cancers

MDPI

Article

Development and External Validation of Deep-Learning-Based Tumor Grading Models in Soft-Tissue Sarcoma Patients Using MR Imaging

Fernando Navarro [1,2,3], Hendrik Dapper [1], Rebecca Asadpour [1], Carolin Knebel [4], Matthew B. Spraker [5], Vincent Schwarze [6], Stephanie K. Schaub [7], Nina A. Mayr [7], Katja Specht [8], Henry C. Woodruff [9,10], Philippe Lambin [9,10], Alexandra S. Gersing [6,11], Matthew J. Nyflot [7,12], Bjoern H. Menze [2,13], Stephanie E. Combs [1,13,14] and Jan C. Peeken [1,10,14,15,*]

[1] Department of Radiation Oncology, Klinikum Rechts der Isar, Technical University of Munich (TUM), Ismaninger Straße 22, 81675 Munich, Germany; fernando.navarro@tum.de (F.N.); hendrik.dapper@mri.tum.de (H.D.); rebecca.asadpour@tum.de (R.A.); stephanie.combs@tum.de (S.E.C.)

[2] Department of Informatics, Technical University of Munich (TUM), Boltzmannstr. 3, 85748 Garching, Germany; bjoern.menze@tum.de

[3] TranslaTUM—Central Institute for Translational Cancer Research, Einsteinstraße 25, 81675 Munich, Germany

[4] Department of Orthopedics and Sports Orthopedics, Klinikum Rechts der Isar, Technical University of Munich (TUM), Ismaninger Straße 22, 81675 Munich, Germany; carolin.knebel@mri.tum.de

[5] Department of Radiation Oncology, Washington University in St. Louis, 4511 Forest Park Ave, St. Louis, MO 63108, USA; mspraker@wustl.edu

[6] Department of Radiology, Grosshadern Campus, Ludwig-Maximilians-University Munich, Marchioninistraße 15, 81377 Munich, Germany; vincent.schwarze@med.uni-muenchen.de (V.S.); alexandra.gersing@tum.de (A.S.G.)

[7] Department of Radiation Oncology, University of Washington, 1959 NE Pacific St, 356043, Seattle, WA 98195, USA; skschaub@uw.edu (S.K.S.); ninamayr@uw.edu (N.A.M.); nyflot@uw.edu (M.J.N.)

[8] Department of Pathology, Technical University of Munich (TUM), Trogerstr. 18, 81675 Munich, Germany; katja.specht@tum.de

[9] Department of Precision Medicine, GROW—School for Oncology and Developmental Biology, Maastricht University, Universiteitssingel 40, 6229 ER Maastricht, The Netherlands; h.woodruff@maastrichtuniversity.nl (H.C.W.); philippe.lambin@maastrichtuniversity.nl (P.L.)

[10] Department of Radiology and Nuclear Imaging, GROW—School for Oncology and Developmental Biology, P. Debyelaan 25, 6229 HX Maastricht, The Netherlands

[11] Department of Radiology, Klinikum rechts der Isar, Technical University of Munich (TUM), Ismaninger Straße 22, 81675 Munich, Germany

[12] Department of Radiology, University of Washington, 4245 Roosevelt Way NE, Seattle, WA 98105, USA

[13] Department for Quantitative Biomedicine, University of Zurich, Winterthurerstrasse 190, CH-8057 Zurich, Switzerland

[14] Institute of Radiation Medicine (IRM), Department of Radiation Sciences (DRS), Ingolstaedter Landstr. 1, 85764 Munich, Germany

[15] Deutsches Konsortium für Translationale Krebsforschung (DKTK), Partner Site, 85764 Munich, Germany

* Correspondence: jan.peeken@tum.de; Tel.: +49-8941404501

Citation: Navarro, F.; Dapper, H.; Asadpour, R.; Knebel, C.; Spraker, M.B.; Schwarze, V.; Schaub, S.K.; Mayr, N.A.; Specht, K.; Woodruff, H.C.; et al. Development and External Validation of Deep-Learning-Based Tumor Grading Models in Soft-Tissue Sarcoma Patients Using MR Imaging. *Cancers* **2021**, *13*, 2866. https://doi.org/10.3390/cancers13122866

Academic Editors: Dirk Clevert and Matthias F. Frölich

Received: 26 April 2021
Accepted: 2 June 2021
Published: 8 June 2021

Simple Summary: In soft-tissue sarcoma (STS) patients, the decision for the optimal treatment modality largely depends on STS size, location, and a pathological measure that assesses tumor aggressiveness called "tumor grading". To determine tumor grading, invasive biopsies are needed before therapy. In previous research studies, quantitative imaging features ("radiomics") have been associated with tumor grading. In this work, we assessed the possibility of predicting tumor grading using an artificial intelligence technique called "deep learning" or "convolutional neural networks". By analyzing either T1-weighted or T2-weighted MRI sequences, non-invasive tumor grading prediction was possible in an independent test patient cohort. The results were comparable to previous research work obtained with radiomics; however, the reproducibility of the contrast-enhanced T1-weighted sequence was improved. The T2-based model was also able to significantly identify patients with a high risk for death after therapy.

Abstract: Background: In patients with soft-tissue sarcomas, tumor grading constitutes a decisive factor to determine the best treatment decision. Tumor grading is obtained by pathological work-up after focal biopsies. Deep learning (DL)-based imaging analysis may pose an alternative way to characterize STS tissue. In this work, we sought to non-invasively differentiate tumor grading into low-grade (G1) and high-grade (G2/G3) STS using DL techniques based on MR-imaging. Methods: Contrast-enhanced T1-weighted fat-saturated (T1FSGd) MRI sequences and fat-saturated T2-weighted (T2FS) sequences were collected from two independent retrospective cohorts (training: 148 patients, testing: 158 patients). Tumor grading was determined following the French Federation of Cancer Centers Sarcoma Group in pre-therapeutic biopsies. DL models were developed using transfer learning based on the DenseNet 161 architecture. Results: The T1FSGd and T2FS-based DL models achieved area under the receiver operator characteristic curve (AUC) values of 0.75 and 0.76 on the test cohort, respectively. T1FSGd achieved the best F1-score of all models (0.90). The T2FS-based DL model was able to significantly risk-stratify for overall survival. Attention maps revealed relevant features within the tumor volume and in border regions. Conclusions: MRI-based DL models are capable of predicting tumor grading with good reproducibility in external validation.

Keywords: deep learning; convolutional neural networks; artificial intelligence; machine learning; soft-tissue sarcomas; tumor grading; MRI

1. Introduction

Soft-tissue sarcomas (STS) constitute a rare cancer type [1]. Risk stratification is primarily performed using tumor location, pathological tumor grading, tumor size, and certain histological subtypes [2]. One of the most decisive factors constitutes tumor grading. Two separate grading systems were originally defined by the French Federation of Cancer Centers Sarcoma Group (FNCLCC) and the National Cancer Institute (NCI) [3,4]. The FNCLCC system, however, showed better predictive values for distant metastases and is used predominantly worldwide [5]. While FNCLCC G1 (termed "low-grade") STS are generally treated with surgery alone, FNCLCC G2/G3 (termed "high-grade") STS require multimodal therapy regimens involving radiotherapy and/or chemotherapy [6–8]. Despite treatment intensifications, the overall outcome remains poor for high-grade STS [9–11].

Quantitative imaging constitutes an alternative method to characterize tissues. In contrast to a focal biopsy sample, image analysis is capable of assessing the whole tumor volume and can enable longitudinal assessment. In recent years, two general analysis methods have been developed which are summarized under the term "radiomics" [12]. First, predefined handcrafted features are extracted by analyzing the tumor's shape, intensity distribution, and texture. Afterwards, machine learning models are applied to predict clinical endpoints [13–16]. Second, approaches such as neural networks can be specifically trained to directly analyze imaging data to make end-to-end predictions [17]. Convolutional neural networks (CNNs) describe a class of architectures that are especially suited for image analysis and that are, among others, often referred to by the terms "deep learning" (DL) or "artificial intelligence" (AI). In DL, the systems can be categorized into supervised, unsupervised, and semi-supervised learning according to their learning strategy. In this work, we used deep supervised learning, where the neural network requires annotations to learn discriminative features directly from the images without need of extra information. Since the input features are the raw images, there is no need for feature extraction or feature selection as in traditional machine learning. Both techniques (traditional machine learning and DL) have been shown to predict prognosis, tumor progression, molecular aberrations, or spatial infiltration in various cancer subtypes [18–24]. Some studies found superior predictive performances using CNNs compared to handcrafted features [25,26]. Radiomics-based approaches also enable localization and segmentation of volumes of interest (VOI) [27,28].

