ELSEVIER

Supervised pattern recognition for the prediction of contrast-enhancement appearance in brain tumors from multivariate magnetic resonance imaging and spectroscopy ☆

Michael C. Lee [*], Sarah J. Nelson

Surbeck Laboratory of Advanced Imaging, Department of Radiology, University of California, UCSF Radiology Box 2532, 1700 4th Street, San Francisco, CA 94143-2532, USA

Summary

Objective: The purpose of this study was to develop a pattern classification algorithm for use in predicting the location of new contrast-enhancement in brain tumor patients using data obtained via multivariate magnetic resonance (MR) imaging from a prior scan. We also explore the use of feature selection or weighting in improving the accuracy of the pattern classifier.

Methods and materials: Contrast-enhanced MR images, perfusion images, diffusion images, and proton spectroscopic imaging data were obtained from 26 patients with glioblastoma multiforme brain tumors, divided into a design set and an unseen test set for verification of results. A *k*-NN algorithm was implemented to classify unknown data based on a set of training data with ground truth derived from post-treatment contrast-enhanced images; the quality of the *k*-NN results was evaluated using a leave-one-out cross-validation method. A genetic algorithm was implemented to select optimal features and feature weights for the *k*-NN algorithm. The binary representation of the weights was varied from 1 to 4 bits. Each individual parameter was thresholded as a simple classification technique, and the results compared with the *k*-NN.

Results: The feature selection *k*-NN was able to achieve a sensitivity of 0.78 ± 0.18 and specificity of 0.79 ± 0.06 on the holdout test data using only 7 of the 38 original

features. Similar results were obtained with non-binary weights, but using a larger number of features. Overfitting was also observed in the higher bit representations. The best single-variable classifier, based on a choline-to-NAA abnormality index computed from spectroscopic data, achieved a sensitivity of 0.79 ± 0.20 and specificity of 0.71 ± 0.11. The k-NN results had lower variation across patients than the single-variable classifiers.

Conclusions: We have demonstrated that the optimized k-NN rule could be used for quantitative analysis of multivariate images, and be applied to a specific clinical research question. Selecting features was found to be useful in improving the accuracy of feature weighting algorithms and improving the comprehensibility of the results. We believe that in addition to lending insight into parameter relevance, such algorithms may be useful in aiding radiological interpretation of complex multimodality datasets.

1. Introduction

Magnetic resonance (MR) imaging has wide uses in the diagnosis, characterization, and planning of treatment for brain tumors. One particularly powerful feature of MR is the ability to exploit several different contrast mechanisms, thereby acquiring multiple views of the same tissue within a single examination. Datasets that are composed of multiple images have been variously described as multivariate, multispectral, or multiparametric. The contrast mechanisms used to generate these images may rely upon different physical or physiological properties, and so by combining these images, it is possible to gain insight into the status of brain tumors and surrounding tissues beyond that which is available from any single image. A state-of-the-art brain tumor examination may include conventional T1- and T2-weighted images as well as physiologically motivated images, such as diffusion images, perfusion images, and magnetic resonance spectroscopic imaging (MRSI). Features derived from diffusion images are thought to provide information on cellular density and the nature of local edema [1,2], while perfusion imaging yields information about cerebral vasculature [3–5]. Proton spectroscopy can be used to measure levels of cerebral metabolites that may be indicative of metabolically active tumor, necrosis and regions with compromised energetics [6–8].

In traditional analysis of multivariate images, these images are presented to radiologists who apply a combination of experience with previous images and prior knowledge of the contrast mechanisms of each image in order to formulate an opinion about the tumor and its borders. However, as the number of images increases, so does the possibility that valuable but complex patterns may exist undiscovered in the data. Additionally, while qualitative visual combination of images may be appropriate for structural images, such as conventional T1- or T2-weighted images, it fails to take advantage of the quantitative nature of diffusion, perfusion, and spectroscopic images. These issues have led many researchers to pursue computational methods for combining information from multivariate MR images.

A large amount of research effort has been applied to the automated classification of *in vivo* proton spectroscopy in brain tumor research. Using machine-learning tools such as linear discriminant analysis, artificial neural networks, and support vector machines, researchers have reported computerized classification results that have compared favorably with histopathology based on biopsy [9–12]. However, there has been somewhat less work on combining imaging data with spectroscopy. When this combination has been performed, the accuracy of the classification has been shown to be superior to using images or spectroscopy alone [13–16].

One approach that can further improve the accuracy of pattern classification algorithms is the use of feature selection or weighting. By using only a subset of features, it is possible to increase generalization accuracy, particularly if some features are irrelevant, noisy, or redundant. The search space may be very large and complex; genetic algorithms are a commonly used method for choosing a near-optimal set of features in a potentially very complicated landscape [17,18].

In this study, we use a k-nearest neighbor algorithm to perform a voxel-by-voxel outcome prediction in brain tumor images. Instead of using biopsy pathology as the ground truth, the goal is to classify voxels based on their predicted radiological outcome. The ground truth for this outcome is defined by images acquired at a followup examination. Our initial feature space will include features from spectroscopic, diffusion, and perfusion imaging; we will study the reduction of this space through the use of a genetic algorithm.

We hypothesize that the pre-radiation images contain information that can be used to predict where new regions of contrast-enhancing (CE) lesion will appear after radiotherapy. We further hypothesize that by applying feature selection, we will be able to improve the accuracy of the classification and gain insight into the relevance of various MR measured indices. We will consider the use of non-binary feature weighting, but with the understanding that this may lead to overfitting and degrade the sensitivity and specificity on test cases. The results of the genetically optimized classifiers will be compared with thresholding on individual variables and the non-optimized classifier.

2. Materials and methods

2.1. Image acquisition

Twenty-six patients (19 male, 7 female; median age 54 years, range 26–75) with a histologically confirmed diagnosis of grade IV gliomas received MRI examinations on a 1.5 T GE Signa Echospeed MR imager (GE Healthcare, Waukesha, WI, USA) within 2 weeks following surgery, but before the initiation of radiotherapy. A second MRI was performed within a month of the conclusion of radiotherapy. The mean inter-exam interval was 69 ± 13 days.

The MR examination included a 3D T1-weighted spoiled-gradient recalled echo (SPGR) sequence (TR/TE = 27/6 ms; flip angle = 40°; 1 mm × 1 mm × 1.5 mm nominal resolution), acquired both with and without gadolinium diethylene-triamine-pentaacetate (Gd-DTPA) contrast agent. Chemical shift imaging was performed using point-resolved spectroscopy (PRESS) volume-selection techniques (TR/TE = 1000/144 ms; 12 × 12 × 8 or 16 × 8 × 8 phase encode steps; 10 mm × 10 mm × 10 mm nominal spatial resolution). For the spectroscopy, water suppression was achieved through the use of spectral-spatial spin-echo pulses [19] and outer volume suppression was performed using very selective suppression pulses [20]. Perfusion imaging was performed by dynamic imaging during the injection of a bolus Gd-DTPA contrast. A power injector was used to infuse a 0.1 mmol/kg bolus of contrast agent at a constant rate of 5 mL/s followed by a 20 mL continuous saline flush. A series of 60 T2*-weighted gradient-echo echo-planar images was acquired (TR/TE = 1000/54 ms, flip angle 35°). The acquisition matrix was 128 × 128 with a 26 cm × 26 cm field-of-view (FOV) and a nominal slice thickness of 3–6 mm, depending on the size and position of the tumor. Diffusion tensor imaging (b = 1000 s/mm^2; 1.5 mm × 1.5 mm × 2.2 mm) was acquired using six gradient encoding directions with a spin-echo echo-planar imaging sequence.

