Universität des Saarlandes

Fachrichtung Mathematik

Preprint Nr. 393

Benchmarking Wilms' Tumor in Multi-Sequence MRI Data: Why Does Current Clinical Practice Fail? Which Popular Segmentation Algorithms Perform Well?

Sabine Müller, Iva Farag, Joachim Weickert,
Yvonne Braun, Jonas Dobberstein, André Lollert,
Andreas Hötker and Norbert Graf

Saarbrücken 2019

Fachrichtung Mathematik
Universität des Saarlandes

Preprint No. 393
submitted: May 10, 2019

Benchmarking Wilms' Tumor in Multi-Sequence MRI Data: Why Does Current Clinical Practice Fail? Which Popular Segmentation Algorithms Perform Well?

Sabine Müller
Department of Pediatric Oncology and Hematology
Saarland University Medical Center
66421 Homburg, Germany

Mathematical Image Analysis Group
Faculty of Mathematics and Computer Science, Campus E1.7
Saarland University, 66041 Saarbrücken, Germany
smueller@mia.uni-saarland.de

Iva Farag
Department of Pediatric Oncology and Hematology
Saarland University Medical Center
66421 Homburg, Germany
ivabaykova@gmail.com

Joachim Weickert
Mathematical Image Analysis Group
Faculty of Mathematics and Computer Science, Campus E1.7
Saarland University, 66041 Saarbrücken, Germany
weickert@mia.uni-saarland.de

Yvonne Braun
Department of Pediatric Oncology and Hematology
Saarland University Medical Center
66421 Homburg, Germany
yvonne.braun@uks.eu

Jonas Dobberstein
Department of Pediatric Oncology and Hematology
Saarland University Medical Center
66421 Homburg, Germany
jonas.sge.92@hotmail.de

André Lollert
Department of Diagnostic and Interventional Radiology
Medical Center of the Johannes Gutenberg University
55131 Mainz, Germany
andre.lollert@unimedizin-mainz.de

Andreas Hötker
Department of Diagnostic Radiology
University Hospital Zürich
Rämistrasse 100, 8091 Zürich, Switzerland
andreas.hoetker@usz.ch

Norbert Graf
Department of Pediatric Oncology and Hematology
Saarland University Medical Center
66421 Homburg, Germany
graf@uks.eu

Edited by
Fachrichtung Mathematik
Universität des Saarlandes
Postfach 15 11 50
66041 Saarbrücken
Germany

Fax: + 49 681 302 4443
e-Mail: preprint@math.uni-sb.de
WWW: http://www.math.uni-sb.de/

Abstract

Wilms' tumor is one of the most frequent solid and malignant tumors in childhood. Accurate segmentation of tumor tissue is a key step during therapy and treatment planning. Since it is difficult to obtain a comprehensive set of tumor data of children, there is no benchmark so far allowing to evaluate the quality of human or computer-based segmentations. The contributions in our paper are threefold: (i) We present the first heterogeneous Wilms' tumor benchmark data set. It contains multi-sequence MRI data sets before and after chemotherapy, along with ground truth annotation, approximated based on the consensus of five human experts. (ii) We analyze human expert annotations and interrater variability. It turns out that the current clinical practice of determining tumor volume is inaccurate and that manual annotations after chemotherapy may differ substantially. (iii) We evaluate six computer-based segmentation methods, ranging from classical approaches to recent deep learning techniques. We show that the best ones offer a quality comparable to human expert annotations.

1 Introduction

Wilms' tumor, or nephroblastoma, accounts for 5 % of all cancers in children and juveniles. It constitutes the most frequent malignant kidney tumor in childhood [32]. About 75% of all patients are younger than five years - with a peak between two and three years [11, 22]. In Europe, diagnosis and therapy follow the guidelines of the International Society of Pediatric Oncology (SIOP) [15, 21]. One of the most important characteristics of this therapy protocol is a preoperative chemotherapy. Clinicians categorize patients as high-, intermediate- or low-risk candidates according to histology, local stage and tumor volume after chemotherapy. Postoperative treatment ranges from no chemotherapy (low risk stage I) up to chemotherapy with irradiation of the tumor bed (high risk, stage II and III).

The most common histological subtypes of regressive and mixed type actually belong to the intermediate risk tumors. However, if, after chemotherapy, these tumors have a volume of more than 500 ml, they are highly malignant and the patients are treated according to the high risk group protocol [14]. In order not to expose children to unnecessary medical burden on the one hand and to maximize their chances of survival on the other, an exact determination of the tumor volume is indispensable.

Current Practices of Segmentation by Human Experts. Radiologists traditionally model the tumor through a time-intensive manual segmentation procedure involving the outlining of the gross tumor volume on nu-

merous two-dimensional imaging "slices". Alternatively, they estimate the tumor volume by measuring three axes of tumor extension and assuming the nephroblastoma to have an ellipsoid shape [14]. Usually both variants are conducted using either computed tomography (CT) or magnetic resonance imaging (MRI) data. The reliability and consistent reproducibility of expert delineations of Wilms' tumors has not been investigated so far.

Computer-based Segmentation Algorithms. One obvious step to avoid the reproducibily problem is to replace human segmentations by automatic ones. Fully-automatic segmentation of Wilms' tumors is a challenging task as these tumors do not show a discriminative texture, might have intensities overlapping with the surrounding tissue, and can be directly attached to the remaining kidney. To the best of our knowledge, there is no method available so far that does not need massive user interaction. Moreover, the scientific literature on computer-based segmentation algorithms for Wilms' tumors is fairly limited and shall be discussed next.

An initial idea for segmentation is to extend user marked seed points in the tumor by region growing based on intensity thresholding [10]. A refined approach is to initialize an active contour inside the tumor and to expand the segmentation according to image intensities and gradients [10]. More recently, a more advanced energy-based method for segmentation of nephroblastoma has been proposed [30]. User-set scribbles are employed to approximate the gray value distributions of tumor and surrounding tissues. The energy is then regularized by an image metric induced by a state-of-the-art edge detection. However, this method still needs user interaction.

In spite of the fact that segmentation is an active research field in image analysis for quite some decades, it is remarkable that many well-established classes of algorithms have not been evaluated in the context of Wilms' tumor segmentation. Moreover, a comparative evaluation of these algorithms is prevented by the fact that there is no public benchmark available. So far the few computer-based algorithms for Wilms' tumor segmentation have been tested on different data sets.

