User-Guided Segmentation of Multi-Modality Medical Imaging Datasets with ITK-SNAP

Paul A. Yushkevich[a], **Artem Pashchinskiy**[a], **Ipek Oguz**[a], **Suyash Mohan**[a], **J. Eric Schmitt**[a], **Joel M. Stein**[a], **Dženan Zuki** [c], **Jared Vicory**[c], **Matthew McCormick**[c], **Natalie Yushkevich**[a], **Nadav Schwartz**[d], **Yang Gao**[e], and **Guido Gerig**[b]

[a]Department of Radiology, University of Pennsylvania, Philadelphia PA

[b]Department of Computer Science and Engineering, NYU Tandon School of Engineering, New York, NY

[c]Kitware, Inc, Carrboro NC

[d]Department of Obstetrics and Gynecology, University of Pennsylvania, Philadelphia PA

[e]Department of Computer Science, University of Utah, Salt Lake City UT

Abstract

ITK-SNAP is an interactive software tool for manual and semi-automatic segmentation of 3D medical images. This paper summarizes major new features added to ITK-SNAP over the last decade. The main focus of the paper is on new features that support semi-automatic segmentation of multi-modality imaging datasets, such as MRI scans acquired using different contrast mechanisms (e.g., T1, T2, FLAIR). The new functionality uses decision forest classifiers trained interactively by the user to transform multiple input image volumes into a foreground/background probability map; this map is then input as the data term to the active contour evolution algorithm, which yields regularized surface representations of the segmented objects of interest. The new functionality is evaluated in the context of high-grade and low-grade glioma segmentation by three expert neuroradiogists and a non-expert on a reference dataset from the MICCAI 2013 Multi-Modal Brain Tumor Segmentation Challenge (BRATS). The accuracy of semi-automatic segmentation is competitive with the top specialized brain tumor segmentation methods evaluated in the BRATS challenge, with most results obtained in ITK-SNAP being more accurate, relative to the BRATS reference manual segmentation, than the second-best performer in the BRATS challenge; and all results being more accurate than the fourth-best performer. Segmentation time is reduced over manual segmentation by 2.5 and 5 times, depending on the rater. Additional experiments in interactive placenta segmentation in 3D fetal ultrasound illustrate the generalizability of the new functionality to a different problem domain.

Keywords

Image segmentation; semi-automatic segmentation; gliomas; software; MRI

1. Introduction

ITK-SNAP is a software tool that provides a graphical user interface for manual and user-guided semi-automatic segmentation of 3D medical imaging datasets. ITK-SNAP was created to address image segmentation problems for which fully automated algorithms are not yet available. Automatic segmentation may be lacking because a given problem has not received sufficient attention from algorithm developers; because the problem is too complex to be solved without human input; or because there is not yet sufficient expert-annotated data to train automated algorithms. Current state-of-the-art automatic medical image segmentation algorithms often use machine learning [23, 29], multi-atlas label fusion [21], or statistical shape priors [20]. These techniques require training data in the form of tens or even hundreds of manually or semi-automatically segmented example images. Furthermore, extending fully automatic techniques to new domains (e.g. from adult to pediatric subjects, from healthy subjects to subjects with extensive pathology, or from one scanner manufacturer to another) requires yet more expert-generated segmentations as additional training data and/or validation data. Subsequently, there is a robust need for expert-guided medical image segmentation approaches that span multiple imaging modalities and application domains.

ITK-SNAP was first developed in the early 2000s to provide an interactive platform for segmenting anatomical structures in 3D images both manually (by painting outlines on 2D cross-sections of a 3D image) and semi-automatically (by manually setting the parameters and initial seeds for two active contour algorithms [8, 38]). Since its introduction, ITK-SNAP became a popular tool, as evidenced by a large number of citations in the scientific literature.[1] We analyzed a sample of 50 articles from 2014 that cited ITK-SNAP, and found that 86% of the articles used ITK-SNAP for image segmentation, 4% used it as an image viewer, and 10% cited ITK-SNAP but did not use it. Among the first group, 42% used ITK-SNAP only for manual segmentation, 42% used the semi-automatic segmentation features, and 16% did not state which approach was used. The relatively modest utilization of semi-automatic segmentation prompted us to extend ITK-SNAP with more advanced and generally applicable semi-automatic segmentation capabilities. These new capabilities focused primarily on two areas: enabling the concurrent use of multiple image channels during semi-automatic segmentation (such as multiple MRI contrast mechanisms, e.g., T1-weighted and FLAIR, as done increasingly in fully automatic segmentation) and leveraging machine learning to identify foreground and background image regions in the input to active contour segmentation (in contrast to the original ITK-SNAP [35], which relies on simple thresholding and edge detection). Introducing these capabilities required not only integrating new algorithms into ITK-SNAP, but also incorporating a large set of new user interface capabilities, for example to support the display and management of multiple image channels.

The present paper serves several aims: (1) to describe the new features introduced in ITK-SNAP software since the original 2006 publication [35]; (2) to demonstrate that the new semi-automatic functionality in ITK-SNAP can be applied to problems where threshold and

---

[1]The number of articles citing the original ITK-SNAP paper [35] reported by Scopus (scopus.com) as of 3/12/2018 was 1,944; the number of citations reported by Google Scholar (scholar.google.com) was 2,822.

edge-based active contour tools are ineffective; (3) to quantitatively compare ITK-SNAP semi-automatic segmentation to state-of-the-art specialized automatic segmentation algorithms in a widely studied problem; (4) to show that semi-automatic segmentation in ITK-SNAP can reduce segmentation time over manual segmentation; and (5) to demonstrate that ITK-SNAP segmentation capabilities can be applied in multiple application domains and imaging modalities.

Towards addressing aims 2, 3 and 4, we apply the new semi-automatic segmentation capabilities in ITK-SNAP to the challenging problem of brain tumor segmentation in multi-modality MRI. This study leverages data from the MICCAI 2013 Multi-Modal Brain Tumor Segmentation Challenge (BRATS) [25], a well characterized benchmark dataset. ITK-SNAP was used by three neuroradiologists as well as one novice user to label a set of 20 glioma cases from multi-modality MRI (pre-contrast T1, post-contrast T1, T2, and FLAIR). We hypothesized that users (both novices and neuroradiologists) would label tumors using ITK-SNAP reliably and in less time than what is required for manual segmentation. Inter-rater and intra-rater evaluation, as well as online evaluation by the BRATS system against reference manual segmentations were performed to quantify segmentation reliability and accuracy. Additionally, towards addressing aims 2, 4, and 5, we apply semi-automatic segmentation in ITK-SNAP to the very different problem of placenta segmentation in the first-trimester in 3D ultrasound images, and evaluate against manual segmentation in terms of accuracy and segmentation time.

2. Materials and Methods

This section describes the main features of ITK-SNAP software. The first part of the section (2.1–2.3) briefly summarizes the software design principles and the user interface features for image navigation, visualization, and manual segmentation. The remainder of the section focuses on the semi-automatic segmentation workflow, including the machine learning approach for reducing multi-modality image information to object/background probability maps and active contour segmentation.

2.1. ITK-SNAP software design principles

The software development for ITK-SNAP is guided by three simple principles: *exclusive focus on segmentation*, *generality of purpose*, and *ease of use.* These principles are applied by the developers to prioritize potential new features. Features are excluded if they do not directly support the needs of image segmentation, if they are tailored exclusively to specific segmentation problems, or if they are largely redundant with existing capabilities. The application of these principles has resulted in a feature set that is relatively contained, as seen in Table 1, which lists the primary features incorporated into the software since 2011. It is possible for new users to learn the primary features of ITK-SNAP in the course of a 90 minute training session.

ITK-SNAP is open-source software and is distributed under the General Public License [17]. It is a cross-platform application written in the C++ language. It leverages the Insight Toolkit (www.itk.org) library for image processing functionality, the Visualization Toolkit (www.vtk.org) for image and surface visualization, and Qt (www.qt.io) for user interface

functionality. The CMake, CTest and CDash tools (cmake.org) are used for cross-platform compilation, automated testing, and posting of compilation and test results to a web-based dashboard. A web portal (www.itksnap.org) provides source code, pre-complied binaries for the Windows, MacOS and Linux platforms, documentation, and user support resources.

2.2. Image Navigation and Visualization in ITK-SNAP

ITK-SNAP allows the user to load image volumes using common 3D medical image formats, including DICOM, NIFTI, MetaImage and NRRD. ITK-SNAP recognizes the information encoded in the image header on the spatial position and orientation of image volumes relative to the scanner physical coordinate system. The first image loaded into ITK-SNAP is designated as the "main image" and all visualization is performed relative to the main image geometry. Additional images can be loaded into ITK-SNAP after the main image, and these images can have different dimensions, resolution, and spatial orientation than the main image.

As shown in Fig. 1, 3D volumes are visualized as three orthogonal slices (cross-sections). The slices are parallel to the axes of the main image. When the main image is acquired non-obliquely, the slices correspond to the axial, coronal and sagittal planes in physical space. The three slices intersect at the center of a single voxel in the main image; the position of this voxel is defined as the "3D cursor" position. Crosshairs displayed on each slice visualize the 3D cursor position. Moving this crosshair in one slice view adjusts the slices visualized in other views. This "linked crosshair" concept provides a convenient way to navigate through 3D volumes, with all three views focused on the same location in the 3D image.

Multiple images loaded in ITK-SNAP can be visualized in three ways: (1) a "tiled" layout, where the coronal, axial and sagittal slice views each display the same slice through all loaded modalities; (2) a "thumbnail" layout, where one modality occupies most of each slice view, while others are shown as small thumbnails, clicking on which switches to that modality; and (3) a "overlay" mode, in which selected images are shown as semiopaque overlays shown on top of the main image and other images.

Images loaded into ITK-SNAP may be scalar images (each voxel holds a single intensity value) or multi-component images, such as RGB color images (each voxel holds a red, green and blue value), displacement fields (each voxel holds a displacement vector, e.g., from deformable registration), diffusion tensor images, or dynamic image sequences. For multi-component images, the user can select between viewing a single selected component, the maximum, average or magnitude of the components. Special visualization modes are provided for RGB color images and displacement fields.

The visualization of individual voxels is controlled by a user-controlled intensity remapping function, which may be linear (e.g., window/level control) or spline-based; and a color map function that maps scalar intensities to display color. Window and level can be set automatically based on the image histogram.

Navigation in 3D image space is accomplished by the repositioning of the crosshairs in the three slice views with the mouse or keyboard, as well as mouse-based zooming and panning.

Multiple redundant user interface widgets are provided to support crosshair repositioning, exact zooming and panning.

The ITK-SNAP state, i.e., the set of images currently loaded, their layout, intensity remapping, color map function, and various other state variables can be saved in lightweight XML format workspace files.

2.2.1.   Segmentation Representation and Visualization—Segmentations are represented in ITK-SNAP as 3D images with the same dimensions and orientation as the main image. Each voxel in the segmentation volume is assigned a discrete integer label, with the label 0 representing the clear label. Segmentations can be loaded and saved using popular 3D image file formats, such as NIFTI. A label description table is maintained internally that assigns a name, color, opacity, and other metadata to each label. The label table may be edited using a "label editor" UI, and saved in the XML format. Segmentations are visualized as color overlays rendered on top of the main and additional image slices; and as surfaces in the 3D render view (Fig. 1). These surfaces can be exported to file using common 3D geometry formats used in 3D printing and 3D visualization (STL, VTK). The 3D render view supports navigation by allowing the user to click on the rendered surfaces to reposition the 3D cursor.

The approach of representing segmentation as discrete label images limits the resolution of the segmentation to that of the main image and disallows partial volume segmentation.[2] However, it simplifies three-dimensional editing of the segmentations, as a change made in one slice view is unambiguously translated into changes in the other slice views. The use of a common representation for both images and segmentations also facilitates analysis.

2.2.2.   Image Registration—ITK-SNAP provides a linear registration mode that makes it possible to correct for subject motion between scans, such as head motion between multiple MRI scans obtained in the same session. A manual registration mode allows the user to rotate and translate images relative to the main image using widgets displayed on top of the orthogonal slices and mouse-based panning. An automatic registration mode can be used to find locally optimal rigid and affine transformations between the main image and a given additional image using common image similarity metric.

2.3.   Manual Segmentation

ITK-SNAP provides simple tools for creating manual segmentation and editing semi-automatic segmentations. The "polygon" tool can be used to draw structure outlines in any of the slice views. Polygons can be edited by moving vertices in the slice plane. Once accepted, the polygon is assigned the current label and integrated into the 3D segmentation volume. The "paintbrush" tool allows quick drawing and touch-up editing using the mouse, with masks of different shape and size. An adaptive paintbrush mask is also provided, wherein only the neighboring voxels similar in intensity to the voxel clicked on by the user are assigned the foreground label. Additionally, the 3D render view provides a "3D scalpel"

---

[2]Note that active contour segmentations of individiual structures generated in the semi-automatic mode can be exported in a way that retains partial volume information

tool that can be used to assign a different label to a part of a structure using a user-specified cut plane.