In STS patients, multiple groups previously demonstrated the possibilities of radiomics and DL to predict patients' prognosis based on MRI, CT, and PET imaging [29–35]. Wang et al. developed radiomic models to differentiate malignant and benign soft-tissue lesions [36]. Further research studies evaluated the differentiation of high-grade from low-grade STS based on MRI and CT imaging scans using radiomic analysis [37–42]. No study has yet analyzed the possibility of DL-based tumor grading prediction.

The scope of this study was to evaluate the potential of DL to predict tumor grading based on pre-therapeutic MRI scans. The value of T2-weighted fat-saturated (T2FS) MRI sequences was compared to contrast-enhanced and fat-saturated T1-weighted (T1FSGd) MRI sequences. All models were externally validated and tested for significant patient risk stratification. Attention maps were generated to evaluate relevant qualitative imaging features and increase explainability of the developed models.

2. Materials and Methods

2.1. Patients

Two independent consecutive patient cohorts from the Technical University of Munich, Munich, Germany (TUM) and the University of Washington, Seattle, WA, USA (UW) were collected retrospectively. Inclusion criteria included: histologically proven STS with available FNCLCC tumor grading information. Exclusion criteria were endoprosthesis-dependent MRI artifacts, previous radiotherapy, primary bone sarcomas, or Ewing sarcomas. Patient records were analyzed for FNCLCC tumor grading and basic patient demographics. The patient cohort with a higher balance between low-grade and high-grade STS was selected for training (TUM). In the training cohort, for each sequence all available patients were included (T1FSGd: 148 patients, T2FS: 130 patients). To allow a better comparison in the test set (UW), all patients that did not have both MRI sequences were excluded (final test set: 158 patients).

See Figure S1 for a patient workflow. In the final patient cohort, no modeling-specific data were missing. Overall survival (OS) was calculated from the initial pathologic diagnosis to the time point of death or the time point of censoring. Data reporting follows the STARD recommendations (Table S1: STARD checklist) [40].

2.2. Image Acquisition, Definition of Volumes of Interest and Preprocessing

Pre-therapeutic MRI scans were analyzed for each included patient. See Table S2 for acquisition parameters and scan planes. For all STS, tumor segmentation was performed using Eclipse 13.0 (Varian Medical Systems, 3100 Hansen Way, Palo Alto, CA 94304, USA), MIM software version 6.6 (MIM Software Inc., 25800 Science Park Dr #180, Beachwood, OH 44122, USA), iplan RT 4.1.2 (Brainlab, Olof-Palme-Straße 9, 81829 Munich, Germany), and 3D Slicer (3D Slicer, version 4.8 stable release). The primary tumor as the VOI was manually segmented by JCP, by adapting existing expert segmentations from RT treatment planning in the TUM cohort. In the UW cohort, segmentation was performed by MBS, MM, JCP, and TC. Edematous changes were not included in the VOI. N4ITK MRI bias field correction was applied to each imaging study using the Slicer3D implementation to compensate for non-uniform intensity caused by field inhomogeneity [43].

2.3. Data Preprocessing

All volumes were resampled to 1 mm^3 isotropic resolution and normalized using z-score normalization. From the 3D VOI, transversal 2D slices were obtained and resized to 224 × 244 before sending the images to the deep neural networks, according to the requirements of the pre-trained architecture for the 2D model. Obtaining transversal slices from one patient allowed us to increase the number of training samples for the deep neural networks. This means that from every patient in the training set we can generate as many training samples as transversal slices are available from the patient tumor. When counting the overall number of training samples, we can then go from hundreds in the original MRI data to thousands after slicing the patient.

2.4. MRI-Based DL Models

We developed DL models to differentiate low-grade (G1) and high-grade (G2/3) STS. For each sequence, a separate DL model was developed: *DL-T1FSGd* and *DL-T2FS*. The base deep learning architecture for this study was based on the ImageNet pre-trained DenseNet-161 described in Figure 1 [44].

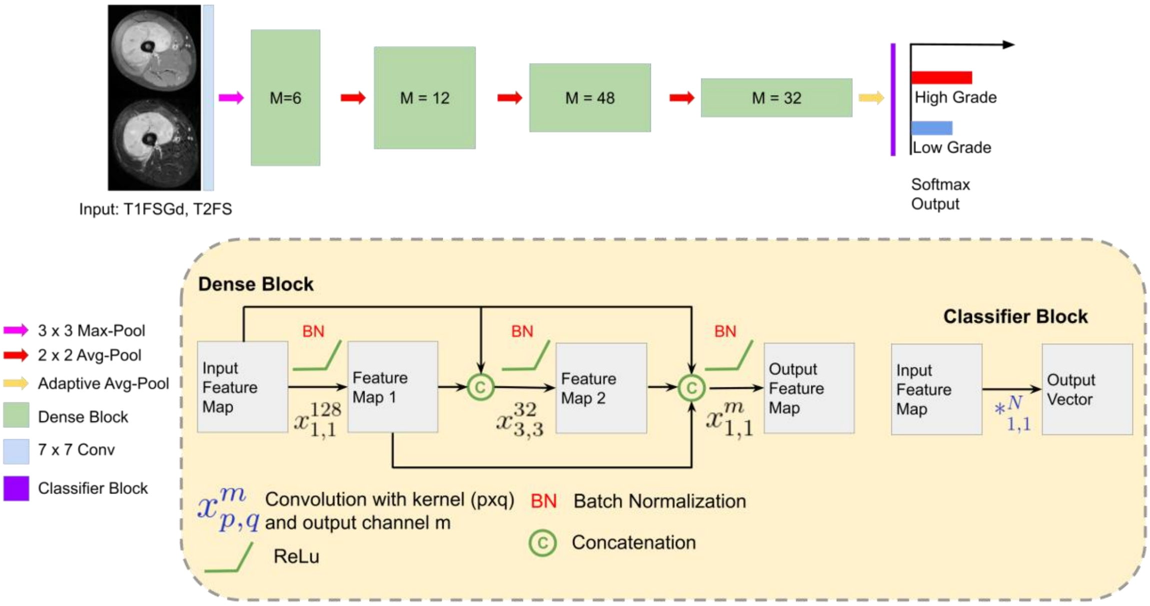

Figure 1. Deep learning strategy: DenseNet 161 architecture for tumor grading in MRI [44]. The network receives the 2D transversal slice from the VOI and outputs the probability of the image for the two classes. In the lower part of the figure, each component of the DenseNet is described. Abbreviations: Avg: Average, T1FSGd: contrast-enhanced and fat-saturated T1-weighted sequence, T2-weighted fat-saturated (T2FS) sequence.

We empirically found that other architectures, including VGGNet [44], ResNet [26], WideResNet [45], AlexNet [46], and CBRNet [47] in 2D and 3D resulted in worse-performing models for our task. DenseNet 161 was the optimal architecture for tumor grading during optimization. Other architectures as well as deeper or shallower pre-trained networks obtained sub-optimal results. Similarly, other approaches such as full MR image without VOI selection, VOI image masked with tumor segmentation, and VOI image and mask as extra channel were tested with inferior performance compared to the proposed approach.

2.5. Optimization of Deep Learning Models

All models were developed in Pytorch with a 12 GB Titan XP [48]. The models were trained with a batch size of 30 and a learning rate of 1×10^{-4} with an ADAM optimizer for 100 epochs. We used early stopping during training, monitoring the validation loss to select the best model. Categorical weighted cross-entropy was used as the loss function. Data augmentation was applied at training time and included vertical and horizontal flip, random rotation, random zoom, elastic transform, and random cropping. Additional training details and the code can be found online (https://github.com/ferchonavarro/SarcomaTumorGrading) (accessed on 4 June 2021).