1.1 Our Contributions

The goal of our paper is to offer solutions to the before mentioned problems in a threefold way:

(i) We establish the first publically available heterogeneous benchmark data set for Wilms' tumors. This data set will be released once the paper is accepted.
It allows clinicians to train their segmentation abilities, and computer

scientists to evaluate their algorithms. Our benchmark consists of multi-sequence MRI data before and after chemotherapy. Ground truth segmentations are approximated by consensus truth of five human experts.

(ii) Based on this benchmark, we scrutinize the widely-used ellipsoid approximation to the tumor volume as well as the interrater variability of manual delineations. Both results will reveal substantial shortcomings of the current standards.

(iii) As a second benchmark application, we evaluate six algorithms w.r.t. their usefulness for Wilms' tumor segmentation. Although most of these segmentation algorithms are popular and time-proven methods in the computer vision community, none of them has been used for Wilms' tumor segmentation yet. Our algorithms include a fully-automatic method based on Chan-Vese active contours [8], a random forest classifier [7], a support vector machine [6], a k-means clustering algorithm [25], and a clustering of superpixels [24]. Since the Wilms' tumor data are necessarily limited, most segmentation methods based on deep learning cannot be applied due to an insufficient amount of training data. One of the few methods that can be used is the U-Net [35], which we are also evaluating.

In computer vision, benchmarking and performance evaluation have established themselves as important triggers for scientific progress in key areas ranging from motion analysis [4, 3, 13] over optimisation algorithms [19] to segmentation methods [27]. Pure benchmarking and performance evaluation have become equally influential in medical image analysis [16, 36], e.g. in registration [31, 38] and various segmentation problems [5, 20, 26, 29]. The authors of these publications typically follow the clear scientific practice not to mix benchmark data with own unpublished algorithms, since this enables a fair comparison and avoids conflicts of interests. We adhere to these standards and refrain from proposing novel algorithms: We focus on evaluating the performance of popular fully-automatic segmentation methods when being applied to Wilms' tumor data.

1.2 Paper Organisation

Section 2 introduces our new multi-sequence benchmark for Wilms' tumor segmentation. We analyze interoperator variability and compare the determined volumes with volume approximations used in clinical practice. The third section evaluates human segmentations, and Section 4 is devoted to the

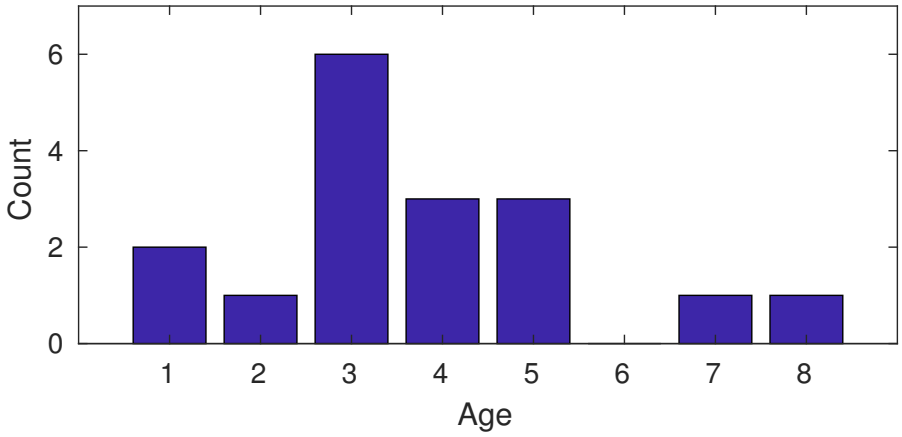


Figure 1: Age distribution of patients whose images are made available anonymously.

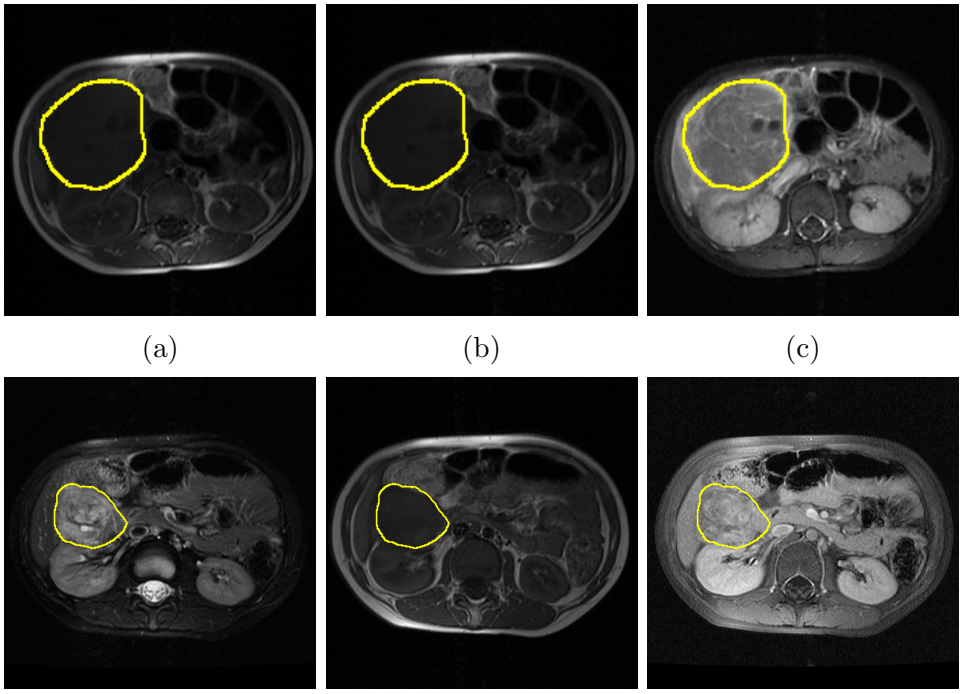
evaluation of computer-based segmentation algorithms. Our conclusions are summarized in Section 5.

2 Benchmark Data

To describe our benchmark data set, we first present details on the acquired MRI data and the chosen method for ground truth approximation. Afterwards, we introduce our error metrics and evaluate the interoperator variability on the proposed data set. In the end, we compare volume variability among human expert raters, ground truth and ellipsoid shapes.

