Int J CARS
DOI 10.1007/s11548-016-1483-3

ORIGINAL ARTICLE

Automated brain tumour detection and segmentation using superpixel-based extremely randomized trees in FLAIR MRI

Mohammadreza Soltaninejad[1] · Guang Yang[2,3] · Tryphon Lambrou[1] · Nigel Allinson[1] · Timothy L. Jones[4] · Thomas R. Barrick[2] · Franklyn A. Howe[2] · Xujiong Ye[1]

Received: 6 May 2016 / Accepted: 31 August 2016
© The Author(s) 2016. This article is published with open access at Springerlink.com

Abstract
Purpose We propose a fully automated method for detection and segmentation of the abnormal tissue associated with brain tumour (tumour core and oedema) from Fluid-Attenuated Inversion Recovery (FLAIR) Magnetic Resonance Imaging (MRI).
Methods The method is based on superpixel technique and classification of each superpixel. A number of novel image features including intensity-based, Gabor textons, fractal analysis and curvatures are calculated from each superpixel within the entire brain area in FLAIR MRI to ensure a robust classification. Extremely randomized trees (ERT) classifier is compared with support vector machine (SVM) to classify each superpixel into tumour and non-tumour.
Results The proposed method is evaluated on two datasets: (1) Our own clinical dataset: 19 MRI FLAIR images of patients with gliomas of grade II to IV, and (2) BRATS 2012 dataset: 30 FLAIR images with 10 low-grade and 20 high-grade gliomas. The experimental results demonstrate the high detection and segmentation performance of the proposed method using ERT classifier. For our own cohort, the average detection sensitivity, balanced error rate and the Dice overlap measure for the segmented tumour against the ground truth are 89.48 %, 6 % and 0.91, respectively, while, for the BRATS dataset, the corresponding evaluation results are 88.09 %, 6 % and 0.88, respectively.
Conclusions This provides a close match to expert delineation across all grades of glioma, leading to a faster and more reproducible method of brain tumour detection and delineation to aid patient management.

Keywords Brain tumour segmentation · Extremely randomized trees · Feature selection · Magnetic resonance imaging · Superpixels · Textons

✉ Mohammadreza Soltaninejad
msoltaninejad@lincoln.ac.uk

Guang Yang
gyang@sgul.ac.uk; g.yang@imperial.ac.uk

Tryphon Lambrou
tlambrou@lincoln.ac.uk

Nigel Allinson
nallinson@lincoln.ac.uk

Timothy L. Jones
timothy.jones@stgeorges.nhs.uk

Thomas R. Barrick
tbarrick@sgul.ac.uk

Franklyn A. Howe
howefa@sgul.ac.uk

Xujiong Ye
xye@lincoln.ac.uk

1 Laboratory of Vision Engineering, School of Computer Science, University of Lincoln, Lincoln LN6 7TS, UK

2 Neurosciences Research Centre, Molecular and Clinical Sciences Institute, St. George's, University of London, London SW17 0RE, UK

3 National Heart and Lung Institute, Imperial College London, London SW7 2AZ, UK

4 Atkinson Morley Department of Neurosurgery, St George's Hospital London, London SW17 0RE, UK

Introduction

Despite improvements in the diagnosis and oncological treatment of primary brain tumours, they remain associated with significant morbidity and a poor overall prognosis. The majority of primary brain tumours originate from glial cells (termed glioma) and are classified by their histopathological appearances using the World Health Organization (WHO)

Published online: 20 September 2016

system into low-grade glioma (LGG) (grades I and II) and high-grade glioma (grade III anaplastic glioma and grade IV glioblastoma). The typical natural history of low-grade glioma is a latent period of growth and infiltration of white matter with subtle neuro-cognitive deficit and seizures in some cases followed by regional change or transformation to a more malignant variant. High-grade glioma may present as a de novo (primary) glioblastoma or as a transformation of a lower-grade tumour (e.g. secondary glioblastoma).

Gliomas typically originate within white matter and exhibit irregular growth patterns along white matter fibres, infiltrating surrounding brain. As a result, they exhibit irregular boundaries that may be visually indistinct on conventional magnetic resonance images. Delineation of the tumour boundary and assessment of tumour size are needed for patient management in terms of treatment planning and monitoring treatment response, and current guidelines incorporate the use of both contrast-enhanced T1-weighted (CE T1w) images and T2-weighted (T2w) / FLAIR images [1,2]. Many low- grade gliomas do not show contrast enhancement; hence, T2w/FLAIR images are used to define the tumour extent and volume. A longitudinal study has shown that LGG volume and growth rate can be used to assess whether patients are at risk with tumours likely to undergo an early malignant transformation [3]. In clinical studies, current Response Assessment in Neurooncology (RANO) criterion simply uses a bidirectional measurement to determine tumour volume for assessing treatment response [4]. Although a full 3D volume measurement may provide a more accurate volume assessment, there is a need for accurate and fully automated methods since manual segmentation (region of interest drawing) around tumour margins on a slice-by-slice basis is time-consuming and can take 12 min or more per tumour, with semiautomatic methods taking 3–5 min [5,6]. T2w/FLAIR images can also be useful to help define the target volumes for radiotherapy planning of high-grade gliomas [2,5]; hence, an automated segmentation that is not subject to operator subjectivity may be beneficial [5]. In this study, we have concentrated on developing and validating an automated method for a single MRI modality, FLAIR, that could be readily translated for clinical use. Future automated methods are likely to incorporate information from multimodal clinical MRI as in the Multimodal Brain Tumor Image Segmentation Benchmark (BRATS) database studies [7–9] and also include perfusion and diffusion imaging to detect tumour tissue subtypes (e.g. necrosis, active tumour, infiltrative tumour, oedema) [10].

However, automated detection and segmentation of brain tumour is a very challenging task due to its high variation in size, shape and appearance (e.g. image uniformity and texture) [11]. Also, typical clinical image acquisition protocols usually lead to higher intraslice resolution than interslice resolution to achieve the balance of good apparent image resolution with adequate signal to noise and restricted scanning time that causes asymmetry in partial-volume effects. High-grade gliomas usually have irregular boundaries which, in some cases, are unclear or discontinuous [12]. Current work on brain tumour segmentation can be categorized into atlas-based [13–15], unsupervised [16–19], hybrid [20–22] and supervised- based approaches [23–26].

In Ref. [20], a hybrid method was proposed for brain tissue detection in MRI images which included seeded region growing segmentation and neural network classification. However, the method was semiautomatic and different parts of tumour need to be pre-selected to initiate the segmentation process. Another method is proposed for detection of multiple sclerosis (MS) lesions in brain MR images which consisted of rule-based, level-set and support vector machines [21]. Rajendran and Dhanasekaran [22] proposed a hybrid method for segmenting the tumour by combining region-based fuzzy clustering and deformable model. However, the method was only applied on a few FLAIR images with fixed parameters.

Supervised learning-based algorithms use training data labelled by experts for segmentation of tumours. Geremia et al. [23] used discriminative random decision forests to classify the voxels of 3D MRI image for segmentation of MS. Wu et al. [24] used superpixel features in a conditional random fields (CRF) framework to detect brain tumours. However, the method was not satisfactory for low-grade tumours segmentation. A method was proposed in [25] which used extremely randomized forest classification considering both appearance and context-based features. Another method was proposed in [26] which used ERT for classification of voxels based on their statistical and textural features, which were extracted using different MRI protocols. In order to reduce the computation time, it was suggested that features were only extracted from a random set of voxels, but this resulted in losing some part of data. In addition, a fixed size neighbourhood for each voxel was used to calculate features.

A number of advanced algorithms [23,27–31] were recently presented in [7] using the BRATS [8,9] organized in conjunction with the international conference on Medical Image Computing and Computer-Assisted Interventions (MICCAI) 2012 and 2013 conference. The methods were based on segmentation of different tumour tissues, i.e. tumour core, oedema, necrosis, using multiprotocol containing FLAIR, T1-weighted (T1w), T1w with contrast and T2-weighted protocols [32].

Despite much effort being devoted to the segmentation problem, brain tumour segmentation remains an ongoing research topic. Very few completely automatic segmentation algorithms have been adopted in the clinic. Recently, only one automated tool has been clinically evaluated [33].

In this study, we investigate a fully automated superpixel-based method for detection and segmentation of the abnormal

tissue associated with brain tumours as defined by the T_2 hyperintensity from Fluid-Attenuated Inversion Recovery (FLAIR) MRI. FLAIR images are routinely acquired as part of standard diagnostic clinical MRI of brain tumours. Delineation of the FLAIR hyperintensity is relevant for assessing low-grade glioma growth [34], defining an abnormality region from which imaging features for tumour classification can be extracted [35], aiding with radiation dose planning [36] and assessing treatment response [37]. Different from the methods in [25] and [26], in which image features were calculated based on each individual voxel and a fixed size neighbour-hood was considered for the feature extraction, in this paper, superpixel partition is firstly calculated which provides accurate boundaries between different tissues, and then image features are extracted from each superpixel. This will not only improve the accuracy of feature calculation, but also increase the speed of computation. We demonstrate the automated method that provides a close match to expert delineation across all grades of glioma and so could provide a faster and more reproducible method of brain tumour delineation to aid patient management. To assess the robustness of the proposed method, the method is also evaluated on the FLAIR protocol of BRATS 2012 annotated training dataset [8,9].

The rest of this paper is organized as follows. The "Method" section describes the proposed method, including superpixel partition, feature extraction, classification and final segmentation. The "Experimental results" section presents the data description and experimental results for the two datasets, followed by "Discussion" and "Conclusion" sections.

Method

Our method consists of four main steps, which are depicted in Fig. 1. After preprocessing, in the superpixel segmentation step, FLAIR image is partitioned into irregular patches with approximately similar size and intensity values. For each superpixel patch, a number of features including statistical, texton and shape features are calculated. This is then followed by feature selection to find the most significant features, based on which each superpixel is classified into tumour and non-tumour using an ERT classifier.

Preprocessing

First, the skull is removed from all the MRI images using FSL [38]. Then, histogram matching algorithm [39] is applied to ensure that all the data have similar dynamic ranges. ITK software [40] is used for this task, and one of the cases is selected as the reference, and then, other MRI FLAIR scan intensities are transformed to match the histogram of the reference image.

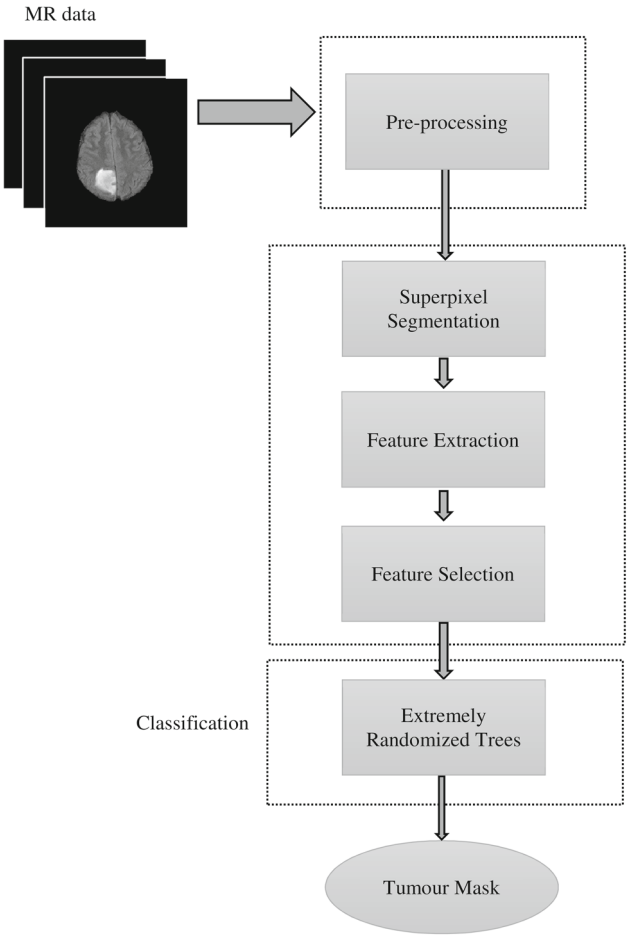


Fig. 1 Flowchart of the proposed method

Superpixel segmentation

The simple linear iterative clustering (SLIC) [41] method is used to partition the image into patches with approximately similar size. SLIC method has a few parameters which are flexible to be tuned by controlling the trade-off between them and boundary adherence. Furthermore, it is computational and memory efficient. Each slice of FLAIR image is gridded into equally sized squares with a user-defined size. The size of grid side for these initial superpixels is considered as S. The geometrical centre of each segment is considered as the superpixel centre. These centre coordinates are then updated in each iteration. The pixels are grouped based on their spatial and intensity distance metrics. The spatial distance d_s between the ith pixel and the jth pixel is calculated as:

$$d_s = \sqrt{(x_j - x_i)^2 + (y_j - y_i)^2} \quad (1)$$

where x and y are the pixel location coordinates. The intensity distance d_c between the two pixels is defined as:

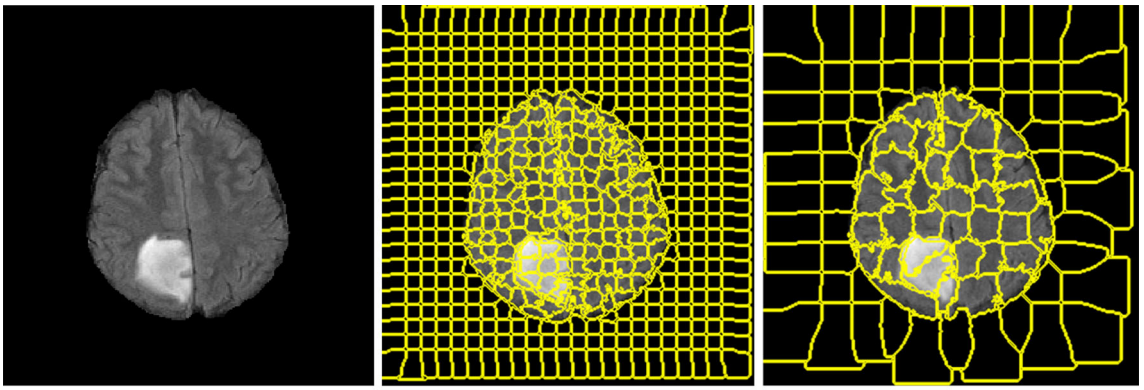


Fig. 2 Example of superpixel segmentation with different window sizes: **a** original MRI FLAIR image with a grade II tumour, **b** superpixel segmentation with $S = 10$ (initial grids 10×10) and $m = 0.2$, **c** superpixel segmentation with $S = 20$ (initial grids 20×20) and $m = 0.2$

$$d_c = \sqrt{(I_j - I_i)^2} \tag{2}$$

where I_i and I_j are the normalized intensity values of the ith and the jth pixel, respectively.

The overall distance measure which is a combination of spatial and intensity distances is then calculated with:

$$D = \sqrt{d_c^2 + \left(\frac{d_s}{S}\right)^2 m^2} \tag{3}$$

where m is a compactness coefficient which determines the flexibility of superpixel boundaries. A higher value of m results in more compact segments and a lower value creates more flexible boundaries. It is noted that, to obtain an optimum compactness coefficient m, the MRI image intensities used in Eq. (2) are normalized to the values of [0, 1]. This is to ensure that both the intensity and space distances are within the same range.

Figure 2 shows MR images acquired with protocol FLAIR containing a grade II tumour which is partitioned to superpixels with two different side sizes, S. The compactness factor m is set to be 0.2 for both sizes. In Fig. 2b, c, the superpixels are extracted with $S = 10$ and $S = 20$, respectively.

Feature extraction and selection

In order to train a robust classifier for the detection and segmentation of brain tumour, different types of features are considered, including intensity statistics, textons and curvature features.

Intensity statistical features

First-order intensity statistics [42] are referred as pixel intensity-based features. They express the distribution of grey levels within the selected region of interest (ROIs) which are the superpixels in our work. For each superpixel, 16 features are calculated which are average, standard deviation, variance, mean of the absolute deviation, median absolute deviation, coefficient of variance, skewness, kurtosis, maximum, minimum, median and mode of the intensity values, central moments, range, interquartile range and entropy.

Texton feature

Brain tissues have complex structures, so the intensity features are not sufficient for accurate segmentation of tumour. Texture features are used to improve the accuracy of segmentation. In this study, the texture features are calculated based on texton analysis. Textons are small elements of the image generated by convolution of the image with a specific filter bank, in which Gabor filter [43] defined as Eq. (4) is used:

$$G(x, y; \theta, \sigma, \lambda, \psi, \gamma) = \exp\left(-\frac{x'^2 + \gamma^2 y'^2}{2\sigma^2}\right) \exp\left(i\left(2\pi \frac{x'}{\lambda} + \psi\right)\right) \tag{4}$$

where σ is the filter size, λ is the wavelength of sinusoid, ψ is the phase shift and γ is the spatial aspect ratio. In Eq. (4), the terms x' and y' are calculated from the spatial orientation of the filter, θ, defined as:

$$\begin{aligned} x' &= x\cos\theta + y\sin\theta \\ y' &= x\cos\theta + y\sin\theta \end{aligned} \tag{5}$$

The values which are set for these parameters will be discussed in "Texton feature parameters" section.

The FLAIR image is convolved with all the N_{FB} filters (i.e. N_{FB} is the number of filters in the filter bank) and a response vector with length of N_{FB} is generated for each pixel. These filter response vectors (the number of vectors is the same as the number of the pixels in the image) are then clustered into

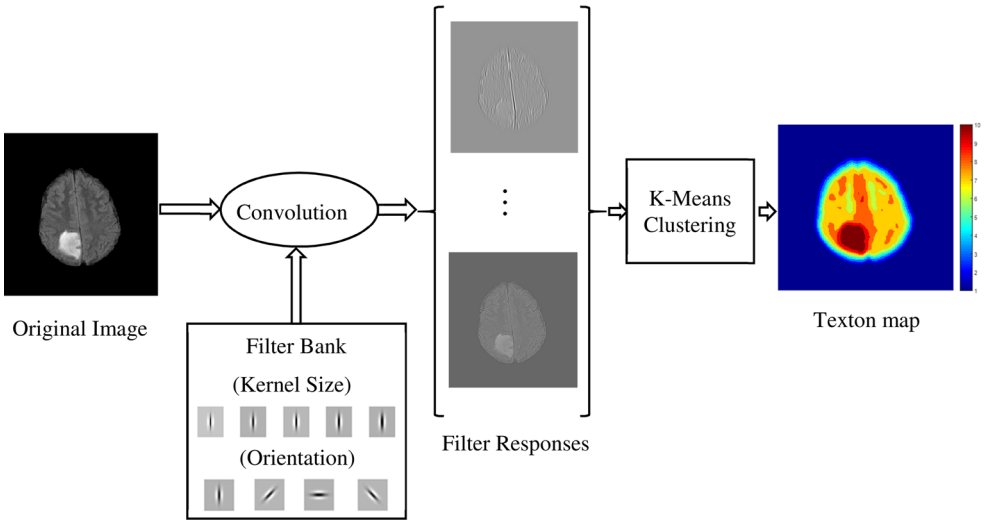


Fig. 3 Procedure of texton feature extraction using Gabor filters applied to a grade II glioma

k clusters using N_{FB}-dimensional k-means clustering. The filter response vectors corresponding to each cluster are considered as the texton of a particular texture class. By assigning the cluster number to each pixel, a texton map of the image is obtained. The procedure of texton map extraction is depicted in Fig. 3. The texton features for each superpixel are then calculated using the histogram of texton map within each superpixel.

Fractal features

A segmentation-based fractal texture analysis method (SFTA) [44] is used to calculate fractal features. The image is firstly decomposed into a set of binary images based on multilevel thresholds computed using Otsu algorithm [45]. The desired number of thresholds n_{t} is defined by the user (in this paper, $n_{\mathrm{t}} = 3$). Then for each binary channel, all the image boundaries are extracted using edge detection [46]. The fractal features are calculated from these binary edge channels which include area, intensity and fractal dimension. Area feature is the number of edge pixels in a superpixel. Intensity feature is the mean intensity of image pixels corresponding to the edge pixels in a superpixel. Fractal dimension represents the complexity of the structure of the image and is calculated from image boundary as:

$$D_0 = \lim_{\varepsilon \to 0} \frac{\log N(\varepsilon)}{\log \varepsilon^{-1}} \tag{6}$$

where $N(\varepsilon)$ denotes the counting of hypercubes of dimension E and length ε. By using box counting algorithm [47], an approximation of fractal distance is obtained from the binary images.

Figure 4 presents a flowchart of fractal analysis. Figure 5 shows fractal features including: area, mean intensity and fractal dimension. Figure 6 illustrates an example of fractal dimension and mean intensity features calculated from healthy and tumour superpixels from one patient data containing a grade IV glioma. It demonstrates a good separation in feature space (mean intensity fractal dimension) for FLAIR images.

Curvature feature

Image curvature is a shape-based feature which is computed by the gradients along x and y directions of an image, namely f_x and f_y. The image normal at pixel $(x,\ y)$ is then calculated as [48]:

$$\hat{N}(x, y) = \frac{1}{\left(f_x^2 + f_y^2\right)^{1/2}} \begin{pmatrix} f_x \\ f_y \end{pmatrix}. \tag{7}$$

The two-dimensional curvature of the image is the divergence of this normal and is calculated as:

$$\mathrm{Curv} = \frac{f_{xx} f_y^2 + f_{yy} f_x^2 - 2 f_{xx} f_x f_y}{\left(f_x^2 + f_y^2\right)^{3/2}} \tag{8}$$

where f_{xx} and f_{yy} are the second derivatives of the image intensity $I(x, y)$. The curvature feature for each superpixel is the average of the curvature values for all the pixels in the superpixel.

In summary, there are in total 28 features calculated for each superpixel, among which there are 5 texton histogram features from 5 clusters and 6 fractal features obtained from 3 thresholded binary images (each binary image provides 3

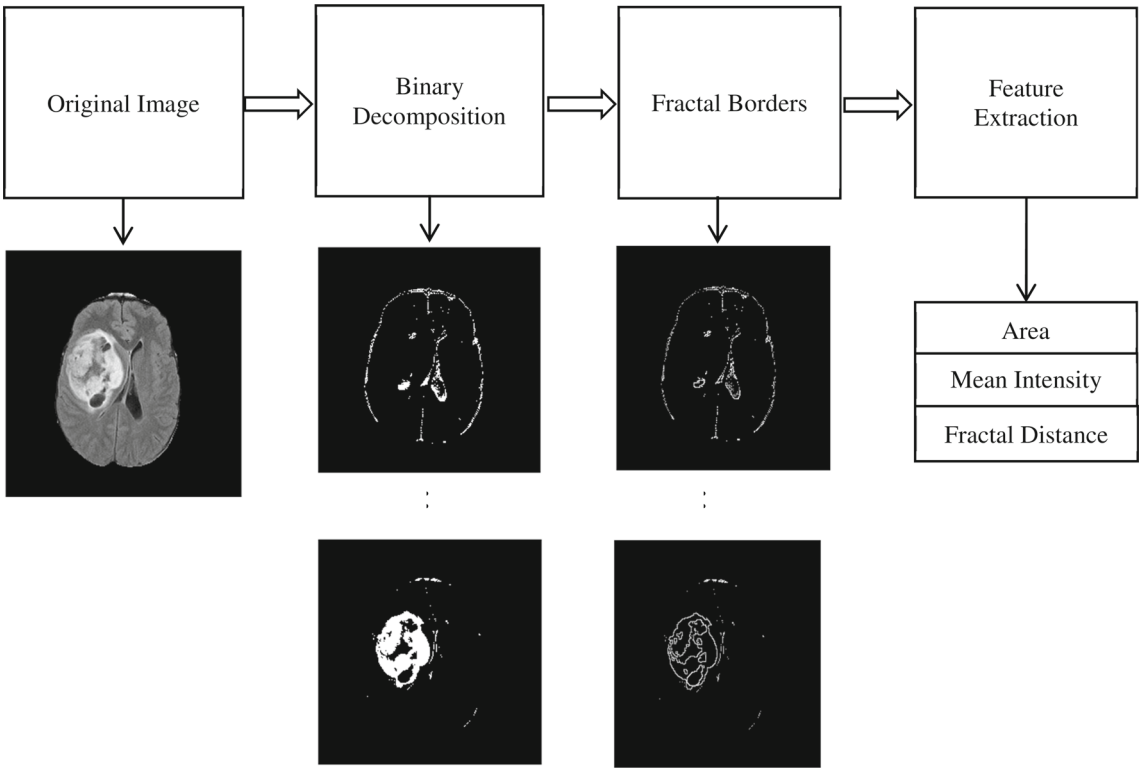


Fig. 4 The flowchart of extracting fractal features from a grade III glioma

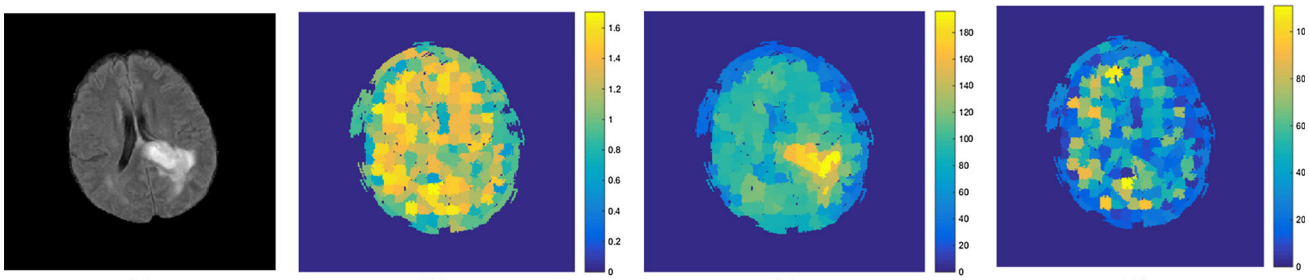


Fig. 5 An example of fractal analysis applied to a grade III glioma to generate superpixel-based fractal feature maps: **a** FLAIR image, **b** area, **c** mean intensity and **d** fractal dimension

fractal features). It is noted that all the features, except the 5 texton histogram features, are normalized to the range of [0,30], and this is to ensure that all the features have similar dynamic ranges and also are close to the textons histogram values. Table 1 shows a list of the features. The details of parameter setting in feature calculation will be discussed in "Selection of parameters" section.

