Clinical Science (2017) **131** 1465–1481
DOI: 10.1042/CS20170051

PORTLAND PRESS

Research Article

Reliability of an automatic classifier for brain enlarged perivascular spaces burden and comparison with human performance

Víctor González-Castro[1], María del C. Valdés Hernández[1], Francesca M. Chappell[1], Paul A. Armitage[2], Stephen Makin[1] and Joanna M. Wardlaw[1]

[1]Department of Neuroimaging Sciences, Centre for Clinical Brain Sciences, University of Edinburgh, 49 Little France Crescent, Edinburgh EH16 4SB, U.K.; [2]Department of Infection, Immunity and Cardiovascular Disease, University of Sheffield, Royal Hallamshire Hospital, Sheffield S10 2JF, United Kingdom

Correspondence: Víctor González-Castro (victor.gonzalez@ed.ac.uk) or María del C. Valdés Hernández (M.Valdes-Hernan@ed.ac.uk)

OPEN ACCESS

In the brain, enlarged perivascular spaces (PVS) relate to cerebral small vessel disease (SVD), poor cognition, inflammation and hypertension. We propose a fully automatic scheme that uses a support vector machine (SVM) to classify the burden of PVS in the basal ganglia (BG) region as low or high. We assess the performance of three different types of descriptors extracted from the BG region in T2-weighted MRI images: (i) statistics obtained from Wavelet transform's coefficients, (ii) local binary patterns and (iii) bag of visual words (BoW) based descriptors characterizing local keypoints obtained from a dense grid with the scale-invariant feature transform (SIFT) characteristics. When the latter were used, the SVM classifier achieved the best accuracy (81.16%). The output from the classifier using the BoW descriptors was compared with visual ratings done by an experienced neuroradiologist (Observer 1) and by a trained image analyst (Observer 2). The agreement and cross-correlation between the classifier and Observer 2 ($\kappa = 0.67$ (0.58–0.76)) were slightly higher than between the classifier and Observer 1 ($\kappa = 0.62$ (0.53–0.72)) and comparable between both the observers ($\kappa = 0.68$ (0.61–0.75)). Finally, three logistic regression models using clinical variables as independent variable and each of the PVS ratings as dependent variable were built to assess how clinically meaningful were the predictions of the classifier. The goodness-of-fit of the model for the classifier was good (area under the curve (AUC) values: 0.93 (model 1), 0.90 (model 2) and 0.92 (model 3)) and slightly better (i.e. AUC values: 0.02 units higher) than that of the model for Observer 2. These results suggest that, although it can be improved, an automatic classifier to assess PVS burden from brain MRI can provide clinically meaningful results close to those from a trained observer.

Introduction

Perivascular spaces (PVS), also known as Virchow–Robin spaces, are fluid-containing spaces that surround the walls of small vessels and capillaries in the brain as they go through the grey or white matter. PVS are microscopic, filled with interstitial fluid and act as drainage pathways for fluid and metabolic wastes from the brain and, when enlarged, are visible in structural MRI sequences [1]. High number of enlarged PVS has been reported to be associated with worse cognition [2], active inflammation in multiple sclerosis plaques [3] or aging [4], depression at older ages [5], Parkinson's disease [6] and cerebral small vessel disease (SVD) [7].

The term SVD refers to a group of pathological processes that affect the small arteries, veins and capillaries of the brain [8]. It is the most common cause of vascular dementia and causes approximately a fifth of the strokes worldwide [9], proven to have significant and strong associations with

Received: 18 January 2017
Revised: 25 April 2017
Accepted: 02 May 2017

Accepted Manuscript Online: 03 May 2017
Version of Record published: 28 June 2017

© 2017 The Author(s). This is an open access article published by Portland Press Limited on behalf of the Biochemical Society and distributed under the Creative Commons Attribution Licence 4.0 (CC BY).

Clinical Science (2017) **131** 1465–1481
DOI: 10.1042/CS20170051

vascular risk factors [10]. A moderate-to-severe burden of PVS in the basal ganglia (BG) is one of the markers of SVD [9], along with lacunes, cerebral microbleeds and white matter hyperintensities (WMH).