2.1 Data Sets

Our image data set consists of 28 multi-sequence MR scans from 17 Wilms' tumor patients (5 male, 12 female), out of which 15 have been acquired from intermediate risk tumor (histological diagnosis: stromal predominant (2), mixed histology (6) or regressive type (7)) and 2 from high risk tumor types (histological diagnosis: blastemal predominant). For eleven patients, we have both data before and after chemotherapy. The remaining ones are missing either data before or after chemotherapy. Fig. 1 shows the age distribution of the children. Only patients with histologically confirmed Wilms' tumors were eligible for inclusion. The MRI sequences before and after chemotherapy for one of these patients are shown in Fig. 2.



(a) (b) (c)

(d) (e) (f)

Figure 2: Example of Wilms' tumor (training data) before ((a)-(c)) and after ((d)-(f)) chemotherapy with experts' consensus truth. From left to right: T2, T1, T1c.

Since it is difficult to obtain a comprehensive and representative set of Wilms' tumor data, the images have been acquired at different centers over the course of several years, using MR scanners from different manufacturers, varying field strength (1.5T and 3T) and implementations of the imaging sequences. The data sets used in our benchmark share the following three MRI settings:

- T2: T_2-weighted images, axial 2D acquisition with 3.6 mm to 9.1 mm slice thickness and inslice-sampling ranging from 0.3 mm to 1.4 mm.
- T1: T_1-weighted images, native image, axial 2D acquisition with 2.5 mm to 9.1 mm slice thickness and inslice-sampling ranging from 0.5 mm to 1.6 mm.
- T1c: T_1-weighted and contrast enhanced (Gadolinium) images, axial 2D acquisition with 1.8 mm to 7.7 mm slice thickness and inslice-sampling ranging from 0.5 mm to 1.6 mm.

Table 1: Image properties before and after chemotherapy. The values in brackets indicate the average occurrence.

	Training Set		**Test Set**	
	Slices	Tumor	Slices	Tumor
Pre-Chemo	$19-55$ (31)	$9-25$ (15)	$26-50$ (35)	$11-28$ (18)
Post-Chemo	$19-44$ (30)	$6-26$ (12)	$29-70$ (54)	$6-23$ (13)

The different MRI sequences were spatially co-registered on the T2 sequence using a rigid transformation. We balanced the number of slices with tumor areas before and after chemotherapy; see Tab. 1. Subtypes are not balanced among the data sets. We deploy all images in NRRD-file format [1]. NRRD stands for "nearly raw raster data" and is a standard file format for storing medical image data, fully anonymized and without sensitive patient information.

2.2 Annotations by Human Experts

The images were manually annotated by five human expert raters coauthoring this publication. Rater-1 and Rater-4 are experienced radiologists with several years of experience in Wilms' tumor analysis. Rater-2 is a physician familiar with Wilms' tumors. Rater-3 is an M.D. student previously trained in MRI imaging with advanced experience in the field. Rater-5 is an experienced oncologist with decades of practice in Wilms' tumor exploration. Segmentations were performed using the MITK software from `www.mitk.org`, and experts outlined tumor structures in T2-sequences in every axial slice.

2.3 Ground Truth Generation

Since the generation of error-free ground truth information for medical images is usually not possible, we rely on expert votes to approximate the tumor area. Majority voting for each voxel has been shown to be useful in several contexts [17, 33]. Unfortunately, this simple approach neither regards variability in quality or performance amongst the human raters nor does it provide guidance as to how many experts should agree before a voxel is labeled as tumor. Hence, we decide to use the STAPLE framework [37] to produce consensus segmentations.
The STAPLE algorithm uses expectation maximization. Let $D_{\mathbf{x},j}$, $j = 1,\ldots,n$ be the expert decisions and $\hat{\mathbf{G}}$ the true consensus segmentation. The performance of each expert is rated on the basis of the sensitivity

Table 2: Estimated quality parameters of each expert before and after chemotherapy. Rater-1: Radiologist, Rater-2: Physician, Rater-3: M.D. student, Rater-4: Radiologist, Rater-5: Oncologist.

	Rater-1	Rater-2	Rater-3	Rater-4	Rater-5
Pre-Chemotherapy					
Sensitivity	0.76	0.71	0.80	0.65	0.58
Specificity	0.65	0.75	0.73	0.82	0.74
Post-Chemotherapy					
Sensitivity	0.78	0.60	0.77	0.70	0.67
Specificity	0.72	0.57	0.75	0.85	0.82

$p_j = \mathtt{Pr}(D_{\mathbf{x},j} = 1 \mid \hat{\mathbf{G}} = 1)$ and the specificity $q_j = \mathtt{Pr}(D_{\mathbf{x},j} = 0 \mid \hat{\mathbf{G}} = 0)$. It iterates between estimating the conditional probability of $\hat{\mathbf{G}}$ in relation to the expert decisions and previous estimates of the performance parameters and estimation of updated reliability parameters. Before chemotherapy, convergence is on average reached with less than 33 iterations. After chemotherapy, the algorithm converged on average after 52 iterations. The estimated quality parameters of each expert are shown in Tab. 2 and indicate high inter-rater variability.
Fig. 3 shows annotations from all five human experts and the final ground truth approximation.

2.4 Error Metrics

We show results in terms of the metrics suggested in [29] and compute *precision* and *recall* as

$$P_{\hat{\mathbf{G}},\mathbf{G}} := \frac{|\hat{\mathbf{G}} \cap \mathbf{G}|}{|\mathbf{G}|}, \qquad R_{\hat{\mathbf{G}},\mathbf{G}} := \frac{|\hat{\mathbf{G}} \cap \mathbf{G}|}{|\hat{\mathbf{G}}|}, \tag{1}$$

where $\hat{\mathbf{G}}$ is the experts' consensus truth, and $\mathbf{G}$ the algorithmic prediction. The harmonic mean of precision and recall is called *Dice score*. It relates the area of a cluster to its voxelwise overlap with the approximated ground truth. The average Dice score determines the overall segmentation accuracy. Another class of error measures evaluates the distance between the segmentation boundaries, i.e. the surface distance. The best known example of this is the Hausdorff distance [34]. It calculates for a given volume the shortest distance to all points on the surface of another volume and vice versa, and finally extracts the maximal distance. However, the return of the maximum