2.2. Data processing

2.2.1. Image and spectral processing

Spectroscopic data were processed using previously described methods [6]. After processing, the spectral amplitudes and linewidths of choline (Cho), creatine (Cr), N-acetyl-aspartate (NAA), lactate (Lac), and lipids (Lip) were estimated. Abnormality indices were derived for the relative levels of Cho to NAA, Cr to NAA, and Cho to Cr (CNI, CrNI, and CCrI, respectively) using a robust linear regression algorithm [9,21]. The first-pass perfusion data were modeled as a modified gamma-variate function, as described previously in the literature [22]. This yielded cerebral blood volume (CBV) as well as curve shape features of peak height, time-to-peak (TTP), peak width (FWHM), recirculation factor, and percent recovery to baseline [23]. Estimates of the pure WM and GM values were derived from a partial-volume model based on the segmented T1-weighted images [22]. Maps of the apparent diffusion coefficient (ADC) and anisotropy were derived from the diffusion images using standard methods [24].

All images were registered to the post-contrast T1-weighted image and then resampled to the resolution of the perfusion images through the use of Fourier interpolation. The registration was performed through the maximization of normalized mutual information using a gradient ascent algorithm [25]. Non-rigid alignment was performed to correct image distortion in the perfusion and diffusion EPI sequences, and consisted of the optimization of control point positions for a grid of B-splines [26]. The importance of this correction has been previously quantified for perfusion images [22].

A significant concern in this study was the normalization and scaling of the data before combining across patients. Therefore, data were expressed in several different ways, and the feature selection and weighting processes performed over the full set of redundant data. A complete listing of the features and their normalization methods is given in Table 1. A total of 38 features were computed for every voxel.

2.2.2. Regions of interest

This study was focused on the growth of contrast-enhancing regions, as observed in T1-weighted images. The region of contrast-enhancement was manually contoured on both the pre-therapy and post-therapy Gd-enhanced T1-weighted SPGR images by research staff using commonly accepted clinical standards for defining contrast-enhance-

Table 1 Features and optimal weights

Feature	1-Bit	2-Bit	3-Bit	4-Bit
Spectral height				
CNI	1	3	7	15
CrNI	0	0	0	0
CCrI	0	3	7	14
Cho[a]	0	0	0	6
Cr[a]	1	3	7	15
NAA[a]	0	0	2	2
Cho[b]	0	0	6	0
Cr[b]	0	0	4	0
NAA[b]	0	0	7	0
Lactate[a]	1	3	7	13
Lipid[b]	1	3	6	12
Cho[c]	1	3	0	10
Cr[c]	0	2	0	0
NAA[c]	1	3	3	15
Diffusion				
FA	0	0	0	13
ADC[b]	0	0	0	13
ADC[c]	0	0	0	0
Perfusion				
CBV[b]	0	0	0	11
CBV[d]	0	0	5	0
Peak height[b]	0	0	0	4
Peak height[d]	0	0	0	0
Recirculation[c]	0	0	0	7
Recirculation[b]	0	0	0	0
Peak width[c]	0	0	6	11
Peak width[b]	1	3	0	10
Spectral widths				
Cho linewidth	0	0	0	0
Cr linewidth	0	0	2	0
NAA linewidth	0	0	0	0
Cho linewidth[c]	0	0	2	4
Cr linewidth[c]	0	0	0	0
NAA linewidth[c]	0	0	2	9
Cho linewidth[b]	0	0	0	6
Cr linewidth[b]	0	0	0	0
NAA linewidth[b]	0	3	6	13

[a] Normalized within each patient to standard error of the noise.
[b] Normalized within each patient to the mean value in normal appearing white matter (NAWM).
[c] Standardized to the mean and standard deviation in normal appearing white matter for each patient.
[d] Scaled linearly such that 0 is the estimated value of normal appearing white matter and 1 is normal appearing gray matter for each patient.

ment, performed under the guidance of and reviewed by an experienced neuroradiologist. All voxels that enhanced at the pre-therapy exam were excluded from further consideration. The contrast-enhancing region is thought to have a very different physiology from the non-enhancing region, and so the evolution of such regions over time is expected to be different from non-enhancing regions. It is therefore likely that including both enhancing and non-enhancing tissue would confound the results and make the selected features more difficult to understand. The region of interest in the study was also limited to the region of overlap between all registered images for a given patient. For example, diffusion and perfusion have limited slice ranges compared with the full T1-weighted volume. The geometry of the spectroscopy PRESS box further limits the extent of overlap.

The ground truth for classification was defined as follows. Voxels within the post-therapy enhancement contour were designated as being *positives*, and voxels outside the contours were designated as *negatives*. A positive was therefore a voxel that was originally non-enhancing but became enhancing, while a negative was a voxel that remained non-enhancing for both scan times. All references to true and false positives and negatives in this study will follow this convention. The overall goal of the study was to optimally utilize the pre-therapy feature vector of each voxel to predict whether that voxel was positive or negative in terms of its post-therapy contrast-enhancement. Informally, we were seeking to answer whether there was information in the pre-therapy images that would have suggested that certain regions were going to become contrast enhancing.

2.2.3. Patient groups
Patients were randomly divided into a *design set* of 15 patients and a *verification set* of 11 patients. The design set was used to optimize the classifier through feature selection or feature weighting, as described in the subsequent sections. The verification set was only used after the classifier development was completed. Unless stated otherwise, all statistics are reported for this holdout verification set. We use the terms design and verification here instead of training and test, since we will be describing the feature selection process as using an internal training and test set. In total, the design set consisted of 60,465 negative examples and 1838 positive examples, while the verification set consisted of 55,526 negative examples and 1102 positive examples.

2.3. Classification method

The classifier used in this study was a *k*-nearest neighbor (*k*-NN) algorithm. Consider a test point with an unknown classification and its *n*-dimensional feature vector. In this study, a 38-dimensional feature vector was associated with each voxel, as described in the previous section. The training data for the classifier consists of a set of training points,

each with a known classification and its own 38-dimensional parameter vector. To classify the unknown test point, the distances between the test point and all training points are computed and sorted. In this study, the distance measure was the Euclidean distance:

$$d(X, Y) = \sqrt{\sum_{i=1}^{38} (X_i - Y_i)^2} \qquad (1)$$

where X is the feature vector of the test point and Y is the feature vector of the training point. The k closest training points then 'vote' on the classification of each unknown test point. In the standard k-NN algorithm, the estimated probability P of the test point being positive was given by the fraction of the k closest points that were themselves positive. In this study, a threshold of 0.5 was used (majority voting).

