Abbreviations: T1w: T1-weighted; AUC: area under the receiver operator characteristic curve; BCA: balanced classification accuracy; PPV: positive predictive value; NPV: negative predictive value.

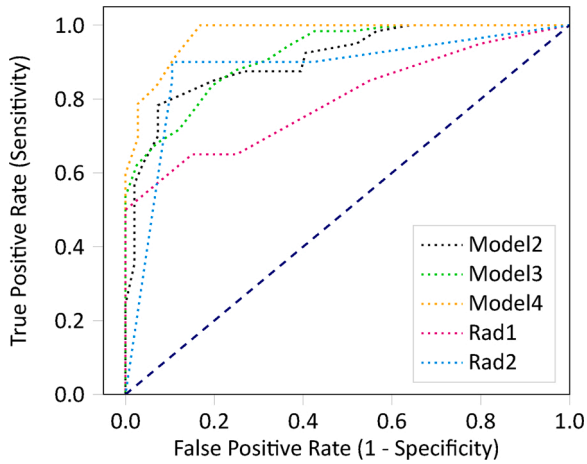

Fig. 4. Receiver operating characteristic curves of the radiomics models based on age and sex (model 2); imaging (model 3); and imaging, age and sex (model 4); and those of the radiologists (Rad1 and Rad2), in the location-matched cohort.

contain subsets of the scans from models 5 and 6.

3.5. Model insight

As the *CTNNB1* mutation status stratification models did not perform well, the model insight analysis was only conducted for the differential diagnosis. The p-values from the Mann-Whitney *U* test between the DTF and non-DTF patients of all features are shown in Table A.3. In the feature importance analysis, 76 T1w-MRI features had significant p-values (5.4×10^{-8} to 4.8×10^{-2}). These included two intensity features (entropy and peak), two shape features (radial distance and volume), and 72 texture features. The p-value of age (1×10^{-11}) was lower than that of all imaging features. The ICC values of all T1w-MRI features are shown in Table A.4. Of the 411 features, 270 (66 %) had an ICC > 0.75 and thus good reliability. Only using these features with a good reliability in model 3 did not alter the performance.

As we are mostly interested in which imaging features define typical DTF, and not age and sex, the patient ranking was conducted for model 3. Of the 203 patients, 104 tumors (24 DTFs, 80 non-DTFs) were always classified correctly by model 3, i.e. in all 100 cross-validation iterations. Nineteen tumors (17 DTFs, 2 non-DTFs) were always classified incorrectly. In Fig. 5, MRI slices of such typical and atypical examples of DTFs are shown.

4. Discussion

This study showed that radiomics based on T1w-MRI can distinguish DTF from STS. Adding T2w or T1w post-contrast MRI did not substantially improve the model. The DTF *CTNNB1* mutation status could not be predicted through radiomics. To our knowledge, this is the first study to evaluate the DTF differential diagnosis and mutation status through an automated radiomics approach.

Age and sex appeared to be strong predictors for the diagnosis of DTF, performing better than T1w-MRI. The combination of imaging, age and sex did not improve the model. This implies that age and sex are sufficient for distinguishing DTF from STS. In line with previous nationwide DTF cohort studies, females represented the majority of our cohort, with a lower median age compared to the median age of the patients from the non-DTF group [2,32]. The relation in our database may however be too strong, and thereby not representative of clinical practice. For example, above 63 years of age, our database included 60 non-DTF and only a single DTF. While the peak incidence of DTF is between 20–40 years, DTF can affect patients of all ages with reported ranges from 2 to 90 years 32]. Simply classifying all tumors in patients above 63 years as non-DTF, regardless of any tumor (imaging) information, is unfeasible. Such a model cannot be applied in the general population, while the model purely based on T1w-MRI imaging, as it does not use any population-based information. Our cohort might be biased due to the focus on MRI and the extremity as a location, while other modalities (e.g. CT or ultrasound) may be used for certain locations or for certain types of patients. Further research should include the expansion of our dataset to make especially the age distribution more representative.