PVS can be better identified on T2-weighted (T2w) MRI, where they appear as linear or dot-like structures with intensities close to those of the cerebrospinal fluid (CSF) and less than 3 mm diameter in cross-section [9]. Therefore, PVS can be potentially quantified. Visual counting and/or manual delineation of PVS can be time-consuming, and the development of computational methods to assess them is challenging, partly due to inconsistencies within the literature regarding PVS diameter and overlap in shape, intensity, location and size with these of lacunes [11]. Recently, [12] and [13] presented computational methods to obtain quantificative measurements of PVS and validated the usefulness of their procedures in clinical research, but both approaches are semi-automatic being, therefore, prone to interobserver variations and could be time-consuming. [14] also proposed a method for quantifying PVS using high-resolution 7T MRI scanners but the use of such field strengths, although providing good spatial resolution and signal-to-noise ratio, has limited clinical use. [15] use a Frangi filter whose parameters are optimized by means of the ordered logit model to enhance the differentiation between PVS and the background, but is unsuitable for images with anisotropic voxels commonly used in clinical settings (e.g. voxel sizes of $0.5 \times 0.5 \times 6$ mm) and still requires the (visual) rating of the PVS.

As an alternative to quantificative measurements, several visual rating scales that provide a qualitative assessment of the burden of PVS have been proposed in recent years. Potter et al. [16] reviewed the ambiguities of these scales and combined their strengths to develop the one that proved to be robust. However, as with any visual recognition process, it is subject to observer bias. Making the PVS rating automatic (e.g. replicating the visual rating scale using image processing and pattern recognition) could potentially overcome these and also the drawbacks that the current methods of PVS segmentation have.

Computer vision and pattern recognition have already been successfully applied to computer-aided diagnosis using MRI [17,18] and for segmentation of brain structures or lesions [19-21]. It has also been used to assess markers of SVD qualitatively. For example, Chen et al. [22] proposed a framework based on multiple instance learning to distinguish between absent/mild compared with moderate/severe SVD in computed tomography (CT) scans.

However, to the best of our knowledge, only two papers have addressed the task of assessing automatically the PVS rating in brain MRI using computer vision and pattern recognition techniques [23,24]. They explored the use of different descriptors for this task, but did not analyse agreement with a human observer other than the one that provided the ground truth ratings or whether the predictors of the classification were clinically meaningful. Moreover, each of these two works evaluated different descriptors to characterize the brain region selected for classifying PVS burden and report similar levels of accuracy for the preferred schemes, albeit having validated the schemes differently (i.e. [23] uses cross-validation and [24] compares results on randomly divided train and test subsets). An overall evaluation of the schemes proposed so far for classifying the burden of PVS from brain MRI is lacking.

In the present paper, we build upon the work presented in [23,24], comparing the performance of the descriptors proposed by both studies for automatically classifying the burden of PVS using a support vector machine (SVM) [25]. We focus on the PVS in the BG, since moderate-to-severe PVS in this region (i.e. ratings 2–4) is a marker of cerebral SVD. We evaluate three different types of descriptors: (i) statistics obtained from Wavelet transform's coefficients [26], (ii) local binary patterns [27] and (iii) bag of visual words (BoW) based descriptors, using keypoints obtained from a dense grid characterized with the scale-invariant feature transform (SIFT) characteristics. Moreover, we validate the results by comparing the predictions made by the automatic method (i.e. the classifier using the descriptors that achieve the best performance) with the ratings from the two observers. Finally, we also investigate the applicability of this classifier to clinical studies, to assess if its outcome is clinically meaningful. The paper is organized as follows: in 'Materials and methods' section, the dataset and proposed methods are explained. 'Results' section introduces the experimental setup and the results of the experiments, which are discussed in section titled 'Discussion'. Finally, the conclusions and possible future lines of work are presented in 'Conclusion and future work' section.

Materials and methods

Subjects and MRI protocol

We used data from 264 patients who gave written informed consents to participate in a study of lacunar stroke mechanisms [28].

The study that provided data for this manuscript [28] included patients with lacunar stroke and/or minor cortical strokes which were clinically evident, and did not consider diabetes, hypertension and other vascular risk factors as criteria for exclusion. However, it excluded patients with other nonvascular neurological disorders, major medical conditions including renal failure, contraindications to MRI, unable to give consent, and those who had haemorrhagic

© 2017 The Author(s). This is an open access article published by Portland Press Limited on behalf of the Biochemical Society and distributed under the Creative Commons Attribution Licence 4.0 (CC BY).