2.4. Training and testing data

Before applying the k-NN algorithm, the size of the training data set was condensed to reduce noise, prevent any single patient from having undue influence, remove the statistical correlations between voxels, and correct the imbalance in the amounts of positive and negative training data. Several methods have been proposed to reduce the training data set (see, for example Ref. [27]), but this data presents a natural method based on separation into patients. We compress the training data such that each of the 15 patients in the design set provides only two training data points, instead of one point per voxel. These two points are the average of the 38 features describing the positive voxels and the negative voxels for that patient. Again, positive voxels are defined as those that converted from non-enhancing to enhancing, while negative voxels remain non-enhancing over the two sets of scans. The 15 patients in the design set thus yielded a training data set of only 60 data points: 30 positive and 30 negative.

During both the leave-one-out validation and final testing on the holdout data, the test data remains the full data set for each patient, in which each voxel is a separate data point. The k-NN classification is therefore applied to every voxel, but the search for nearest neighbors occurs in the much smaller compressed training data set, consisting of only 30 positive and 30 negative points. Clearly, the full 38-dimensional space is very sparsely populated by this reduced training set; among the goals of feature extraction and feature selection are the reduction of this space, thereby increasing the density of points in meaningful regions. We describe this technique in the next section.

2.5. Genetic optimization for feature weighting

The k-NN algorithm is known to be sensitive to parameter scaling and the presence of irrelevant or noisy features. To improve classification accuracy, we apply *feature weighting*, in which the features are scaled so that their numerical range better matches their relevance [17,28]. Feature selection is implicitly included by allowing the weight of a feature to be zero. Feature weights were included in classification by introducing a set of weights w_i to Eq. (1), yielding:

$$d(X, Y) = \sqrt{\sum_{i=1}^{38} (w_i X_i - w_i Y_i)^2}. \qquad (2)$$

In this study, optimal feature weights were found through the use of a genetic algorithm. Genetic algorithms have been shown to be very effective at locating high quality, if not globally optimal, solutions in combinatorial optimization problems or cases where the structure of the objective function is not amenable to deterministic optimization methods. A schematic example of the interplay between the genetic operations and the k-NN algorithm is given in Fig. 1.

The CHC (Cross-generational elitist selection, Heterogeneous recombination, and Cataclysmic mutation) algorithm was used as the particular genetic optimization scheme in this study [29]. A candidate solution was defined by a 38-dimensional vector of weights. Each weight was expressed as an m-bit binary integer using Gray coding [30]. In this study, four values of m were considered: 1, 2, 3, and 4. Pure feature selection is performed with the 1-bit optimization. For 2-, 3-, and 4-bit representations, feature selection was implicit; some features were not considered during classification because they will have a weight of zero. For each candidate solution, a chromosome was defined as the concatenated string of $38m$ binary bits; in the CHC genetic algorithm, the order of weights in the chromosome was irrelevant. Increasing the number of bits used in representing the weights increases the range of weights available for each feature, and also increases the dimensionality of the search space. A population of 50 chromosomes was used, as is typical in the CHC algorithm. The initial population was randomly initialized with each bit having a 50% chance of being active.

At each iteration, offspring are produced through random mating and recombination, without any selective bias. Pairs of parents chromosomes are randomly selected without replacement for mating. Half uniform crossover (HUX) is used for recombina-
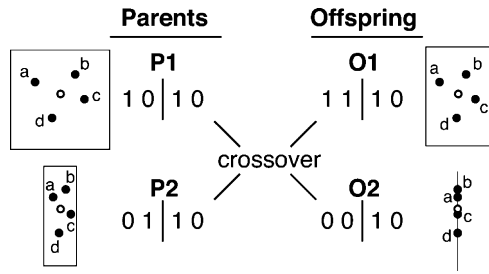
Figure 1 A schematic example of the interaction between the genetic search and the k-NN algorithm. Consider a set of training points, a—d in a two-dimensional feature space, and an unknown test point, denoted by an open circle. For the k-NN algorithm, we are interested in the relative distances between points a—d and the test point. We begin with two parent chromosomes, with two bits representing the weight for each of the two features. The first parent P1 has a chromosome 1010, corresponds to weights of (3,3) in Gray code, while the second parent P2 has bits 0110, corresponding to (1,3). Consequently, P2 has its horizontal axis shrunken by a factor of 3 compared with P1. When ordered from nearest to farthest from the test point, the test points in P1 are d, b, c, and a while in P2, the order is c, a, b, and d. These two parent chromosomes mate using HUX, or half uniform crossover, swapping a single bit (in HUX we swap half the non-matching bits, rounding up as needed; in this schematic example, exactly 1-bit is swapped). O1 is a copy of P1, with the second bit swapped, while O2 is a swapped copy of P2. The order of points in the offspring are now different: c, b, d, and a in O1, while in O2 the order is c, a, d, then b. In this way, the genetic algorithm is able to sample the search space, altering the measured distances and thus altering the outcome of the k-NN algorithm.

tion between the two parents, swapping exactly half of the differing bits. The only provision to allow mating is that the two parents must be a Hamming distance of d apart; parents that do not meet this criteria are not returned to the pool of potential parents. The value of d is initialized to the one fourth the total number of bits ($38m/4$ in this study), and is reduced by one whenever no mating events occur in a generation. At each generation, a total of n offspring are generated, where $0 \leq n \leq 50$. When d is unity but no further offspring ($n = 0$) are produced during mating, the population has converged to a very low-diversity state. At that point, 'cataclysmic' mutation occurs, in which the best individual is retained intact, and all other members of the population are copies of this individual but mutated with a probability of mutation of 0.35 per bit. The value of the Hamming threshold d is then reset to its initial value and offspring produced as before.

After the n offspring are generated, the cost function is evaluated on each new chromosome; this is described in detail in the next section. The members of the previous generation and of all their

offspring (total of $50 + n$ chromosomes) are sorted based on the cost function, and the 50 best chromosomes are retained. These chromosomes are then moved to the new generation, where the reproduction process begins anew.

This algorithm combines a very rapid and aggressive search combined with highly disruptive crossover and mutation events to prevent premature convergence, and has been previously shown to be robust for feature subset selection [31]. An example of the evolution of solutions is shown in Fig. 2. The candidate solutions were allowed to evolve for 1000 generations. The optimization was implemented in C and parallelized using the MPICH library (Argonne National Laboratories), on a grid-enabled cluster of 24 Intel Xeon processors (Intel Corporation, Santa Clara, CA) running at 2.8 GHz and required less than 10 h to complete the full 1000 generations. Note that the genetic algorithm itself remained a serial algorithm, with the hardware parallelization used for accelerating the computation rather than modifying the search, as in a true parallel genetic algorithm.

2.6. Cost function

In order to sort the chromosome by their quality, it is necessary to define a cost function. The goal of the genetic algorithm is to minimize this cost function. To evaluate the cost function for a single chromosome, leave-one-out verification was performed on a patient-by-patient basis. For N design patients, a single patient p was selected for testing, and the remaining N-1 patients were retained as training data. Every voxel in patient p was classified using the k-nearest neighbor algorithm for a pre-determined set of k values. Again, we emphasize that the data compression only affects N-1 patients in the training data, i.e. the data in which the nearest neighbors are found, but not the test data. The k-NN is applied to every voxel in the patient p. The sensitivity, specificity, and area under the ROC curve (Az) were then computed for patient p at a given k value. The sensitivity and specificity were defined in the usual manner, i.e.

$$\text{sens} = \frac{TP}{TP + FN} \tag{3}$$

$$\text{spec} = \frac{TN}{TN + FP} \tag{4}$$

where TP, TN, FP, and FN represent true positive, true negative, false positive, and false negative, respectively. The Az was computed non-parametrically through an analogy to the Wilcoxon statistic and represents the fraction of times the positive example was ranked higher by the k-NN algorithm [32].