Feature selection

Feature selection step is used not only to increase the computation speed, but also to remove redundant features which may cause more classification error. In this paper, we employ the Minimum Redundancy Maximum Relevance (mRMR) feature selection technique proposed by [49]. mRMR is an efficient technique for subset selection of features, which selects more relevant features by removing the irrelevant ones. Mutual information is used for identifying the similarity between features. For features, f_i, in feature set S, the maximum relevance is obtained between features and class c by maximizing the following:

$$\max D(S,c), \quad D = \frac{1}{|S|}\sum\nolimits_{f_i \in S} I_{\mathrm{M}}(f_i;c) \tag{9}$$

where I_{M} is mutual information between feature f_i and the class c. Minimum redundancy is calculated from:

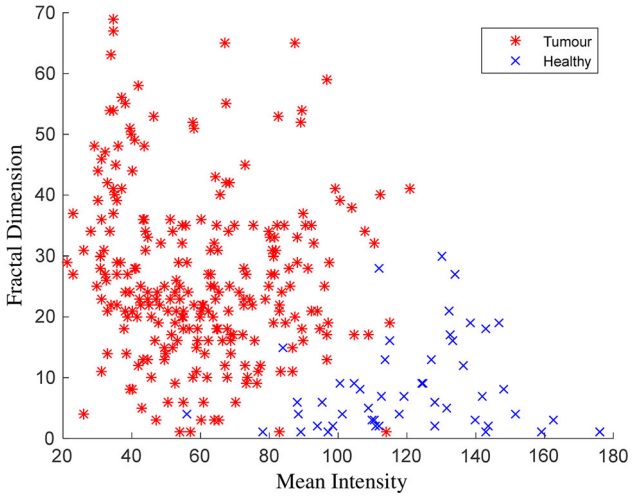


Fig. 6 Fractal dimension versus mean intensity for healthy and tumour superpixels calculated from one FLAIR MRI data with grade IV glioma

Table 1 Total number of features calculated from MRI FLAIR image

Feature name	Number of features
Statistical 1st order	16
Texton histogram	5
Fractal	6
Curvature	1
Total	28

$$\min R(s), \quad R = \frac{1}{|S|^2} \sum\nolimits_{f_i, f_i \in S} I_{\mathrm{M}}\left(f_i, f_j\right). \tag{10}$$

The feature selection is performed on the entire feature vector, and it is based on leave-one-out cross-validation using voting scheme. For each case, using cross-validation, the best N_{FEA} features were selected. It is noted that, if one feature is selected by one case, the feature will get one vote. For all the features voted by all the cases, the top N_{FEA} features with highest scores will be chosen as the final features. The selected features will be used in the classification stage to classify each superpixel into tumour or non-tumour.

Extremely randomized tree-based classification of superpixels

In order to tackle the problem of extremely imbalanced data in our dataset, ERT classifier [50] is used to categorize each superpixel into tumour or normal brain tissue and to improve the accuracy of the minority class (e.g. tumour). Like random forests (RF) [51], ERT is an ensemble technique which uses multiple decision trees. For both methods, each node of the tree includes a set of training examples and the predictor. Splitting starts from the root node and will continue at every node. The procedure is performed based on the feature representation and allocating the partitions to the right and left nodes. Tree grows until a specified tree depth. During the bagging process and at each attribute split, a random subset of features is used. In RF, by generating large number of trees, the most popular class is voted [52].

ERT is an extension of RF in which a further randomization stage is added for selecting the cut-points alongside with randomized selection of attributes like in RF. In this technique, the splits of attributes and cut-points are selected randomly. Each tree is determined by $t \epsilon \{1 \ldots T\}$ in which T is the number of randomized trees. For a given data point x and dataset D_{train}, a feature vector is represented by $f(x, D_{\mathrm{train}})$. To classify the class c of the data, for an n-dimensional feature representation, each tree learns a weak predictor of $p_t(c|f(x, D_{\mathrm{train}}))$.

In the testing process, for an unseen data point, x', the probability of belonging to a class c is calculated by the average of probabilities on all the trees:

$$p\left(c|f\left(x', D\right)\right) = \frac{1}{T} \sum_{t=1}^{T} p_t\left(c|f\left(x', D\right)\right) \tag{11}$$

The structures of randomized trees are independent of training sample outputs. The parameters should be selected and tuned for the specific case. In our method, there are 20 extra trees in the ensemble and five attributes, which are equal to the number of selected features, are selected to perform the random splits. Tree depth is chosen to be 15 and the minimum number of samples for splitting a node is 2 as this is a classification task. Setting these parameters will be discussed in "Extremely randomized trees parameters" section.

After the ERT, each superpixel is then classified into tumour or non-tumour candidates. For all the tumour superpixels, a 3D connected component analysis [53] is then used to obtain 3D connected superpixel regions. Each small superpixel region in which the total number of voxels in the region is less than a pre-defined threshold (i.e. 100) is regarded as a false-positive (FP) region and removed from the tumour candidates. The remaining tumour superpixel regions are the segmented tumour.

Experimental results

Two experiments were carried out in this section. In the first experiment, our own clinical dataset is used for training and validation of the algorithm. In the second experiment, the method is further validated on the publicly available MICCAI BRATS 2012 dataset [7–9] to assess the robustness of the method. The following subsections, including data description, parameters selection and comparative experi-

mental results are focused on our own data cohort, while the next subsection presents the evaluation results on MICCAI BRATS 2012 clinical training dataset.

Data description

We acquired patient data using a GE Signa Horizon LX 1.5 T MRI system (GE Healthcare, Milwaukee, WI, USA) equipped with a maximum field gradient strength of 22 mT/m and using a quadrature head coil. The MRI sequence used in this study is FLAIR which is acquired in the axial plane with a field of view (FOV) $240 \times 240\,\text{mm}^2$, matrix size 256×256 and 5 mm slice thickness with no slice gap. In particular, the following sequence is used: FLAIR (TE = 133 ms, TR = 9000 ms, inversion time 2200 ms).

A cohort consisting of 19 patients entered retrospectively into our study, each with a brain tumour, who has been imaged with the FLAIR MRI sequences. The dataset consists of 6 grade II tumours, 3 grade III tumours and 10 grade IV tumours. Each patient has a histological gold standard of tumour grading. Figure 7 shows some examples of the manual segmentations for different tumour grades in FLAIR images. Patient ages at the time of scanning ranged from 22 to 73 (mean 54) and consisted of 7 females and 12 males.

Selection of parameters

Statistical features are calculated directly from the intensity values of the pixels within the superpixels, and they are non-parametric. Parameter setting is required to calculate texton and fractal features. For texton features, parameters of Gabor filter bank and the number of clusters in k-means clustering need to be determined. For the ERT classifier, an optimum number of trees should be selected for an accurate and fast classification. In this study, the parameters are determined through the training stage, in which a total number of 6 patients' data are randomly selected including 2 grade II, 1 grade III and 3 grade IV. In the following section, the process of these parameters selection is explained in detail.

Superpixel parameters

To investigate the effect of compactness factor, m, defined in Eq. (3), on the superpixels boundaries, we apply different values from 0 to 1 and inspect the results visually. The intensity values of the FLAIR voxels within the brain are normalized to the range of [0, 1]. A compactness factor $m = 1$ results in more rigid boundaries, while $m = 0$ produces very flexible boundaries but increases the variation and irregularity of the superpixels shapes. An example of this parameter is shown in Fig. 8. By visually inspecting the superpixel boundaries and area, the value of $m = 0.2$ is chosen, which presents coherent boundaries.

To select an appropriate superpixel size, different initial window side sizes are considered in the partitioning stage. The compactness factor is fixed to $m = 0.2$ for all the experiments. Then, the superpixels which have more than 0.9 overlap with the manual segmentation mask are selected

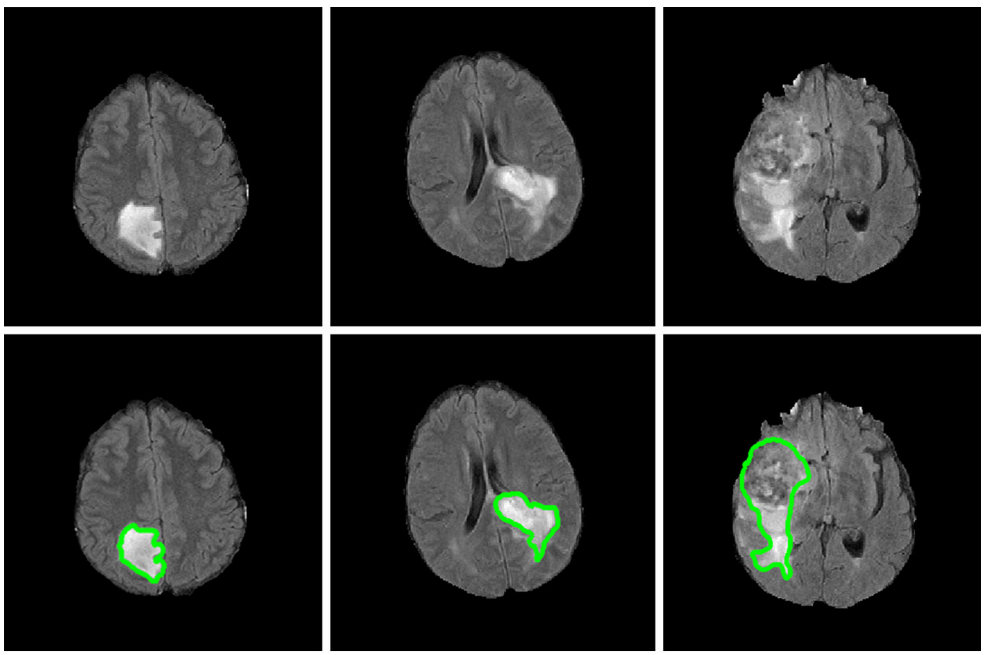


Fig. 7 FLAIR images with different tumour grades in *upper row* and their ground-truth manual segmentation of the FLAIR hyperintensity in the *lower row*. Tumour grades are: **a** grade II, **b** grade III and **c** grade IV

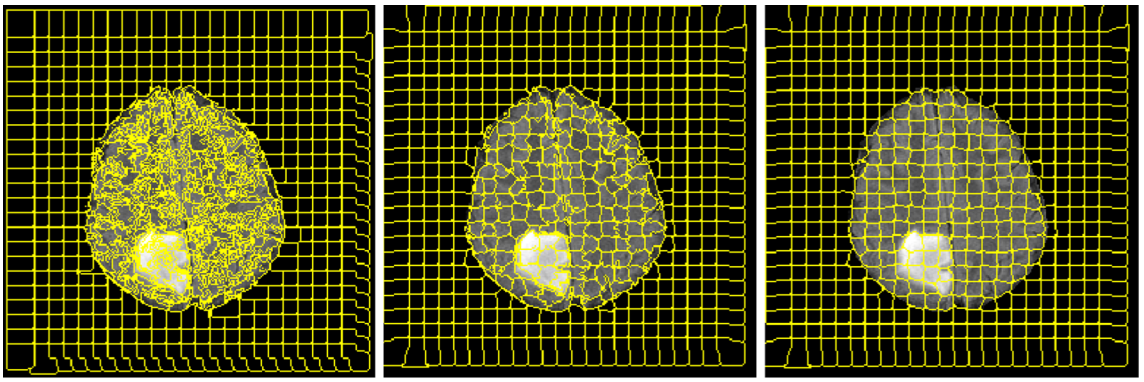


Fig. 8 Superpixel segmentation with $S = 10$ and different compactness: **a** $m = 0$, **b** $m = 0.2$ and **c** $m = 0.5$

Table 2 Examples of the impact of different initial superpixel side sizes, S, on the segmentation accuracy of the tumour in FLAIR images with compactness factor $m = 0.2$

Superpixel side size	4	6	8	10	15	20
Dice overlap	0.98	0.96	0.92	0.85	0.73	0.56

and the Dice measure is used for assessing the performance of superpixel segmentation. The experiment ran on the selected training images from different tumour grades. The results are presented in

Table 2, which shows that increasing the superpixel size results in less segmentation accuracy. A superpixel size of $S = 6$ is chosen which has a good performance and also contains sufficient information within the superpixel for texture feature calculation.