When applying the manual segmentation tools, users select the active label used to perform the drawing/painting operation as well as the way in which operations will affect existing labels. For the latter, the users may select to paint over all existing labels, only the clear label, or only a specific selected label. This allows the user to "protect" previously drawn labels during segmentation and editing.

2.4. Semi-Automatic Segmentation

Semi-automatic workflow proceeds in five stages, which are detailed in the sub-sections below.

1. The users specifies the region of interest (ROI) in which to perform semi-automatic segmentation;

2. The user uses one of several available *presegmentation* modes to transform the input image volumes into a single synthetic image volume called the *speed* image. In most presegmentation modes, the speed image represents the difference between the probability that a voxel belongs to the object of interest and the probability that a voxel belongs to the image background.

3. The user places one or more initial contours inside of the object of interest.

4. The contours evolve in a manner governed by the speed image and a shape regularization term.

5. The semi-automatic segmentation result is incorporated into the main ITK-SNAP segmentation volume.

2.4.1. ROI Selection for Semi-Automatic Segmentation—The first stage of semi-automatic segmentation involves defining the rectilinear image domain in which segmentation will be performed. It is often desirable for this domain to be smaller than the whole main image, so as to reduce computational and memory demands of the segmentation algorithm. It is also desirable for the images input to the active contour segmentation algorithm to have approximately isotropic voxel size, as noted in Subsection 2.4.3.

In the ROI selection stage, the user defines the corners of a rectilinear ROI that contains the object of interest, and can optionally set the voxel size for the ROI to be different from that of the main image. All images are then cropped and resampled to the space of the user-selected ROI. All subsequent operations are performed on these cropped and resampled images. We note that while the multiple image layers loaded in a single ITK-SNAP session may be in different voxel spaces (i.e., have different resolution and orientation from each other), they are brought into a common voxel space for the purpose of semi-automatic segmentation.

2.4.2. Intensity-Based Presegmentation—For each voxel x in the ROI, multiple intensity values may be available, e.g., if the user applies segmentation to multiple co-

registered image volumes or to a multi-component image. Presegmentation reduces all the image intensity values available at a voxel to a single scalar value. The resulting scalar image $g(x)$ is called the *speed image.* ITK-SNAP offers four presegmentation modes: supervised classification, unsupervised classification, soft thresholding, and edge detection.

- In the *supervised classification* mode, presegmentation estimates the probability $P_{\text{obj}}(x)$ of the voxel x belonging to the object of interest and the probability $P_{\text{bkg}}(x)$ of it belonging to the background. These probabilities are estimated by training a random forest classifier [6, 10] on a set of example voxels identified by the user via painting operations. The probabilities are estimated using all available image intensity values at x. The speed image has range $[-1,1]$ and is defined as the difference between object and background probabilities, $g(x) = P_{\text{obj}}(x) - P_{bkg}(x)$. Fig. 2 shows ITK-SNAP in this presegmentation mode.

- In the *unsupervised classification* mode, the speed image also estimates the difference between object and background probabilities at each voxel. These probabilities are also estimated using all available image intensity values at each voxel. However, this estimation is obtained without training data using a Gaussian mixture model and the Expectation-Maximization (EM) algorithm [11]. The user specifies the number of distinct tissue classes in the ROI, and the initial parameters for each class are randomly seeded using the k-means++ algorithm [2].

- In the *soft thresholding* mode, the speed image is also of the form $g(x) = P_{\text{obj}}(x) - P_{bkg}(x)$, but the foreground and background probabilities are estimated in a more rudimentary way. A soft binary threshold function with user-supplied upper and lower threshold values is applied to a single image intensity component selected by the user. Intensity values between the lower and upper thresholds are assigned positive speed values, and values outside the thresholds map to negative speed values. The soft thresholding mode corresponds to the "region competition" segmentation approach developed by Zhu and Yuille [37], and its implementation within ITK-SNAP is described and evaluated in [35].

- In the *edge attraction* mode, the speed image has the range $[0,1]$ and the speed image is derived from the gradient magnitude of a single image intensity component selected by the user. Large gradient magnitude values (strong edges) are mapped to small speed values, and vice versa. The edge attraction mode corresponds to the "geodesic active contours" approach described by Caselles [7, 8], and its implementation within ITK-SNAP is described in [35].

The remainder of this section focuses on the supervised classification mode, which is used in all evaluation experiments in this paper. In this mode, the user specifies examples of k 2 tissue classes present in the segmentation ROI. Examples are specified by painting brushstrokes in one or more orthogonal slice views (the polygon tool can also be used). Each voxel painted by the user is treated as a separate example $\{F_j, y_j\}$ for training the random decision forest classifier, where F_j denotes the vector of features associated with the j-th example voxel and $y_j \in \{1,\ldots, k\}$ is its tissue class.

By default, the feature vector F_j consists of all the image intensity values available for the j-th voxel (i.e., all the components of all the images). However, the feature vector can also be made to include all the intensity values in the rectangular patch centered on the j-th voxel, with the patch size set by the user. Including neighboring intensities as additional features makes it possible for the classifier to learn more complex intensity patterns that separate different tissue classes. As illustrated in Fig. 3, patch features make it possible to differentiate between image regions based on texture. The feature vector F_j may also be made to include the coordinates of the j-th voxel as features. This makes it possible to differentiate between image regions that have identical intensity characteristics but distinct locations in the image, as illustrated in Fig. 3.

The random forest algorithm [6, 10] is applied to the training data. The algorithm trains an ensemble of decision tree classifiers. Each decision tree is trained using a random bootstrap sample of the training data, and a random sample of the features [6, 10]. The number of decision trees and the depth of each decision tree are user-adjustable parameters (defaulting to 50 and 30, respectively).

After training, the random forest classifier is applied to all voxels in the ROI. For each voxel x, the feature vector F_x is constructed, and each decision tree in the ensemble is applied to F_x, resulting in a set of posterior probability values $P_{x,l}^{t}$, where $l \in \{1,\ldots, k\}$ is the class index and $t \in \{1,\ldots, T\}$ is the tree index. The ensemble posterior probability for voxel x and label l is computed as $P_{x,l} = \Sigma_{t=1}^{T} P_{x,l}^{t}$.

For the computation of $P_{obj}(x)$ and $P_{bkg}(x)$, the user tags one or more segmentation labels as corresponding to the object of interest, and the remaining labels are assigned to the background. Then we set $P_{obj}(x) = \Sigma_{l \in obj} P_{x,l}$ and $P_{bkg}(x) = \Sigma_{l \in bkg} P_{x,l}$.

The user interface for the supervised classification mode is lightweight. It includes a set of buttons for selecting classes (labels) for painting examples, a button to train/retrain the classifier, a button to clear all examples, and a list of defined classes in which labels can be tagged as object or background. A separate window allows the user to specify how the feature vectors are constructed, to set classifier parameters, and to export and import examples.

The user interface is also highly responsive in order to allow interactive modification of the training data. During presegmentation, the random forest is applied selectively to the input image volume, so that only the slices visible to the user are classified. This is much faster than applying the classifier to the whole image. If the user moves the 3D crosshair (thus changing which slices are shown in the three ITK-SNAP views), the classification is recomputed on the fly. This allows the user to repeatedly retrain the classifier until a desired classification is accomplished. For example, if a particular area of the object of interest is mislabeled, the user can paint some voxels in that area with the object label and retrain the classifier. The availability of undo/redo functionality for the painting operations also speeds up classifier training.

2.4.3.   Active Contour Segmentation—In the supervised classification, unsupervised classification, and soft thresholding modes, positive values of the speed image correspond to parts of the image that have higher probability of being the object than the background. However, simply thresholding the speed image at 0 usually fails to provide a satisfactory segmentation. Firstly, there may be multiple objects of interest in the image (e.g., left and right caudate, or multiple lesions) that need to be assigned different labels; and there may be parts of the background that have nearly identical appearance to the object of interest. Secondly, the speed image may be noisy due to imaging noise and due to the fact that presegmentation is applied independently to each voxel. In ITK-SNAP, presegmentation is followed by a more geometric *active contour segmentation* step, in which seeds are placed inside of the specific object of interest and grown in a way that balances adherence to the speed image with a geometric regularization term [37, 8, 28, 34].

The active contour evolution algorithm implementation in ITK-SNAP was described previously in [35], and we only provide a brief summary here for completeness. Let t be time, and let C_t be a smooth contour in $\mathbb{R}^3$, i.e., there exists a continuous, smooth function $\phi_t : \mathbb{R}^3 \to \mathbb{R}$ such that $C_t = \{x \in \mathbb{R}^3 : \phi(x) = 0\}$. The contour evolves according to the differential equation

$$\frac{dC_t}{dt} = [g(C_t) + \alpha \cdot \kappa_{C_t}] \cdot N_{C_t}, \quad (1)$$

where $g(x)$ is the speed function, κ_{C_t} is the mean curvature of the contour C_t, and N_{C_t} is the unit outward normal to the contour C_t, and α is a scalar parameter set by the user. Under this evolution equation, the contour expands into regions where the speed function is positive (and contracts where the speed function is negative), while also contracting at points where curvature is high. The evolution equation (1) corresponds to the variational gradient descent of an energy function that maximizes the energy

$$E[C] = \int_c g(x)dx - \int_{\mathbb{R}^3 \setminus \mathscr{C}} g(x)dx + \alpha \int_C dA,$$

where $\mathscr{C}$ denotes the interior of the contour C, and dA is the element of area. Numerically, the contour evolution (1) is solved using the level set method [28], which expresses all terms of (1) in terms of the function ϕ_t and uses a robust finite difference scheme to approximate derivatives. An efficient extreme narrow banding method that only computes ϕ_t at a set of nodes adjacent to the zero contour [34] is used for computational efficiency. The requirement for approximately isotropic voxels in active contour segmentation stems from the fact that the surface normal and mean curvature of C are approximated from the partial derivatives of ϕ, and the approximation is inaccurate when voxels have large aspect ratios (e.g., 1:2 or greater).

The workflow for active contour segmentation consists of selecting an active label for the segmentation (in the supervised classification mode, this label is pre-populated as the first "object" tissue class); placing spherical seeds in the ROI; and supervising the active contour

evolution. The user can choose to advance the evolution by fixed step size, or continuously, until pressing "stop". The evolving contour is visualized in real time in 2D slices and in 3D if enabled by the user.

As the last step, the user "accepts" the segmentation. The active contour segmentation is then resampled into the space of the ITK-SNAP main image and integrated into the overall segmentation image.

2.4.4. Segmentation of Multiple Structures in Supervised Classification Mode

—The supervised classification mode can be used to define multiple tissue classes in the image (e.g., edema, non-enhancing tumor core, enhancing tumor core, necrosis, healthy tissue, etc. in the case of glioblastomas), whereas the active contour segmentation only segments a single object at a time. To facilitate the segmentation of all relevant objects in the image, ITK-SNAP retains the classifier training data after active contour segmentation is completed. To segment additional objects in the ROI, the user re-enters the semi-automatic segmentation mode, assigns a different combination of tissue classes as object and background, and applies active contour evolution; all without having to re-train the random forest classifier.

3. Experiments and Results

The overall goal of the evaluation experiments is to show that ITK-SNAP can be used to perform complex image segmentation tasks in multi-modality image data quickly and reliably.

3.1. Brain Tumor Segmentation in Multi-Contrast MRI

The primary evaluation is carried out in the context of semi-automatic segmentation of high-grade and low-grade gliomas in multi-contrast MRI from the 2013 MICCAI Brain Tumor Segmentation (BRATS) challenge [25], BRATS challenge data has been used to evaluate dozens of brain tumor segmentation algorithms, so it offers a well-established benchmark for evaluating ITK-SNAP segmentation performance. The reliability of "ground truth" manual segmentation in BRATS data is also known [25, Figure 5].

Our evaluation uses data from the 25-subject "leaderboard" subset of the 2013 BRATS dataset [25]. For each patient, four MRI scans are provided: precontrast Tl-weighted, T2-weighted and FLAIR scans, as well as a gadolinium contrast enhanced Tl-weighted scan (TICE). All four scans are co-registered by BRATS organizers and resampled to 1mm × 1mm × 1mm resolution. Most high-grade gliomas have four distinct tissue classes: edema, which appears bright on T2 and FLAIR; enhancing tumor core (EC), which appears bright on T1CE; non-enhancing tumor core (NEC), which is abnormal in T2 but appears as normal gray/white matter in T1CE; and necrosis, which appears dark in T1. However not all classes are present in all subjects and appearance can be variable. Low-grade gliomas typically do not have EC or necrosis.