To estimate the clinical value of our model, we compared the performance with the assessment of two radiologists. The model based on imaging performed similar to the radiologists. The model combining age, sex and imaging features, using the same dataset as the radiologist, had a higher AUC than the musculoskeletal radiologists. However this model may suffer from the selection bias as mentioned in the previous section. The agreement between the radiologists was intermediate, indicating observer dependence in the prediction. The radiomics model is observer independent, assuming the segmentation is reproducible as indicated by the high DSC and ICC, and will always give the same prediction on the same image.

The DTF differential diagnosis is highly important for treatment decisions, but difficult on imaging due to its rarity, while using invasive biopsies brings risks such as tumor growth. The use of our T1w-MRI radiomics model may therefore aid early recognition and diagnosis of DTF, thus shortening the diagnostic delay by enabling direct referral to an STS expertise center. Since all routine MRI protocols include a T1w-MRI, our radiomics method is generalizable, feasible and applicable for use in daily clinical practice. After further model optimization, it may serve as a quick, non-invasive, and low-cost alternative for a biopsy, currently limited to extremities due to the used dataset.

Additionally, we investigated the predictive value of sequences other than T1w-MRI. The number of available sequences was however limited

Table 5
Performance of the random forest multilabel radiomics models for the DTF *CTNNB1* mutation stratification based on; model 4: T1w imaging features, age and sex; model 5: T1w + T2w imaging features; and model 6: T1w + T1w post-contrast imaging features. Model 4 was evaluated for a single class (S45 F, T41A, and WT) or the overall performance (All). Outcomes are presented with the 95% confidence interval.

	Model 4 - S45 F T1w + age + sex	Model 4 - T41A T1w + age + sex	Model 4 - WT T1w + age + sex	Model 4 - All T1w + age + sex	Model 5 - All T1w + T2w	Model 5 - All T1w + T1w post-contrast
AUC	0.61 [0.44, 0.77]	0.56 [0.43, 0.68]	0.74 [0.60, 0.87]	0.63 [0.54, 0.72]	0.63 [0.53, 0.72]	0.60 [0.50, 0.69]
BCA	0.48 [0.35, 0.61]	0.53 [0.42, 0.64]	0.65 [0.54, 0.75]	0.56 [0.47, 0.64]	0.57 [0.48, 0.66]	0.53 [0.44, 0.61]
Sensitivity	0.15 [<0, 0.37]	0.49 [0.27, 0.71]	0.56 [0.35, 0.77]	NA	NA	NA
Specificity	0.83 [0.67, 0.98]	0.59 [0.41, 0.76]	0.72 [0.55, 0.89]	NA	NA	NA
NPV	0.76 [0.70, 0.82]	0.65 [0.53, 0.77]	0.73 [0.64, 0.82]	NA	NA	NA
PPV	0.17 [<0, 0.45]	0.42 [0.28, 0.56]	0.59 [0.40, 0.77]	NA	NA	NA

*Abbreviations: T1w: T1-weighted MRI; T2w: T2-weighted MRI; AUC: area under the receiver operator characteristic curve; BCA: balanced classification accuracy; PPV: positive predictive value; NPV: negative predictive value; WT: wild-type, NA: not applicable.