Texton feature parameters

For the direction of Gabor filters, six settings from the range: [0°, 30°, 45°, 60°, 90°, 120°] are chosen. These degrees cover the whole space of the region with a reasonable step. Although adding more orientations seems to include more details to the features, it will also increase the computation time and may add redundant information which may affect the classification accuracy.

The maximum and minimum values for size and wavelength coefficients are selected empirically in conjunction with visual inspection. For the size values under the 0.3, filtered images are very close to the original image, while for the values above the 1.5, the images are intensively blurred. Therefore, the kernel sizes are selected within this range with the increment of 0.3, i.e. [0.3, 0.6, 0.9, 1.2, 1.5]. Wavelength coefficients are selected empirically by visual inspection of the filters in the range of [0.8, 1.0, 1.2, 1.5].

As discussed in "Texton feature" section, the texton map is created by applying k-means clustering to different filter responses. A key question in using k-means clustering is to determine the cluster (texton) number k. However, it is not straightforward to provide an accurate number of the structures presented in the image. Theoretically, for the texton generation, with the increasing number of clusters, more specific texton differences between clusters could be extracted. However, a large k may result in overclassification and also in increasing computational cost. In our experiment, the number of clusters (textons) ($k = 5$) is chosen empirically according to the number of tissues that may be present in the FLAIR images, i.e. grey matter, white matter, tumour, oedema and other tissue types.

Fractal feature parameters

Different threshold levels for fractal feature extraction have been examined. The accuracy of superpixel classification using fractal features only is a measure to assess the effect of number of threshold level. As shown in Fig. 9, after increasing $n_t = 3$ levels of threshold, which creates 6 binary channels, the overlap measure does not increase significantly. On the other hand, increasing each level will add 6 more features (each binary channel has 3 fractal features) to the feature vector. This makes the classification more complicated and also increases the computation time for both fractal feature calculation and classification. Therefore, the optimum level of threshold $n_t = 3$ is chosen for the segmentation of oedema and tumour core.

Extremely randomized trees parameters

Implementation of ERT was performed in MATLAB 2015b using the open- source code provided in [54] which is based on the method by Geurts et al. [50]. To assess the impact of ERT parameters on the classification performance, the experiment ran on the selected training images with different sizes of trees. The maximum depth of the trees for the

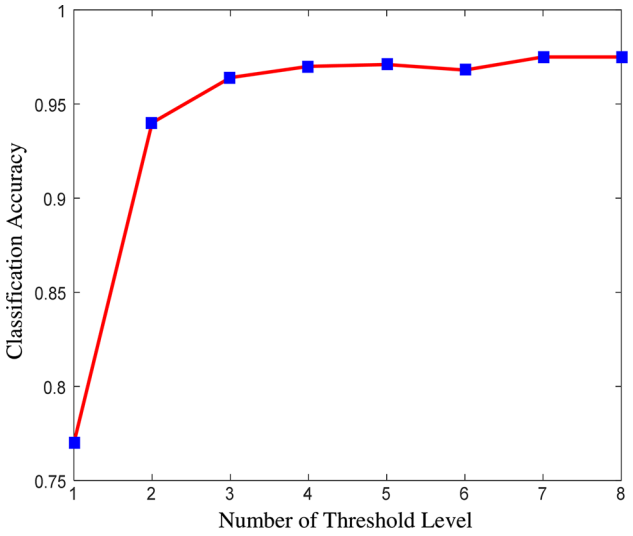


Fig. 9 Effect of number of threshold levels on the classification accuracy

Table 3 Impact of the number of trees on ERT classifier accuracy

Number of trees	5	10	20	50	100
Classification accuracy (%)	92.35	97.86	98.22	98.28	98.28

ERT is set to 15. Minimum sample size, $n_{\min}$, for splitting a node is selected to be 2 as according to [50] $n_{\min} = 2$ is an optimal value for most classification tasks. The number of attributes for random split is considered as 5 which is equal to the number of selected features after applying the mRMR feature reduction. As shown in Table 3, by adding more than 20 trees to the ERT, there is no significant improvement for the classifier accuracy. In addition, increasing the number of trees will increase the computation time. Therefore, in our experiment, the size of 20 trees is used for the ERT classifier.

Comparative experimental results

Our automated method is compared with the manual annotation provided by an expert. Dice similarity score [55], which calculates the overlap of segmented area and manual segmentation, is used to quantitatively evaluate the proposed method. The Dice overlap measure ranges from 0 to 1. The lower value represents lower overlap, while 1 demonstrates a complete overlap.

In the classification stage, leave-one-out validation is performed on single-channel MR FLAIR data. The brain MR images are partitioned into superpixels based on Eq. (3) using the initial window side size of $S = 6$ pixels and the compactness factor $m = 0.2$. All the superpixels inside the brain area are used for classification. Based on the manual annotation, superpixels are split into two classes: normal tissue and brain tumour including tumour core and oedema. Superpixel with at least 50 % of tumour pixels in manual annotation is considered as a tumour superpixel. The remaining superpixels are labelled as normal. The model is trained based on these two labels. During the testing stage, the trained model is then applied and labels are assigned to all the superpixels inside the brain. The ERT classifier is compared with support vector machine (SVM) [56] for the classification of superpixels. The tumour area is obtained by grouping the superpixels related to tumour class.

In total, five features are used after mRMR feature selection, which are the normalized mean intensity, fractal dimension, two texton channels and mean curvature within the superpixel. It is noted that, though ERT can be directly used as feature selection and classification, to ensure a fair comparison between the ERT and SVM classifiers, the same feature set is considered in this study.

Evaluations have been carried out qualitatively by visual inspection and quantitatively using three classification measures for the detection and the Dice overlap measure for the segmentation. It is noted that, for the standard four classification measures (accuracy, precision, sensitivity, specificity), both accuracy and specificity will give very high values due to the highly imbalanced nature of our data. Therefore, to properly evaluate the classification performance, only precision and sensitivity are considered.

Table 4 presents the evaluation measures for SVM and ERT, respectively. It can be seen that, ERT produces a better classification performance, compared to that of SVM, with an overall classification precision of 87.86 %, sensitivity of 89.48 % and BER of 6 % for ERT, and of 83.59 %, 87.82 % and 7 % for SVM, respectively.

The Dice score overlap measure of the individual patient comparing the ground truth with the segmented tumour masks using both SVM and ERT is plotted in Fig. 10. It can be seen that the overlap ratio using the ERT-based method is much better than that of SVM-based for all the three tumour grades, with mean and standard deviation Dice score of 0.91 ± 0.04 for ERT-based and 0.87 ± 0.05 for SVM-based.

Figure 11 shows comparison results of Dice score overlap measure (mean and standard deviation) for SVM versus ERT for different tumour grade types from II to IV. The results show that using ERT classifier increases the segmentation accuracy for all grades of tumour type. There is an evident difference between segmentation overlap measures for different tumour grades using SVM classifier. The result is not satisfactory for grade II (with mean overlap of 0.81), compared with the other two grades (with mean overlap of 0.90), while the segmentation results based on ERT classifiers are consistent for all tumour grade types, with mean overlap of 0.91.

Table 4 Comparison evaluation on superpixel classification using SVM-based and ERT-based classifier, respectively, on the 5 features selected using mRMR

Case no.	Grade	SVM			ERT		
		Precision (%)	Sensitivity (%)	BER	Precision (%)	Sensitivity (%)	BER
1	II	62.71	97.33	0.02	69.85	97.45	0.02
2	II	58.65	98.14	0.02	90.24	98.65	0.01
3	II	72.55	98.41	0.02	74.21	99.12	0.01
4	II	68.53	94.88	0.03	70.24	96.05	0.02
5	II	76.33	55.64	0.23	78.43	56.32	0.22
6	II	75.83	73.45	0.14	85.63	71.32	0.15
7	III	84.75	98.75	0.01	86.07	99.35	0.01
8	III	88.54	83.32	0.09	90.78	85.64	0.08
9	III	88.92	98.11	0.01	91.44	98.67	0.01
10	IV	95.22	83.25	0.09	97.44	89.03	0.06
11	IV	93.45	88.53	0.07	96.57	91.65	0.05
12	IV	81.55	73.98	0.14	84.33	75.92	0.13
13	IV	80.35	92.68	0.04	82.53	95.73	0.03
14	IV	90.12	92.51	0.04	91.32	95.87	0.03
15	IV	93.42	93.76	0.04	96.78	94.02	0.03
16	IV	87.45	83.06	0.09	90.21	84.15	0.08
17	IV	95.34	87.75	0.06	96.81	91.87	0.04
18	IV	98.33	82.56	0.09	98.43	85.33	0.08
19	IV	96.21	92.51	0.05	98.12	94.03	0.04
Mean	All	83.59	87.82	0.07	87.86	89.48	0.06
STD	All	11.76	11.09	0.06	9.27	11.23	0.06

BER balanced error rate

The classification is performed for tumour including oedema and active tumour core versus normal brain tissue

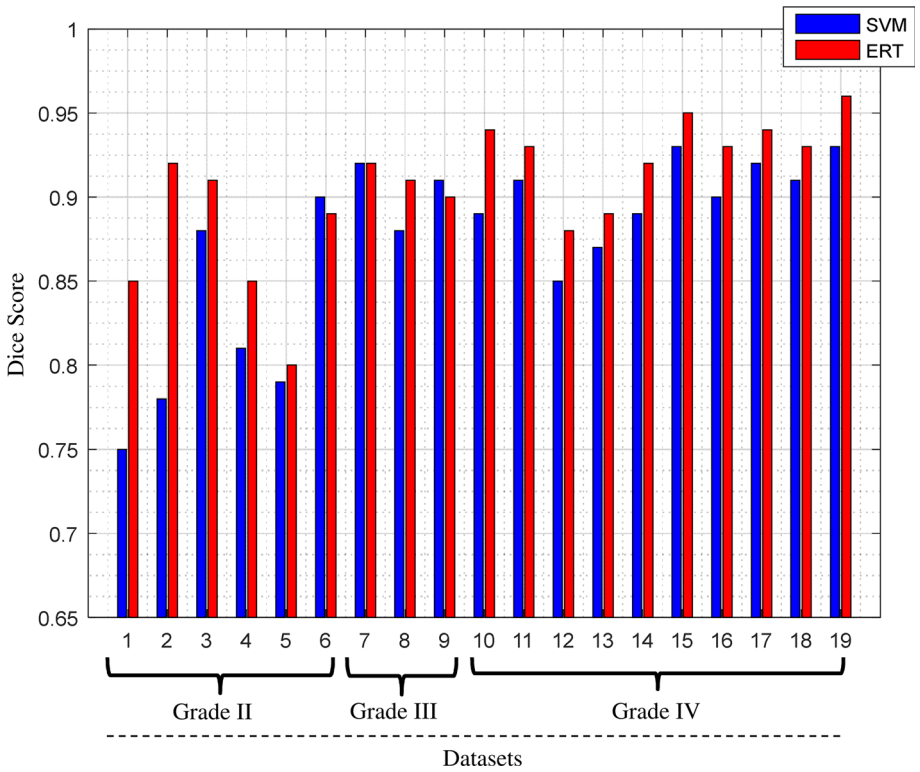


Fig. 10 Comparison of Dice score overlap measure of SVM versus ERT for all our clinical patient data (19 scans). Dice score in vertical axis starts from 0.65 for better illustration