2.6. Evaluation Strategy

To evaluate the performance, reproducibility, and generalizability of the MRI-based DL models, stratified 5-fold cross-validation with 3 repetitions was performed, producing 15 DL models per image modality. For training and validation of the DL models, the TUM patient cohort was used (referred to as “training cohort”). All 15 models were externally tested using the UW cohort (referred to as “testing cohort”). During inference time, to obtain the tumor grading prediction per patient, the average of all 15 models and all transversal slices in the VOI was computed. Finally, the soft-max activation function converted the average predicted values into the probabilities of low-grade and high-grade STS.

2.7. Interpretability of DL Models

Visualization of attention maps is shown together with the model probabilities to further gain insights on the model predictions for tumor grading. The attention maps were obtained from gradient-weighted class activation maps (Grad-CAM) [49].

2.8. Comparison to Baseline Models

To compare the clinical relevance of the developed models, we compared the DL-based models to regression models using clinical features (TNM T-stage, TNM-N-stage, TNM M-stage, Age) (*Clinical*), tumor volume (*Tumor-Volume*), and the combination of clinical features and tumor volume (*Clinical-Volume-Combined*). The same aforementioned strategy was used for model evaluation.

2.9. Statistical Analysis

Statistical analysis and modeling were performed using Python 3.6. Model performances were characterized using calibration curves, receiver operating characteristic curves (ROC), and additional classification metrics. In addition, 95% confidence intervals were generated using 1000-fold bootstrapping. Kaplan–Meier survival curves were used to analyze model-based stratification for OS in the test set. The maximum argument from the probabilities was used to split patients into low-risk and high-risk patients. Statistical significance was tested using the log-rank test. Bonferroni correction was performed in cases of multiple testing as specified. A p-value below 0.05 was regarded as significant.

3. Results

3.1. Patient Characteristics, Histology, and VOI Definition

Overall, patient demographics were similar (Table 1). However, the distribution of histology subtypes and patients' age was significantly different between both cohorts ($p < 0.001$, $p = 0.03$) (Table S3). Moreover, the training cohort consisted of 35.1% low-grade and 64.9% high-grade STS. The testing cohort showed a more uneven distribution with 14.5% low-grade and 85.5% high-grade STS.

Table 1. Patient demographics and outcome.

Institution	**TUM**	**UW**	***p*-Value** [1]
Total Patients	**148 p**	**158 p**	
Location			1
Extremity or trunk	141/148 p (95.2%)	154/158 p (97.4%)	
Abdomen/retroperitoneal	5/148 p (3.3%)	2/158 p (1.3%)	
Thorax	1/148 p (0.6%)	0/158 p (0%)	
Head and neck	1/148 p (0.6%)	2/158 p (1.3%)	
Age	57.29 ± 17.48	53.91 ± 15.40	0.04 *
Gender			
Female	69/148 p (46.6%)	95/158 p (60.2%)	0.2
Male	79/148 p (53.4%)	63/158 p (39.8%)	
T-Stage			
1	25/148 p (16.8%)	28/158 (17.7%)	0.88
2	123/148 p (83.2%)	130/158 p (83.3%)	
a	13/148 p (8.7%)	6/158 p (3.7%)	0.09
b	135/148 p (91.3%)	152/158 p (96.3%)	
M-Stage			
0	140/148 p (94.6%)	153/158 p (96.8%)	0.40
1	8/148 p (5.4%)	5/158 p (3.2%)	

Table 1. *Cont.*

Institution	TUM	UW	*p*-Value [1]
Total Patients	**148 p**	**158 p**	
N-Stage			
0	145/148 p (98%)	158/158 p (100%)	0.11
1	3/148 p (2%)	0/158 p (0%)	
Grading [2]			0.16
1	52/148 p (35.1%)	25/158 p (15.8%)	
2	36/148 p (24.4%)	53/158 p (33.6%)	
3	60/148 p (40.5%)	80/158 p (50.6%)	
Tumor volume	294.52 ± 442.07	320.0 ± 487.04	0.42
AJCC-Stage [3]			0.47
IA	10/148 p (6.7%)	5/158 p (3.1%)	
IB	42/148 p (28.3%)	20/158 p (12.6%)	
IIA	11/148 p (7.4%)	23/158 p (14.5%)	
IIB	5/148 p (3.3%)	37/158 p (23.4%)	
III	72/148 p (48.6%)	68/158 p (43.0%)	
IV	8/148 p (5.4%)	5/158 p (3.16%)	
Median OS	37.37 mo	45.8 mo	0.25
Available imaging			
T1FsGd	148	158	
T2FS	130	158	

Abbreviations: *: *p*-value < 0.05, AJCC: American Joint Committee on Cancer and the International Union for Cancer Control, m: median, p: patients, r: range, RT: radiation therapy. [1] Wilcoxon rank-sum test for continuous and ordinal variables, Fisher's exact test for nominal variables, log-rank test for comparison of survival times. Corrected for multiple testing by Bonferroni correction ("*p*-value adjusted"). [2] According to the French Federation of Cancer Centers Sarcoma Group (FNCLCC). [3] Following AJCC staging system version 7 [50].

3.2. Classification Performance

The results shown in Figure 2 describe the ROC curves and AUCs for the baseline models and DL-based models classifying patients as low or high-grade STS in the independent test set. It can be observed that for the baseline models (Clinical, *Tumor-Volume*, *Clinical-Volume-Combined*) the obtained AUCs were 0.54, 0.59, and 0.57, respectively. In contrast, the developed DL-based models achieved AUC values of 0.75 and 0.76 for DL-T1FSGd and DL-T2FS, respectively. Table 2 depicts additional classification metrics. All models showed good precision of at least 0.87. DL-T1FSGd classified with the best accuracy of 0.83. This was also reflected by the best sensitivity value of 0.91 but with a suboptimal specificity of 0.40. Delta-T2FS had a better specificity of 0.72 but with the cost of a worse sensitivity value of 0.62, leading to a total accuracy of 0.64. In terms of the less imbalance-biased metric, F1-Score Delta-T1FSGd achieved the best result (0.90). See Figure S2 for calibration curves.

Table 2. Classification metrics for the test set. In bold, the best result among all models for each metric is marked.

	Precision	Sensitivity	Specificity	F1-Score	Accuracy
Clinical	0.87	0.69	0.44	0.77	0.65
Tumor Volume	0.89	0.74	0.52	0.81	0.70
Clinical-Volume-Combined	0.89	0.54	0.64	0.67	0.56
DL-T1FsGd	0.89	**0.91**	0.40	**0.90**	**0.83**
DL-T2Fs	**0.92**	0.62	**0.72**	0.74	0.64