(a) generation 1

(b) generation 15
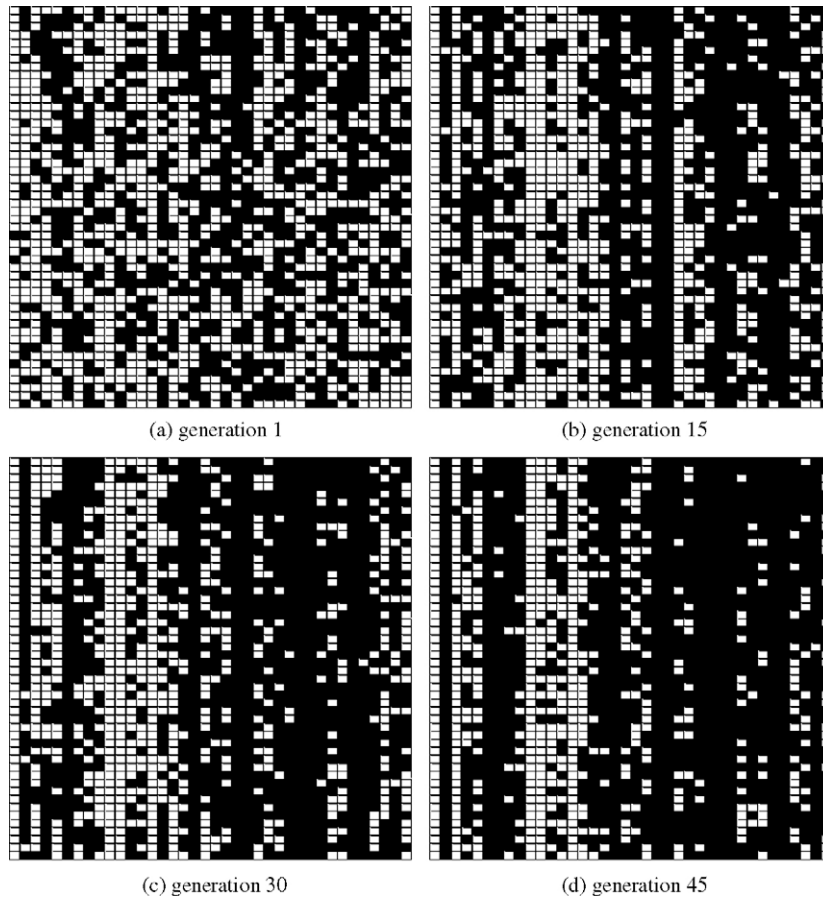
(c) generation 30

(d) generation 45

Figure 2   Schematic of the evolution of solutions in the 1-bit (feature selection) optimization. Each of the 50 rows represents a separate chromosomes, sorted from best (top) to worst (bottom), while each column represents a different parameter. A white square indicates that a specific parameter is active in that chromosome. The identity of these features can be seen in Table 1.

Following this leave-one-out classification of all the voxels from all the design patients, the sensitivity, specificity, and Az were computed separately for each patient. The mean μ and standard deviation σ across patients was computed for these statistics. The cost function for each value of k was then defined as

$$c_{mean}(k) = (1 - \mu_{sens}(k))^2 + (1 - \mu_{spec}(k))^2$$
$$+ 0.5|\mu_{sens}(k) - \mu_{spec}(k)|$$
$$+ 0.01(1 - \mu_{Az}(k))^2 \qquad (5)$$
$$c_{S.D.}(k) = 0.05\sigma_{sens}(k) + 0.05\sigma_{spec}(k) \qquad (6)$$
$$c_{overall}(k) = c_{mean}(k) + c_{S.D.}(k) + 0.002c_{mask}(k) \qquad (7)$$

where c_{mask} is the number of non-zero feature weights included. A separate value of the cost function was computed for each value of k, and the final cost for that chromosome was the minimum across all k values. Note that k is not a parameter in the genetic optimization, but is exhaustively sampled during evaluation of every chromosome; a single chromosome has a single optimal k that is applied to all patients in the leave-one-out testing. Every chromo-

some may have a different k value. It is important to note that a cost function is computed over the whole group of patients, not for any individual patient.

2.7. Single-variable thresholding

An alternative classification scheme would be to simply threshold a single feature, such that any voxels with a value greater than or equal to the threshold would receive a positive prediction, and those below the threshold would receive a negative prediction. Similarly, the classification can be inverted, so that values lower than the threshold receive a positive prediction. This thresholding classifier was first applied to the design data to select an optimal threshold. For each of the 38 features, the minimum and maximum values across all patients was determined. This range was used to create a list of 1000 equally spaced thresholds, which were applied to all design patients. The cost function in Eqs. (5)—(7) were then evaluated for all 1000 threshold values, and the optimum threshold and minimum costs were selected. The 38 features were then sorted according
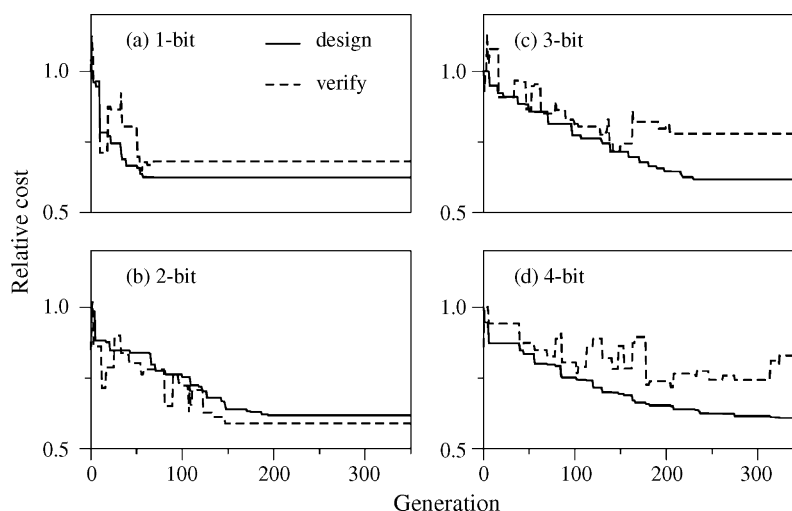
Figure 3 Retrospective view of the relative cost when the best member of each generation is applied to the design and to the verification data. Shown are results for the (a) 1-bit, (b) 2-bit, (c) 3-bit, and (d) 4-bit representations.

to their costs, and the five features with the five lowest costs were selected for further study. For these five features, the optimum thresholds determined above were applied to the verification data, and receiver operating characteristic (ROC) curves and related statistics computed and compared with the k-NN results. Note that for a given threshold, a single cost function is calculated across all patients, and that for each feature, a single optimum threshold is chosen to be applied to all patients.