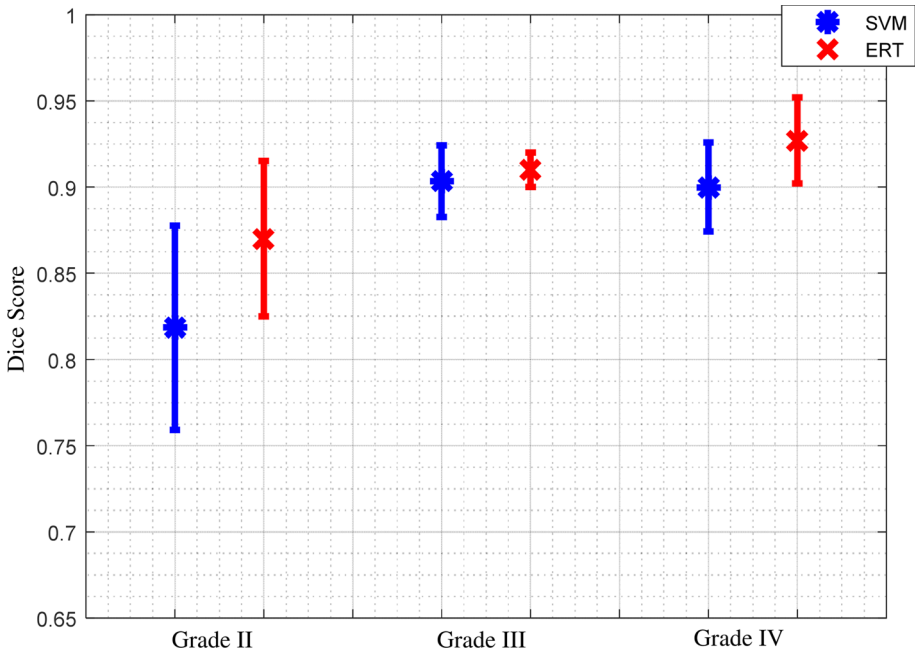


Fig. 11 Comparison between average and standard deviation of Dice score overlap measure for SVM versus ERT for different tumour grade types II to IV

Table 5 Statistical parameters of the Wilcoxon signed-rank test

	p	Z value
Dice	<0.001	−3.826
Precision	<0.001	−3.823
Sensitivity	0.001	−3.340

The Wilcoxon signed-rank test is employed to determine whether there are any differences in both the segmentation measure of Dice overlap and classification measures of precision and sensitivity, obtained using the two different classifiers (i.e. SVM and ERT), at 99 % confidence level, with 19 subjects. Our analysis, based on the p and z values of the statistical test, suggests that there is a statistically significant improvement in the segmentation measures of Dice overlap and in the classification measures of precision and sensitivity, when using the ERT classifier instead of the SVM. Table 5 shows the statistical parameters of our analysis.

Figure 12 shows examples of segmentation results for ERT and SVM methods overlaid on the manual annotation. Both SVM- and ERT-based methods obtained satisfactory results for the detection and segmentation of different tumour types, with ERT-based method providing slightly better results than that from SVM. Figure 13 shows examples of much better detection and segmentation results obtained from ERT-based methods, compared to that from SVM. Most of the false-positive superpixels from SVM (e.g. Figs. 12c4 and 13c1) can be effectively eliminated using ERT, while some tumour superpixels which are wrongly classified to the normal brain tissues by using SVM (e.g. Fig. 13c2, c3) can be correctly classified as tumour by using the ERT, demonstrating the higher sensitivity of the ERT. Comparison examples of segmentation for grade II tumour in the first row of both Figs. 12 and 13 illustrate that the segmented tumour boundary from ERT (d1) is closer to the manual annotation, compared to that of SVM (c1).

Evaluation on BRATS 2012 dataset

To assess the robustness of our method, we further validate the method on a publicly available BRATS 2012 clinical training dataset [8,9]. In this section, the data are described and the segmentation results are presented and discussed.

BRATS 2012 dataset description

The BRATS 2012 annotated clinical training dataset is used which consists of multicontrast MR scans of 30 glioma patients (e.g. 10 low grade and 20 high grade) [7–9]. It should be noted that the BRATS 2012 clinical training datasets are similar to that of BRATS 2013. For those training set, the ground truths are provided by a trained human expert [7]. For each patient data, T1, T2, FLAIR and post-gadolinium T1 MR images are available. Data were acquired from multicentres and using different scanners with different field strengths (1.5 T and 3T). In this study, only FLAIR images are used to evaluate our method.

Experimental results

The majority of parameters tuned for our own clinical dataset, including compactness coefficient for superpixel segmen-

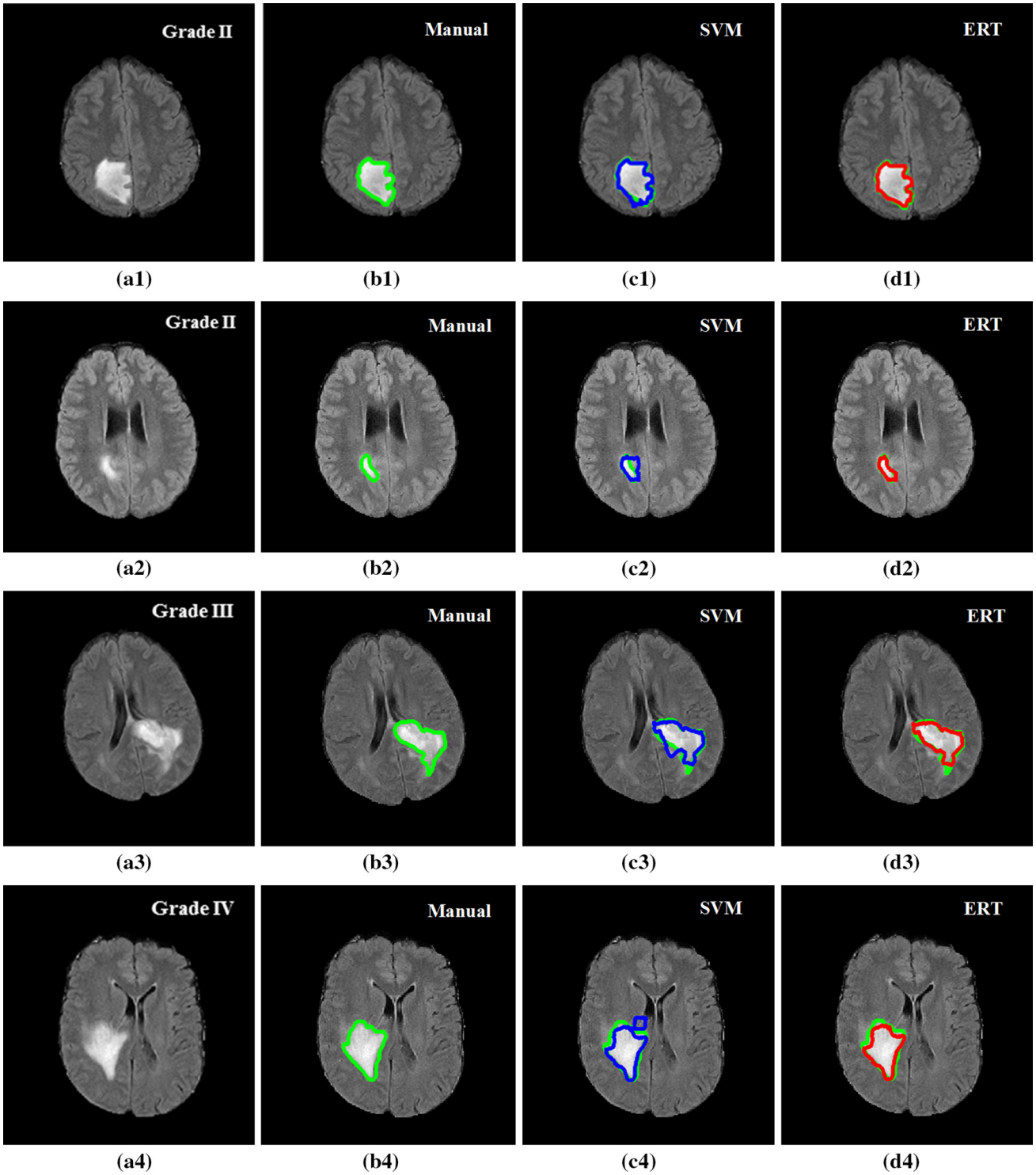


Fig. 12 Examples of segmentation results overlay on manual segmentation (*green*). FLAIR image with tumour grade II (*a1*), grade II (**a2**), grade III (**a3**) and grade IV (**a4**); **b1**–**b4** manual segmentation; **c1**–**c4** results using SVM; and **d1**–**d4** results using ERT. Both SVM- and ERT-based methods obtained satisfactory results for the segmentation of different tumour types, with ERT-based method providing slightly better results than that from SVM

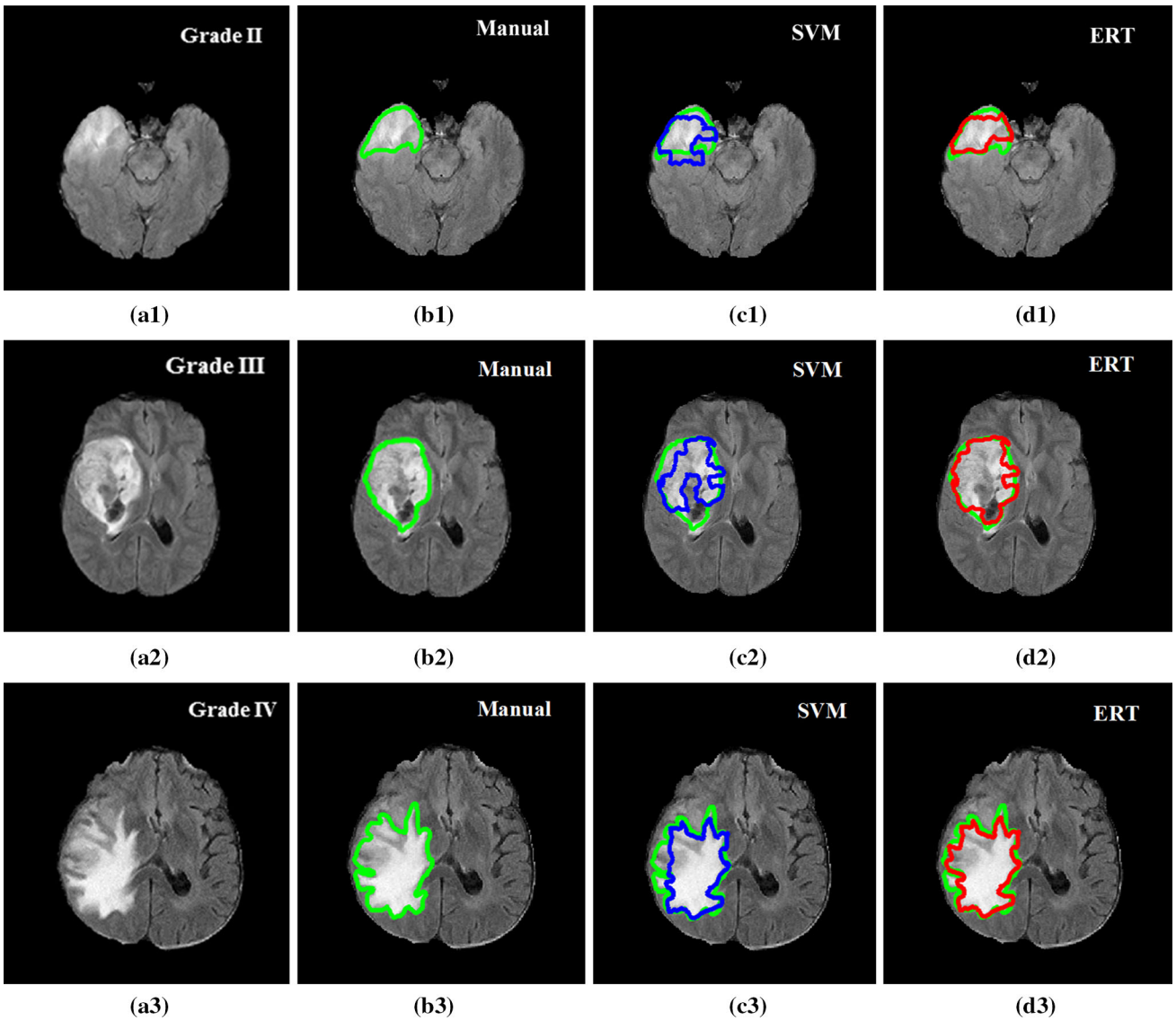


Fig. 13 Examples of good detection and segmentation results obtained from ERT-based methods. FLAIR image with tumour grade II (**a1**), grade III (**a2**), grade IV (**a3**); **b1**–**b3** manual segmentation; **c1**–**c3** results using SVM; and **d1**–**d3** results using ERT. Most of the false-positive superpixels from SVM (e.g. (**c1**) and (**c3**)) can be effectively eliminated using ERT, while some tumour superpixels which are wrongly classified to the normal brain tissues by using SVM (e.g. (**c2**)) can be correctly classified as tumour by using the ERT

tation, fractal features, and number of clusters for texton generation, are directly used in the BRATS dataset. All the parameters for both ERT and SVM classifiers are the same. However, only superpixel size and filter size used for Gabor filter defined in Eq. (4) are slightly adjusted, e.g. for superpixel size, instead of using size of 6 in our own dataset, size of 5 is used in the BRATS dataset, while a smaller range of filter size (e.g. [0.3 0.5 0.8 1.1 1.4]) is used for the Gabor filter bank in texton feature extraction. This is due to the different image sizes and resolutions between the two datasets. All the five features selected using mRMR are also used in BRATS dataset for the classification of each superpixel.