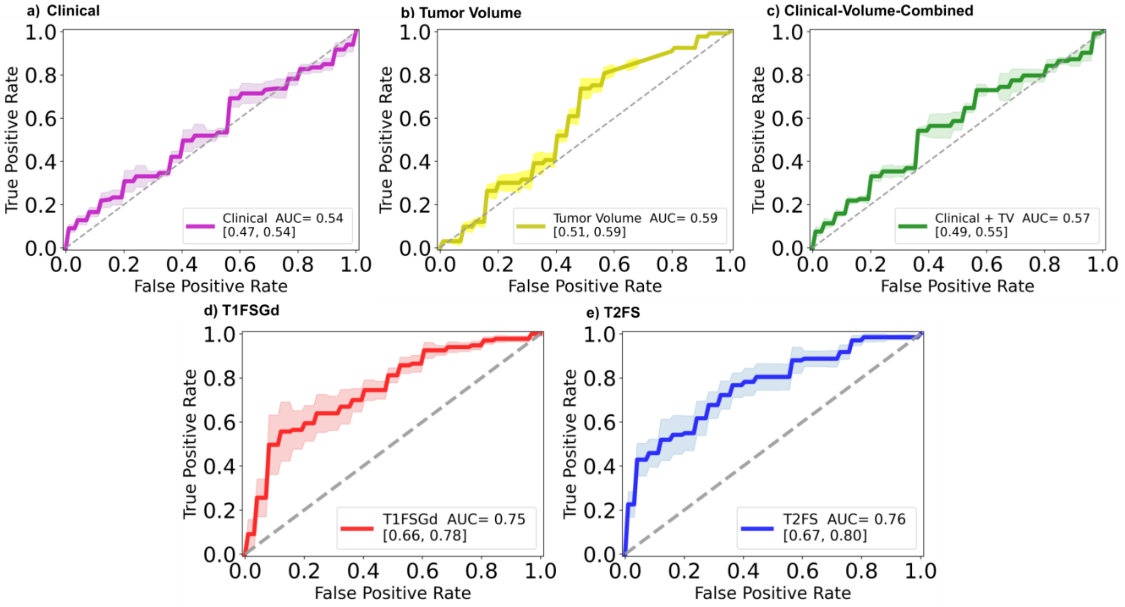

Figure 2. Predictive performance of MRI-based DL models. Receiver operator characteristic curves (ROC) and the respective area under the curve (AUC) values depicting the performance of the prediction models (**a**) *Clinical*, (**b**) *Tumor-Volume*, (**c**) *Clinical-Volume-Combined*, (**d**) *DL-T1FSGd*, and (**e**) *DL-T2FS*.

3.3. Patient Risk Stratification

We used the classification of the developed DL-based grading models for dichotomization of the patient cohort into low-risk and high-risk patients to evaluate the stratification performance for OS. In Figure 3, the Kaplan Meier (KM) survival curves and results of the log-rank test for the baseline models (*Clinical*, *Tumor-Volume*, *Clinical-Volume-Combined*), the ground truth tumor grading stratification (*Grading*), and the DL-based models are shown. *Clinical* and *Grading* achieved significant patient stratification (*p*-value = 0.028 and 0.02, respectively). We also found that both DL-based models separated survival curves into low-risk and high-risk patients. However, only the *DL-T2FS* achieved significant patient stratification (*p*-value = 0.045).

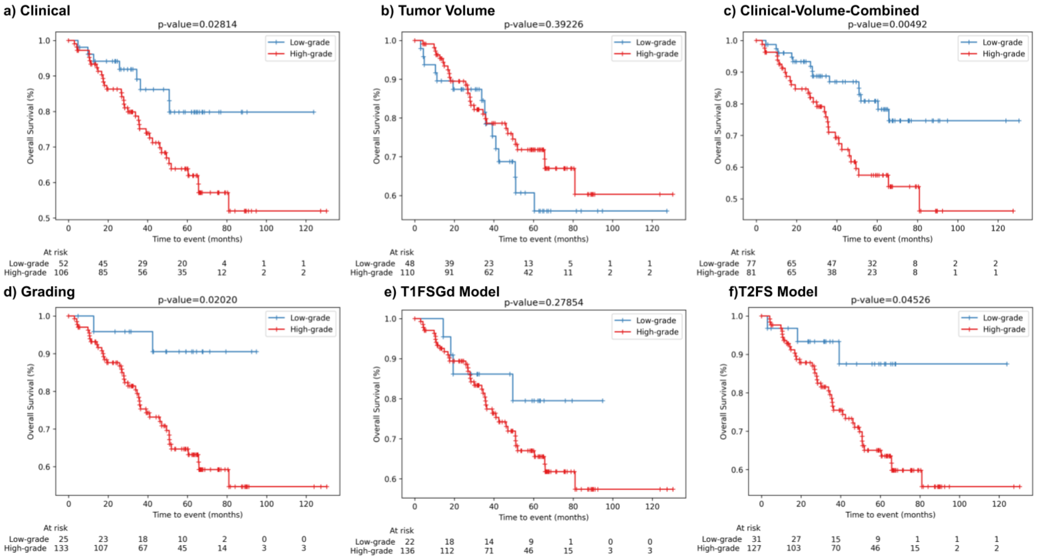

Figure 3. Patient risk stratification in the independent test set. Kaplan Meier survival curves for patients' overall survival displaying risk stratification of the developed models on the test cohort. (**a**) *Clinical*, (**b**) *Tumor-Volume*, (**c**) *Clinical-Volume-Combined*, (**d**) *Grading* (low-grade vs. high-grade), (**e**) *DL-T1FSGd*, and (**f**) *DL-T2FS*. Depicted *p*-values describe the results of the log-rank test.

3.4. Prediction Visualization and Model Interpretability

Figures 4 and 5 depict representative attention maps for the *DL-T1FSGd* and *DL-T2FS* models. Four general patterns can be observed: (1) in many cases, the largest area of activation was present within the tumor volume depicting the tumors' "texture" (e.g., Figure 5a,b); (2) the second most frequent activation was seen in border areas of the tumor focusing on good or bad confinement (e.g., Figure 4a); (3) within border areas the interfaces of tumor to bone and tumor to vessel were frequently represented (Figure 4b,c); (4) in a small number of false cases the network failed to locate the tumor on the cropped image, focusing instead on normal anatomy or air (Figure 5d). Patterns 1–3 were often seen in parallel on the same slice (e.g., Figure 4d) or on different slices of the same tumor.