The BRATS leaderboard dataset includes 21 scans of patients with high-grade gliomas and 4 scans of patients with low-grade gliomas. These include 15 cases (11 high-grade, 4 low-

grade) that were used for off-site evaluation both in BRATS 2012 and BRATS 2013, and 10 additional cases that were used only in the 2013 challenge (Figure 4). The 15-case subset (subset B in Figure 4) was used to compare 20 tumor segmentation methods in the report on BRATS 2012 and 2013 by Menze et al. [25].

Gliomas were segmented in ITK-SNAP using the following protocol. In the supervised classification presegmentation mode, seven tissue classes were available: edema, EC, NEC, necrosis, normal brain tissue, cerebrospinal fluid (CSF), air/background. In most segmentations the first five classes were marked by the raters, and CSF was marked when tumors were adjacent to the CSF. Active contour segmentation was performed repeatedly, starting from the whole tumor and working inwards, as illustrated in Fig. 5. First, the combined tumor region (edema+NEC+EC+necrosis) is segmented; then (NEC+EC +necrosis); then (EC+necrosis); and finally necrosis only. This sequence takes advantage of the fact that in most tumors necrosis lies within the EC, which is within the NEC, which in turn is within the edema, and minimizes the need to label structures with holes.

The whole BRATS 2013 leaderboard dataset was segmented twice by a non-expert rater (AP) who had no previous experience with image segmentation or brain tumor segmentation. After studying the BRATS manual segmentation protocol [22] and ITK-SNAP tutorials, this rater practiced on a set of 20 cases with available segmentations from the BRATS "training" subset for about one week. The rater then segmented the 25-case leaderboard dataset sequentially over the course of 10 days. Following a one-month delay, the rater segmented each dataset again. The total segmentation time (from loading a workspace file in ITK-SNAP to saving final segmentation) was recorded for each segmentation attempt. The strokes used to train the random forest classifier were also saved as an image volume.

Additional segmentation was performed independently by three expert neuroradiologists (JES, JMS, SM) who had no prior experience with ITK-SNAP. The neuroradiologists performed segmentation in a subset of 20 cases (16 high-grade, 4 low-grade), designated as subset C in Figure 4. Subset C includes the 15-case subset B used for the comparison methods in the BRATS 2012 and 2013 challenges [25, Figure 7]. A smaller subset of 5 images (3 high-grade, 2 low-grade) were segmented twice by each neuroradiologist after a delay of at least two weeks. The neuroradiologists attended a two-hour training session from ITK-SNAP developers, watched ITK-SNAP training videos online, and practiced on the 20-subject training subset until they felt comfortable with the tool and the segmentation protocol.

Segmentations performed by different raters, as well as repeat segmentations by the same rater, were compared in terms of the Dice similarity coefficient [12] and volume. These measurements were conducted in a manner consistent with evaluations in the BRATS challenge [25]. Specifically, Dice coefficient was computed and reported for the "complete tumor" (edema+NEC+EC+necrosis), "tumor core" (NEC+EC+necrosis) and "enhancing core" (EC) for high-grade gliomas and "complete tumor" and "tumor core" for low-grade gliomas. Dice coefficient is a measure of relative overlap between segmentations, defined as

the ratio of the volume of the intersection between two segmentations to the average volume of the two segmentations, and ranging between 0 and 1.

3.1.1. Results: Intra-Rater and Inter-Rater Reliability—The intra-rater reliability for the non-expert rater is summarized in Table 2. There is a substantial difference between the mean and median Dice coefficient, driven in part by zero Dice coefficient for one subject (high-grade 137), for which the rater labeled a completely different part of the image as tumor. The intra-rater reliability for low-grade gliomas is much lower than for high-grade gliomas, particularly for the tumor core. Overall, the median *intra-rater* Dice coefficient for our non-expert rater in subset B (0.92 for complete tumor, 0.81 for tumor core, 0.78 for enhancing region) compares favorably with the *inter-rater* reliability of the ground truth BRATS manual segmentation reported in [25, Figure 5] for the same set of cases (0.85 for complete tumor, 0.84 for tumor core, 0.72 for enhancing region). Mean intra-rater Dice for our non-expert rater (0.83 for complete tumor, 0.63 for tumor core, 0.64 for enhancing region) compares less favorably with [25, Figure 5] (0.81 for complete tumor, 0.77 for tumor core, 0.68 for enhancing region), which is likely driven by the lower performance of the non-expert rater on low-grade cases and the outlier high-grade case 137.

Average intra-rater reliability for each of the three radiologists and average inter-rater reliability between all pairs of radiologists are reported in Table 3. Intra-rater reliability is consistently higher than inter-rater reliability, as would be expected, since inter-rater disagreements may include differences in the in-terpretation of underlying anatomy, while intra-rater disagreements, in principle reflect difficulty in applying a given set of anatomical rules consistently. Additionally, Table 4 compares the average inter-rater reliability of three radiologists using ITK-SNAP and the average inter-rater reliability of three raters who provided reference manual segmentation in [25] in the same set of images. Compared with [25], the ITK-SNAP inter-rater reliability is lower for high-grade cases and higher for low-grade cases; the average over all cases is lower for ITK-SNAP, particularly for the tumor core.

The intra-class correlation coefficients (ICC, [30]) in Table 3 display a wide range, and higher ICC values do not always correspond to higher inter-rater and intra-rater Dice coefficient. Figure 6 uses Bland-Altman plots [5] to plot the between-rater and within-rater disagreement in volume. Large range of ICC values is likely explained by very different ranges of volume for the different regions. For example, for the enhancing tumor region, the between-rater and within-rater error is approximately the same in absolute terms, but the range of volumes is very small for the set of 3 high-grade cases in which intra-rater ICC is computed, resulting in a very low ICC (0.1).

3.1.2. Results: Comparison to BRATS Reference Segmentation—
Segmentations by the three experts and the non-expert were uploaded to the online BRATS evaluation system, which reports the Dice coefficient between each segmentation and the BRATS reference segmentation, which is a consensus segmentation derived from combining multiple manual segmentations [25]. Table 5 reports the mean and median overlap for each rater on the set of 20 cases that were segmented by all four raters (Subset C in Figure 4). The non-expert's agreement with the BRATS reference is generally on par with the experts for

the high-grade gliomas, but lower than that of the experts on the low-grade gliomas. Combining across low-grade and high-grade gliomas, Expert 2 has the highest agreement with the BRATS reference of all four raters.

Table 6 estimates how the ITK-SNAP segmentations by the four raters compare to the 20 brain tumor segmentation techniques evaluated in the BRATS 2012 and 2013 challenges [25, Figure 7]. The "Rank" columns of Table 6 indicates where the segmentation by each expert would rank with respect to the 20 methods evaluated in [25, Figure 7]. The best overall ranking is achieved by Expert 2, with Experts 1, 3 and the non-expert having very similar ranking profiles. However, for low-grade gliomas, Expert 3 has the best ranking. If the Dice coefficient across all three regions is averaged, then ITK-SNAP segmentation by experts 2 and 3 comes out in the first place relative to the 20 methods in [25, Figure 7] and in the second place for Expert 1 and the non-expert.

The BRATS online evaluation system is open and new methods are continually added. The online system ranks methods based on average Dice coefficient relative to the BRATS reference segmentation across the 25-case leaderboard dataset (Subset A in Figure 4). As of April 2017, the segmentation produced by the non-expert was given overall rank 5 of 45 by the online system. The dataset combining segmentations by Expert 2 for subset C and non-expert segmentation for the remaining 5 leaderboard cases, was given overall rank of 4 of 45.

3.1.3.   Segmentation Effort—The distribution of segmentation times for each rater, separated by tumor grade, is plotted in Figure 7. The mean segmentation time for the non-expert across all cases was 12.3 min, while the mean segmentation for the three experts was higher: 24.1 min, 24.3 min, and 16.2 min, respectively. We used the Wilcoxon signed rank test to determine whether the differences in segmentation time between pairs of raters were statistically significant. Significant differences were found between Expert 1 and non-expert ($p < 0.001$), Expert 2 and non-expert ($p < 0.001$), Expert 2 and Expert 3 ($p = 0.005$); a trend toward significance was found between Expert 2 and Expert 3 ($p < 0.06$).

Figure 8 plots the number of training voxels labeled by each rater for each tissue class during supervised classification. The average number of training voxels per case was greatest for Expert 1, followed by the non-expert, Expert 2, and Expert 3. Differences in the number of training voxels were statistically significant on the Wilcoxon signed rank test between all pairs of raters ($p < 0.01$), except between Expert 2 and the non-expert.

For the non-expert rater, who has the most training data available for analysis, and whose training examples were consistently drawn on disjoint slices, we estimated the number of training strokes by performing 1-voxel morphological erosion on each x, y and z slice, and counting the number of connected components. The number of training strokes correlated significantly with segmentation time ($R = 0.51$, $p = 0.01$) and with Dice coefficient on BRATS evaluation ($R = 0.52$, $p = 0.01$) . These correlations remained significant if only high-grade gliomas were considered. However, segmentation time was not significantly correlated with Dice coefficient on BRATS evaluation ($R = 0.25$, $p = 0.26$).

3.2. Placenta Segmentation

A secondary evaluation of ITK-SNAP semi-automatic segmentation was performed in the context of segmenting the placenta in 3D ultrasound (3DUS) images taken during the first trimester of pregnancy. The placenta is difficult to segment in 3DUS because of non-uniform intensity within the placenta, high levels of noise, and weak intensity contrast between the placenta and surrounding uterine tissue. A dataset of 25 first-trimester (11–14 weeks) 3DUS scans acquired with the GE Voluson E8 ultrasound machines was used in this study. Images were exported in cartesian format with isotropic resolution, ranging from 0.34mm × 0.34mm × 0.34mm to 0.61mm × 0.61mm × 0.61mm.

Manual segmentation was performed in ITK-SNAP in collaboration between a non-expert rater (NY) and an expert obstetrician with over 10 years of experience in prenatal ultrasound imaging and placenta segmentation (NS). The expert supervised training for the non-expert and inspected and approved each segmentation. The non-expert traced the placental outline in approximately every fifth slice in all three orthogonal slice planes. The resulting sparse segmentations were reviewed by the expert, and if needed, sent back to the non-expert rater for correction. Sparse segmentations were interpolated to create a smooth 3D placental volume, which was reviewed and corrected if necessary. Although segmentation time was not recorded, it took over ten hours of total segmentation time per case with the correction and editing.

A semi-automatic placenta segmentation protocol was developed using a subset of 13 "training" images for which he had access to manual segmentations. Images were first processed using the $3 \times 3 \times 3$ voxel median filter to reduce speckle noise. The supervised classification mode in ITK-SNAP was used to define examples of seven tissue classes: placenta, anterior uterine muscle, posterior muscle, amniotic fluid, fetus, fat and other. The feature set for supervised classification included both patch intensity features (with patch radius 2, i.e., $5 \times 5 \times 5$ voxel patches) and coordinate features. The coordinate features helped define placental boundaries in regions where there was little or no contrast between the placenta and the adjacent structures. An example of placenta presegmentation is shown in Fig. 9.

A non-expert rater (AP) applied this protocol to a "testing" set of 12 placenta 3DUS images for which manual segmentation were available. The mean Dice coefficient between the manual and semi-automatic segmentation was 0.88±0.04, and the median Dice coefficient was 0.89. The average segmentation time using the ITK-SNAP semi-automatic protocol was 27.8 ± 7.0 min per case.

4. Discussion

4.1. Brain Tumor Segmentation

The evaluation in BRATS data illustrates the ability of ITK-SNAP to label complex multi-label structures quickly and effectively. Semi-automatic segmentation by the non-expert rater required only 12.5 min per case on average, compared with 60 min per case for the manual segmentation used to generate the BRATS ground truth [25]. Experts took longer to perform segmentation (from 16 to 24 min, on average), and experts 2 and 3 had better overall

accuracy relative to the BRATS reference than the non-expert. ITK-SNAP segmentation by Experts 2 and 3 (Table 6) was very competitive with the 20 methods evaluated in [25], achieving first and second ranks in almost all categories in Table 6 Segmentation by the non-expert and expert 1 was also competitive, ranking just behind the top methods in [25]. When new methodology development since 2013 is taken into account, tumor segmentation by ITK-SNAP remains competitive, ranking 4th or 5th (depending on the rater) in the online BRATS leaderboard evaluation system.