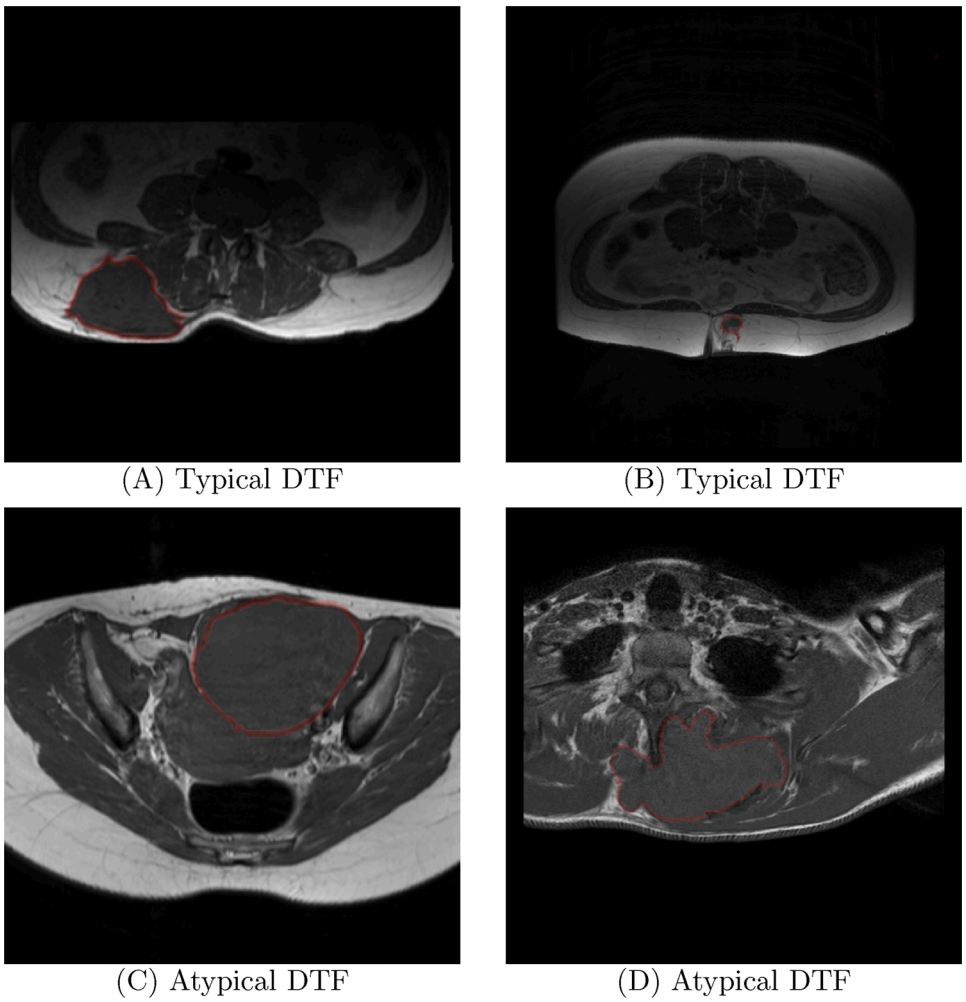

Fig. 5. The typical examples (A and B) are two cases always classified correctly by the T1-weighted (T1w) imaging model; the atypical examples (C and D) are two cases always classified incorrectly by the T1w imaging model.

due to the multicenter imaging dataset. Although T2w-MRI is often used to correlate DTF signal intensity with prognosis or response to therapy [33–36], in the current study T2w-MRI added little predictive value to the T1w-MRI, similar to the T1w post-contrast MRI. This may however be attributed to the fact that these sequences were only available for a subset of the patients. Our cohort contained too few patients with PD, DCE, or DWI sequences to be analyzed. However, there is little to no indication of the added value of these sequences in DTF [37–39].

The second aim of this study was to predict the DTF *CTNNB1* mutation status. Our radiomics model was not able to stratify the *CTNNB1* mutation type, which is in line with the absence of literature linking DTF MRI appearance to the *CTNNB1* mutation.

The current study enclosed several limitations. First, due to the rarity of DTF, the DTF sample size was limited and possibly too small for the mutation stratification model to learn from. This also resulted in little statistical power for the mutation analysis, as shown by the large width of our confidence intervals, and for the comparison with the radiologists in the differential diagnosis. Besides primary tumors, the DTF cohort contained also recurrent tumors. As this number was low, and to our knowledge, there are no indications that recurrent DTF appear different on MRI than primary DTF, the expected influence is small. Within the DTF cohort, the WT group was relatively large and might have been subjected to incorrect allocation, as Sanger Sequencing is not always sensitive enough to detect all mutations [11]. The results of the *CTNBB1* mutation status stratification showed a strong bias towards the majority classes, which may be attributed to the class imbalance. Although we exploited commonly used imbalanced learning strategies such as resampling and ensembling. other strategies may improve the performance. Second, only extremity DTFs were included for comparison with STS. This was due to the limited availability of MRI in non-extremity soft tissue tumors. However, this is not representative for the entire DTF population, which also occurs frequently in the abdominal wall and trunk [3]. Third, the current radiomics approach requires manual annotations. While accurate, this process is also time consuming and subject to some observer variability as indicated by our DSC, and thus limits the transition to clinical practice. Automatic segmentation methods, for example deep learning, may help to overcome these limitations [40]. Lastly, the dataset originated from 68 different scanners, which resulted in substantial heterogeneity in the acquisition protocols. The lack of standard imaging parameters can be problematic as these can affect the appearance of the tumor and thus the radiomics performance. However, our method was successfully able to create diagnostic models despite these differences. As these models were trained on a variety of imaging protocols, there is an increased chance that the reported performance can be reproduced in a routine clinical setting when using other MRI scanners. Using a single-scanner with dedicated tumor protocols may improve the model performance, but will limit the generalizability.