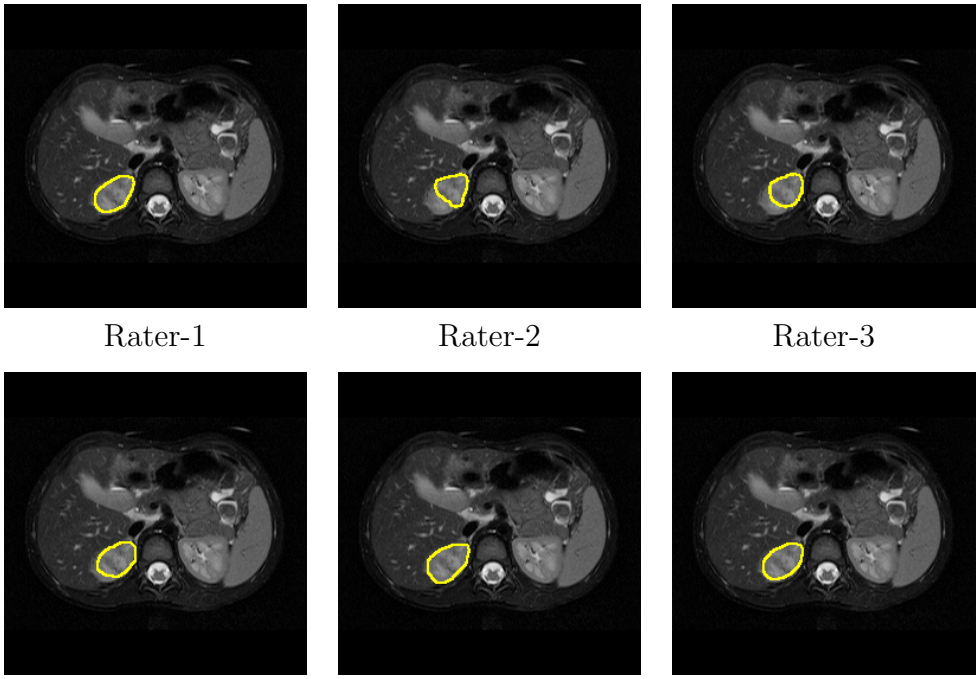


Figure 3: Example annotations by human expert raters. Rater-1: Radiologist, Rater-2: Physician, Rater-3: M.D. student, Rater-4: Radiologist, Rater-5: Oncologist.

over all surface distances makes the Hausdorff measurement very susceptible to small remote subregions in either ground truth or segmentation result. In the evaluation of the fully automated methods, predictions with few false-positive areas - which only marginally influence the overall quality of the segmentation - can also dramatically influence Hausdorff's overall result. Therefore, we refrain from evaluating this error measure. It is not conclusive in our scenario.

3 Evaluation of Human Expert Segmentations

3.1 Accuracy

3.1.1 Interoperator Variability

We calculate the interoperator variability using all 28 data sets of all 17 patients. In order to do so, we compute the disagreement of the outlined volume marked by each physician with each volume outline prepared by each of the

Table 3: Interoperator variability before and after chemotherapy in terms of Dice score. Rater-1: Radiologist, Rater-2: Physician, Rater-3: M.D. student, Rater-4: Radiologist, Rater-5: Oncologist.

	Rater-1	Rater-2	Rater-3	Rater-4
Pre-Chemotherapy				
Rater-2	0.85 ± 0.13			
Rater-3	0.89 ± 0.11	0.89 ± 0.08		
Rater-4	0.85 ± 0.13	0.90 ± 0.05	0.88 ± 0.08	
Rater-5	0.83 ± 0.13	0.89 ± 0.05	0.87 ± 0.07	0.89 ± 0.05
Post-Chemotherapy				
Rater-2	0.63 ± 0.37			
Rater-3	0.83 ± 0.24	0.65 ± 0.37		
Rater-4	0.84 ± 0.10	0.65 ± 0.36	0.80 ± 0.24	
Rater-5	0.84 ± 0.10	0.64 ± 0.35	0.80 ± 0.24	0.89 ± 0.05

other four clinicians for the same data set. This process was repeated for each patient to provide a data set comprising the average disagreement between the five contours for each data set. We also divide the data sets based on their acquisition time relative to chemotherapy, i.e. before and after chemotherapy. Tab. 3 shows the interoperator variability in terms of Dice score before chemotherapy and after chemotherapy, respectively. Before chemotherapy, the average Dice score between human experts shows their agreement on average with 0.87 ± 0.09 on tumor areas. After chemotherapy, when tumor tissues are barely visible, the average Dice score between human expert raters drops to 0.78 ± 0.24 indicating a high inter-rater variability. Especially after chemotherapy, Rater-2 seems to be the bottleneck in agreement of the human experts. Therefore, we also computed the average Dice scores excluding this annotator. It turns out that average Dice score and standard deviation between human expert raters before chemotherapy slightly decreases to 0.87 ± 0.1. After chemotherapy it improves to 0.83 ± 0.17, but still shows a high variability.

We also evaluated our expert annotations with McNemar's statistical test [28] :

$$\chi^2 = \frac{\mid b - c \mid^2}{b + c} \,. \tag{2}$$

This χ^2-test for paired nominal data, based on the contingency matrix of these samples, provides information on whether there is a statistically signif-

Table 4: Confusion matrix for McNemar's statistical test.

a: no / no	b: yes / no
c: no / yes	d: yes / yes

icant difference. We calculated the corresponding matrix according to Tab. 4. Here, the first entry per field refers to the first expert to be compared and the second to the other. For example, field b means that the first of the two has labeled a pixel as tumor and the other as non-tumor region. Furthermore, a significance level of $\alpha = 0.05$ corresponds approximately to a fiducial level of $\chi^2 = 3.8415$.
The results in Tab. 5 highlight the differences in expert annotations analogous to our previous analyses: All results of McNemar's test reject the null hypothesis that the annotations are similar with high values for all rater combinations. Unfortunately, it is not possible to compare these test results before and after chemotherapy: On the one hand, the tumor shrinks during therapy, and on the other hand, the resolution of the images is usually not the same. Both result in a different number of pixels in the contingency matrix.