A caveat in the comparison with [25] is that most of the methods participating in the BRATS challenges were either fully automatic or required minimal seeding to initialize [25], whereas ITK-SNAP is a highly interactive method that required raters to spend 10–15 minutes training the classifier. Indeed, the fact that intra-rater reliability of ITK-SNAP is in the 0.7–0.9 Dice coefficient range for most combinations of region/grade (Tables 2 and 3) indicates the significant impact that user input plays in determining the ITK-SNAP segmentation result. This can be viewed as both a drawback and a strength. The drawback is operator bias: just as in manual segmentation, results produced by one user may not be reproduced well by another user. The strength, compared to fully automatic segmentation, is that the expert is largely in control of the anatomical definitions applied during segmentation. Differences between the experts' segmentations are driven, at least in part, by differences in their interpretations of the underlying anatomy. As demonstrated in a typical example shown in Figure 10, the experts and the non-expert frequently disagreed as to what constituted different parts of the tumor. Crucially, the neuroradiologists did not study and try to mimic the manual segmentation protocol used in [25], but instead applied their individual understanding of what constitutes the edema, enhancing and non-enhancing core, and necrotic components of the tumors.

The successful methods in BRATS were specialized to the problem of brain tumor segmentation and incorporated machine learning classifiers trained on a dataset of 30 annotated cases. Menze et al. [25] conclude that "a majority of the top ranking algorithms relied on a discriminative learning approach, where low-level image features were generated in a first step, and a discriminative classifier was applied in a second step, transforming local features into class probabilities with MRF regularization to produce the final set of segmentations". By contrast, ITK-SNAP is a general-purpose tool, and the Random Forest [6, 10] classifiers used in ITK-SNAP were trained using the input images themselves, without a need for a separate annotated dataset. This suggests that ITK-SNAP is a viable alternative to manual segmentation in "novel" problems where annotated training data has not yet been generated.

The inter-rater reliability between the experts using ITK-SNAP was below that of the inter-rater reliability of BRATS reference segmentation [25, Figure 5] for high-grade tumors (Table 4). However, the BRATS segmentation itself was not purely manual: for example, segmentation was performed on every second or third slice and interpolated; and Gaussian smoothing was performed following the segmentation [22]. Local threshold-based segmentation was used to segment the enhancing region and the necrosis [22]. Interpolation, smoothing and threshold-based painting operations can result in more consistent segmentations between raters and result in higher inter-rater Dice coefficient than purely

manual segmentation might produce. Furthermore, BRATS manual segmentation followed a detailed and procedural protocol [22] that involved tracing edema in T2-weighted MRI and tumor sub-regions on T1CE MRI [22]. The manual segmentation in BRATS also had access to native resolution MRI scans. By contrast, the segmentation in ITK-SNAP used information from all modalities concurrently, and worked with images resampled to a common 1mm^3 isotropic resolution. The three radiologists defined the tissue classes in the image based on their individual interpretation of anatomy, which likely resulted in greater divergence between experts than in the BRATS reference segmentation.

4.2. Placenta Segmentation

In the placenta segmentation problem, the accuracy of semi-automatic segmentation in ITK-SNAP relative to manual segmentation (average Dice coefficient of 0.88 ± 0.04) compared favorably with previously published work on semi-automatic segmentation of the placenta in 11–13 week 3DUS scans using the random walker algorithm [9, 33], where the average Dice coefficient of 0.86 ± 0.06 is reported; and with fully automatic multi-atlas segmentation results, where Dice coefficient of 0.83 ± 0.05 is reported [26]. However, the segmentation time in ITK-SNAP (27.8 ± 7.0) was several times greater than the average time of 3.6 min reported for the random walker segmentation [33]. One possible factor explaining the difference in time is that the initialization for the random walker algorithm is performed in a manner favorable for placenta identification: in slices taken radially around the placenta; whereas in ITK-SNAP the placenta is viewed in orthogonal planes that may make placental anatomy harder to interpret.

4.3. Conclusions from the ITK-SNAP Evaluation

The main conclusion of the evaluation in brain tumor and placenta segmentation is that while ITK-SNAP does not unequivocally improve on existing segmentation solutions in these domains, it offers segmentation performance competitive with the state of the art, without being specially tuned, optimized, or trained for these specific problems. The ability of a general-purpose semi-automatic segmentation tool to compete with highly specialized and/or heavily trained approaches in these two problems suggests that ITK-SNAP can be an effective tool in a broad range of complex image segmentation problems, including problems for which there currently do not exist specialized solutions or annotated data for training machine learning and multi-atlas based techniques. Even in problems where automated algorithms have been developed, they may not always be available in the form of ready-to-use software, making ITK-SNAP a preferred segmentation solution.

4.4. ITK-SNAP in Relation to Other Medical Image Analysis Software

A number of mature medical image analysis software tools provide automatic and semi-automatic segmentation capabilities. General-purpose segmentation functionality is offered in 3D Slicer [19, 15], ImageJ [1, 4], BioImage Suite [13] and MIPAV [24]. Additionally, a large number of domain-specific tools that support segmentation exists, such as FreeSurfer [16], FSL [31] or SPM [3] provide excellent automated tools for brain MRI segmentation, but they are not general-purpose segmentation tools. To our knowledge, ITK-SNAP is the first interactive image segmentation tool to combine Random Forest classification and active contour segmentation in a single interactive segmentation workflow. It is also the first to

extend such a workflow to concurrent semi-automatic segmentation of multiple image modalities. In the 2D microscopy segmentation domain, the tool Ilastik [32], which also leverages Random Forest classifiers trained interactively by the user on multi-channel images offers similar scope and functionality to ITK-SNAP. 3D Slicer is the general-purpose 3D medical image analysis tool that arguably comes closest in offering the semi-automated segmentation functionality in ITK-SNAP. However, 3D Slicer follows a distinct design philosophy from ITK-SNAP that emphasizes "breadth of functionality [and] extensibility" [15, p. 1324] that "separates Slicer from such task-oriented packages as ITK-Snap" [15, p. 1324]. For example, 3D Slicer offers extensive functionality for image filtering, non-linear registration, surface model visualization and editing, diffusion MRI processing and tractography, and image-guided surgery [15]. By contrast, ITK-SNAP focuses specifically on the task of image segmentation and eschews functionality not directly related to segmentation. 3D Slicer is a much larger tool that ITK-SNAP, with over 570,000 lines of C++ and Python code (not counting the 90+ extensions that do not form the core of the software). By contrast, ITK-SNAP has 190,000 lines of code and does not currently offer extensions. 3D Slicer offers several built-in options for image segmentation, such as tissue classification based on the expectation-maximization algorithm [27], a simple region growing algorithm[3], a competitive region growing algorithm "GrowCut" [14], a more recent Fast GrowCut algorithm [36], and active contour segmentation via robust statistics [18]. The extensions of 3D Slicer supporting general-purpose segmentation include a tool for watershed segmentation[4], and a tool called the "Segmentation Wizard" [5]. The above modules, with the exception of the Segmentation Wizard, take only a single-modality image as the input. Semi-automatic GBM segmentation using 3D Slicer was evaluated in [14]; the authors determined that the GrowCut module was the most suitable of the available tools, and was shown to have good overlap with manual segmentations (88% Dice coefficient) while reducing segmentation time by 39% over manual segmentation. However, direct comparison of the quantitative results in [14] and the current paper cannot be made for several reasons: (1) in [14], segmentation was performed in a single modality (contrast-enhanced T1-MRI) and only using a single anatomical label (tumor vs. no tumor); (2) in [14], the segmented tumors were considerably smaller than in the BRATS challenge (average tumor volume 27.7cm^3, as opposed to 82.4cm^3 for the complete tumor and 60.1 cm^3 for the tumor core in the current paper); (3) the comparison of semi-automatic and manual segmentation in [14] was between segmentations performed by the *same rater*, whereas in the current paper, we report accuracy relative to the BRATS consensus reference segmentation derived by different raters.

4.5. Limitations

A limitation of ITK-SNAP is that it is a general-purpose image segmentation tool, and thus not optimized for any specific segmentation problem. As the result, it is likely to produce segmentation results inferior to tools that are specialized. However, as noted in the Introduction, expert-guided segmentation is needed both to create and train new specialized

---

[3]https://www.slicer.org/wiki/Documentation/4.8/Modules/SimpleRegionGrowingSegmentation
[4]https://www.slicer.org/wiki/Documentation/4.4/Extensions/Wasp
[5]https://www.slicer.org/wiki/Documentation/Nightly/Extensions/SegmentationWizard

tools, and to address existing segmentation problems for which specialized tools have not yet been developed. A particular limitation of the active contour algorithm implemented in ITK-SNAP is that for thin tissue layers (e.g., fewer than 3–4 voxels thick), the active contour may not flow into the tissue layer due to the strong action of the curvature term in (1) and due to the inherent limitation of the level set method in representing parallel contours separated by short distances. These limitations can be overcome by reducing the weight of the curvature term and/or by super-sampling the input image (effectively making the tissue layer thicker in units of voxels). However, these mitigation strategies add time to segmentation and may result in leakage and poor segmentation due to under-regularization. In the tumor segmentation problem, we took advantage of the adjacency of thin layers composing the tumors by performing segmentation in a nested fashion, so that thin layers never had to be segmented directly. Such a nesting strategy may not always be feasible in other applications. In particular, the active contour algorithm as implemented in ITK-SNAP is not well suited for the segmentation of thin tubular structures, such as vessels.

A limitation of the evaluation in this paper is that ITK-SNAP was compared to manual segmentation and, in the case of brain tumors, to specialized brain tumor segmentation algorithms, *but not to other general-purpose interactive image segmentation tools.* It would have been difficult to conduct such an evaluation in a manner that is free of real or perceived bias, since in the current study, the developers of ITK-SNAP participated in the evaluation (e.g., by training the neuroradiologists) while the developers of other interactive open-source tools did not. However, the fact that the brain tumor evaluation was performed on a publicly available BRATS challenge dataset will allow the developers of other tools to carry out similar evaluations independently and to compare their segmentation results to those reported in this paper in a direct way.

Another limitation of the evaluations performed in this paper is the use of Dice coefficient to characterize segmentation accuracy. Dice coefficient provides only one aspect of segmentation accuracy and the value is correlated with the shape of the segmented object, e.g., the same displacement will result in a much larger reduction in Dice overlap for thin shapes than for blob-like shapes. Boundary distance metrics are frequently reported alongside Dice coefficients when evaluating segmentation, but in this paper, such data were unavailable to us for the BRATS challenge because the underlying reference manual segmentations are not publicly available.

Information Sharing Statement

The ITK-SNAP software (source code and binaries) are hosted on Source-Forge at https:// sourceforge.net/projects/itk-snap, and can also be accessed through the ITK-SNAP website www.itksnap.org and the Neuroimaging Tools & Resources Collaboratory (NITRC) at https://www.nitrc.org/projects/itk-snap.

The BRATS challenge data were obtained from the Swiss Institute for Computer Assisted Surgery (SICAS) Medical Image Repository at https://www.smir.ch/BRATS/Start2013. The ITK-SNAP segmentations of the BRATS datasets, as well as the placenta ultrasound scans, manual segmentations, and ITK-SNAP segmentations are available in the ITK-SNAP

NITRC repository https://www.nitrc.org/projects/itk-snap under the package "nein2018". The R language statistical analysis scripts and spreadsheets used to generate the tables and figures in this paper are hosted on Github at https://github.com/pyushkevich/nein2018_stats.

Acknowledgement

References

[1]. Abramoff M, Magelhaes P, Ram S, 2004 Image Processing with ImageJ. Biophotonics International 11 (7), 36–42.

[2]. Arthur D, Vassilvitskii S, 2007 K-means++: The advantages of careful seeding. In: Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms. Society for Industrial and Applied Mathematics, pp. 1027–1035.

[3]. Ashburner J, 10 2009 Computational anatomy with the SPM software. Magn Reson Imaging 27 (8), 1163–74. [PubMed: 19249168]

[4]. Barboriak DP, Padua AO, York GE, Macfall JR, 6 2005 Creation of DICOM–aware applications using ImageJ. J Digit Imaging 18 (2), 91–9. [PubMed: 15827831]

[5]. Bland J, Altman D, 2007 Agreement between methods of measurement with multiple observations per individual. J Biopharm Stat 17 (4), 571–582. [PubMed: 17613642]

[6]. Breiman L, 2001 Random forests. Machine learning 45 (1), 5–32.

[7]. Caselles V, Catte F, Coll T, Dibos F, 1993 A geometric model for active contours. Numerische Mathematik 66, 1–31.

[8]. Caselles V, Kimmel R, Sapiro G, 1997 Geodesic active contours. Int J Comput Vision 22, 61–79.