Clinical Science (2017) **131** 1465–1481
DOI: 10.1042/CS20170051

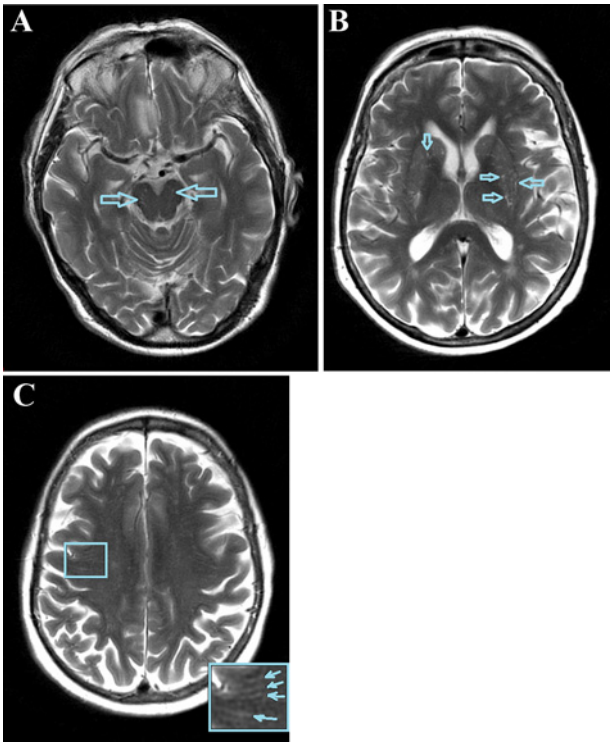

Figure 1. Anatomical regions and appearance of PVS.

Example of the anatomical regions where the PVS (arrowed) are rated: midbrain, BG and CS (from (**A**) to (**C**) respectively). Note the longitudinal appearance in the CS in axial view ((**C**) inset).

stroke or whose symptoms resolved within 24 h (i.e. transient ischaemic attack). It was approved by the Lothian Ethics of Medical Research Committee (REC 09/81101/54) and the NHS Lothian R&D Office (2009/W/NEU/14) and was conducted according to the principles expressed in the Declaration of Helsinki.

Brain MRI was conducted at baseline (i.e. there was a maximum of 8 days between the stroke and the scan) on a 1.5 tesla GE Signa LX clinical scanner (General Electric, Milwaukee, WI), equipped with a self-shielding gradient set and manufacturer supplied eight-channel phased array heal coil. For our analyses, we used the T2w images, acquired with TE 147 ms, TR 9002 ms, field of view 240 × 240 mm, acquisition matrix 256 × 256, slice thickness: 5 mm, 1 mm interslice gap and voxel size 0.469 × 0.469 × 6 mm. The reconstructed image size (in voxels) is 512 × 512 × 28. For tissue segmentation, diffusion-weighted and structural T1-weighted (T1w), T2w and gradient echo, acquired as specified in [28] were also used.

PVS visual rating scale

The visual rating scale proposed by Potter et al. [16] was used for assessing the burden of PVS in the sample. It rates the PVS separately in three major anatomical brain regions, i.e. midbrain, BG and centrum semiovale (CS) shown in Figure 1–using T2w MRI. The rating is done separately for left and right hemispheres, but a combined score that represents the average of the PVS burden is given.

In each of these anatomical regions, the rating can be 0 (no PVS), 1 (mild; 1–10 PVS), 2 (moderate; 11–20 PVS), 3 (frequent; 21–40 PVS) or 4 (severe; >40 PVS)[1].

All visual ratings were made by two observers: a neuroradiologist (Observer 1) with more than 25 years of experience who participated in the development of the scale and a trained image analyst (Observer 2). The ratings were done blind to all clinical information, each other's results and any intermediate or final computational results.

In the present paper, we focus only on the PVS in the BG, since moderate to severe PVS in this region (i.e. ratings 2–4) is a marker of cerebral SVD, which has been associated with cognitive decline [10], vascular dementia and stroke [1]. An example of each of the ratings for the BG is shown in Figure 2. We dichotomize the BG PVS scores into two

[1]http://www.sbirc.ed.ac.uk/documents/epvs-rating-scale-user-guide.pdf.