Table 6 presents the evaluation measures for SVM and ERT, respectively. It can be seen that ERT produces a slightly better classification performance, compared to that of SVM, with an overall classification precision of 89.09 %, sensitivity of 88.09 % and BER of 6 % for ERT and of 83.79, 82.72 and 9 % for SVM, respectively.

The Dice overlap ratio between the ground truth from manual annotation and the segmented tumour using ERT and SVM classifiers for the BRATS dataset is presented in Table 7. It can be seen that the overlap ratio using the ERT-based method is much better than that of SVM-based for all

Table 6 Comparison evaluation on superpixel classification using SVM-based and ERT-based classifiers, respectively, on BRATS 2012 dataset using 5 features selected by mRMR

Case no.	Grade/ID	SVM			ERT		
		Precision (%)	Sensitivity (%)	BER	Precision (%)	Sensitivity (%)	BER
1	LG-01	87.68	89.43	0.06	91.84	88.18	0.06
2	LG-02	96.98	88.60	0.06	99.02	92.63	0.04
3	LG-04	75.59	81.95	0.10	78.40	90.67	0.05
4	LG-06	84.57	87.42	0.07	92.15	90.05	0.05
5	LG-08	90.95	83.54	0.09	93.11	91.05	0.05
6	LG-11	89.91	82.67	0.09	91.41	86.78	0.07
7	LG-12	91.42	83.19	0.09	92.18	84.19	0.08
8	LG-13	74.48	79.19	0.11	79.28	85.86	0.08
9	LG-14	83.17	80.37	0.10	88.03	82.58	0.09
10	LG-15	76.15	80.60	0.10	82.64	89.29	0.06
11	HG-01	92.77	92.55	0.04	98.47	95.91	0.03
12	HG-02	83.51	82.15	0.09	90.45	88.62	0.06
13	HG-03	85.46	79.59	0.11	91.31	88.68	0.06
14	HG-04	94.08	89.30	0.06	98.69	90.96	0.05
15	HG-05	78.96	72.06	0.14	83.16	77.70	0.12
16	HG-06	81.54	74.77	0.13	93.13	90.32	0.05
17	HG-07	75.48	79.60	0.11	83.16	87.81	0.07
18	HG-08	87.87	90.58	0.05	89.21	93.88	0.04
19	HG-09	84.78	87.04	0.07	87.56	90.35	0.05
20	HG-10	67.77	65.63	0.18	73.17	71.84	0.15
21	HG-11	90.53	85.68	0.08	92.39	90.21	0.05
22	HG-12	88.58	86.82	0.07	92.08	89.36	0.06
23	HG-13	80.10	84.35	0.08	88.64	89.23	0.06
24	HG-14	84.74	87.99	0.07	88.80	91.76	0.05
25	HG-22	78.21	80.75	0.10	88.79	92.83	0.04
26	HG-24	82.50	85.14	0.08	88.87	87.98	0.07
27	HG-25	82.23	86.08	0.07	90.95	88.16	0.06
28	HG-26	84.41	82.60	0.09	91.71	89.84	0.06
29	HG-27	77.16	72.67	0.14	80.93	75.54	0.13
30	HG-22	82.10	79.19	0.11	93.09	90.42	0.05
Mean	All	83.79	82.72	0.09	89.09	88.09	0.06
STD	All	6.63	5.95	0.03	6.00	5.22	0.03

BER balanced error rate
The classification is performed for tumours including oedema and active tumour core versus normal brain tissue

Table 7 Comparison results for Dice overlap ratio between manual annotation and the automated segmentation using SVM and ERT for BRATS 2012 dataset (30 scans)

Case no.	Grade/ID	Dice	
		SVM	ERT
1	LG-01	0.85	0.89
2	LG-02	0.93	0.95
3	LG-04	0.78	0.87
4	LG-06	0.84	0.91
5	LG-08	0.88	0.92
6	LG-11	0.86	0.89
7	LG-12	0.88	0.92
8	LG-13	0.75	0.81
9	LG-14	0.80	0.84
10	LG-15	0.78	0.88
11	HG-01	0.89	0.92
12	HG-02	0.83	0.88
13	HG-03	0.82	0.91
14	HG-04	0.90	0.92
15	HG-05	0.74	0.78
16	HG-06	0.79	0.91
17	HG-07	0.78	0.85
18	HG-08	0.89	0.91
19	HG-09	0.86	0.89
20	HG-10	0.65	0.71
21	HG-11	0.87	0.92
22	HG-12	0.88	0.91
23	HG-13	0.81	0.89
24	HG-14	0.86	0.90
25	HG-15	0.78	0.91
26	HG-22	0.84	0.88
27	HG-24	0.85	0.89
28	HG-25	0.84	0.90
29	HG-26	0.75	0.79
30	HG-27	0.81	0.91
Mean	All	0.83	0.88
STD	All	0.06	0.05

the three tumour grades, with mean Dice score of 0.88 for ERT-based and 0.83 for SVM-based.

Figures 14 and 15 show examples of segmentation results for ERT and SVM methods overlaid on the manual annotations for high-grade tumour (Fig. 14) and low-grade tumour (Fig. 15). Both SVM- and ERT-based methods obtained satisfactory results for the detection and segmentation of different tumour types, with ERT-based method providing slightly better results than that from SVM. Most of the false-positive superpixels from SVM (e.g. Figs. 14c2 and 15c3) can be effectively eliminated using ERT, while some tumour superpixels which are wrongly classified to the normal brain tissues by using SVM (e.g. Fig. 15c2) can be correctly classified as tumour by using the ERT, demonstrating the higher sensitivity of the ERT. Comparison examples of segmentation for both high-grade and low-grade tumours in Figs. 14 and 15 illustrate that the segmented tumour boundary from ERT is closer to the manual annotation compared to that of SVM.

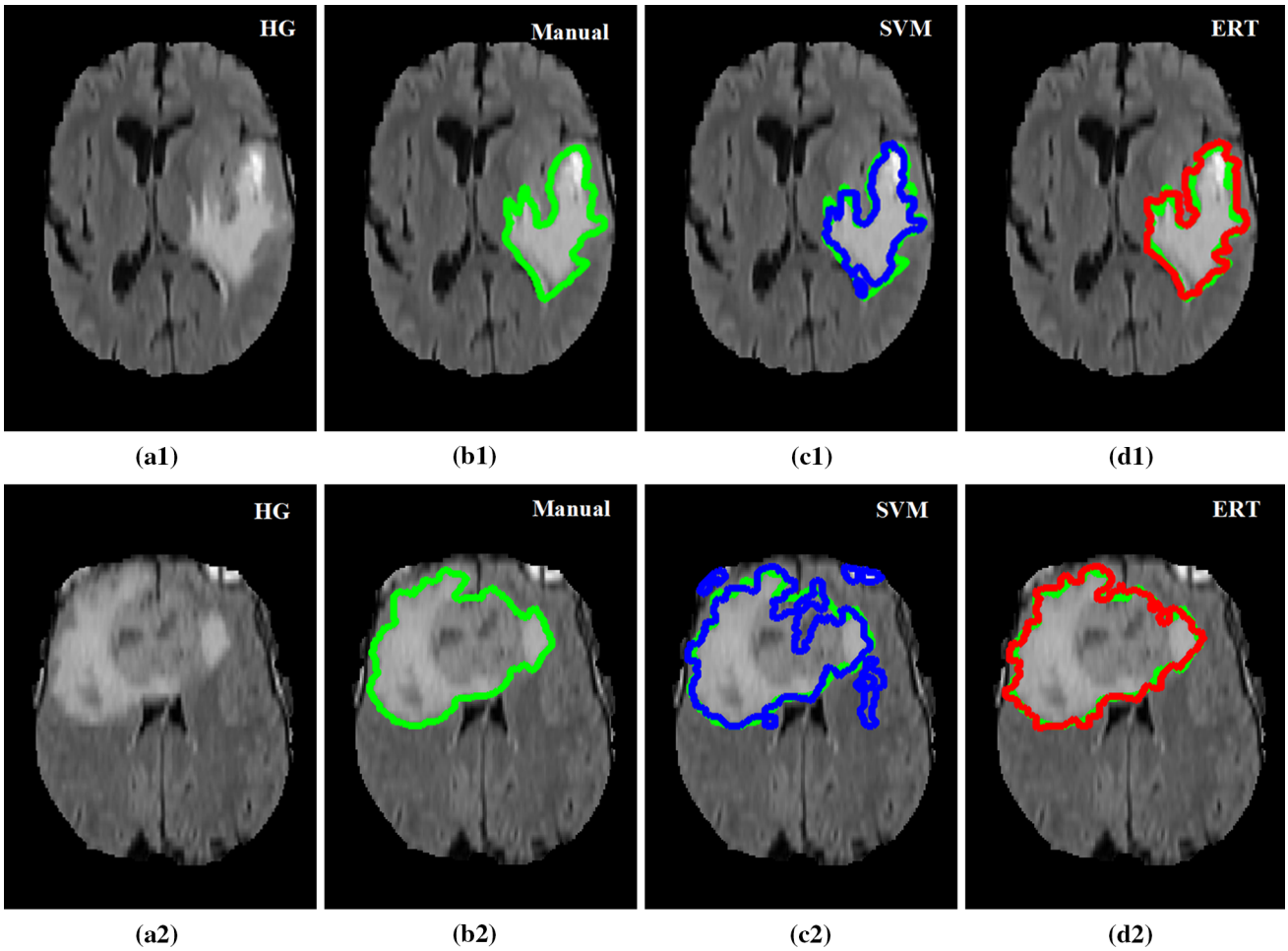


Fig. 14 Examples of detection and segmentation results obtained from ERT-based methods on BRATS 2012 data. FLAIR image with high-grade tumour Case HG-01 (**a1**), HG-15 (**a2**); **b1**–**b2** manual segmentation; **c1**–**c2** results using SVM; and **d1**–**d2** results using ERT

Discussion

Discussion of applying our method to BRATS dataset

The BRATS clinical training dataset is used to further evaluate the robustness of the method. As discussed in "Experimental results" section, the majority of the parameters are the same as those optimized for our own clinical data. The overall average and standard deviation of Dice score overlap measures for all our 19 patient data and 30 BRATS 2012 dataset using both ERT-based and SVM-based methods are shown in Fig. 16. The results show that using the state-of-the art ERT for classification of superpixels results in more accurate and robust segmentation compared to that of SVM classifier. For our own clinical dataset, the Dice score overlap measure for ERT-based segmentation is 0.91 ± 0.04, while for SVM-based method, it is 0.87 ± 0.05. For BRATS 2012 dataset, the score overlap measure for ERT-based segmentation is 0.88 ± 0.05, while for SVM-based method, it is 0.83 $\pm$ 0.06. It can be seen that the mean Dice scores obtained from BRATS training dataset are closer to that from our own clinical dataset; this suggests robustness of the method.