[9]. Collins SL, Stevenson GN, Noble JA, Impey L, 2 2013 Rapid calculation of standardized placental volume at 11 to 13 weeks and the prediction of small for gestational age babies. Ultrasound Med Biol 39 (2), 253–60. [PubMed: 23219036]

[10]. Criminisi A, Shotton J, Konukoglu E, 2012 Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. Foundations and Trends in Computer Graphics and Vision 7 (2–3), 81–227.

[11]. Dempster AP, Laird NM, Rubin DB, 1977 Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society. Series B (Methodological) 39 (1), 1–38.

[12]. Dice LR, 7 1945 Measures of the amount of ecologic association between species. Ecology 26 (3), 297–302.

[13]. Duncan JS, Papademetris X, Yang J, Jackowski M, Zeng X, Staib LH, 2004 Geometric strategies for neuroanatomic analysis from MRI. Neuroimage 23 Suppl 1, S34–45. [PubMed: 15501099]

[14]. Egger J, Kapur T, Fedorov A, Pieper S, Miller JV, Veeraraghavan H, Freisleben B, Golby AJ, Nimsky C, Kikinis R, 2013 GBM volumetry using the 3D Slicer medical image computing platform. Sci Rep 3, 1364. [PubMed: 23455483]

[15]. Fedorov A, Beichel R, Kalpathy-Cramer J, Finet J, Fillion-Robin JC, Pujol S, Bauer C, Jennings D, Fennessy F, Sonka M, Buatti J, Aylward S, Miller JV, Pieper S, Kikinis R, 11 2012 3D Slicer as an image computing platform for the Quantitative Imaging Network. Magn Reson Imaging 30 (9), 1323–41. [PubMed: 22770690]

[16]. Fischl B, Salat DH, Busa E, Albert M, Dieterich M, Haselgrove C, van der Kouwe A, Killiany R, Kennedy D, Klaveness S, Montillo A, Makris N, Rosen B, Dale AM, 1 2002 Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. Neuron 33 (3), 341–55. [PubMed: 11832223]

[17]. Free Software Foundation June 2007 GNU General Public License, version 3. http://www.gnu.org/licenses/gpl.html, last retrieved 2017-03-25.

[18]. Gao Y, Kikinis R, Bouix S, Shenton M, Tannenbaum A, 8 2012 A 3D interactive multi-object segmentation tool using local robust statistics driven active contours. Med Image Anal 16 (6), 1216–27. [PubMed: 22831773]

[19]. Gering D, Nabavi A, Kikinis R, Hata N, Odonnell L, Grimson WEL, Jolesz F, Black P, III, W. W, 6 2001 An integrated visualization system for surgical planning and guidance using image fusion and an open MR. Journal of Magnetic Resonance Imaging 13, 967–975. [PubMed: 11382961]

[20]. Heimann T, Meinzer H-P, 8 2009 Statistical shape models for 3D medical image segmentation: a review. Med Image Anal 13 (4), 543–63. [PubMed: 19525140]

[21]. Iglesias JE, Sabuncu MR, 8 2015 Multi-atlas segmentation of biomedical images: A survey. Med Image Anal 24 (1), 205–19. [PubMed: 26201875]

[22]. Jakab A, 2012 Segmenting brain tumors with the Slicer 3D software. Tech. rep., Technical Report.

[23]. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, van der Laak JAWM, van Ginneken B, Sánchez CI, 12 2017 A survey on deep learning in medical image analysis. Med Image Anal 42, 60–88. [PubMed: 28778026]

[24]. McAuliffe MJ, Lalonde FM, McGarry D, Gandler W, Csaky K, Trus BL, 2001 Medical image processing, analysis & visualization in clinical research. In: CBMS '01: Proceedings of the Fourteenth IEEE Symposium on Computer-Based Medical Systems. IEEE Computer Society, Washington, DC, USA, p. 381.

[25]. Menze BH, Jakab A, Bauer S, Kalpathy-Cramer J, Farahani K, Kirby J, Burren Y, Porz N, Slotboom J, Wiest R, Lanczi L, Gerstner E, Weber M-A, Arbel T, Avants BB, Ayache N, Buendia P, Collins DL, Cordier N, Corso JJ, Criminisi A, Das T, Delingette H, Demiralp Ç, Durst CR, Dojat M, Doyle S, Festa J, Forbes F, Geremia E, Glocker B, Golland P, Guo X, Hamamci A, Iftekharuddin KM, Jena R, John NM, Konukoglu E, Lashkari D, Mariz JA, Meier R, Pereira S, Precup D, Price SJ, Raviv TR, Reza SMS, Ryan M, Sarikaya D, Schwartz L, Shin H-C, Shotton J, Silva CA, Sousa N, Subbanna NK, Szekely G, Taylor TJ, Thomas OM, Tustison NJ, Unal G, Vasseur F, Wintermark M, Ye DH, Zhao L, Zhao B, Zikic D, Prastawa M, Reyes M, Van Leemput K, 10 2015 The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Trans Med Imaging 34 (10), 1993–2024. [PubMed: 25494501]

[26]. Oguz I, Pouch AM, Yushkevich N, Wang H, Gee JC, Schwartz N, Yushkevich PA, 2016 Automated placenta segmentation from 3D ultrasound images. In: MICCAI Workshop on Perinatal, Preterm and Paediatric Image analysis (PIPPI).

[27]. Pohl KM, Bouix S, Nakamura M, Rohlfing T, McCarley RW, Kikinis R, Grimson WEL, Shenton ME, Wells WM, 9 2007 A hierarchical algorithm for MR brain image parcellation. IEEE Trans Med Imaging 26 (9), 1201–12. [PubMed: 17896593]

[28]. Sethian JA, 1999 Level set methods and fast marching methods. Cambridge University Press.

[29]. Shen D, Wu G, Suk H-I, 6 2017 Deep learning in medical image analysis. Annu Rev Biomed Eng 19, 221–248. [PubMed: 28301734]

[30]. Shrout P, Fleiss J, 1979 Intraclass correlations: uses in assessing rater reliability. Psychol Bull 86, 420–428. [PubMed: 18839484]

[31]. Smith SM, Jenkinson M, Woolrich MW, Beckmann CF, Behrens TEJ, Johansen-Berg H, Bannister PR, Luca MD, Drobnjak I, Flitney DE, Niazy RK, Saunders J, Vickers J, Zhang Y, Stefano ND, Brady JM, Matthews PM, 2004 Advances in functional and structural MR image analysis and implementation as FSL. Neuroimage 23 Suppl 1, S208–S219. [PubMed: 15501092]

[32]. Sommer C, Straehle C, Kothe U, Hamprecht FA, 2011 ilastik: Interactive learning and segmentation toolkit. In: Biomedical Imaging: From Nano to Macro, 2011 IEEE International Symposium on. IEEE, pp. 230–233.

[33]. Stevenson GN, Collins SL, Ding J, Impey L, Noble JA, 12 2015 3-D ultrasound segmentation of the placenta using the random walker algorithm: Reliability and agreement. Ultrasound Med Biol 41 (12), 3182–93. [PubMed: 26341043]

[34]. Whitaker RT, 1998 A level-set approach to 3D reconstruction from range data. Int. J. Comput. Vision 29 (3), 203–231.

[35]. Yushkevich PA, Piven J, Hazlett HC, Smith RG, Ho S, Gee JC, Gerig G, 7 2006 User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. Neuroimage 31 (3), 1116–1128. [PubMed: 16545965]

[36]. Zhu L, Kolesov I, Gao Y, Kikinis R, Tannenbaum A, 2014 An effective interactive medical image segmentation method using fast growcut. In: MICCAI workshop on interactive medical image computing.

[37]. Zhu S, Yuille A, 1995 Region competition: Unifying snakes, region growing, and Bayes/MDL for multi-band image segmentation. In: International Conference on Computer Vision (ICCV'95). pp. 416–423. URL citeseer.nj.nec.com/zhu95region.html

[38]. Zhu SC, Yuille A, 1996 Region competition: Unifying snakes, region growing, and Bayes/MDL for multiband image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 18 (9), 884–900.

[39]. Zukic D, McCormick M, Gerig G, Yushkevich P, 2016 RLEImage: run-length encoded memory compression scheme for an itk::Image. Insight Journal (published online). URL http://hdl.handle.net/10380/3562