© 2017 The Author(s). This is an open access article published by Portland Press Limited on behalf of the Biochemical Society and distributed under the Creative Commons Attribution Licence 4.0 (CC BY).

Clinical Science (2017) **131** 1465–1481
DOI: 10.1042/CS20170051

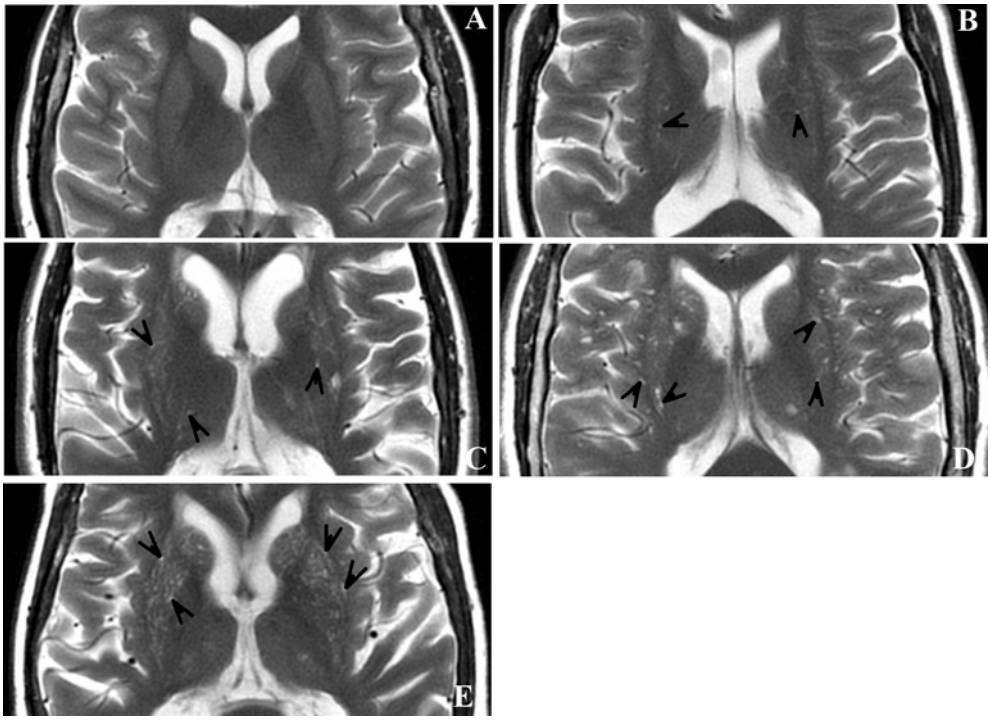

Figure 2. Basal ganglia PVS ratings using Potter et al. [16] scale.

Example for the PVS ratings in the BG, from 0 (none) to 4 (many) ((**A**–**E**) respectively) with black arrowheads pointing to some of the PVS.

classes as per [1], scores 0–1 (i.e. none or mild PVS burden) and scores 2–4 (i.e. moderate to severe), to be our classes 0 and 1 respectively.

Image preprocessing

The guidelines for the visual rating of PVS according to this scale states that the rating should be done on the slice with the highest number of PVS, so as to minimize inconsistencies and intra-/interobserver variations due to interslice variations in PVS visibility, varying number of PVS on different slices and double counting of linear PVS [16]. In case of the BG region, this slice should be chosen among the slices with at least one characteristic BG structure, as indicated by [12]. A pipeline to extract the BG region and find the axial slice (from the BG) with the highest number of PVS for each subject, was developed.

The first step of this pipeline is to automatically segment the intracranial volume and CSF on the T1w images. This was achieved using optiBET [29] and FMRIB software library (FSL)-FAST [30] respectively. The second step is to, also automatically, extract all subcortical structures, which was achieved using other tools from the same FSL as is described in [28]. Thereafter, from the slices that contained BG structures, we selected those in which the total area of these structures was more than 5% of the intracranial area defined on the slice.

On each of the BG slices initially selected, a polygon enclosing the BG, internal and external capsules and thalami was automatically drawn by joining anatomical points in the insular cortex, the closest points to them in the lateral ventricles (frontal and occipital horns) and the intercept of the genu of the corpus callosum with the septum; and subtracting from it the region occupied by the CSF. These steps are illustrated in Figure 3.