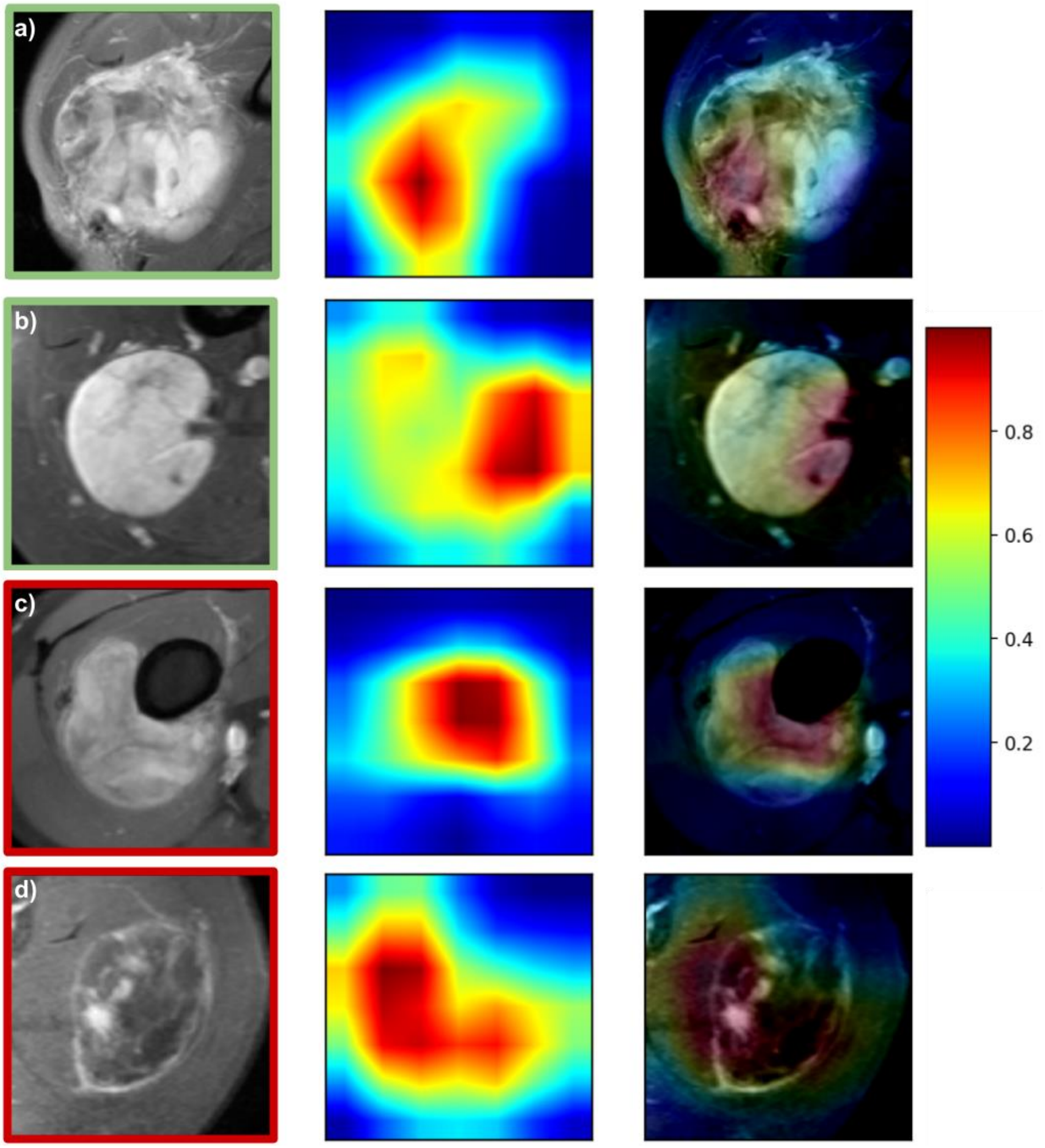

Figure 4. Attention maps of the *DL-T1FSGd* model. Green and red squares around images denote correct and false predictions, respectively: (**a**) correct prediction with 87% probability: high-grade (G2) synovial sarcoma—focus on tumor texture and tumor-tissue border with low confinement; (**b**) correct prediction with 95% probability: low-grade (G1) myxoid liposarcoma—focus on tumor-vessel interface with good confinement. (**c**) False prediction with 87% probability: low-grade (G1) myxoid liposarcoma—focus on tumor-bone interface. (**d**) False prediction with 98% probability: high-grade (G3) pleomorphic sarcoma—focus on central tumor parts and tumor-tissue border, low in-plane resolution.

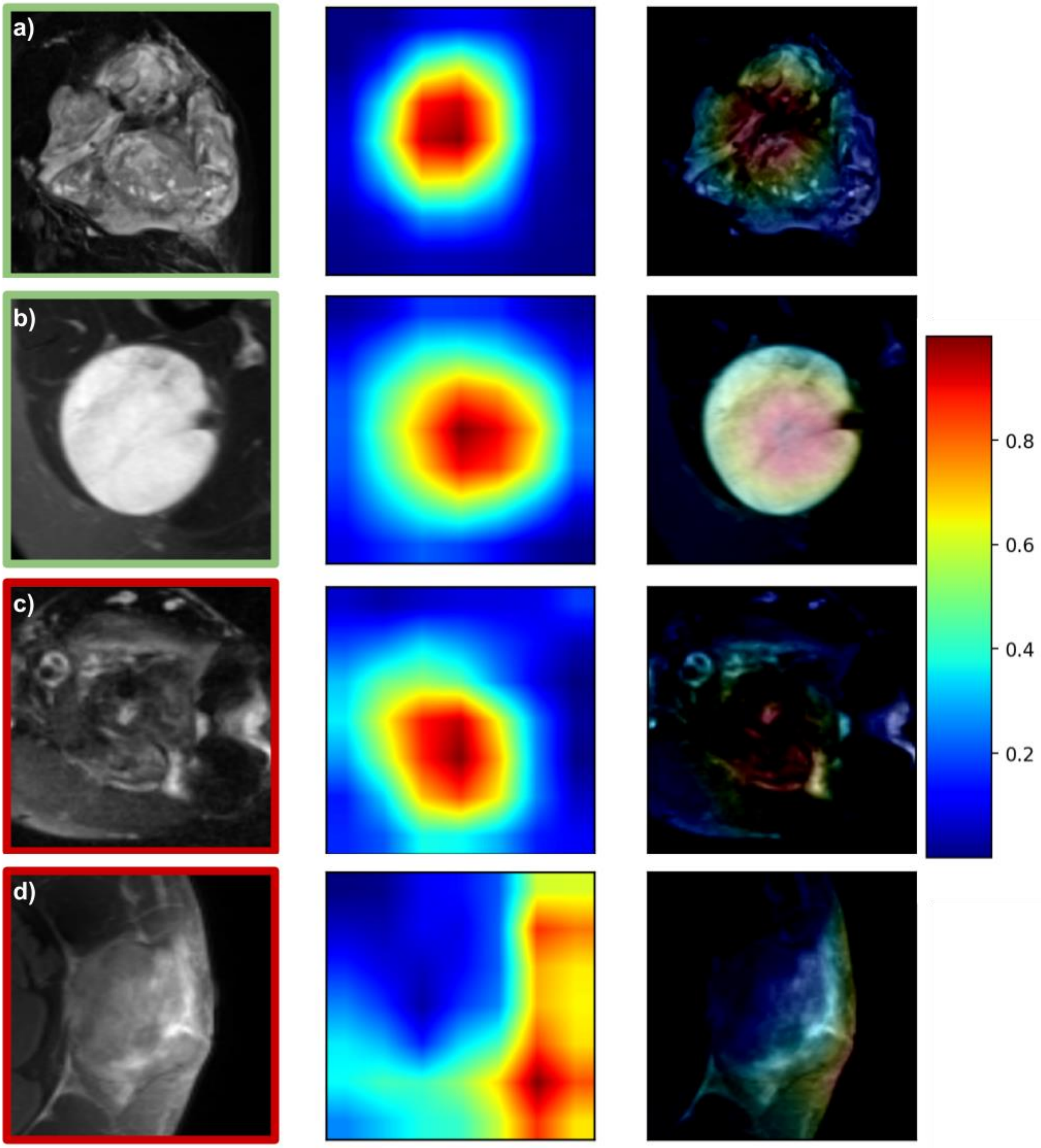

Figure 5. Attention maps of the *DL-T2FS* model. Green and red squares around images denote correct and false predictions, respectively: (**a**) correct prediction with 99% probability: high-grade (G3) spindle cell sarcoma—focus on tumor texture; (**b**) correct prediction with 97% probability: low-grade (G1) myxoid liposarcoma—focus on tumor texture; (**c**) false prediction with 52% probability: low-grade (G1) myofibrosarcoma—focus on tumor texture; (**d**) false prediction with 97% probability: high-grade (G3) pleomorphic sarcoma—trunk location, focus on air.

4. Discussion

In this work, we developed DL-based tumor grading models based on two distinct MRI sequences. The T2FS-based DL model achieved the best predictive performance in an independent testing cohort, comparable to a previously published radiomic model. The contrast-enhanced T1-based model achieved a better performance than a previously published model. The T2-based model was able to significantly risk-stratify STS patients for overall survival. Attention maps confirmed tumor-specific features within and surrounding the tumor volume.

In a previous study, we used similar patient cohorts to develop and externally test radiomics-based tumor grading models [37]. For T2FS with AUC values of 0.76 (DL) and 0.78 (handcrafted features), the predictive performance was comparable, although with a slightly higher performance of the handcrafted feature model in the test set. For T1FSGd, the DL model showed a higher performance with an AUC of 0.75 while the handcrafted feature model achieved an AUC of 0.69. It should be noted that both cohorts have been expanded since then. The skewed proportion of low-grade and high-grade STS, however, remained similar. The patient numbers in the training set size were enlarged by 6% and 20%, and in the test set by 53% and 53% for T2FS and T1FSGd, respectively. The training set size also played a role in the comparison of our DL models. To allow direct comparability, we selected only patients for the test set that had both imaging studies available. *DL-T2FS* had a 12% smaller training sample number than the *DL-T1FSGd* model. *DL-T2FS* achieved a higher AUC but worse classification performance (e.g., F1-Score) than *DL-T1FSGd*. A previous study, however, showed a correlation between DL model performance and training sample size on a logarithmic scale [51]. Thus, for significant model performance improvements, much larger differences in training size would be beneficial and a large impact of the small differences in training size is rather unlikely.