Table 5: Interoperator variability before and after chemotherapy in terms of McNemar's test averaged on all data sets. Rater-1: Radiologist, Rater-2: Physician, Rater-3: M.D. student, Rater-4: Radiologist, Rater-5: Oncologist.

	Rater-1	Rater-2	Rater-3	Rater-4
Pre-Chemotherapy				
Rater-2	11169			
Rater-3	7594	7158		
Rater-4	15840	4683	8636	
Rater-5	19106	7195	10538	5916
Post-Chemotherapy				
Rater-2	10898			
Rater-3	2423	10152		
Rater-4	11668	11765	9187	
Rater-5	13733	14598	11293	1502

3.1.2 Deviation from Ground Truth

The average Dice score before chemotherapy of human experts in comparison to ground truth is 0.93 ± 0.05. After chemotherapy, the contrast of tumor regions is usually lower and the tumor outlines are more ambiguous. Consequently, human experts agree less on tumor areas. The average Dice score decreases to 0.85, and variability increases dramatically to 0.16.

3.2 Volume Variability

Tumor expansion after preoperative chemotherapy is an important metric used to categorize patients as high-, intermediate- or low-risk candidates. High-risk patients receive an additional postoperative chemotherapy aligned with an irradiation. Therefore, an accurate determination of tumor volume is critical. The *clinical volume* equals the volume information used in therapy and treatment planning. It approximates the tumor by an ellipsoid shape. It is computed as $\texttt{width} \times \texttt{height} \times \texttt{depth} \times 0.524$ [14], where `width`, `height` and `depth` of tumor denote the maximal expansion of tumor tissue on MR images. Note that the volume of the largest ellipsoid that fits in a cuboid is $\pi/6 \approx 0.524$ times the cuboid volume. Starting with the assumption that the true tumor volume is found through the consensus of our five human experts, we compare human expert annotations and clinical volumes in terms of percental volume differences in relation to the ground truth volume before and after chemotherapy, respectively. It turns out that clinical volumes differ before chemotherapy on average by 22.62 ± 16.12 %, and after chemotherapy by 35.07 ± 41.01 % from the ground truth volumes. Before and after chemotherapy, clinical volumes are on average smaller than the ground truth volume, i.e. 85.71 % before and 92.86 % after chemotherapy. In contrast, human experts differ before chemotherapy on average by 10.58 ± 5.90 % and after chemotherapy by 25.98 ± 34.57 % from the ground truth volume.
This shows that assuming an ellipsoid shape for Wilms' tumors is an erroneous oversimplification, and human expert annotations are helpful to determine tumor volumes more precisely.

4 Evaluation of Segmentation Algorithms

In the following, we conduct example evaluations on our new benchmark data with six fully-automatic methods:

- Chan-Vese active contours [8] with two level sets.
- K-means clustering [25] with intensities.

- Entropy Rate Superpixel Segmentation [24].
- Classification with a support vector machine [6] with intensities and HOG-features [9].
- Random-forest classification [7], either with intensities or HOG-features [9].
- Segmentation with a U-Net [35].

To guarantee a fair evaluation, we equally split the data sets in training and test data, each containing seven data sets before and after chemotherapy. For each segmentation approach we include information from all modalities. The slice thickness of up to $9.1mm$ inhibits 3D segmentations, so we perform all segmentations in 2D slices. Let us now sketch each of the evaluated segmentation approaches.

4.1 Chan-Vese Active Contours

We consider a cubic data domain $\Omega \subset \mathbb{R}^3$ and a volumetric data set $\boldsymbol{f} : \Omega \to \mathbb{R}^3$. In our setting, the co-domain describes the different MRI modalities T2, T1, and T1c. Then a segmentation of $\boldsymbol{f}$ by means of the Chan-Vese active contour model [8] minimizes the cost function

$$\begin{aligned} E(\boldsymbol{u}, C) = \lambda_{\mathtt{in}} \int_{C_{\mathtt{in}}} \|\boldsymbol{u}_{\mathtt{in}} - \boldsymbol{f}\|^2 d\mathbf{x} \\ + \lambda_{\mathtt{out}} \int_{C_{\mathtt{out}}} \|\boldsymbol{u}_{\mathtt{out}} - \boldsymbol{f}\|^2 d\mathbf{x} + \nu\, \ell(C) \end{aligned} \tag{3}$$

where the data domain Ω is split in two regions C_{in} and C_{out}. The function $\boldsymbol{f}$ is approximated by a piecewise constant function where $\boldsymbol{u}_{\mathtt{in}}$ and $\boldsymbol{u}_{\mathtt{out}}$ are the arithmetic means of $\boldsymbol{f}$ inside and outside the segment boundaries C, respectively. The positive weights $\lambda_{\mathtt{in}}$ and $\lambda_{\mathtt{out}}$ control the influence of each region to the final partitioning, $\|\,.\,\|$ denotes the Euclidean norm in $\mathbb{R}^3$, and C are the segment boundaries with a (Hausdorff) length of $\ell(C)$. This length is weighted with a parameter $\nu > 0$.

4.2 K-means Clustering

K-means clustering [25] is a vector quantization method that partitions n observations into k clusters. Data points are assigned to cluster centers, prototypes of corresponding classes, with minimal Euclidean distance. In our application, we want to split the observations into two classes, tumor

and non-tumor points.
Given a set of data points $\boldsymbol{f} : \Omega \to D$ with $D \subset \mathbb{R}^3$ and $\Omega \subset \mathbb{R}^3$, k-means minimizes

$$\begin{aligned} E(D_1, D_2) &= \int_{D_1} \|\xi - \boldsymbol{u_1}\|^2 d\xi + \int_{D_2} \|\xi - \boldsymbol{u_2}\|^2 d\xi \\ D &= D_1 \cup D_2, \quad D_1 \cap D_2 = \emptyset, \end{aligned} \tag{4}$$

where $\boldsymbol{u_1}$ and $\boldsymbol{u_2}$ are the arithmetic means of both classes. In this case, k-means clustering is equivalent to Otsu's method [23].

4.3 Support Vector Machine

Support Vector Machines [6] are based on the concept of hyperplanes in a multidimensional space, separating between sets of objects having different classes, e.g. tumor and non-tumor points. In our application, we use a five-fold cross validation to find optimized hyperparameters. Training was performed using MATLAB (`www.mathworks.com/products/matlab`) and the problem was solved via Sequential Minimal Optimization [12]. Furthermore, we used Gaussian-like kernels and the classification error, i.e. the weighted fraction of misclassified observations, as loss function.