Future work should firstly focus on the prospective validation of our findings. Although we did use a multicenter imaging dataset and performed a rigorous cross-validation experiment strictly separating training from testing data, we did not validate our model on an independent, external dataset. Afterwards, the radiomics model could be used to predict clinical outcomes of DTF receiving active surveillance or

systemic treatment.

5. Conclusions

Our radiomics approach is capable of distinguishing DTF from non-DTF tumors on T1w-MRI, and can potentially aid diagnosis and shorten diagnostic delay. The performance of the model was similar to that of two experienced musculoskeletal radiologists. The model was not able to predict *CTNNB1* mutation status of DTF tumors. Further optimization and external validation of the model is needed to incorporate radiomics in clinical practice.

Sources of funding for research

This study was financed by the Stichting Coolsingel (reference number 567), a Dutch non-profit foundation.

CRediT authorship contribution statement

Milea J.M. Timbergen: Conceptualization, Resources, Investigation, Funding acquisition, Writing - original draft, Data curation. **Martijn P.A. Starmans:** Conceptualization, Methodology, Data curation, Resources, Software, Validation, Visualization, Formal analysis, Writing - original draft. **Guillaume A. Padmos:** Resources, Investigation, Writing - review & editing. **Dirk J. Grünhagen:** Conceptualization, Supervision, Writing - review & editing. **Geert J.L.H. van Leenders:** Resources, Investigation, Writing - review & editing. **D.F. Hanff:** Resources, Writing - review & editing. **Cornelis Verhoef:** Conceptualization, Supervision, Funding acquisition, Writing - review & editing. **Wiro J. Niessen:** Methodology, Software, Writing - review & editing. **Stefan Sleijfer:** Conceptualization, Supervision, Writing - review & editing. **Stefan Klein:** Conceptualization, Methodology, Formal analysis, Software, Supervision, Writing - review & editing. **Jacob J. Visser:** Conceptualization, Data curation, Investigation, Methodology, Project administration, Supervision, Validation, Writing - review & editing.

Declaration of Competing Interest

Wiro J. Niessen is founder, scientific lead and stock holder of Quantib BV. The other authors do not declare any conflicts of interest.

Acknowledgements

Martijn P.A. Starmans acknowledges funding from the research program STRaTeGy (project number 14929-14930), which is (partly) financed by the Netherlands Organisation for Scientific Research (NWO). This work was partially carried out on the Dutch national e-infrastructure with the support of SURF Cooperative.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi: https://doi.org/10.1016/j.ejrad.2020.109266.