As previously mentioned, other authors have evaluated tumor grading prediction using MRI-based radiomics [38–42]. However, only one study validated their models in an external testing cohort [40]. In this study, Yan et al. used a training cohort of 109 patients to develop radiomic models based on T2FS and T1-weighted MRI sequences (without contrast-enhancement). In the 70-patient test set, both models achieved predictive performances with AUCs of 0.645 (T2FS) and 0.641 (T1). Combining both features significantly increased the performance up to an AUC of 0.829. In contrast, our study used contrast-enhanced fat-saturated T1-weighted MRI scans. Both developed models had better performances than the single sequence models but were inferior to the combined model, although in a similar range. Interestingly, an additive benefit following a combination of the radiomic feature sets of both sequences (T1FSGd and T2FS) could not be observed in our previous study. However, the testing cohort was significantly smaller, increasing the chance-based risk of falsely optimistic or pessimistic results. Moreover, it had a more balanced distribution of low-grade and high-grade STS. This may also explain the lack of significant patient risk stratification of the combined model by Yan et al. Still, combining multiple imaging modalities for DL models remains a promising approach.

The attention map analysis gave insights into the functioning of the DL models. This allows a certain amount of semantic explainability which cannot be derived for models based on handcrafted features. In many cases, the DL model focused on the internal texture structures of the STS. This may correspond to features implemented in the previously published radiomic models that were always restricted to the gross tumor volume as VOI. At the same time, it may represent semantic imaging features, such as necrosis, that have previously been linked with tumor grading [52]. Interestingly, our DL models also regularly focused on tumor-surrounding tissue, reflecting, e.g., the confinement of the tumor-tissue border. In accordance, another semantic feature, "peritumoral enhancement", has previously been described as being correlated with tumor grading [52]. Further work is needed to evaluate associations between attention maps and semantic features.

In a select number of cases the model did not correctly locate the tumor but instead focused on unrelated areas (e.g., air), restricting a reliable prediction. These cases predominantly occurred in rare anatomic locations (e.g., location at the trunk in Figure 5d) and constitute a limitation when extending the cropped images beyond the VOI. By increasing future training sample sizes, or excluding rare anatomical sites, the resulting models might learn to perform better. By providing the attention maps alongside each prediction, the physician could directly assess the technical reliability of the prediction. A future direction to use attention maps could be to objectively identify regions of concern for a high risk of positive margins as well as potential internal sub-volumes of high-grade histology that might inform design of future risk-adaptive, precision clinical trials for spatial intensification of therapies.

As in many STS studies, a large plethora of histologies was combined to achieve significantly large patient cohorts. As these different subtypes stem from different mesenchymal tissue types, one could speculate that histology-specific models may be more effective in predicting histology-specific tumor grading. Sub-cohorts of patients with relevant histological groups such as pleomorphic sarcomas or with dominant myxoid or fibrous matrix comprise only 19–56 patients in the training set. Previous research demonstrated a significant decrease in classification performance below 100 samples [53]. This would further be aggravated by the cross-validation approach, a low event-rate, and missing imaging scans. As a consequence, no histology-specific models could be effectively trained using the underlying patient cohort. We are currently working on extending our international collaborations to allow histology-specific models in the future.

The cohorts used in this study were retrospectively gathered from two different medical centers. For the training cohort, patients were treated in the department of radiation oncology and the department of orthopedic surgery which led to a relatively high number of low-grade STS. The testing set was derived only from a radiation oncology department, leading to an overall lower number of low-grade STS. Overly aggressive STS with a metastatic state may thus be underrepresented at first diagnosis. As a consequence, in the future, non-invasive grading models should be tested in less biased cohorts.

This work bears several limitations. Both study cohorts were collected retrospectively, constituting a reason for a potential source of bias as described above [54]. Due to the multicentric setting, the patient cohorts presented a large technical heterogeneity, including different imaging protocols and MRI scanner types. Despite this heterogeneity, successful reproduction of CNN models was possible, showing effective generalizability. In our work, we compared the performance of two imaging sequences. To ensure a maximum amount of information, we used all available imaging studies per sequence leading to slightly different sizes of the training set. Relative underrepresentation in the training set of T2FS may have impaired a better classification performance. Moreover, due to sequence availability from both centers, our analysis was restricted to only fat-saturated MRI sequences. As fat-dependent signals constitute important semantic features, other sequences such as T2-weighted could provide complementary information.

5. Conclusions

In conclusion, we demonstrated that both MRI-based DL models were able to classify tumor grading in soft-tissue sarcoma patients. Attention maps can provide insight into semantic imaging features relevant for model classification and can function as a valuable tool for patient-specific quality assurance. Further investigation is warranted to establish imaging-based biomarkers for non-invasive STS characterization.

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/cancers13122866/s1, Figure S1 Patients Workflow, Figure S2 Calibration Curves, Table S1 STARD Checkliste, Table S2 MRI acquisition parameters, Table S3 Histologies of Soft-Tissue Sarcomas.

Author Contributions: Conceptualization, J.C.P., F.N., H.C.W., P.L., M.J.N., A.S.G., and S.E.C.; methodology, F.N., J.C.P.; software, F.N.; validation, M.J.N.; formal analysis, J.C.P., F.N., R.A., H.D., M.B.S., K.S., V.S.; investigation, J.C.P., F.N.; resources, M.J.N., N.A.M., S.E.C., B.H.M.; data curation, J.C.P., H.D., R.A., F.N.; writing—original draft preparation, F.N., J.C.P.; writing—review and editing, M.J.N., N.A.M., C.K., A.S.G., H.D., H.C.W., P.L., S.E.C., S.K.S., V.S.; visualization, F.N., J.C.P.; supervision, H.C.W., P.L., M.J.N., S.E.C., B.H.M.; project administration, J.C.P.; funding acquisition, J.C.P., M.J.N., S.E.C., B.H.M. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by physician-scientist program of the medical faculty of the Technical University of Munich and the Helmholtz Zentrum Muenchen (JCP). We also gratefully acknowledge the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—GRK for the financial support (FN). The authors gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki and approved by the Institutional Review Board (or Ethics Committee) of the Technical University of Munich (protocol code 466/16s and date of approval: 28 March 2019).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data presented in this study are available on request from the corresponding author dependent on ethics board approval. The data are not publicly available due to data protection legislation.

Acknowledgments: We sincerely thank Chapman T and Macomber M for segmentations of VOIs at UW.

Conflicts of Interest: PL reports, within the submitted work, grants/sponsored research agreements from Radiomics SA; PL has minority shares in the company Radiomics SA, he is co-inventor of two issued patents with royalties on radiomics (PCT/NL2014/050248, PCT/NL2014/050728) licensed to Radiomics SA, one non-patented invention (software) licensed to Radiomics SA, and two non-issues, non-licensed patents on Deep Learning—Radiomics (n° 17123816, PCT/NL/2020/050794). PL confirms that none of the above entities or funding was involved in the preparation of this paper. HW has minority shares in the company Radiomics SA. The other authors declare no potential conflicts of interest.