4.4 Random-forest Classification

Ensemble methods employ a finite set of different learning algorithms to get better predictive performance than using a single learning algorithm. Random forests [7] are ensemble approaches for classification combining a group of decision trees. A single tree is highly sensitive to noise, while the average of many decorrelated trees is not. Training all decision trees of a random forest on the same training data would result in strongly correlated trees. Bagging (bootstrap aggregation) generates new training sets $\mathbf{K}$ by sampling from the original training set $\mathbf{Y}$ uniformly and with replacement. In this way, decision trees are decorrelated by using different training data. Additionally, random forests use feature bagging, i.e. features are randomly sampled for each decision tree [18]. To estimate how well the results can be generalized, we use 2-fold-cross validation, i.e. we train two sets of models.

4.5 Entropy Rate Superpixel Segmentation

The method of Liu et al. [24] formulates the superpixel segmentation problem as maximization of the entropy rate of cuts in the graph. Optimizing this

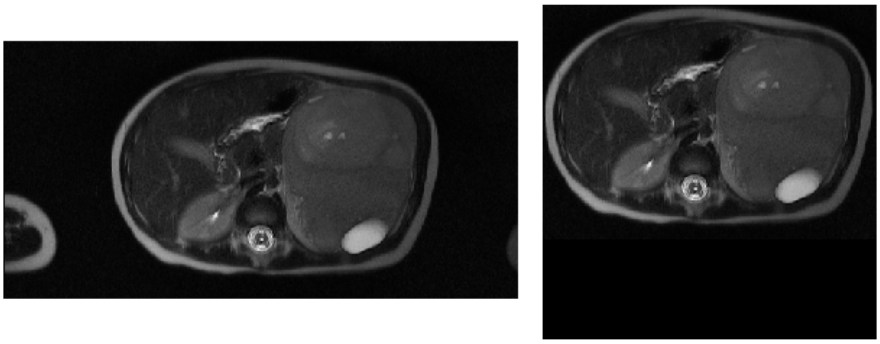
entropy rate encourages the clustering of compact and homogeneous regions, which also favors the superpixels to overlap with only one single object on the perceptual boundaries.
This technique starts with each pixel being considered as a separate cluster. Clusters are then gradually merged into larger superpixels. In this way, during segmentation, a hierarchy of superpixels is created until finally only one superpixel, the image itself, is left. In our case we want to segment a tumor, i.e. we use the hierarchy of superpixels to divide the image into three groups: tumor, body and background. Unfortunately we do not know in advance which superpixel contains which class. This objective function is optimized with a greedy algorithm.

4.6 U-Net

In many areas of medical image processing, deep learning and especially convolutional neural networks (CNN) have proven to be very powerful tools. Within these, the U-net architecture [35] is one of the standard CNNs in the field of medical image segmentation. It learns segmentation in an end-to-end setting and only needs a few training examples. Since our benchmark consists of real clinical data, they are available in different resolutions. Some of them also contain other parts of the body, e.g. the arms. Therefore, the amount of non-tumor areas outweighs the tumor areas substantially, such that it becomes necessary to balance the classes. This is done in three steps: First we determine the connected components, i.e. connected parts of the body, and remove everything except the largest one. Then we determine the maximum extent of the existing object and extract this part to a new, smaller image; see Fig. 4. This is then rescaled to a size of 512×512 pixels. We use the implementation presented in [2] to solve our segmentation problem and set up the network with batch size 5 and 50 epochs.

4.7 Results

In Tab. 6 we present the mean precision, recall and Dice score over the 14 test data sets of the different segmentation algorithms. Since the Chan-Vese method is region-based, it suffers from the fact that the visual appearance of Wilms' tumors can be highly heterogeneous. Our experiments show that intensities are an important feature to identify tumor areas, resulting in high precision values for the pixel-based classifiers k-means clustering and random forests. However, spatial information is essential as intensities of a tumor can overlap with those of the surrounding tissue. Accordingly, the pixel-based methods suffer from low recall. Using HOG-features in addition to intensities



(a) (b)

Figure 4: Exemplary pre-processing step for the U-Net. (a) Original image containing abdomen and extremities, (b) Image after pre-processing.

improves k-means clustering after chemotherapy, SVM classification as well as random forests both before and after chemotherapy.
The results of the superpixel-based method are unexpectedly poor both before and after chemotherapy. The optimum number of superpixels depends strongly on the image and it is also difficult to identify the respective segments. We could not find a parameter set that worked on all data sets. Deep learning methods usually require a large amount of training data. The U-net used here deviates from this paradigm and can also be trained with smaller amounts of data. Tab. 6 shows that it gives a high mean recall, but a low mean precision. This indicates that although the network can recognize the basic structure of the nephroblastoma, it is not able to distinguish it from similar tissue.
Overall, segmentation with random forests provides the best results before chemotherapy, but is also the leading approach after chemotherapy, yielding the highest quality measures. Therefore, we suggest random forests trained on HOG-features as well as intensities as the baseline method for this benchmark data set. Since the tumor volume after chemotherapy is decisive for postoperative treatment planning, it is currently the optimal method for this purpose. The segmentation quality lies within the variability of human experts.
In order to ensure spatial consistency, we also apply Chan-Vese active contours on the predicted probabilities of the random forest. It turns out that predictions of this method lack too much global information and the result-

Table 6: Results on the proposed benchmark data set (test data). k-means: k-means clustering, CV: Chan-Vese active contours, RF: Random Forest Classification, SVM: Support Vector Machine, INT: Intensity values, HOG: HOG-features, PP: Post-processing. Best results are depicted in bold face.

Method	Dice Score	Precision	Recall
Pre-Chemotherapy			
CV [8]	0.57	0.48	0.69
k-means [25] (INT)	0.53	0.76	0.41
Superpixel [24]	0.41	0.33	0.56
SVM [6] (INT + HOG [9])	0.71	0.71	0.72
RF [7] (INT + HOG [9])	**0.92**	**0.92**	0.91
U-net [35]	0.64	0.49	**0.94**
Post-Chemotherapy			
CV [8]	0.41	0.32	0.58
k-means [25] (INT)	0.35	0.50	0.27
Superpixel [24]	0.41	0.29	0.68
SVM [6] (INT + HOG [9])	0.68	0.69	0.67
RF [7] (INT + HOG [9])	**0.81**	**0.73**	**0.92**
U-net [35]	0.30	0.25	0.61

ing segmentation loses quality. These observations highlight the challenges in the data set.