[40]. Zukic D, Vicory J, McCormick M, Wisse L, Gerig G, Yushkevich P, Aylward S, 2016 ND morphological contour interpolation. Insight Journal (published online). URL http://hdl.handle.net/10380/3563
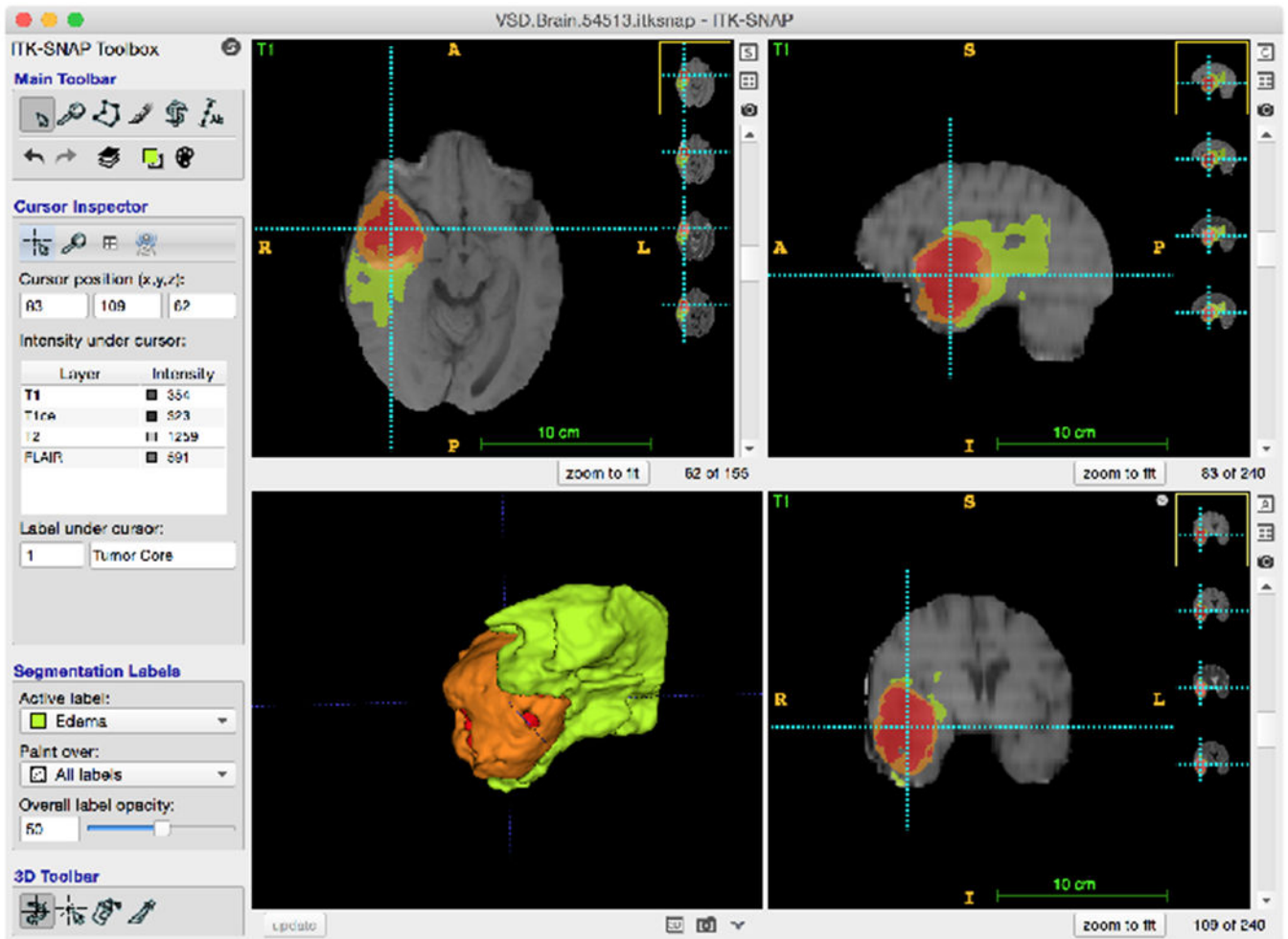
Figure 1:
Screen shot of the ITK-SNAP user interface after completed brain tumor segmentation. Three orthogonal slices through the T1-weighted MRI scan are shown, with segmentation overlaid in color. A 3D rendering of the segmentation appears in the lower left quadrant. Small thumbnails in the top and bottom right quadrants represent other MRI scans loaded in ITK-SNAP (T2-weighted, FLAIR, contrast-enhanced T1).
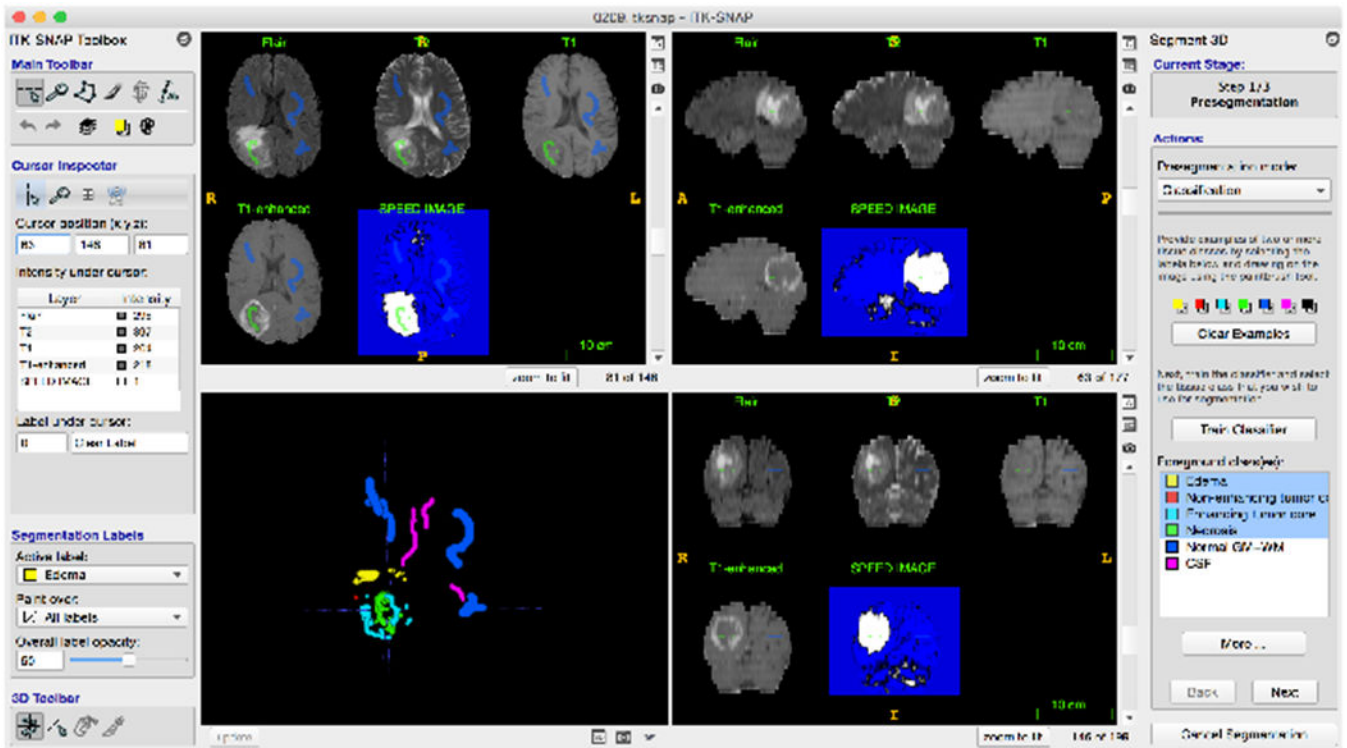
Figure 2:
Screen shot of the ITK-SNAP user interface during brain tumor presegmentation (supervised classification mode). Axial, sagittal and coronal slices through four MRI modalities and the *speed image* are in the top left, top right and bottom right quadrants of the user interface, respectively. The speed image (blue-to-white color map) has range between −1 (blue) and 1 (white), with positive values indicating higher probability that a voxel belongs to the object of interest and negative values indicating higher probability of a voxel belonging to the background. The object of interest in this example consists of all tissue classes composing the complete tumor: edema, active tumor, enhancing tumor core and necrosis. The lower left quadrant shows a 3D rendering of the samples used for training the random forest classifier. Some of the samples (necrosis: green, normal gray and white matter: blue) are also seen as color overlays in the axial, sagittal and coronal views.
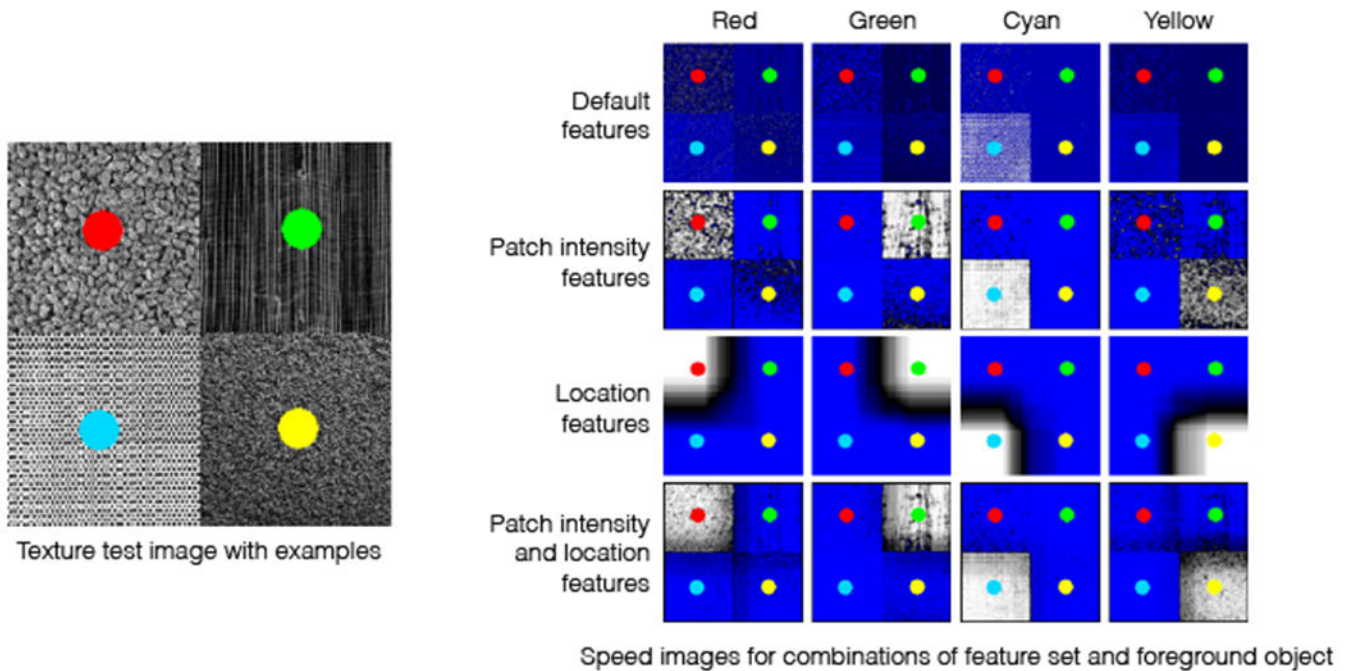
Texture test image with examples

Speed images for combinations of feature set and foreground object
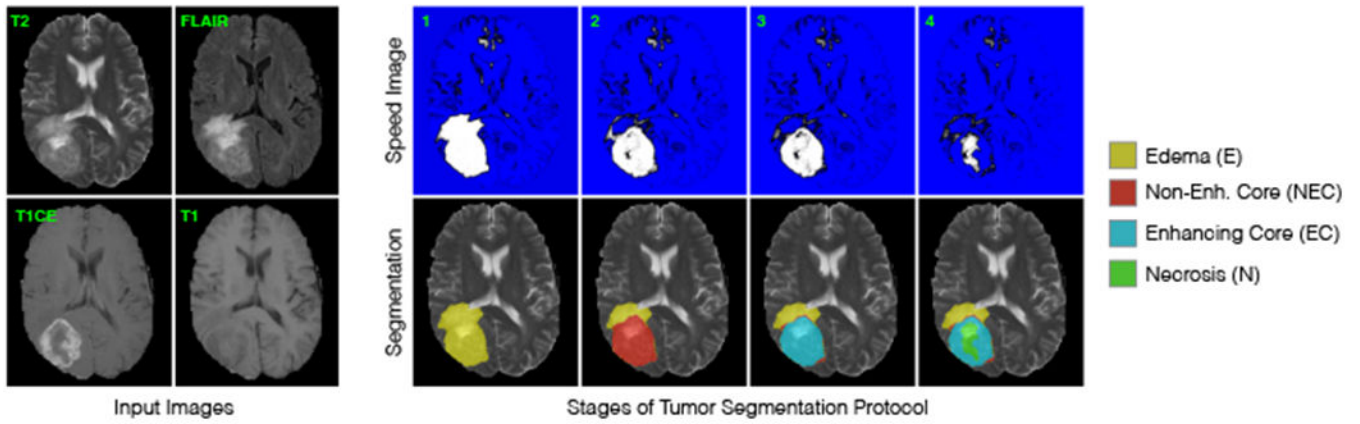
Figure 3:

Illustration of the "patch intensity" and "location" features available in the ITK-SNAP supervised classification presegmentation mode. A test 2D image on the left consists of four regions with similar mean intensity but different texture. A set of example voxels has been marked in each region (red, green, cyan and yellow circles) The 4×4 grid on the right consists of speed images generated by training a Random Forest classifier using the four circles as the training data under different conditions. The rows in the grid correspond to different set of features used to train the classifier: the default features (the intensity of each voxel serving as its only feature), patch intensity features (the set of intensities in the $5 \times 5 \times 5$ patch around the voxel used as features), location features (the coordinates of each voxel used as features), and patch and location features combined. The columns in the grid correspond to different objects (red, green, cyan and yellow) being selected as the object of interest. The addition of patch features improves the ability of the classifier to discriminate regions based on texture, while the location features allow imposition of geometrical constraints into the segmentation

Subset	N	High Grade																					Low Grade			
		116	117	119	120	130	131	133	134	135	136	137	201	202	203	204	205	206	207	208	209	210	103	105	109	116
A: BRATS 2013 Leaderboard	25	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
B: BRATS 2012/13 Evaluation	15	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓											✓	✓	✓	✓
C: Neuroradiologists' Inter-Rater	20	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓		✓		✓		✓		✓	✓	✓	✓
D: Neuroradiologists' Intra-Rater	5						✓	✓		✓													✓		✓	

Figure 4:

Composition of the different subsets of the BRATS 2012/2013 data referenced in this paper. **A:** the "Leaderboard" dataset provided for off-site evaluation in BRATS 2013. The online BRATS system (virtualskeleton.ch) continues to use this dataset for evaluating and ranking segmentation methods. **B:** the subset of 15 cases used for off-site evaluation in both BRATS 2012 and BRATS 2013. It served as the primary dataset for the comparison of 20 algorithms from the two challenges in the BRATS evaluation paper by Menze et al. [25, Figure 7]. **C:** the subset of 20 cases that was segmented using ITK-SNAP by all three neuroradiologists in this study. **D:** the subset of five cases segmented twice by the three neurorad iologists.

Stage	Object	Background	Label
1	E+NEC+EC+N	GM/WM+CSF	E
2	NEC+EC+N	E+GM/WM+CSF	NEC
3	EC+N	NEC+E+GM/WM+CSF	EC
4	N	EC+NEC+E+GM/WM+CSF	N

Figure 5:

Sequence of segmentation used by the tumor segmentation protocol. The input images for the example tumor dataset are shown on the left. The columns on the right show the speed images and the segmentations obtained in the four stages of the segmentation. Segmentation is performed proceeding from the largest object inwards, so that during each segmentation stage, the object being segmented does not have holes. The table in the bottom right describes how the different tissue classes in the image are assigned to the object of interest and background during each stage, as well as what label is assigned to the result of active contour segmentation during each stage. For example, in stage 1, the four tissue classes comprising the complete tumor are assigned to the object of interest, while healthy appearing gray/white matter and CSF are assigned to the background. This yields a speed image that is positive in the complete tumor and negative in the healthy tissue. After applying active contour segmentation to this speed image, the result is assigned the edema label.