References

1. Gutierrez, J.C.; Perez, E.A.; Franceschi, D.; Moffat, F.L.; Livingstone, A.S.; Koniaris, L.G. Outcomes for soft-tissue sarcoma in 8249 cases from a large state cancer registry. *J. Surg. Res.* **2007**, *141*, 105–114. [CrossRef]
2. Callegaro, D.; Miceli, R.; Bonvalot, S.; Ferguson, P.; Strauss, D.C.; Levy, A.; Griffin, A.; Hayes, A.J.; Stacchiotti, S.; Pechoux, C.L.; et al. Development and external validation of two nomograms to predict overall survival and occurrence of distant metastases in adults after surgical resection of localised soft-tissue sarcomas of the extremities: A retrospective analysis. *Lancet Oncol.* **2016**, *17*, 671–680. [CrossRef]
3. Costa, J.; Wesley, R.A.; Glatstein, E.; Rosenberg, S.A. The grading of soft tissue sarcomas. Results of a clinicohistopathologic correlation in a series of 163 cases. *Cancer* **1984**, *53*, 530–541. [CrossRef]
4. Trojani, M.; Contesso, G.; Coindre, J.M.; Rouesse, J.; Bui, N.B.; de Mascarel, A.; Goussot, J.F.; David, M.; Bonichon, F.; Lagarde, C. Soft-tissue sarcomas of adults; study of pathological prognostic variables and definition of a histopathological grading system. *Int. J. Cancer* **1984**, *33*, 37–42. [CrossRef]
5. Guillou, L.; Coindre, J.M.; Bonichon, F.; Nguyen, B.B.; Terrier, P.; Collin, F.; Vilain, M.O.; Mandard, A.M.; Le Doussal, V.; Leroux, A.; et al. Comparative study of the National Cancer Institute and French Federation of Cancer Centers Sarcoma Group grading systems in a population of 410 adult patients with soft tissue sarcoma. *J. Clin. Oncol.* **1997**, *15*, 350–362. [CrossRef] [PubMed]
6. Gerrand, C.H.; Rankin, K. The treatment of soft-tissue sarcomas of the extremities. Prospective randomized evaluations of (1) limb-sparing surgery plus radiation therapy compared with amputation and (2) the role of adjuvant chemotherapy. *Class. Pap. Orthop.* **2014**, 483–484. [CrossRef]
7. Koshy, M.; Rich, S.; Mohiuddin, M. Improved survival with radiation therapy in high grade soft tissue sarcomas of the extremities: A SEER analysis. *Int. J. Radiat. Oncol. Biol. Phys.* **2013**, *77*, 1–15. [CrossRef] [PubMed]
8. O'Connor, J.M.; Chacón, M.; Petracci, F.E.; Chacón, R.D. Adjuvant chemotherapy in soft tissue sarcoma (STS): A meta-analysis of published data. *J. Clin. Oncol.* **2008**, *26*, 10526. [CrossRef]
9. Alektiar, K.M.; Brennan, M.F.; Healey, J.H.; Singer, S. Impact of intensity-modulated radiation therapy on local control in primary soft-tissue sarcoma of the extremity. *J. Clin. Oncol.* **2008**, *26*, 3440–3444. [CrossRef] [PubMed]

10. Muehlhofer, H.M.L.; Schlossmacher, B.; Lenze, U.; Lenze, F.; Burgkart, R.; Gersing, A.S.; Peeken, J.C.; Combs, S.E.; Von Eisenhart-Rothe, R.; Knebel, C. Oncological outcome and prognostic factors of surgery for soft tissue sarcoma after neoadjuvant or adjuvant radiation therapy: A retrospective analysis over 15 years. *Anticancer Res.* **2021**, *41*, 359–368. [CrossRef] [PubMed]
11. Peeken, J.C.; Knie, C.; Kessel, K.A.; Habermehl, D.; Kampfer, S.; Dapper, H.; Devecka, M.; Von Eisenhart-rothe, R.; Specht, K.; Weichert, W.; et al. Neoadjuvant image-guided helical intensity modulated radiotherapy of extremity sarcomas—A single center experience. *Radiat. Oncol.* **2019**, *14*, 4–11. [CrossRef]
12. Nyflot, M.J.; Thammasorn, P.; Wootton, L.S.; Ford, E.C.; Chaovalitwongse, W.A. Deep learning for patient-specific quality assurance: Identifying errors in radiotherapy delivery by radiomic analysis of gamma images with convolutional neural networks. *Med. Phys.* **2019**, *46*, 456–464. [CrossRef] [PubMed]
13. Peeken, J.C.; Nüsslin, F.; Combs, S.E. "Radio-oncomics"—The potential of radiomics in radiation oncology. *Strahlenther. Onkol.* **2017**, *193*, 767–779. [CrossRef] [PubMed]
14. Peeken, J.C.; Wiestler, B.; Combs, S.E. The potential of radiomics in clinical application. In *Image Guided Radiooncology*; Debus, J., Schober, O., Kiessling, F., Eds.; Springer: Berlin/Heidelberg, Germany, 2020.
15. Peeken, J.C.; Bernhofer, M.; Wiestler, B.; Goldberg, T.; Cremers, D.; Rost, B.; Wilkens, J.J.; Combs, S.E.; Nüsslin, F. Radiomics in radiooncology—Challenging the medical physicist. *Phys. Med.* **2018**, *48*, 27–36. [CrossRef] [PubMed]
16. Lambin, P.; Rios-Velazquez, E.; Leijenaar, R.; Carvalho, S.; van Stiphout, R.G.P.M.; Granton, P.; Zegers, C.M.L.; Gillies, R.; Boellard, R.; Dekker, A.; et al. Radiomics: Extracting more information from medical images using advanced feature analysis. *Eur. J. Cancer* **2012**, *48*, 441–446. [CrossRef] [PubMed]
17. Litjens, G.; Kooi, T.; Bejnordi, B.E.; Setio, A.A.A.; Ciompi, F.; Ghafoorian, M.; van der Laak, J.A.W.M.; van Ginneken, B.; Sánchez, C.I. A survey on deep learning in medical image analysis. *Med. Image Anal.* **2017**, *42*, 60–88. [CrossRef]
18. Aerts, H.J.W.L.; Velazquez, E.R.; Leijenaar, R.T.H.; Parmar, C.; Grossmann, P.; Carvalho, S.; Cavalho, S.; Bussink, J.; Monshouwer, R.; Haibe-Kains, B.; et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat. Commun.* **2014**, *5*, 4006. [CrossRef]
19. Rios Velazquez, E.; Parmar, C.; Liu, Y.; Coroller, T.P.; Cruz, G.; Stringfield, O.; Ye, Z.; Makrigiorgos, M.; Fennessy, F.; Mak, R.H.; et al. Somatic mutations drive distinct imaging phenotypes in lung cancer. *Cancer Res.* **2017**, *77*, 3922–3930. [CrossRef]
20. Diehn, M.; Nardini, C.; Wang, D.S.; McGovern, S.; Jayaraman, M.; Liang, Y.; Aldape, K.; Cha, S.; Kuo, M.D. Identification of noninvasive imaging surrogates for brain tumor gene-expression modules. *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 5213–5218. [CrossRef]
21. Peeken, J.C.; Shouman, M.A.; Kroenke, M.; Rauscher, I.; Maurer, T.; Gschwend, J.E.; Eiber, M.; Combs, S.E. A CT-based radiomics model to detect prostate cancer lymph node metastases in PSMA radioguided surgery patients. *Eur. J. Nucl. Med. Mol. Imaging* **2020**, *47*, 2968–2977. [CrossRef]
22. Starke, S.; Leger, S.; Zwanenburg, A.; Leger, K.; Lohaus, F.; Linge, A.; Schreiber, A.; Kalinauskaite, G.; Tinhofer, I.; Guberina, N.; et al. 2D and 3D convolutional neural networks for outcome modelling of locally advanced head and neck squamous cell carcinoma. *Sci. Rep.* **2020**, *10*, 15625. [CrossRef]
23. Lang, D.M.; Peeken, J.C.; Combs, S.E.; Wilkens, J.J.; Bartzsch, S. Deep learning based hpv status prediction for oropharyngeal cancer patients. *Cancers* **2021**, *13*, 786. [CrossRef] [PubMed]
24. Peeken, J.C.; Molina-Romero, M.; Diehl, C.; Menze, B.H.; Straube, C.; Meyer, B.; Zimmer, C.; Wiestler, B.; Combs, S.E. Deep learning derived tumor infiltration maps for personalized target definition in Glioblastoma radiotherapy. *Radiother. Oncol.* **2019**, *138*, 166–172. [CrossRef]
25. Truhn, D.; Schrading, S.; Haarburger, C.; Schneider, H.; Merhof, D.; Kuhl, C. Radiomic versus convolutional neural networks analysis for classification of contrast-enhancing lesions at multiparametric breast MRI. *Radiology* **2019**, *290*, 290–297. [CrossRef]
26. Thammasorn, P.; Chaovalitwongse, W.A.; Hippe, D.S.; Wootton, L.S.; Ford, E.C.; Spraker, M.B.; Combs, S.E.; Peeken, J.C.; Nyflot, M.J. Nearest neighbor-based strategy to optimize multi-view triplet network for classification of small-sample medical imaging data. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**. [CrossRef]
27. Navarro, F.; Shit, S.; Ezhov, I.; Paetzold, J.; Gafita, A.; Peeken, J.C.; Combs, S.E.; Menze, B.H. Shape-aware complementary-task learning for multi-organ segmentation. In Proceedings of the MLMI Workshop 2019 Held in Conjunction with MICCAI 2019, Shenzhen, China, 13 October 2019; pp. 620–627.
28. Navarro, F.; Sekuboyina, A.; Waldmannstetter, D.; Peeken, J.C.; Combs, S.E.; Menze, B.H. Deep reinforcement learning for organ localization in CT. *Proc. Mach. Learn. Res.* **2020**, *121*, 544–554.
29. Spraker, M.B.; Wootton, L.S.; Hippe, D.S.; Ball, K.C.; Peeken, J.C.; Macomber, M.W.; Chapman, T.R.; Hoff, M.; Kim, E.Y.; Pollack, S.M.; et al. MRI radiomic features are independently associated with overall survival in soft tissue sarcoma. *Adv. Radiat. Oncol.* **2019**, *4*, 413–421. [CrossRef]
30. Peeken, J.C.; Bernhofer, M.; Spraker, M.B.; Pfeiffer, D.; Devecka, M.; Thamer, A.; Shouman, M.A.; Ott, A.; Nüsslin, F.; Mayr, N.A.; et al. CT-based radiomic features predict tumor grading and have prognostic value in patients with soft tissue sarcomas treated with neoadjuvant radiation therapy. *Radiother. Oncol.* **2019**, *135*, 187–196. [CrossRef]
31. Vallieres, M.; Kumar, A.; Sultanem, K.; El Naqa, I. FDG-PET image-derived features can determine HPV status in head-and-neck cancer. *Int. J. Radiat. Oncol.* **2013**, *87*, S467. [CrossRef]
32. Crombé, A.; Fadli, D.; Buy, X.; Italiano, A.; Saut, O.; Kind, M. High-grade soft-tissue sarcomas: Can optimizing dynamic contrast-enhanced MRI postprocessing improve prognostic radiomics models? *J. Magn. Reson. Imaging* **2020**. [CrossRef]