5 Conclusions

We have proposed the first multi-sequence benchmark for segmentation of Wilms' tumors. In spite of the fact that such a data set involving tumors in children is necessarily limited in size, its amount of information is rich: There are multi-sequence MRT images for all patients, and for eleven patients both pre- and post chemotherapy images. That is supplemented by manual annotations by five independent human experts, as well as histological diagnoses. Our benchmark allows several important conclusions:
We have demonstrated that human expert annotations suffer from a large interoperator variability especially after pre-operative chemotherapy. Further-

more, we have shown that the popular tumor volume determination based on ellipsoid shapes tends to be highly erroneous.
Our data set also allowed to evaluate six computer-based algorithms. At this time, all fully-automatic segmentations apart from random forests undersegment the tumor volume compared to human expert raters. Thus, their precision is insufficient, especially after chemotherapy. Our experiments indicate that segmentation with random forests [7] is the most appropriate tool for Wilms' tumors. Its results lie within the variability of the contouring performed by human expert raters on the same data. Moreover, it offers the advantage that it is much faster than a full segmentation by human experts. In our ongoing research, we plan to include more anatomical knowledge into our segmentation strategies and to constantly enlarge the number of available data sets. It is our hope that our benchmark data set for segmentation of nephroblastoma will stimulate a growing interest in this research field which is challenging both from a medical and a computer vision viewpoint. Most importantly, we are confident that the resulting progress will help to maximize the survival chances of the affected children.

Conflicts of Interest

We do not have conflicts of interest.

Acknowledgment

This work was partially funded by the European Union's seventh framework program under the project Computational Horizons in Cancer (grant agreement No 600841). J.W. has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement no. 741215, ERC Advanced Grant INCOVID).

References

[1] NRRD: Nearly Raw Raster Data. `http://teem.sourceforge.net/nrrd/index.html`, accessed: 2019-04-25

[2] Akeret, J., Chang, C., Lucchi, A., Refregier, A.: Radio frequency interference mitigation using deep convolutional neural networks. Astronomy and Computing 18, 35–39 (2017), [doi:10.1016/j.ascom.2017.01.002]

[3] Baker, S., Scharstein, D., Lewis, J., Roth, S., Black, M.J., Szeliski, R.: A database and evaluation methodology for optical flow. International Journal of Computer Vision 92(1), 1–31 (2011), [doi:10.1007/s11263-010-0390-2]

[4] Barron, J.L., Fleet, D.J., Beauchemin, S.S.: Performance of optical flow techniques. International Journal of Computer Vision 12(1), 43–77 (1994), [doi:10.1007/BF01420984]

[5] Bernard, O., Bosch, J.G., Heyde, B., Alessandrini, M., Barbosa, D., Camarasu-Pop, S., Cervenansky, F., Valette, S., Mirea, O., Bernier, M., et al.: Standardized evaluation system for left ventricular segmentation algorithms in 3D echocardiography. IEEE Transactions on Medical Imaging 35(4), 967–977 (2016), [doi:10.1109/TMI.2015.2503890]

[6] Boser, B.E., Guyon, I.M., Vapnik, V.N.: A training algorithm for optimal margin classifiers. In: Proc. 1992 Fifth Annual Workshop on Computational Learning Theory. pp. 144–152. ACM, New York, NY, USA (1992), [doi:10.1145/130385.130401]

[7] Breiman, L.: Random Forests. Machine Learning 45(1), 5–32 (2001), [doi:10.1023/A:1010933404324]

[8] Chan, T.F., Sandberg, B.Y., Vese, L.A.: Active contours without edges for vector-valued images. Journal of Visual Communication and Image Representation 11(2), 130–141 (2000), [doi:10.1006/jvci.1999.0442]

[9] Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proc. 2005 IEEE Conference on Computer Vision and Pattern Recognition. pp. 886–893. San Diego, CA (Jun 2005), [doi:10.1109/CVPR.2005.177]

[10] David, R., Graf, N., Karatzanis, I., Stenzhorn, H., Manikis, G.C., Sakkalis, V., Stamatakos, G.S., Marias, K.: Clinical evaluation of DoctorEye platform in nephroblastoma. In: Proc. 5th International Advanced Research Workshop on In Silico Oncology and Cancer Investigation. pp. 1–4. IEEE (2012)

[11] Davidoff, A.M.: Wilms' tumor. Current Opinion in Pediatrics 21(3), 357–364 (2009), [doi:10.1097/MOP.0b013e32832b323a]

[12] Fan, R.E., Chen, P.H., Lin, C.J.: Working set selection using second order information for training support vector machines. Journal of Machine Learning Research 6(Dec), 1889–1918 (Dec 2005)

[13] Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the KITTI vision benchmark suite. In: Proc. 2012 IEEE Conference on Computer Vision and Pattern Recognition. pp. 3354–3361. IEEE (2012), [doi:10.1109/CVPR.2012.6248074]

[14] Graf, N., Reinhard, H., Semler, J.O.: SIOP 2001/GPOH Therapieoptimierungsstudie zur Behandlung von Kindern und Jugendlichen mit einem Nephroblastom. `http://www.kinderkrebsinfo.de` (2003)

[15] Graf, N., Tournade, M.F., de Kraker, J.: The role of preoperative chemotherapy in the management of Wilms' tumor: The SIOP studies. Urologic Clinics of North America 27(3), 443–454 (2000), [doi:10.1016/S0094-0143(05)70092-6]

[16] Hanbury, A., Müller, H., Langs, G.: Cloud-Based Benchmarking of Medical Image Analysis. Springer (2017), [doi:10.1007/978-3-319-49644-3]

[17] Heckemann, R.A., Hajnal, J.V., Aljabar, P., Rueckert, D., Hammers, A.: Automatic anatomical brain MRI segmentation combining label propagation and decision fusion. NeuroImage 33(1), 115–126 (Oct 2006), [doi:10.1016/j.neuroimage.2006.05.061]