Figure 6:

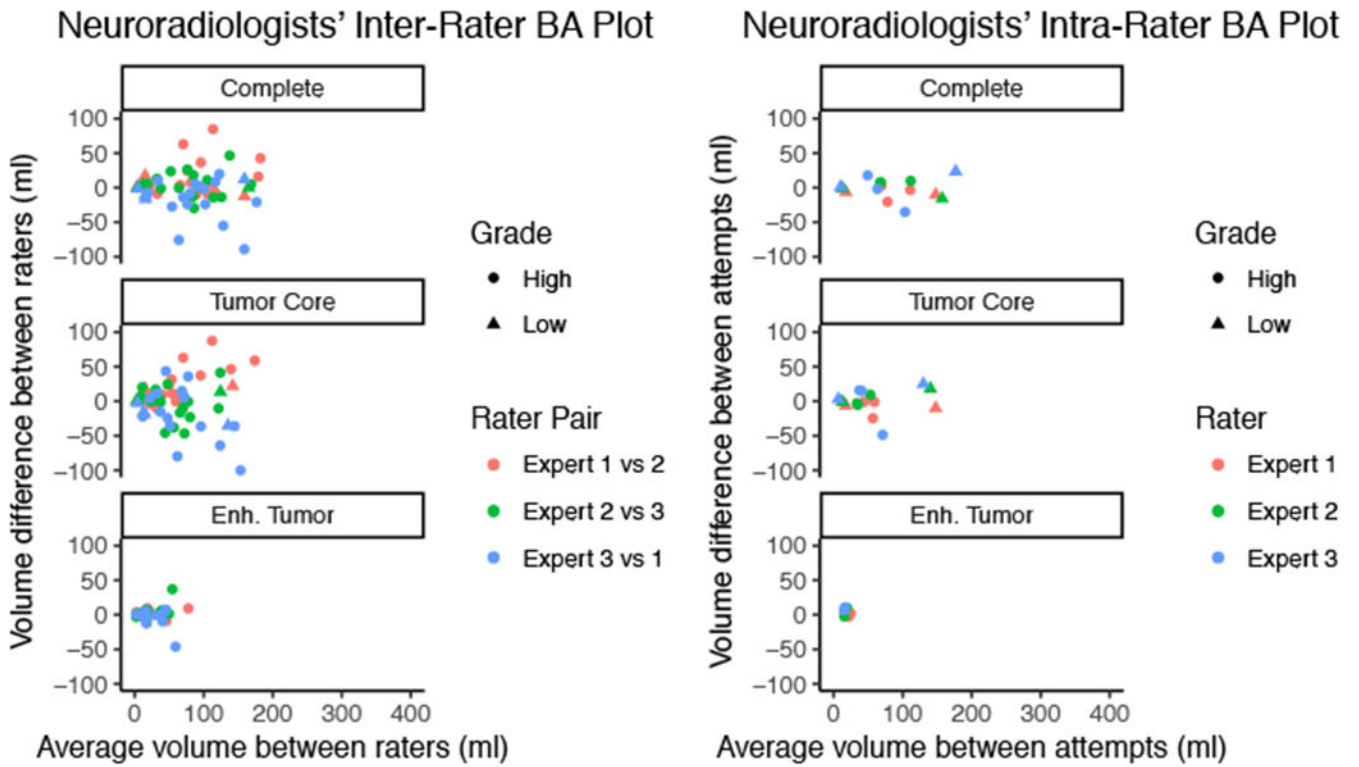Bland-Altman plots showing agreement in segmentation volume between attempts by different pairs of neuroradiologists (inter-rater plot, left) and different attempts by the same neuroradiologist (intra-rater plot, right). The average volume in both attempts is plotted on the horizontal axis, and the difference in volume between attempts is plotted on the vertical axis. All sub-plots have aspect ratio of 1.
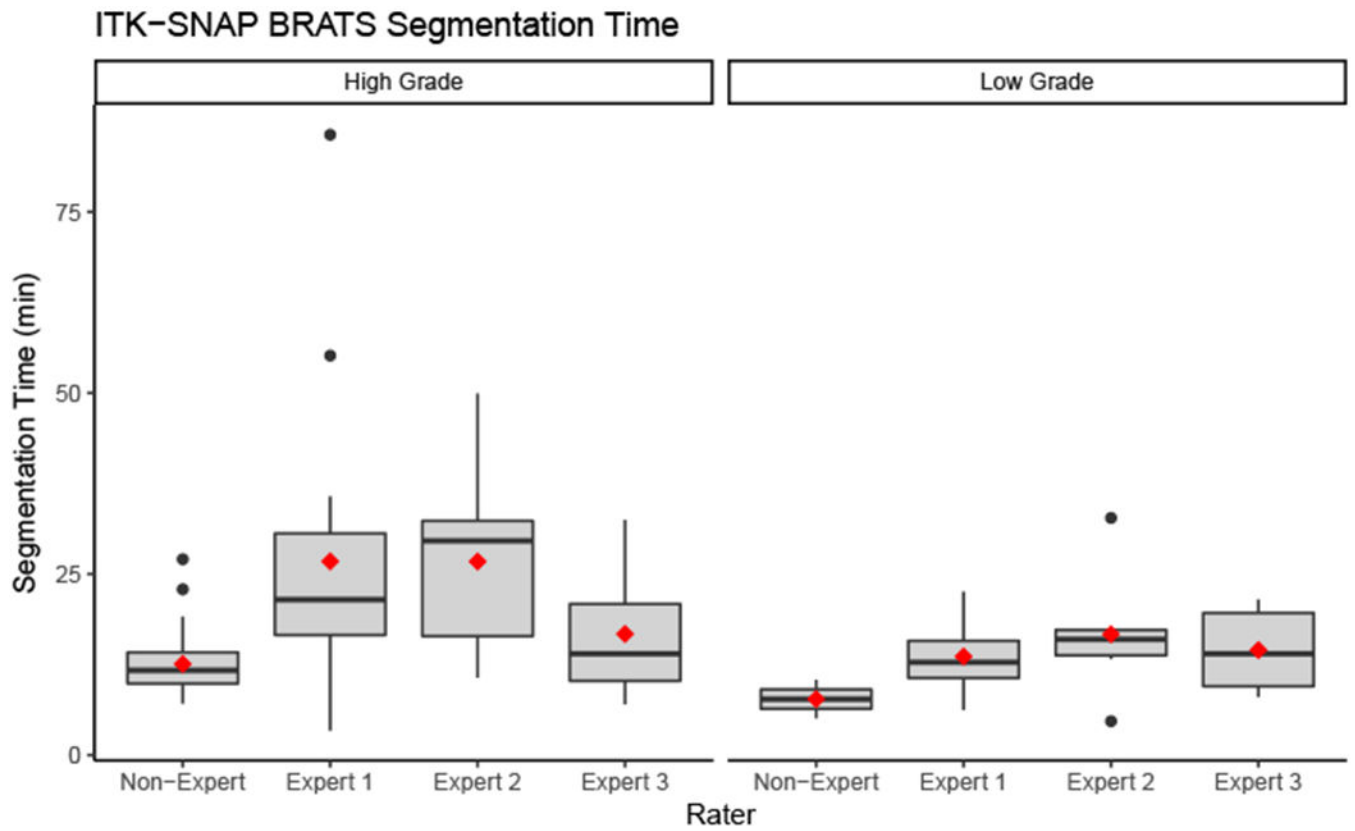
Figure 7:
Distribution of ITK-SNAP segmentation time for each rater, by tumor grade.
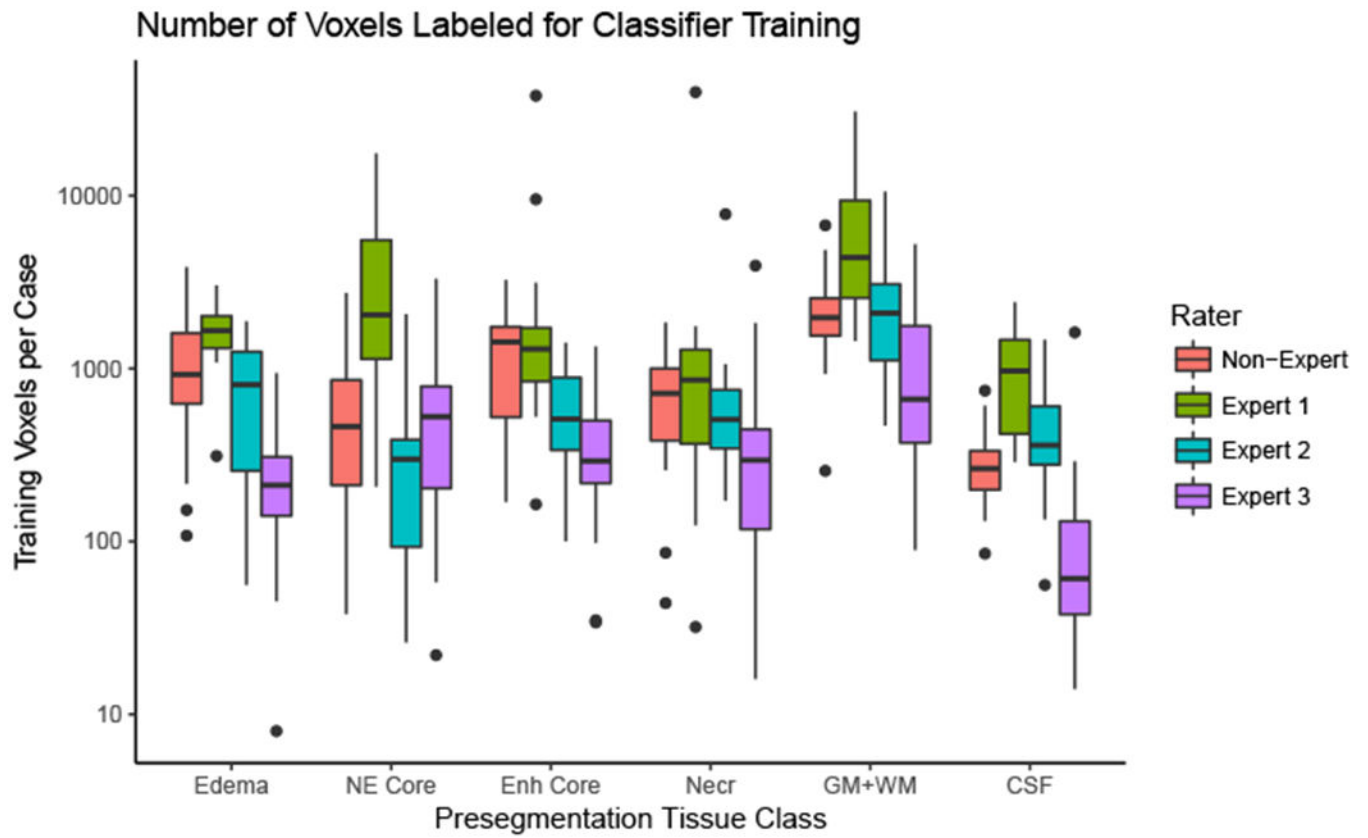
Figure 8:
Number of voxels labeled for each tissue class by each rater. The number of voxels is plotted on a logarithmic scale.
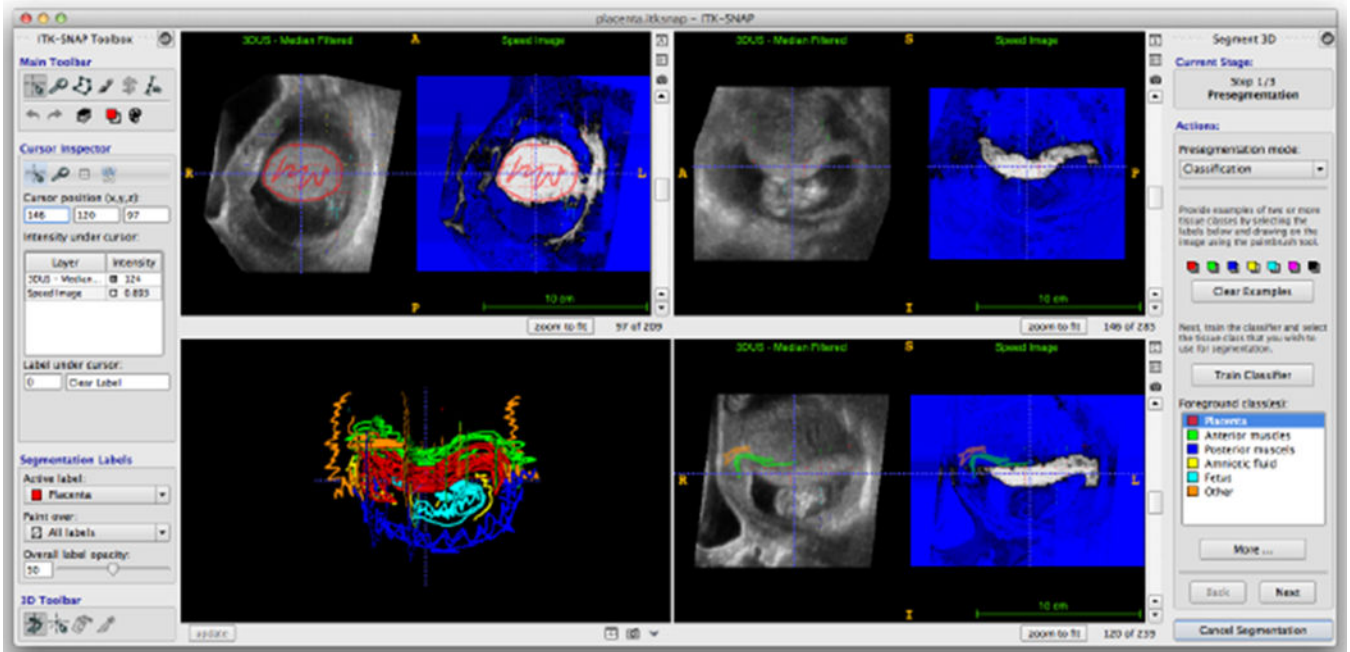
Figure 9:
ITK-SNAP during placenta presegmentation using the supervised classification mode. The orthogonal slice views show a median-filtered 3DUS image of the placenta and the fetus and the speed image. The 3D view shows the samples traced for training the classifier.
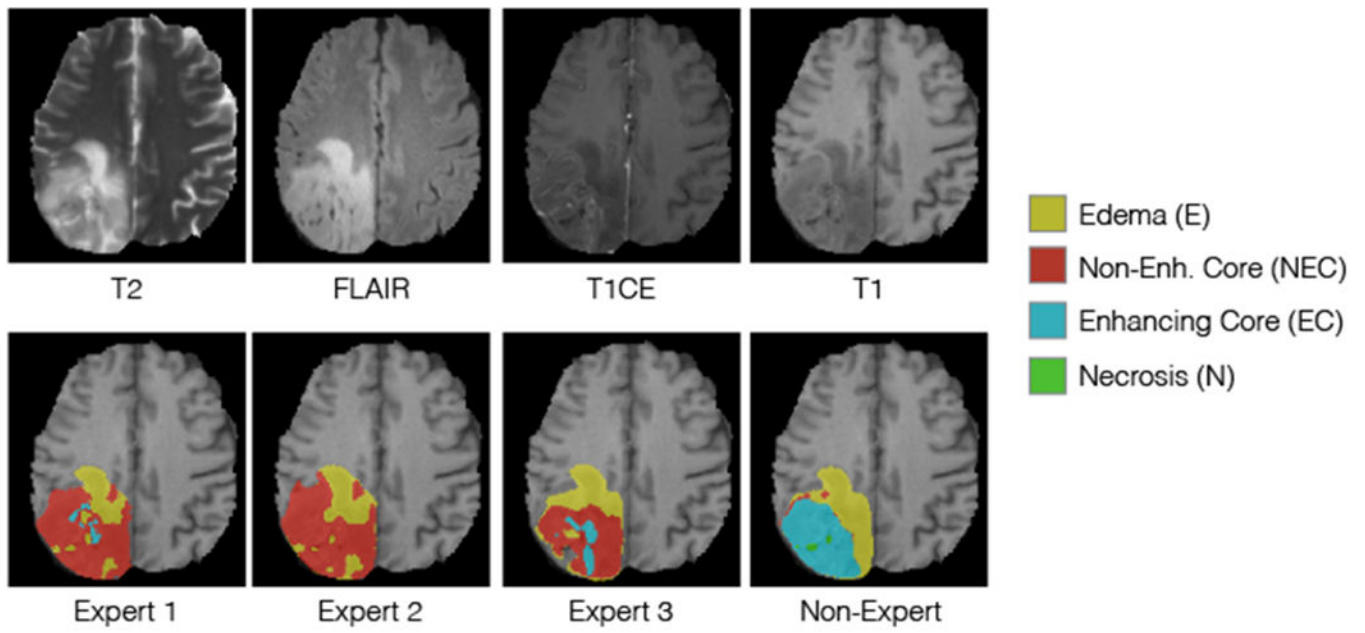
Figure 10:
Example high-grade case demonstrating areas of disagreement between raters. To a considerable degree, disagreement is driven by differences in the anatomical interpretation of complex cases (e.g., what constitutes enhancement, what constitutes edema vs. tumor), as opposed to fine-scale disagreements in tracing of agreed-upon boundaries.