References

[1] C.D.M. Fletcher, WHO Classification of Tumours of Soft Tissue and Bone, IARC, Lyon, 2013.
[2] D.L. van Broekhoven, D.J. Grunhagen, M.A. den Bakker, T. van Dalen, C. Verhoef, Time trends in the incidence and treatment of extra-abdominal and abdominal aggressive fibromatosis: a population-based study, Ann. Surg. Oncol. 22 (9) (2015) 2817–2823.
[3] J.J. Reitamo, P. Hayry, E. Nykyri, E. Saxen, The desmoid tumor. I. Incidence, sex-, age- and anatomical distribution in the Finnish population, Am. J. Clin. Pathol. 77 (6) (1982) 665–673.
[4] M. Braschi-Amirfarzan, A.R. Keraliya, K.M. Krajewski, et al., Role of imaging in management of desmoid-type fibromatosis: a primer for radiologists, Radiographics 36 (3) (2016) 767–782.
[5] E.A. Walker, M.E. Fenton, J.S. Salesky, M.D. Murphey, Magnetic resonance imaging of benign soft tissue neoplasms in adults, Radiol. Clin. North Am. 49 (6) (2011) 1197–1217, vi.
[6] T.L. Ng, A.M. Gown, T.S. Barry, et al., Nuclear beta-catenin in mesenchymal tumors, Mod. Pathol. 18 (1) (2005) 68–74.
[7] M.V. Enzo, P. Cattelan, M. Rastrelli, et al., Growth rate and myofibroblast differentiation of desmoid fibroblast-like cells are modulated by TGF-β signaling, Histochem. Cell Biol. 151 (2) (2019) 145–160.
[8] C. Colombo, R. Miceli, A.J. Lazar, et al., CTNNB1 45F mutation is a molecular prognosticator of increased postoperative primary desmoid tumor recurrence: an independent, multicenter validation study, Cancer 119 (20) (2013) 3696–3702.
[9] A.J. Lazar, D. Tuvin, S. Hajibashi, et al., Specific mutations in the beta-catenin gene (CTNNB1) correlate with local recurrence in sporadic desmoid tumors, Am. J. Pathol. 173 (5) (2008) 1518–1527.
[10] D.L. van Broekhoven, C. Verhoef, D.J. Grunhagen, et al., Prognostic value of CTNNB1 gene mutation in primary sporadic aggressive fibromatosis, Ann. Surg. Oncol. 22 (5) (2015) 1464–1470.
[11] A.M. Crago, J. Chmielecki, M. Rosenberg, et al., Near universal detection of alterations in CTNNB1 and Wnt pathway regulators in desmoid-type fibromatosis by whole-exome sequencing and genomic analysis, Genes Chromosomes Cancer 54 (10) (2015) 606–615.
[12] M.J.M. Timbergen, C. Colombo, M. Renckens, et al., The prognostic role of beta-catenin mutations in desmoid-type fibromatosis undergoing resection only: a meta-analysis of individual patient data, Ann. Surg. (2019) (online and ahead of print).
[13] Desmoid Tumor Working Group, The management of desmoid tumours: a joint global consensus-based guideline approach for adult and paediatric patients, Eur. J. Cancer 127 (2020) 96–107.
[14] M.P.A. Starmans, S.R. van der Voort, J.M.C. Tovar, J.F. Veenland, S. Klein, W. J. Niessen, Radiomics: data mining using quantitative medical image features. Handbook of Medical Image Computing and Computer Assisted Intervention, Elsevier, 2020, pp. 429–456.
[15] A.M. Rutman, M.D. Kuo, Radiogenomics: creating a link between molecular diagnostics and diagnostic imaging, Eur. J. Radiol. 70 (2) (2009) 232–241.
[16] M.A. Mazurowski, Radiogenomics: what it is and why it is important, J. Am. Coll. Radiol. 12 (8) (2015) 862–866.
[17] Z. Bodalal, S. Trebeschi, T.D.L. Nguyen-Kim, W. Schats, R. Beets-Tan, Radiogenomics: bridging imaging and genomics, Abdom. Radiol. (NY) 44 (6) (2019) 1960–1984.
[18] S. Otero, E.C. Moskovic, D.C. Strauss, et al., Desmoid-type fibromatosis, Clin. Radiol. 70 (9) (2015) 1038–1045.
[19] H.G. Smith, D. Tzanis, C. Messiou, et al., The management of soft tissue tumours of the abdominal wall, Eur. J. Surg. Oncol. 43 (9) (2017) 1647–1655.