A comparison of our proposed method on BRATS 2012 clinical dataset with the best scores in the challenges [7] is presented in Table 8. As shown in Table 8, method in Tustison et al. [27] which was the winner of on-site BRATS 2013 challenge was performed on the challenge data. Though datasets might be different, the best on-site score could provide a comparable reference using BRATS dataset. Also, comparing our method to the method by Reza and Iftekharuddin [29] which has the best result for the training set of the BRATS multiprotocol dataset (this is the same dataset used in our evaluation; however, we only use FLAIR protocol), our method has achieved the average Dice overlap of 0.88 which is closer to that of 0.92 by Reza's method. As discussed in "Experimental results" section, to assess the robustness of our method, the similar optimum parameters and the same five features tuned for our own clinical dataset are directly applied to the BRATS dataset. In particular, our algorithm is trained on 1.5T data from a signal centre, whereas the BRATS

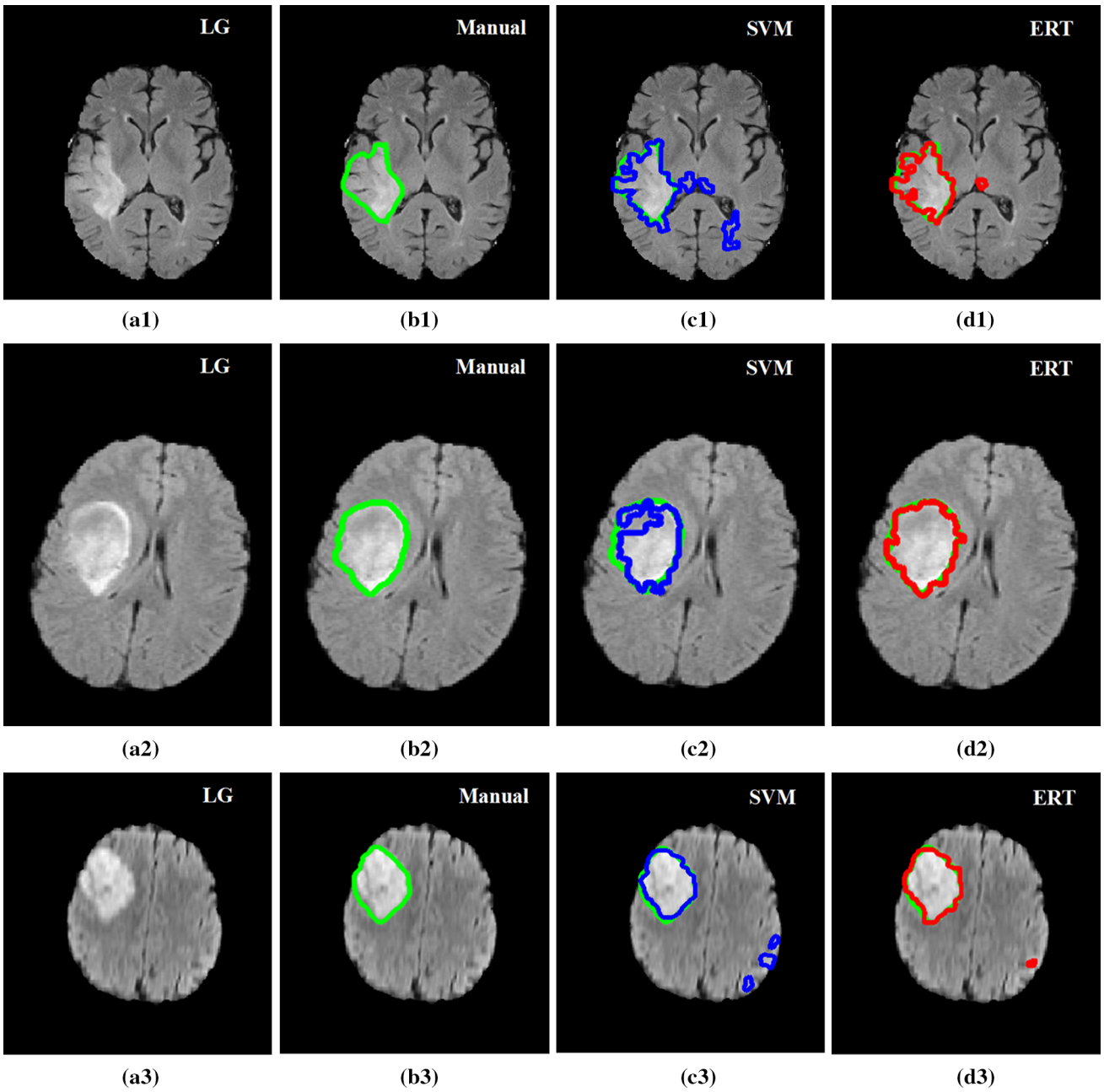


Fig. 15 Examples of detection and segmentation results obtained from ERT-based methods on BRATS 2012 data. FLAIR image with low-grade tumour Case LG-04 (**a1**), LG-11 (**a2**) and LG-12 (**a3**); **b1–b3** manual segmentation; **c1–c3** results using SVM; and **d1–d3** results using ERT

data contain multicentre data from 1.5T and 3T MRI scanners and will likely contain variability of image features and contrast that would not be accounted for within our current optimization and training phase.

Discussion of our method

FLAIR images are routinely acquired in clinical practice as part of standard diagnostic clinical MRI of brain tumours. Our experimental results shown in Tables 4, 5 and Fig. 10 demonstrate high performance of automated detection and segmentation of the brain tumour oedema and core regions in FLAIR MRI. The method was also further validated on BRATS 2012 training dataset (FLAIR) with the similar model parameters and features tuned for our own clinical dataset; good results shown in Tables 6 and 7 suggest the robustness of our method.

Selecting an appropriate superpixel size is critical for increasing the overall segmentation accuracy within an optimum calculation speed. Large superpixel size can ensure

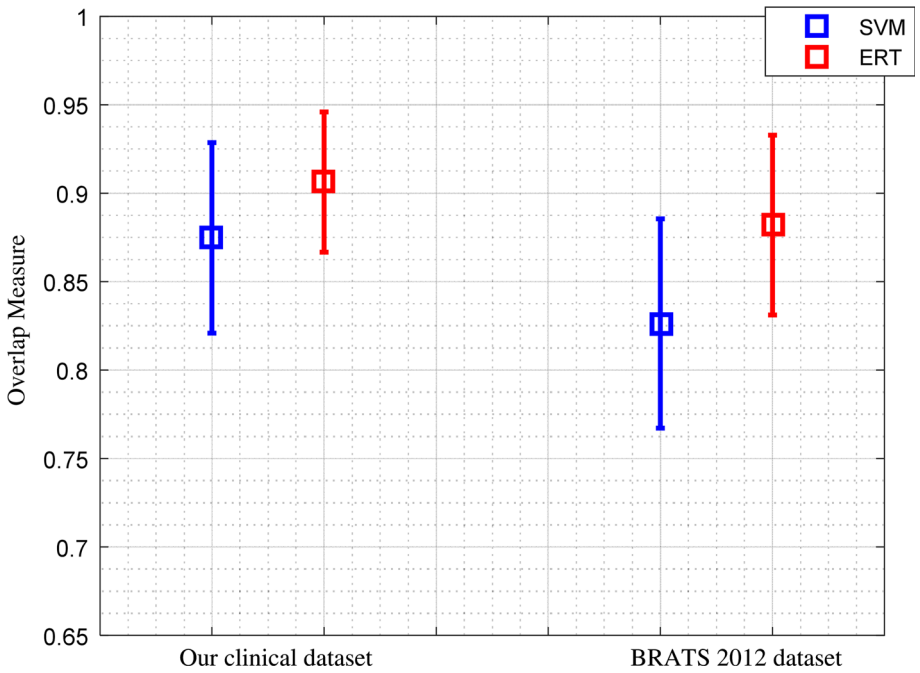


Fig. 16 Comparison of the average and standard deviation of Dice score overlap measures for SVM versus ERT for all 19 data scans in our dataset and 30 clinical scans in BRATS 2012 dataset

Table 8 Comparison with other related methods using BRATS dataset (MICCAI 2012)

Method	Description	Comment	Whole tumour (dice)
Tustison et al. [27]	Random forests (ANTs/ANTsR package)	Best MICCAI 2013 on-site	0.87
Reza and Iftekharuddin [29]	Random forests + texture features	Best on training MICCAI 2013	0.92
Our method	ERT + supervoxels	Training MICCAI 2012	0.88

Our method and Reza and Iftekharuddin [29] are performed on BRATS clinical training data and the other work (Tustison et al. [27]) is performed on BRATS challenge data

fast computation and may provide sufficient information for feature extraction such as stable texture features. However, large size of superpixel may contain more than one class of pixels which leads to inaccurate feature calculation (such as small areas of calcification or haemorrhage), and it is also not suitable for small-sized lesions. While small size of superpixel has higher probability of purely containing one class of pixels, it is preferred for small lesion segmentation. However, they may not have enough pixels for calculating stable features, and the computation time for generating the small size partitions is very high. In this study, the size of superpixel is obtained through exhausted parametric searching during the training stage. An optimization algorithm such as genetic algorithm can be explored to effectively find an optimum superpixel size which provides a good trade-off between computation time and segmentation accuracy.

Another important parameter in superpixel segmentation stage is the compactness factor. Higher value of this parameter leads to more rigid partitions which are more stable and usually less noisy, i.e. holes or sparse separated pixels. However, the segmentation may not follow the tissue boundaries very well, especially in the cases where there are no sharp or clear boundaries. While, lower compactness values result in more flexible and accurate boundaries, but the segmentation may produce more isolated and disconnected pixels. They also may generate very narrow superpixels which are not appropriate for texture analysis. In our current study, the compactness factor is determined through visual inspection. Optimization methods need to be investigated to obtain an optimum compactness factor which provides a good trade-off between noise and flexibility.

For the comparison of our method on BRATS data, we refer to the work published in [7] which used these data in MICCAI challenge. However, some of their methods are assessed on the training dataset, while others are on the separate testing dataset. Due to the fact that our current study is based on binary classification (i.e. tumour including oedema and active tumour core versus normal brain tissue) using single FLAIR protocol, it is difficult to have a direct comparison with the current published methods on BRATS data. However, our results which are in the same range of other methods and are close to the best segmentation of whole tumour demonstrate the promise of the method.

Although currently we have only evaluated our segmentation algorithm for FLAIR images, it should be straightforward to apply the same superpixel methodology to contrast-enhanced T1w images and determine the signal intensity and higher-order features that best segment the contrast-enhancing region of high-grade gliomas. In fact, we are currently working on this direction.

In this study, we note the importance of the preprocessing step, namely MRI histogram normalization. This is of particular importance when the method is applied to BRATS dataset, whose data are from multicentres and different scanners.

In the current study, we also note in Fig. 13a2 that small hypointense spots in the FLAIR (and corresponding T1w) may be calcifications, and the hypointense FLAIR region, which is excluded by the SVM method (Fig. 13c2) but included in the ERT analysis (Fig. 13d2), is haemorrhagic since there is hyperintensity in the T1w MRI. This is a limitation of the current single modality analysis if these regions need to be separately specified. Future studies extending our method to multimodal data are planned. This will include the segmentation of different tissue subtypes (e.g. necrosis, active tumour, infiltrative tumour, oedema) by incorporating information from multimodal clinical MRI, including perfusion and diffusion imaging.

For our own clinical dataset, the ground truths were provided based on one expert's manual annotation. There may have some errors in the manual annotations, which may include intratumoural bleeding or calcification in the tumour (e.g. in Fig. 13b2). When those annotations are used to train the model, it may lead to some errors in the final segmentation. Also, our current clinical dataset mainly contains general cases, such as different tumour grades from a wide range of patient ages (patient ages at the time of scanning ranged from 22 to 73). In the future, we will look into more complicated cases, such as calcification, intratumoural bleeding or elderly patients with white matter disease, which are clinically very important to distinguish against.

Conclusion

This paper proposed a fully automated method for the detection and segmentation of brain tumour from FLAIR MRI images, by calculating Gabor texton feature, fractal analysis, curvature and statistical intensity features from superpixels. ERT is then used to classify each superpixel into tumour or healthy brain tissue. The formation of superpixel by grouping voxels with similar properties and extracting features from superpixels can not only improve the accuracy of feature extraction, especially for the superpixels near the boundaries between different tissues, but also significantly reduce the computation time, compared to voxel-based feature calculation and classification. The experimental results demonstrate the high detection and segmentation performance of the proposed method using ERT classifier, with average sensitivity of 89.48 %, BER of 6 % and Dice overlap ratio of 0.91. To assess the robustness of the method, the method was further evaluated on BRATS 2012 dataset, which results in similar good performances of 88.09 %, 6 % and 0.88, respectively. This provides a close match to expert delineation across all grades of glioma, leading to a faster and more reproducible method of brain tumour delineation to aid patient management.

Acknowledgments This research was supported by European FP7 collaborative Project "MyHealthAvatar" (600929). Dr. Yang was sponsored by Cancer Research UK (C1459/A13303). Data were obtained during the EU FP7 "eTUMOUR" Project (LSHC-CT-2004-503094).

Compliance with ethical standards

Conflict of interest The authors declare they have no conflict of interest with regard to the work presented.

Ethical approval Ethical approval to conduct this study was obtained from Local Committee of St George's Hospital, London.