Table 1:

Major new features added in recent versions of ITK-SNAP. For most features, the corresponding section in this paper or external reference is indicated.

| Functionality	Version	Section
Semi-automatic segmentation of multi-component images and multiple image volumes	3.0	2.2
Semi-automatic segmentation using unsupervised learning (Gaussian mixture models)	3.0	
New Qt-based GUI with project file support and numerous UI improvements	3.0	2.2
Full support for the DICOM image format	3.0	
Semi-automatic segmentation using supervised learning (Random Forests [6])	3.2	2.4.2
4D (3D+time) image support	3.2	2.2
Line and text annotations (to improve collaborative segmentation)	3.4	
Reduced memory footprint using run length encoding of segmentations in memory	3.4	[39]
Images with different size, resolution and orientation may be loaded and segmented together	3.6	2.2
Manual and automated linear image registration	3.6	2.2.2
Interpolation of manual segmentation between slices	3.6	[40]

Table 2:

Intra-rater reliability of glioma segmentation by the non-expert rater. The left portion of the table shows average intra-rater Dice coefficient and intra-class correlation coefficient (ICC) for all 25 cases segmented by the non-expert. The right portion shows the same metrics for Subset B in Figure 4, allowing comparison with inter-rater reliability of the BRATS reference manual segmentation reported in [25, Figure 5] .

Grade	Region	Full Leaderboard ($n = 25$)			Subset B ($n = 15$)		
		Dice Coefficient		ICC	Dice Coefficient		ICC
		Mean (SD)	Median		Mean (SD)	Median	
High	Complete	0.88 (0.20)	0.93	0.97	0.83 (0.28)	0.92	0.98
	Tumor Core	0.79 (0.22)	0.85	0.94	0.73 (0.28)	0.84	0.97
	Enh. Core	0.75 (0.26)	0.83	0.82	0.64 (0.34)	0.78	0.80
Low	Complete	0.81 (0.16)	0.83	1.00	0.81 (0.16)	0.83	1.00
	Tumor Core	0.36 (0.21)	0.34	0.55	0.36 (0.21)	0.34	0.55
All	Complete	0.87 (0.20)	0.93	0.98	0.83 (0.25)	0.92	0.99
	Tumor Core	0.72 (0.27)	0.84	0.85	0.63 (0.31)	0.81	0.84
	Enh. Core	0.75 (0.26)	0.83	0.83	0.64 (0.34)	0.78	0.81

Table 3:

Summary statistics for inter-rater and intra-rater reliability of the three neuroradiologist experts. Intra-rater ICC is not reported for high-grade and low-grade gliomas separately because of the small dataset (3 and 2 cases, respectively). † : Only three cases available for estimating ICC.

Grade	Region	Inter-Rater ($n = 20$)			Intra-Rater ($n = 5$)		
		Dice Coefficient		ICC	Dice Coefficient		ICC
		Mean (SD)	Median		Mean (SD)	Median	
High	Complete	0.84 (0.12)	0.86	0.78	0.85 (0.07)	0.87	
	Tumor Core	0.71 (0.19)	0.75	0.68	0.77 (0.09)	0.76	
	Enh. Core	0.70 (0.21)	0.76	0.86	0.78 (0.11)	0.82	
Low	Complete	0.76 (0.18)	0.79	0.99	0.90 (0.06)	0.92	
	Tumor Core	0.66 (0.20)	0.71	0.97	0.85 (0.12)	0.89	
All	Complete	0.82 (0.13)	0.84	0.87	0.87 (0.07)	0.88	0.97
	Tumor Core	0.70 (0.19)	0.75	0.77	0.80 (0.10)	0.81	0.93
	Enh. Core	0.70 (0.21)	0.76	0.86	0.78 (0.11)	0.82	0.10†

Table 4:

Comparison of mean and median inter-rater Dice coefficient between the three experts in the current study and manual segmentation by three raters in [25]. The comparison is in the same set of cases (subset B in Figure 4, $n = 15$).

Grade	Region	ITK-SNAP Inter-Rater Dice Coefficient		Menze et al. Inter-Rater Dice Coefficient	
		Mean (SD)	Median	Mean (SD)	Median
High	Complete	0.82 (0.12)	0.83	**0.85** (0.09)	0.87
	Tumor Core	0.67 (0.20)	0.71	**0.81** (0.19)	0.87
	Enh. Core	0.64 (0.23)	0.70	**0.68** (0.17)	0.71
Low	Complete	**0.76** (0.18)	0.79	0.71 (0.16)	0.68
	Tumor Core	**0.66** (0.20)	0.71	0.64 (0.27)	0.78
All	Complete	0.80 (0.14)	0.82	**0.81** (0.13)	0.85
	Tumor Core	0.67 (0.20)	0.71	**0.77** (0.22)	0.84
	Enh. Core	0.64 (0.23)	0.70	**0.68** (0.17)	0.71

Table 5:

Evaluation of the ITK-SNAP segmentation by the three experts and the non-expert against the BRATS challenge reference segmentation on Subset C (Figure 4), which was segmented by all four raters. Highest mean Dice coefficient for each region is shown in bold.

Grade	Region	Dice Coefficient between ITK-SNAP Segmentation and BRATS Reference Segmentation							
		Expert 1		Expert 2		Expert 3		Non-Expert	
		Mean (SD)	Median	Mean (SD)	Median	Mean (SD)	Median	Mean (SD)	Median
High	Complete	0.82 (0.13)	0.85	**0.89** (0.04)	0.89	0.81 (0.09)	0.84	0.86 (0.11)	0.88
	Tumor Core	0.63 (0.20)	0.72	0.74 (0.19)	0.80	0.62 (0.20)	0.67	**0.74** (0.20)	0.78
	Enh. Core	0.69 (0.24)	0.77	**0.71** (0.22)	0.81	0.68 (0.13)	0.69	0.66 (0.27)	0.76
Low	Complete	0.67 (0.28)	0.69	0.82 (0.12)	0.83	**0.83** (0.11)	0.82	0.71 (0.33)	0.82
	Tumor Core	0.43 (0.36)	0.41	0.52 (0.37)	0.56	**0.60** (0.41)	0.75	0.31 (0.42)	0.16
All	Complete	0.79 (0.17)	0.84	**0.88** (0.07)	0.89	0.81 (0.09)	0.84	0.83 (0.17)	0.88
	Tumor Core	0.59 (0.25)	0.64	**0.69** (0.24)	0.76	0.62 (0.24)	0.68	0.65 (0.30)	0.76
	Enh. Core	0.69 (0.24)	0.77	**0.71** (0.22)	0.81	0.68 (0.13)	0.69	0.66 (0.27)	0.76

Table 6:

Agreement of ITK-SNAP segmentation by the three experts and the non-expert with the BRATS challenge reference segmentation, ranked against the 20 methods evaluated in the BRATS 2012/2013 challenges [25, Figure 7]. The "rank" columns give the rank of each mean Dice coefficient measurement relative to the Dice coefficients for the 20 methods listed in [25, Figure 7]. Ranks are computed separately for each region/grade combination. Ties, occurring when mean Dice coefficient is within rounding error in [25, Figure 7], are indicated as follows: †: two-way tie, ‡: three-way tie.

| Grade | Region | ITK-SNAP Segmentation vis-a-vis 20 Methods in the BRATS 2012/2013 Challenges | | | | | | | |
| | | Expert 1 | | Expert 2 | | Expert 3 | | Non-Expert | |
		Mean (SD)	Rank	Mean (SD)	Rank	Mean (SD)	Rank	Mean (SD)	Rank
High	Complete	0.83 (0.12)	2	**0.89** (0.03)	1	0.80 (0.09)	3†	0.84 (0.13)	1†
	Tumor Core	0.65 (0.20)	4	**0.74** (0.20)	2	0.62 (0.21)	5	0.70 (0.22)	2
	Enh. Core	0.63 (0.27)	1	**0.65** (0.24)	1	0.64 (0.13)	1	0.59 (0.30)	2†
Low	Complete	0.67 (0.28)	5†	0.82 (0.12)	1	**0.83** (0.11)	1	0.71 (0.33)	3‡
	Tumor Core	0.43 (0.36)	5	0.52 (0.37)	5	**0.60** (0.41)	1†	0.31 (0.42)	15
All	Complete	0.79 (0.18)	2	**0.87** (0.07)	1	0.80 (0.09)	2	0.81 (0.20)	2
	Tumor Core	0.59 (0.26)	4	**0.68** (0.26)	2	0.62 (0.26)	4	0.60 (0.33)	4
	Enh. Core	0.63 (0.27)	1	**0.65** (0.24)	1	0.64 (0.13)	1	0.59 (0.30)	2†