[20] K.H. Zou, S.K. Warfield, A. Bharatha, et al., Statistical validation of image segmentation quality based on a spatial overlap index1: scientific reports, Acad. Radiol. 11 (2) (2004) 178–189.
[21] S. Klein, M. Staring, K. Murphy, M.A. Viergever, J.P. Pluim, Elastix: a toolbox for intensity-based medical image registration, IEEE Trans. Med. Imaging 29 (1) (2010) 196–205.
[22] M.P.A. Starmans, S.R. van der Voort, M. Vos, et al., Fully automatic construction of optimal radiomics workflows, European Conference of Radiology (ECR) (2019).
[23] M.P.A. Starmans, Workflow for Optimal Radiomics Classification (WORC), 2018. https://github.com/MStarmans91/WORC.
[24] M. Vos, M.P.A. Starmans, M.J.M. Timbergen, et al., Radiomics approach to distinguish between well differentiated liposarcomas and lipomas on MRI, Br. J. Surg. 106 (13) (2019) 1800–1809.
[25] D.J. Hand, R.J. Till, A simple generalisation of the area under the ROC curve for multiple class classification problems, Mach. Learn. 45 (2) (2001) 171–186.
[26] R.V. Marinescu, N.P. Oxtoby, A.L. Young, et al., TADPOLE Challenge: Prediction of Longitudinal Evolution in Alzheimer's Disease, arXiv preprint arXiv:180503909, 2018.
[27] C. Nadeau, Y. Bengio, Inference for the Generalization Error, 2000, pp. 307–313.
[28] S.A. Macskassy, F. Provost, S. Rosset, ROC confidence bands: an empirical evaluation, Proceedings of the 22nd International Conference on Machine Learning: ACM (2005) 537–544.
[29] M.P.A. Starmans, DMRadiomics, 2020. http://doi.org/10.5281/zenodo.4017191. (accessed September 7, 2020).
[30] T.K. Koo, M.Y. Li, A guideline of selecting and reporting intraclass correlation coefficients for reliability research, J. Chiropr. Med. 15 (2) (2016) 155–163.
[31] E.R. DeLong, D.M. DeLong, D.L. Clarke-Pearson, Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach, Biometrics 44 (3) (1988) 837–845.
[32] N. Penel, A. Le Cesne, S. Bonvalot, et al., Surgical versus non-surgical approach in primary desmoid-type fibromatosis patients: a nationwide prospective cohort from the French Sarcoma Group, Eur. J. Cancer 83 (2017) 125–131.
[33] P.A. Gondim Teixeira, A. Chanson, J.L. Verhaeghe, et al., Correlation between tumor growth and hormonal therapy with MR signal characteristics of desmoid-type fibromatosis: a preliminary study, Diagn. Interv. Imaging 100 (1) (2019) 47–55.
[34] G. Castellazzi, D. Vanel, A. Le Cesne, et al., Can the MRI signal of aggressive fibromatosis be used to predict its behavior? Eur. J. Radiol. 69 (2) (2009) 222–229.
[35] P.J. Sheth, S. Del Moral, B.A. Wilky, et al., Desmoid fibromatosis: MRI features of response to systemic therapy, Skeletal Radiol. 45 (10) (2016) 1365–1373.
[36] M.R. Cassidy, R.A. Lefkowitz, N. Long, et al., Association of MRI T2 signal intensity with desmoid tumor progression during active observation: a retrospective cohort study, Ann. Surg. 271 (4) (2018) 748–755.

[37] N. Tuncbilek, H.M. Karakas, O.O. Okten, Dynamic contrast enhanced MRI in the differential diagnosis of soft tissue tumors, Eur. J. Radiol. 53 (3) (2005) 500–505.
[38] K. Oka, T. Yakushiji, H. Sato, et al., Usefulness of diffusion-weighted imaging for differentiating between desmoid tumors and malignant soft tissue tumors, J. Magn. Reson. Imaging 33 (1) (2011) 189–193.
[39] M. Khanna, S. Ramanathan, A.S. Kambal, et al., Multi-parametric (mp) MRI for the diagnosis of abdominal wall desmoid tumors, Eur. J. Radiol. 92 (2017) 103–110.
[40] G. Litjens, T. Kooi, B.E. Bejnordi, et al., A Survey on Deep Learning in Medical Image Analysis, 42, 2017, pp. 60–88.