Informed consent Informed consent was obtained from all individual participants in the study.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

1. Wen PY, Macdonald DR, Reardon DA, Cloughesy TF, Sorensen AG, Galanis E, DeGroot J, Wick W, Gilbert MR, Lassman AB, Tsien C, Mikkelsen T, Wong ET, Chamberlain MC, Stupp R, Lamborn KR, Vogelbaum MA, van den Bent MJ, Chang SM (2010) Updated response assessment criteria for high-grade gliomas: response assessment in Neuro-Oncology Working Group. JCO 28:1963–1972. doi:10.1200/JCO.2009.26.3541
2. Niyazi M, Brada M, Chalmers AJ, Combs SE, Erridge SC, Fiorentino A, Grosu AL, Lagerwaard FJ, Minniti G, Mirimanoff R-O, Ricardi U, Short SC, Weber DC, Belka C (2016) ESTRO-ACROP guideline "target delineation of glioblastomas.". Radiother Oncol 118:35–42. doi:10.1016/j.radonc.2015.12.003
3. Rees J, Watt H, Jäger HR, Benton C, Tozer D, Tofts P, Waldman A (2009) Volumes and growth rates of untreated adult low-grade gliomas indicate risk of early malignant transformation. Eur J Radiol 72:54–64. doi:10.1016/j.ejrad.2008.06.013
4. Eisele SC, Wen PY, Lee EQ (2016) Assessment of brain tumor response: RANO and its offspring. Curr Treat Options Oncol 17:35. doi:10.1007/s11864-016-0413-5
5. Aslian H, Sadeghi M, Mahdavi SR, Babapour Mofrad F, Astarakee M, Khaledi N, Fadavi P (2013) Magnetic resonance imaging-based target volume delineation in radiation therapy treatment planning for brain tumors using localized region-based active contour. Int

J Radiat Oncol Biol Phys 87:195–201. doi:10.1016/j.ijrobp.2013.04.049
6. Odland A, Server A, Saxhaug C, Breivik B, Groote R, Vardal J, Larsson C, Bjørnerud A (2015) Volumetric glioma quantification: comparison of manual and semi-automatic tumor segmentation for the quantification of tumor growth. Acta Radiol 56:1396–1403. doi:10.1177/0284185114554822
7. 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 MA, 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 HC, 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, Leemput KV (2015) The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Trans Med Imaging 34:1993–2024. doi:10.1109/TMI.2014.2377694
8. BRATS:: The virtual skeleton database project. https://www.smir.ch/BRATS/Start2012. Accessed 3 Jul 2016
9. Kistler M, Bonaretti S, Pfahrer M, Niklaus R, Büchler P (2013) The virtual skeleton database: an open access repository for biomedical research and collaboration. J Med Internet Res 15:e245. doi:10.2196/jmir.2930
10. Sauwen N, Sima DM, Van Cauter S, Veraart J, Leemans A, Maes F, Himmelreich U, Van Huffel S (2015) Hierarchical non-negative matrix factorization to characterize brain tumor heterogeneity using multi-parametric MRI. NMR Biomed 28:1599–1624. doi:10.1002/nbm.3413
11. Patel MR, Tse V (2004) Diagnosis and staging of brain tumors. Semin Roentgenol 39:347–360
12. Bauer S, Wiest R, Nolte L-P, Reyes M (2013) A survey of MRI-based medical image analysis for brain tumor studies. Phys Med Biol 58:R97–R129. doi:10.1088/0031-9155/58/13/R97
13. Aljabar P, Heckemann RA, Hammers A, Hajnal JV, Rueckert D (2009) Multi-atlas based segmentation of brain images: atlas selection and its effect on accuracy. NeuroImage 46:726–738. doi:10.1016/j.neuroimage.2009.02.018
14. Park MTM, Pipitone J, Baer LH, Winterburn JL, Shah Y, Chavez S, Schira MM, Lobaugh NJ, Lerch JP, Voineskos AN, Chakravarty MM (2014) Derivation of high-resolution MRI atlases of the human cerebellum at 3T and segmentation using multiple automatically generated templates. NeuroImage 95:217–231. doi:10.1016/j.neuroimage.2014.03.037
15. Bauer S, Seiler C, Bardyn T, Buechler P, Reyes M (2010) Atlas-based segmentation of brain tumor images using a Markov random field-based tumor growth model and non-rigid registration. In: Annual international conference of the IEEE engineering in medicine and biology 2010, pp 4080–4083. doi:10.1109/IEMBS.2010.5627302
16. Szilágyi L, Lefkovits L, Benyó B (2015) Automatic brain tumor segmentation in multispectral MRI volumes using a fuzzy c-means cascade algorithm. In: 2015 12th international conference on fuzzy systems and knowledge discovery (FSKD), pp 285–291
17. Mei PA, de Carvalho Carneiro C, Fraser SJ, Min LL, Reis F (2015) Analysis of neoplastic lesions in magnetic resonance imaging using self-organizing maps. J Neurol Sci 359:78–83. doi:10.1016/j.jns.2015.10.032
18. Subbanna N, Precup D, Arbel T (2014) Iterative multilevel MRF leveraging context and voxel information for brain tumour segmentation in MRI. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition, pp 400–405
19. Juan-Albarracín J, Fuster-Garcia E, Manjón JV, Robles M, Aparici F, Martí-Bonmatí L, García-Gómez JM (2015) Automated glioblastoma segmentation based on a multiparametric structured unsupervised classification. PLoS One 10:e0125143. doi:10.1371/journal.pone.0125143
20. Jafari M, Kasaei S (2011) Automatic brain tissue detection in MRI images using seeded region growing segmentation and neural network classification. Aust J Basic Appl Sci 5:1066–1079
21. Yamamoto D, Arimura H, Kakeda S, Magome T, Yamashita Y, Toyofuku F, Ohki M, Higashida Y, Korogi Y (2010) Computer-aided detection of multiple sclerosis lesions in brain magnetic resonance images: False positive reduction scheme consisted of rule-based, level set method, and support vector machine. Comput Med Imaging Graph 34:404–413. doi:10.1016/j.compmedimag.2010.02.001
22. Rajendran A, Dhanasekaran R (2012) Fuzzy clustering and deformable model for tumor segmentation on MRI brain image: a combined approach. Procedia Eng 30:327–333. doi:10.1016/j.proeng.2012.01.868
23. Geremia E, Clatz O, Menze BH, Konukoglu E, Criminisi A, Ayache N (2011) Spatial decision forests for MS lesion segmentation in multi-channel magnetic resonance images. Neuroimage 57:378–390. doi:10.1016/j.neuroimage.2011.03.080
24. Wu W, Chen AYC, Zhao L, Corso JJ (2013) Brain tumor detection and segmentation in a CRF (conditional random fields) framework with pixel-pairwise affinity and superpixel-level features. Int J CARS 9:241–253. doi:10.1007/s11548-013-0922-7
25. Pinto A, Pereira S, Correia H, Oliveira J, Rasteiro DMLD, Silva CA (2015) Brain tumour segmentation based on extremely randomized forest with high-level features. In: 2015 37th annual international conference of the IEEE engineering in medicine and biology society (EMBC), pp 3037–3040
26. Gotz M, Weber C, Blocher J, Stieltjes B, Meinzer H, Maier-Hein K (2014) Extremely randomized trees based brain tumor segmentation. In: Proceeding of BRATS challenge-MICCAI, pp 006–011
27. Tustison N, Wintermark M, Durst C, Avants B (2013) ANTs and arboles. In: Proceedings of NCI-MICCAI BRATS, pp 47–50
28. Zhao L, Sarikaya D, Corso JJ (2013) Automatic brain tumor segmentation with MRF on supervoxels. In: Proceedings of NCI-MICCAI BRATS, pp 51–54
29. Reza S, Iftekharuddin KM (2013) Multi-class abnormal brain tissue segmentation using texture features. In: Proceedings of NCI-MICCAI BRATS, pp 38–42
30. Festa J, Pereira S, Mariz JA, Sousa N, Silva CA (2013) Automatic brain tumor segmentation of multi-sequence MR images using random decision forests. In: Proceedings of NCI-MICCAI BRATS, pp 23–26
31. Zikic D, Glocker B, Konukoglu E, Criminisi A, Demiralp C, Shotton J, Thomas OM, Das T, Jena R, Price SJ (2012) Decision forests for tissue-specific segmentation of high-grade gliomas in multi-channel MR. In: Ayache N, Delingette H, Golland P, Mori K (eds) Medical image computing and computer-assisted intervention—MICCAI 2012. Springer, Berlin Heidelberg, pp 369–376
32. Crimi A, Menze B, Maier O, Reyes M, Handels H (2016) Brainlesion: glioma, multiple sclerosis, stroke and traumatic brain injuries. Springer, Cham
33. Meier R, Knecht U, Loosli T, Bauer S, Slotboom J, Wiest R, Reyes M (2016) Clinical evaluation of a fully-automatic segmentation method for longitudinal brain tumor volumetry. Sci Rep 6:23376. doi:10.1038/srep23376
34. Law M, Young RJ, Babb JS, Peccerelli N, Chheang S, Gruber ML, Miller DC, Golfinos JG, Zagzag D, Johnson G (2008) Gliomas: predicting time to progression or survival with cerebral blood volume measurements at dynamic susceptibility-weighted contrast-enhanced perfusion MR imaging. Radiology 247:490–498. doi:10.1148/radiol.2472070898

35. Itakura H, Achrol AS, Mitchell LA, Loya JJ, Liu T, Westbroek EM, Feroze AH, Rodriguez S, Echegaray S, Azad TD, Yeom KW, Napel S, Rubin DL, Chang SD, Harsh GR, Gevaert O (2015) Magnetic resonance image features identify glioblastoma phenotypic subtypes with distinct molecular pathway activities. Sci Transl Med 7:303ra138. doi:10.1126/scitranslmed.aaa7582
36. Stall B, Zach L, Ning H, Ondos J, Arora B, Shankavaram U, Miller RW, Citrin D, Camphausen K (2010) Comparison of T2 and FLAIR imaging for target delineation in high grade gliomas. Radiat Oncol 5:5. doi:10.1186/1748-717X-5-5
37. Cho KH, Choi JH, Kim JY, Lee SH, Yoo H, Shin KH, Kim TH, Moon SH, Lee SH, Park HC (2012) Volumetric response evaluation after intensity modulated radiotherapy in patients with supratentorial gliomas. Technol Cancer Res Treat 11:41–48
38. Jenkinson M, Beckmann CF, Behrens TEJ, Woolrich MW, Smith SM (2012) FSL. Neuroimage 62:782–790. doi:10.1016/j.neuroimage.2011.09.015
39. Nyúl LG, Udupa JK, Zhang X (2000) New variants of a method of MRI scale standardization. IEEE Trans Med Imaging 19:143–150. doi:10.1109/42.836373
40. ITK—Segmentation & Registration Toolkit. https://itk.org/. Accessed 8 May 2016
41. Achanta R, Shaji A, Smith K, Lucchi A, Fua P, Süsstrunk S (2012) SLIC superpixels compared to state-of-the-art superpixel methods. IEEE Trans Pattern Anal Mach Intell 34:2274–2282. doi:10.1109/TPAMI.2012.120
42. Jain AK (1989) Fundamentals of digital image processing. Prentice Hall, Upper Saddle River
43. Henriksen JJ (2007) 3D surface tracking and approximation using Gabor filters. South Denmark University
44. Costa AF, Humpire-Mamani G, Traina AJM (2012) An Efficient Algorithm for Fractal Analysis of Textures. In: 2012 25th SIBGRAPI conference on graphics, patterns and images (SIBGRAPI), pp 39–46
45. Liao P-S, Chen T-S, Chung P-C (2001) A fast algorithm for multilevel thresholding. J Inf Sci Eng 17:713–727
46. Canny J (1986) A computational approach to edge detection. IEEE Trans Pattern Anal Mach Intell PAMI 8:679–698. doi:10.1109/TPAMI.1986.4767851
47. Schroeder MR (2009) Fractals, chaos, power laws: minutes from an infinite paradise. Dover Publications Inc., Mineola
48. Arridge SR A note on image and curvature. http://www0.cs.ucl.ac.uk/staff/S.Arridge/teaching/ndsp/curvature.pdf
49. Peng H, Long F, Ding C (2005) Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell 27:1226–1238. doi:10.1109/TPAMI.2005.159
50. Geurts P, Ernst D, Wehenkel L (2006) Extremely randomized trees. Mach Learn 63:3–42. doi:10.1007/s10994-006-6226-1
51. Liaw A, Wiener M (2002) Classification and regression by randomForest. R news 2:18–22
52. Breiman L (2001) Random forests. Mach Learn 45:5–32. doi:10.1023/A:1010933404324
53. Gonzalez RC, Woods RE (2002) Digital image processing. Prentice Hall, Upper Saddle River
54. Taormina R MATLAB_ExtraTrees-File Exchange—MATLAB Central. http://uk.mathworks.com/matlabcentral/fileexchange/47372-rtaormina-matlab-extratrees. Accessed 16 Feb 2016
55. Crum WR, Camara O, Hill DLG (2006) Generalized overlap measures for evaluation and validation in medical image analysis. IEEE Trans Med Imaging 25:1451–1461. doi:10.1109/TMI.2006.880587
56. Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M, Haussler D (2000) Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics 16:906–914. doi:10.1093/bioinformatics/16.10.906