3. Results

3.1. Genetic optimization of feature weights

3.1.1. Effect of bit representation
The effect of changing the number of bits used to represent weights in the optimization process can be seen in Fig. 3 and in Table 2. To produce the

results in Fig. 3, the state of the genetic algorithm at each generation was recorded. Each of these intermediate weight vectors were then applied to the holdout verification set. Note that these results are only available in retrospect, since during training, the genetic algorithm is allowed to go to convergence on the design data only, with no knowledge of the verification data. We emphasize that all other results presented for the optimized algorithms use the final result only, as the intermediate (and potentially better) solutions cannot be known in advance.

When the best member of each generation is applied to the holdout verification set, differing behaviors are also observed. Note that in Fig. 3, the costs have been normalized to emphasize the similarities and differences in trends between the design and verification sets. As expected, the number of generations required for convergence increases with the number of bits used to represent weights, with the 1-bit optimization converging in fewer than 100 generations. The improvement in

Table 2 Comparison of classifiers

Classifier	n_{dim}	Mean ± standard deviation			
		Sensitivity	Specificity	|Sens − spec|	Az
k-NN (no opt)	38	0.73 ± 0.16	0.75 ± 0.07	0.15 ± 0.14	0.77 ± 0.10
k-NN (1-bit)	7	0.78 ± 0.18	0.79 ± 0.06	0.16 ± 0.10	0.80 ± 0.08
k-NN (2-bit)	11	0.79 ± 0.14	0.78 ± 0.06	0.13 ± 0.13	0.81 ± 0.07
k-NN (3-bit)	17	0.78 ± 0.15	0.78 ± 0.06	0.15 ± 0.12	0.79 ± 0.06
k-NN (4-bit)	23	0.72 ± 0.22	0.80 ± 0.07	0.22 ± 0.14	0.78 ± 0.12
CNI	1	0.79 ± 0.20	0.71 ± 0.11	0.19 ± 0.17	0.86 ± 0.10
CrNI	1	0.66 ± 0.36	0.76 ± 0.11	0.28 ± 0.27	0.78 ± 0.21
NAA	1	0.77 ± 0.25	0.67 ± 0.07	0.24 ± 0.15	0.79 ± 0.12
Lactate	1	0.48 ± 0.34	0.74 ± 0.05	0.32 ± 0.23	0.61 ± 0.31
Lipid	1	0.52 ± 0.30	0.68 ± 0.03	0.23 ± 0.23	0.61 ± 0.23

the verification cost function is not monotonic in any of the four representations, while the CHC algorithm ensures that the design cost function is monotonic. Using pure feature selection (1-bit), the best possible result on the verification data is reached at generation 54, at which point the cost function has improved by 35% below the initial state. The genetic algorithm then continued to evolve new solutions for 14 generations before converging. The final result upon convergence, when applied to the verification data, was a 32% improvement below the initial state, or within 5% of the best solution. Similarly, the verification results for the 2-bit representation closely match the trend for the design results. The verification results for the 3- and 4-bit algorithms also begin by following the same trend as the design results; however, the design and verification cost trends begin to diverge well before convergence is achieved. The 4-bit algorithm achieved its best result on the verification data at iteration 205, at a cost 17% below the initial state, and then continued to evolve for 119 generations. The final result was only 4% below the initial state. This is displayed graphically in Fig. 3(d), where the 4-bit solution improves significantly but then this improvement is lost as the algorithm continues. This suggests a noticeable degree of overfitting in the 3- and 4-bit representations.

The overfitting problem can also be observed by comparing the final results on the design data with the final results on the verification data. The 1-bit representation achieved a sensitivity and specificity of 0.85 ± 0.14 and 0.84 ± 0.10 on the design data and a sensitivity and specificity of 0.78 ± 0.79 and 0.75 ± 0.07 on the verification data. As expected, there is some level of generalization error, reflected in the discrepancy between the design and verification data. The 4-bit representation was able to achieve a sensitivity of 0.86 ± 0.10 and specificity of 0.87 ± 0.06 on the design data. However, on the verification data, it was only able to achieve a sensitivity and specificity of 0.72 ± 0.22 and 0.80 ± 0.07. Thus, despite a slight improvement in the design results, the increased bit representation resulted in a decline in the sensitivity of the classifier on verification data, and only a marginal improvement in specificity. As seen in Table 2, the 2-bit representation yielded the best value for Az and sensitivity, and also showed the least overfitting. Moreover, Fig. 3 shows that there was less overfitting in this 2-bit representation than the 3- or 4-bit representations.

3.1.2. Optimal feature weights
The optimal weights determined by the genetic algorithm are shown in Table 1. Across all 4-bit

representations, the predominant feature types are the spectral peak amplitudes. As seen in Table 2, the five best single-variable threshold classifiers are all spectral features. All features present in the 1-bit feature selection method were also present in one normalized form or another in the 2-, 3-, and 4-bit representations. There was no clear preference for a normalization method, and in the 2-, 3-, and 4-bit optimizations, the same feature was sometimes included twice, using two different normalization methods. After optimization, the k-NN classification required 7, 11, 17, and 23 features for the 1-, 2-, 3-, 4-bit representations. In the following sections, we will focus primarily on the 1-bit feature selection result, as its results were comparable to the higher bit representations while requiring the fewest number of features.

3.2. Comparison of classifiers

3.2.1. Sensitivity and specificity
Means and standard deviations of the sensitivity, specificity, and Az for all the classifiers are given in Table 2. Note the improvement in sensitivity when moving from the original, non-optimized k-NN classifier to the optimized k-NN classifiers. The optimized k-NN classifiers (1-, 2-, and 3-bit) are the only classifiers in which both the mean sensitivity and specificity are greater that 0.75. Additionally, the minimum sensitivity and specificity across patients (worst-case performance) was higher in the feature selection (1-bit) k-NN classifier than any single-variable classifier. This suggests that the feature selection k-NN demonstrated greater consistency across patients; this is also reflected in the low standard deviation across patients. It is also apparent that the optimized k-NN yielded a smaller discrepancy between sensitivity and specificity than any of the single-variable classifiers. This implies that the probability of correctly classifying a negative voxel is similar to the probability of correctly classifying a positive voxel.

An illustrative example of this variation can be observed in Fig. 4. Shown are ROC curves for thresholding on CNI using two patients from the design set, i.e. the set from which the optimal threshold was derived. Both ROC curves exhibit similar and high Az values: 0.91 for patient 1 and 0.85 for patient 2. For illustrative purposes, we choose optimal operating thresholds for each patient separately based on the criteria of minimizing Eq. (5). These optimum thresholds are 0.80 for patient 1 and 1.32 for patient 2; their positions are plotted on the two ROC curves. Clearly, a compromise between these two thresholds must be chosen, and the result will be suboptimal for both patients. In contrast, a common
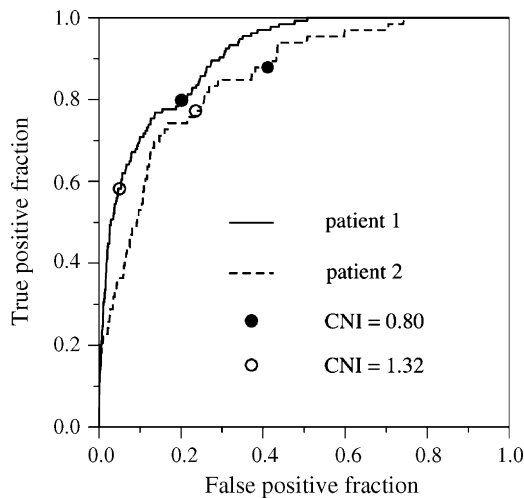
NN and CNI thresholding are comparable; very little of the new contrast-enhancement is outside of either contour. However, the specificity of the k-NN algorithm is higher in both cases, with much less non-enhancing tissue contained within the contours. Note, however, that in both these cases, and indeed most patients, an optimal threshold for CNI could have been chosen which would have been very competitive with the k-NN results. This corresponds to choosing the ideal points on ROC curves, as in Fig. 4. However, as seen in Section 3.2.1, selection of a single optimal threshold is difficult due to inter-patient variation.