[18] Ho, T.K.: The random subspace method for constructing decision forests. IEEE Transactions on Pattern Analysis and Machine Intelligence 20(8), 832–844 (1998), [doi:10.1109/34.709601]

[19] Kappes, J., Andres, B., Hamprecht, F., Schnorr, C., Nowozin, S., Batra, D., Kim, S., Kausler, B., Lellmann, J., Komodakis, N., et al.: A comparative study of modern inference techniques for discrete energy minimization problems. International Journal of Computer Vision 115(2), 155–184 (2015), [doi:10.1007/s11263-015-0809-x]

[20] Karim, R., Bhagirath, P., Claus, P., Housden, R.J., Chen, Z., Karimaghaloo, Z., Sohn, H.M., Rodríguez, L.L., Vera, S., Albà, X., et al.: Evaluation of state-of-the-art segmentation algorithms for left ventricle infarct from late Gadolinium enhancement MR images. Medical Image Analysis 30, 95–107 (Jan 2016), [doi:10.1016/j.media.2016.01.004]

[21] Kaste, S.C., Dome, J.S., Babyn, P.S., Graf, N.M., Grundy, P., Godzinski, J., Levitt, G.A., Jenkinson, H.: Wilms tumour: prognostic factors, staging, therapy and late effects. Pediatric Radiology 38(1), 2–17 (2008), [doi:10.1007/s00247-007-0687-7]

[22] Kim, S., Chung, D.H.: Pediatric solid malignancies: neuroblastoma and Wilms' tumor. Surgical Clinics of North America 86(2), 469–487 (2006), [doi:10.1016/j.suc.2005.12.008]

[23] Liu, D., Yu, J.: Otsu method and k-means. In: Proc. 2009 IEEE Ninth International Conference on Hybrid Intelligent Systems. vol. 1, pp. 344–349. IEEE (2009), [doi:10.1109/HIS.2009.74]

[24] Liu, M.Y., Tuzel, O., Ramalingam, S., Chellappa, R.: Entropy rate superpixel segmentation. In: Proc. 2014 IEEE Conference on Computer Vision and Pattern Recognition. pp. 2097–2104. IEEE, Providence, RI, USA (2011), [doi:10.1109/CVPR.2011.5995323]

[25] Lloyd, S.: Least squares quantization in PCM. IEEE Transactions on Information Theory 28(2), 129–137 (Mar 1982), [doi:10.1109/TIT.1982.1056489]

[26] Maier, O., Menze, B.H., von der Gablentz, J., Häni, L., Heinrich, M.P., Liebrand, M., Winzeck, S., Basit, A., Bentley, P., Chen, L., et al.: Isles 2015-a public evaluation benchmark for ischemic stroke lesion segmentation from multispectral MRI. Medical Image Analysis 35, 250–269 (Jan 2017), [doi:10.1016/j.media.2016.07.009]

[27] Martin, D., Fowlkes, C., Tal, D., Malik, J.: A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In: Proc. 2001 Eighth IEEE International Conference on Computer Vision. vol. 2, pp. 416–423. IEEE (2001), [doi:10.1109/ICCV.2001.937655]

[28] McNemar, Q.: Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika 12(2), 153–157 (1947)

[29] Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al.: The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Transactions on Medical Imaging 34(10), 1993 (2015), [doi:10.1109/TMI.2014.2377694]

[30] Müller, S., Ochs, P., Weickert, J., Graf, N.: Robust interactive multilabel segmentation with an advanced edge detector. In: Andres, B., Rosenhahn, B. (eds.) Pattern Recognition, Lecture Notes in Computer Science, vol. 9796, pp. 117–128. Springer, Cham, Switzerland (2016), [doi:10.1007/978-3-319-45886-1_10]

[31] Murphy, K., Van Ginneken, B., Reinhardt, J.M., Kabus, S., Ding, K., Deng, X., Cao, K., Du, K., Christensen, G.E., Garcia, V., et al.: Evaluation of registration methods on thoracic CT: the EMPIRE10 challenge. IEEE Transactions on Medical Imaging 30(11), 1901–1920 (May 2011), [doi:10.1109/TMI.2011.2158349]

[32] Pastore, G., Znaor, A., Spreafico, F., Graf, N., Pritchard-Jones, K., Steliarova-Foucher, E.: Malignant renal tumours incidence and survival in European children (1978–1997): Report from the Automated Childhood Cancer Information System project. European Journal of Cancer 42(13), 2103–2114 (2006), [doi:10.1016/j.ejca.2006.05.010]

[33] Raykar, V.C., Yu, S., Zhao, L.H., Jerebko, A., Florin, C., Valadez, G.H., Bogoni, L., Moy, L.: Supervised learning from multiple experts: whom to trust when everyone lies a bit. In: Proc. 2009 26th Annual International Conference on Machine Learning. pp. 889–896. ACM, New York, NY, USA (2009), [doi:10.1145/1553374.1553488]

[34] Rockafellar, R.T., Wets, R.J.B.: Variational Analysis, vol. 317. Springer Science & Business Media (2009)

[35] Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention. Lecture Notes in Computer Science, vol. 9351, pp. 234–241. Springer (2015), [doi:10.1007/978-3-319-24574-4_28]

[36] Wang, C.W., Huang, C.T., Lee, J.H., Li, C.H., Chang, S.W., Siao, M.J., Lai, T.M., Ibragimov, B., Vrtovec, T., Ronneberger, O., et al.: A benchmark for comparison of dental radiography analysis algorithms. Medical Image Analysis 31, 63–76 (Jul 2016), [doi:10.1016/j.media.2016.02.004]

[37] Warfield, S.K., Zou, K.H., Wells, W.M.: Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. IEEE Transactions on Medical Imaging 23(7), 903–921 (2004), [doi:10.1109/TMI.2004.828354]

[38] Xu, Z., Lee, C.P., Heinrich, M.P., Modat, M., Rueckert, D., Ourselin, S., Abramson, R.G., Landman, B.A.: Evaluation of six registration methods for the human abdomen on clinically acquired CT. IEEE Transactions on Biomedical Engineering 63(8), 1563–1572 (Jun 2016), [doi:10.1109/TBME.2016.2574816]

List of Figures

List of Tables