33. Crombé, A.; Périer, C.; Kind, M.; De Senneville, B.D.; Le Loarer, F.; Italiano, A.; Buy, X.; Saut, O. T2-based MRI Delta-radiomics improve response prediction in soft-tissue sarcomas treated by neoadjuvant chemotherapy. *J. Magn. Reson. Imaging* **2018**. [CrossRef]
34. Crombé, A.; Le Loarer, F.; Sitbon, M.; Italiano, A.; Stoeckle, E.; Buy, X.; Kind, M. Can radiomics improve the prediction of metastatic relapse of myxoid/round cell liposarcomas? *Eur. Radiol.* **2020**, *30*, 2413–2424. [CrossRef]
35. Peeken, J.C.; Neumann, J.; Asadpour, R.; Leonhardt, Y.; Moreira, J.R.; Hippe, D.S.; Klymenko, O.; Foreman, S.C.; Von Schacky, C.E.; Spraker, M.B.; et al. Prognostic assessment in high-grade soft-tissue sarcoma patients: A comparison of semantic image analysis and radiomics. *Cancers* **2021**, *13*, 1929. [CrossRef] [PubMed]
36. Wang, H.; Nie, P.; Wang, Y.; Xu, W.; Duan, S.; Chen, H.; Hao, D.; Liu, J. Radiomics nomogram for differentiating between benign and malignant soft-tissue masses of the extremities. *J. Magn. Reson. Imaging* **2019**. [CrossRef] [PubMed]
37. Peeken, J.C.; Spraker, M.B.; Knebel, C.; Dapper, H.; Pfeiffer, D.; Devecka, M.; Thamer, A.; Shouman, M.A.; Ott, A.; von Eisenhart-Rothe, R.; et al. Tumor grading of soft tissue sarcomas using MRI-based radiomics. *EBioMedicine* **2019**, *48*, 332–340. [CrossRef]
38. Zhang, Y.; Zhu, Y.; Shi, X.; Tao, J.; Cui, J.; Dai, Y.; Zheng, M.; Wang, S. Soft tissue sarcomas: Preoperative predictive histopathological grading based on radiomics of MRI. *Acad. Radiol.* **2018**, *26*, 1262–1268. [CrossRef]
39. Corino, V.D.A.; Montin, E.; Messina, A.; Casali, P.G.; Gronchi, A.; Marchianò, A.; Mainardi, L.T. Radiomic analysis of soft tissues sarcomas can distinguish intermediate from high-grade lesions. *J. Magn. Reson. Imaging* **2017**. [CrossRef] [PubMed]
40. Yan, R.; Hao, D.; Li, J.; Liu, J.; Hou, F.; Chen, H.; Duan, L.; Huang, C.; Wang, H.; Yu, T. Magnetic resonance imaging-based radiomics nomogram for prediction of the histopathological grade of soft tissue sarcomas: A two-center study. *J. Magn. Reson. Imaging* **2021**. [CrossRef] [PubMed]
41. Xu, W.; Hao, D.; Hou, F.; Zhang, D.; Wang, H. Soft tissue sarcoma: Preoperative MRI-based radiomics and machine learning may be accurate predictors of histopathologic grade. *Am. J. Roentgenol.* **2020**, *215*, 963–969. [CrossRef]
42. Wang, H.; Chen, H.; Duan, S.; Hao, D.; Liu, J. Radiomics and machine learning with multiparametric preoperative MRI may accurately predict the histopathological grades of soft tissue sarcomas. *J. Magn. Reson. Imaging* **2020**, *51*, 791–797. [CrossRef]
43. Tustison, N.J.; Gee, J.C. N4ITK: Nick's N3 ITK implementation for MRI bias field correction. *Insight J.* **2009**, 1–8. Available online: http://hdl.handle.net/10380/3053 (accessed on 4 June 2017).
44. Huang, G.; Liu, Z.; van der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks Gao. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2017, Honolulu, HI, USA, 21 July–26 July 2017; pp. 4700–4708.
45. Zagoruyko, S.; Komodakis, N. Wide residual networks. *arXiv* **2016**, arXiv:1605.07146.
46. Krizhevsky, A.; Sutskever, I.; Hinton, G. *ImageNet Classification with Deep Convolutional Neural Networks*; Chapman and Hall/CRC: Boca Raton, FL, USA, 2012; ISBN 9780429143793.
47. Raghu, M.; Zhang, C.; Kleinberg, J.; Bengio, S. Transfusion: Understanding transfer learning for medical imaging. *arXiv* **2019**, arXiv:1902.07208.
48. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. PyTorch: An imperative style, high-performance deep learning library. *arXiv* **2019**, arXiv:1912.01703.
49. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22 October–29 October 2017; pp. 618–626.
50. Edge, S.B.; Compton, C.C. The American joint committee on cancer: The 7th edition of the AJCC cancer staging manual and the future of TNM. *Ann. Surg. Oncol.* **2010**, *17*, 1471–1474. [CrossRef] [PubMed]
51. Sun, C.; Shrivastava, A.; Singh, S.; Gupta, A. Revisiting unreasonable effectiveness of data in deep learning era. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22 October–29 October 2017; pp. 843–852.
52. Crombé, A.; Marcellin, P.J.; Buy, X.; Stoeckle, E.; Brouste, V.; Italiano, A.; Le Loarer, F.; Kind, M. Soft-tissue sarcomas: Assessment of MRI features correlating with histologic grade and patient outcome. *Radiology* **2019**, *291*, 710–721. [CrossRef] [PubMed]
53. Cho, J.; Lee, K.; Shin, E.; Choy, G.; Do, S. How much data is needed to train a medical image deep learning system to achieve necessary high accuracy? *arXiv* **2015**, arXiv:1511.06348.
54. Sica, G.T. Bias in Research Studies. *Radiology* **2006**, *238*, 780–789. [CrossRef] [PubMed]