Figure 4 Example ROC curves from two design patients when thresholded using CNI. Also shown are two candidate-operating points. Note that each threshold is optimal for one patient (in terms of minimizing the distance to the top left corner) but not for the other patient.

4. Discussion

threshold of 0.5 for the k-NN classifier results in consistent sensitivity and specificity values across patients.

3.2.2. Prediction maps
Examples of the classification results are shown in Fig. 5. In both cases, the sensitivities of the 1-bit k-

In this study, we have demonstrated the use of a genetic algorithm to design a k-nearest neighbor classifier. By reducing the dimensionality of the data or rescaling the feature space, the genetic algorithm is able to improve the performance of the classifier. In this way, it was possible to construct a classifier that performs better than any single-variable and generalizes well to previously unseen data.

It is important to note that the goal of the classifier is not explicitly to detect the presence of tumor. The goal is to predict the appearance of contrast-enhancement; this is a related but distinct
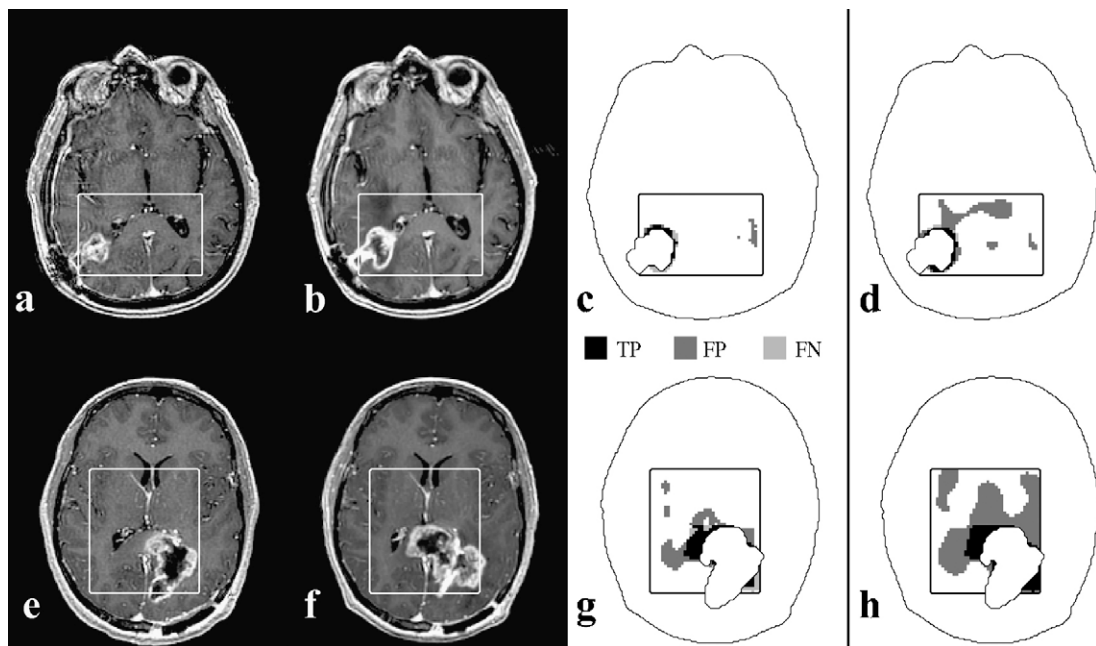

Figure 5 Example classification of a design patient, showing the (a) pre-therapy image and (b) post-therapy image with region of interest indicated; (c) feature-selection k-NN result and (d) thresholding on CNI using the optimal threshold derived from the design patients with contour showing the region of pre-existing contrast-enhancement and resection cavity for which no prediction was made (■, true positive; ■, false positive; ■, false negative). These results are also for a verification patient (d—f).

problem. The growth of contrast-enhancement in this study should also not be confused with physical growth of the bulk tumor. Instead, this study is concerned with conversion of apparently normal regions of tissue into regions of macroscopic contrast-enhancement. The appearance of new post-therapy contrast-enhancement suggests several hypotheses about the pre-therapy state of the tissue. It may suggest microscopic infiltration of tumor cells into an apparently normal region, a hallmark of gliomas. During the inter-exam interval, these cells may grow to such an extent as to cause a breakdown in the blood—brain barrier, resulting in contrast-enhancement. New contrast-enhancement may also occur in regions where blood vessels and tissues are particularly susceptible to radiation. Prediction of new contrast-enhancement may therefore also identify tumor subregions that are being severely damaged by radiotherapy. In that case, the post-therapy contrast-enhancement may not be indicative of tumor growth, but rather, response to therapy. The machine-learning approach adopted in this study is designed only to predict the radiological outcome of individual voxels, but does not seek to specifically decide between these possible causes for the outcome. Nevertheless, by examining the features used in classification, it is possible to gain some insight into both the classification algorithms as well as the underlying physiology.

The best single-variable classifiers were CNI, CrNI, NAA, lactate, and lipid. Tumor cell proliferation is generally associated with elevated choline and reduced NAA and these changes motivated the development of CNI as a marker of tumor [9]. A change in NAA would also be reflected in a change in CrNI. Lactate is a marker of anaerobic metabolism, which may be related to tumor hypoxia. The presence of lactate therefore suggests that a tumor has outgrown its blood supply. The presence of lipids suggests cell death, and is often observed in regions of contrast-enhancement. Abnormalities in any of these five variables are well known to be indicative of tumor presence. These features therefore suggest that early regions of growth in contrast-enhancement correspond to tumor that is already present in the pre-therapy examination. If machine-learning-based classification methods are able to combine these features to predict tumor growth on a local, voxel-by-voxel basis, such methods may prove useful in surgical or radiotherapy planning.

One problem observed with the single-variable classifiers is the variation in optimum thresholds across patients. For example, the high Az of CNI suggests that by properly adjusting the thresholds it is possible to very accurately delineate regions at risk of contrast-enhancement. However, this alone is not sufficient, as the correct threshold cannot be definitively known *a priori*, but must be determined from a set of design or training data. The CNI algorithm was developed to provide a quantitative and objective measure of spectroscopic abnormality, and has been shown to have a sensitivity of 90% and specificity of 86% in predicting tumor based upon correlation with image guided biopsy [21]. This study has shown that it is possible to add a further level of cross-patient robustness and numerical consistency by adding in additional features through machine-learning techniques. This is particularly true for the metabolic peaks heights and areas, where normalization to a well-defined standard is a difficult task currently under study.

While instance-based learning algorithms, such as the k-NN, do not attempt to induce generalized rules from the training data, it is still possible to evaluate the validity of the feature selection and feature weighting by comparing the selected features with prior knowledge of the problem domain. In this case, several spectral features were consistently chosen, regardless of bitwise representations. These include the derived index CNI, as well as the five metabolites generally visible on long-echo proton spectroscopy: choline, creatine, NAA, lactate, and lipids. As describe earlier, it has been well reported in the literature that spectroscopy provides a highly accurate means of identifying the presence of tumor. Contrast-enhancement at the post-therapy examination is commonly assumed to represent clinical progression, and has been found to be most likely to occur in regions that were metabolically abnormal in the pre-treatment examination. Also identified by feature selection was the width of the peak in the perfusion data. Peak recovery to baseline was also included with non-binary feature weighting. Both these features are believed to be related to vessel leakiness and tortuosity. Small-scale leakage of contrast agent below the level that is detected in the post-gadolinium T1-weighted images may correspond to macroscopic contrast-enhancement at a later time point. Neovasculature is known to be both tortuous and leaky in regions of tumor and it is reasonable that increased peak width and decreased percent recovery are observed in the pre-therapy examination predicts subsequent larger scale leakage of the contrast agent.

In addition to clinical research issues related to contrast-enhancement in gliomas, this study is also largely concerned with topics in machine learning for computer-aided diagnosis. The problem of overfitting can be very severe when aggressive optimi-

zation strategies are used in feature weighting. The approach we have adopted is patterned after the idea proposed by Kohavi et al., in which the problem of overfitting is mitigated by reducing the representational power of the optimization algorithm [33]. Kohavi et al. found that in a nearest neighbor algorithm, allowing features weights to take on more than two different non-zero values failed to significantly improve accuracy on a test set. In this study, using different bitwise representations, we allowed 1, 3, 7, and 15 non-zero weights, and found little difference between the 1-, 2-, and 3-bit representations. Given this, we elected to perform most comparisons using the 1-bit representation, which required only 7 features. There was further evidence of overfitting in the 3- and 4-bit representations, as seen by retrospectively applying the intermediate weights to the test data. In these cases, a second strategy such as early stopping may be useful in improving accuracy. Such an approach has previously been applied for *k*-NN algorithms by Loughrey and Cunningham, and has been widely used in backpropagation training of artificial neural networks [34]. This study further emphasizes the well known importance of taking measures to prevent overfitting the data, and also the use of a completely separate set of data for testing and verification. Use of leave-one-out cross-validation only within the feature selection or feature weighting phase will still result in a biased outcome, and does not guarantee generalizability [35].

A major assumption of this study is that it was possible to create a voxel-by-voxel correspondence between the pre-therapy and post-therapy images. As discussed earlier, this study is concerned with conversion of tissue, and not bulk growth of the solid tumor. It is therefore important that bulk tumor growth and corresponding tissue shift during the inter-exam interval is minimal. The images observed in Fig. 5 are typical of the images used in this study. Some tissue shift is inevitable, but in this study, we did not perceive this tissue shift to be a significant problem. Other studies have specifically modeled the growth and infiltration of brain tumors in a mathematical sense [36]. These models have more recently been combined with prior knowledge available from diffusion tensor imaging studies [37]. The results of the present study suggests that information from other MR images may also be useful in modeling tumor growth.

One issue of concern in machine-learning studies is the independence of samples in the training data. While other researchers have found success in treating each voxel in an image as a separate instance in training [38], others have used prior knowledge of the acquisition procedure to reduce correlations between samples by attempting to choose only the most uncorrelated voxels [14]. The approach used in this study was to condense information from the entire ROI in a single patient, so that correlations between samples in a single patient are largely removed, and each patient contributes equally to the training data. This has the disadvantage of eliminating large amounts of data and any important information that may have been in the spatial distribution of this data. Alternative methods of reducing the training data have been proposed, and may be useful in selecting out the most useful, if not the most uncorrelated, data [27,39,40].

This study complements previous work in using combined MRI and spectroscopy features in classifying brain tumors [13−16]. We have chosen to apply similar methods of supervised pattern recognition and computer-aided diagnosis to study a very specific question in glioma imaging, thereby relating machine-learning techniques with image segmentation and clinical prediction. The results of these studies all suggest the potential for computerized methods in understanding which combination of multivariate features may be relevant in addressing specific imaging problems.

5. Conclusions

Techniques of pattern recognition and evolutionary computing have previously been used in computer-aided diagnosis applications, such as in mammography and stroke imaging. The primary contribution of this work was to apply these methods to a new problem in glioma imaging: the identification of regions at risk for developing contrast-enhancement. We have applied techniques of feature-selection and weighting developed in the machine learning and artificial intelligence fields with current techniques in MR imaging to select relevant features and use them to perform this predictive analysis. Our key finding from the tumor biology perspective is that the features in the pre-treatment scan that appear to be most relevant in predicting formation of new contrast-enhancement in the subsequent examination are metabolic features of tumor in non-enhancing tissue. This implies that new contrast-enhancement is not necessarily new tumor. We believe that in addition to aiding radiological interpretation of complex multimodality datasets, such methods can lend insight into parameter relevance and therefore be of use in both the machine learning and clinical research communities.

Acknowledgements

References

[1] Tien RD, Felsberg GJ, Friedman H, Brown M, MacFall J. MR imaging of high-grade cerebral gliomas: value of diffusion-weighted echoplanar pulse sequences. AJR Am J Roentgenol 1994;162:671—7.

[2] Krabbe K, Gideon P, Wagn P, Hansen U, Thomsen C, Madsen F. MR diffusion imaging of human intracranial tumours. Neuroradiology 1997;39:483—9.

[3] Aronen HJ, Gazit IE, Louis DN, Buchbinder BR, Pardo FS, Weisskoff RM, et al. Cerebral blood volume maps of gliomas: comparison with tumor grade and histologic findings. Radiology 1994;191:41—51.

[4] Maeda M, Itoh S, Kimura H, Iwasaki T, Hayashi N, Yamamoto K, et al. Tumor vascularity in the brain: evaluation with dynamic susceptibility-contrast MR imaging. Radiology 1993;189:233—8.

[5] Knopp EA, Cha S, Johnson G, Mazumdar A, Golfinos JG, Zagzag D, et al. Glial neoplasms: dynamic contrast-enhanced T2*-weighted MR imaging. Radiology 1999;211:791—8.

[6] Nelson SJ. Analysis of volume MRI and MR spectroscopic imaging data for the evaluation of patients with brain tumors. Magn Reson Med 2001;46:228—39.

[7] Ott D, Hennig J, Ernst T. Human brain tumors: assessment with in vivo proton MR spectroscopy. Radiology 1993;186:745—52.

[8] Negendank WG, Sauter R, Brown TR, Evelhoch JL, Falini A, Gotsis ED, et al. Proton magnetic resonance spectroscopy in patients with glial tumors: a multicenter study. J Neurosurg 1996;84:449—58.

[9] McKnight TR, Noworolski SM, Vigneron DB, Nelson SJ. An automated technique for the quantitative assessment of 3D-MRSI data from patients with glioma. J Magn Reson Imaging 2001;13:167—77.

[10] Tate AR, Griffiths JR, Martinez-Perez I, Moreno A, Barba I, Cabanas ME, et al. Towards a method for automated classification of 1H MRS spectra from brain tumours. NMR Biomed 1998;11:177—91.

[11] Usenius JP, Tuohimetsa S, Vainio P, Ala-Korpela M, Hiltunen Y, Kauppinen RA. Automated classification of human brain tumours by neural network analysis using in vivo H-1 magnetic resonance spectroscopic metabolite phenotypes. Neuroreport 1996;7:1597—600.

[12] Preul MC, Caramanos Z, Leblanc R, Villemure JG, Arnold DL. Using pattern analysis of in vivo proton MRSI data to improve the diagnosis and surgical management of patients with brain tumors. NMR Biomed 1998;11:192—200.

[13] Devos A, Simonetti AW, van der Graaf M, Lukas L, Suykens JA, Vanhamme L, et al. The use of multivariate MR imaging intensities versus metabolic data from MR spectroscopic imaging for brain tumour classification. J Magn Reson 2005;173:218—28.

[14] Simonetti AW, Melssen WJ, Szabo de Edelenyi F, van Asten JJ, Heerschap A, Buydens LM. Combination of feature-reduced MR spectroscopic and MR imaging data for improved brain tumor classification. NMR Biomed 2005;18:34—43.

[15] Simonetti AW, Melssen WJ, van der Graaf M, Postma GJ, Heerschap A, Buydens LM. A chemometric approach for brain tumor classification using magnetic resonance imaging and spectroscopy. Anal Chem 2003;75:5352—61.

[16] Szabo de Edelenyi F, Rubin C, Esteve F, Grand S, Decorps M, Lefournier V, et al. A new approach for analyzing proton magnetic resonance spectroscopic images of brain tumors: nosologic images. Nat Med 2000;6:1287—9.

[17] Siedlecki W, Sklansky J. A note on genetic algorithms for large-scale feature-selection. Pattern Recognit Lett 1989;10:335—47.

[18] Sahiner B, Chan HP, Wei DT, Petrick N, Helvie MA, Adler DD, et al. Image feature selection by a genetic algorithm: application to classification of mass and normal breast tissue. Med Phys 1996;23:1671—84.

[19] Star-Lack J, Nelson SJ, Kurhanewicz J, Huang LR, Vigneron DB. Improved water and lipid suppression for 3D PRESS CSI using RF band selective inversion with gradient dephasing (BASING). Magn Reson Med 1997;38:311—21.

[20] Tran TK, Vigneron DB, Sailasuta N, Tropp J, Le Roux P, Kurhanewicz J, et al. Very selective suppression pulses for clinical MRSI studies of brain and prostate cancer. Magn Reson Med 2000;43:23—33.

[21] McKnight TR, von dem Bussche MH, Vigneron DB, Lu Y, Berger MS, McDermott MW, et al. Histopathological validation of a three-dimensional magnetic resonance spectroscopy index as a predictor of tumor presence. J Neurosurg 2002;97:794—802.

[22] Lee MC, Cha S, Chang SM, Nelson SJ. Partial-volume model for determining white matter and gray matter cerebral blood volume for analysis of gliomas. J Magn Reson Imaging 2006;23:257—66.

[23] Lupo JM, Cha S, Chang SM, Nelson SJ. Dynamic susceptibility-weighted perfusion imaging of high-grade gliomas: characterization of spatial heterogeneity. AJNR Am J Neuroradiol 2006;26:1446—54.

[24] Basser PJ, Pierpaoli C. Microstructural and physiological features of tissues elucidated by quantitative-diffusion-tensor MRI. J Magn Reson Ser B 1996;111:209—19.

[25] Studholme C, Hill DL, Hawkes DJ. An overlap invariant entropy measure of 3D medical image alignment. Pattern Recognit 1999;32:71—86.

[26] Rueckert D, Sonoda LI, Hayes C, Hill DL, Leach MO, Hawkes DJ. Nonrigid registration using free-form deformations: application to breast MR images. IEEE Trans Med Imaging 1999;18:712—21.

[27] Wilson DR, Martinez TR. Reduction techniques for instance-based learning algorithms. Mach Learn 2000;38:257—86.

[28] Wettschereck D, Aha DW, Mohri T. A review and empirical evaluation of feature weighting methods for a class of lazy learning algorithms. Artif Intell Rev 1997;11:273—314.

[29] Eshelman L. The CHC adaptive search algorithm: how to have a safe search when engaging in nontraditional genetic recombination. In: Rawlins GJE, editor. Foundations of Genetic Algorithms I. San Mateo, CA, USA: Morgan Kaufmann; 1991. p. 265—83.

[30] Mathias KE, Whitley LD. Transforming the search space with Gray coding. In: Schaffer JD, editor. Proceedings of the IEEE Conference on Evolutionary Computation, IEEE World Congress on Computational Intelligence. New York: IEEE Press; 1994. p. 513—8.

[31] Guerra-Salcedo C, Whitley LD. Genetic search for feature subset selection: a comparison between CHC and GENESIS. In: Koza JR, Banzhaf W, Chellapilla K, Deb K, Dorigo M, Fogel DB, et al., editors. Genetic Programming 1998: Proceedings

of the Third Annual Conference. San Mateo, CA: Morgan Kaufmann; 1998. p. 504—9.

[32] Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982;143:29—36.

[33] Kohavi R, Langley P, Yun Y. The utility of feature-weighting in nearest-neighbor algorithms. In: Van Someren M, Widermer G, editors. Poster: Ninth European Conference on Machine Learning, 1997.

[34] Loughrey J, Cunningham P. Overfitting in wrapper-based feature subset selection: the harder you try the worse it gets. In: Bramer M, Coenen F, Allen T, editors. Research and Development in Intelligent Systems XXI. London, UK: Springer-Verlag; 2005. p. 33—43.

[35] Li Q, Doi K. Reduction of bias and variance for evaluation of computer-aided diagnostic schemes. Med Phys 2006;33 :868—75.

[36] Swanson KR, Bridge C, Murray JD, Alvord Jr EC. Virtual and real brain tumors: using mathematical modeling to quantify glioma growth and invasion. J Neurol Sci 2003;216:1—10.

[37] Jbabdi S, Mandonnet E, Duffau H, Capelle L, Swanson KR, Pelegrini-Issac M, et al. Simulation of anisotropic growth of low-grade gliomas using diffusion tensor imaging. Magn Reson Med 2005;54:616—24.

[38] Gottrup C, Thomsen K, Locht P, Wu O, Sorensen AG, Koroshetz WJ, et al. Applying instance-based techniques to prediction of final outcome in acute stroke. Artif Intell Med 2005;33:223—36.

[39] Tomek I. Experiment with edited nearest-neighbor rule. IEEE Trans Syst Man Cybern 1976;6:448—52.

[40] Kuncheva LI. Editing for the K-nearest neighbors rule by a genetic algorithm. Pattern Recognit Lett 1995;16: 809—14.