From this subset of slices, the slice where our classifier operated was selected after applying contrast-limited adaptive histogram equalization (CLAHE) [31] to the polygonal regions, thresholding them to 0.43 times the maximum intensity level [11,12] (Figure 3D), and counting the number of thresholded hyperintense regions on each candidate slice with area between 3 and 15 times the in-plane voxel dimensions [12]. Although this procedure overestimates the number of PVS in the presence of other features of SVD markers (e.g. small lesions and lacunes) [11], it provides a good estimate of the number of PVS on each candidate axial slice, so as to select the one with more PVS.

© 2017 The Author(s). This is an open access article published by Portland Press Limited on behalf of the Biochemical Society and distributed under the Creative Commons Attribution Licence 4.0 (CC BY).

Clinical Science (2017) **131** 1465–1481
DOI: 10.1042/CS20170051

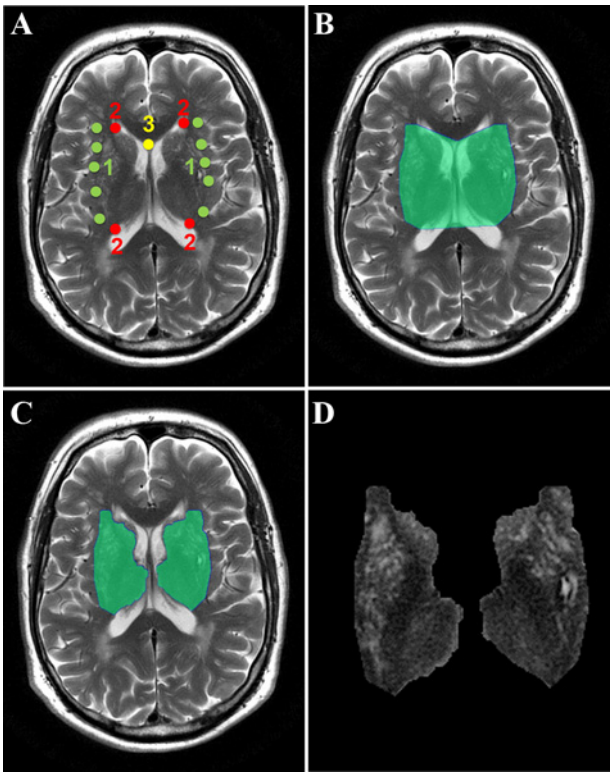

Figure 3. Graphical representation of the steps for delineation of the basal ganglia region

Steps of the BG segmentation: (**A**) Detection of the vertices in the insular cortex (1), lateral ventricles (2) and genu (3); (**B**) creation of the polygon; (**C**) subtraction of the CSF from the BG polygonal region and (**D**) segmented BG region.

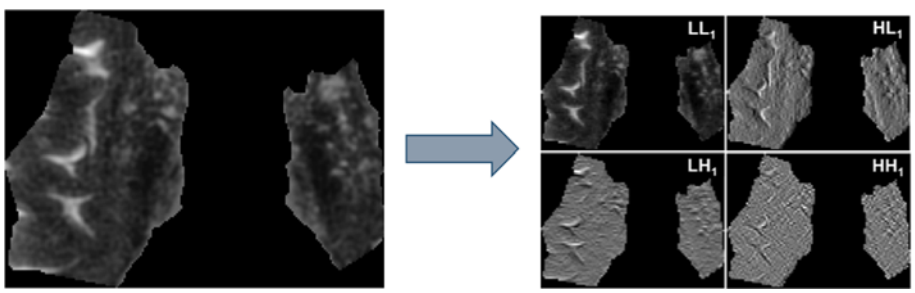

Figure 4. First-level DWT decomposition of the BG region from a T2w image.

Example showing the graphical representation of the values of the four matrices of coefficients (i.e. LL, LH, HL and HH) (right hand side) obtained after applying the discrete wavelet transform (DWT) to the region shown at the left hand side.

Descriptors

Descriptors based on the Wavelet transform

The information represented by spatial frequencies has often been used for texture description with successful results [32]. Due to its frequency domain localization capability, we have applied the discrete Wavelet transform (DWT) to each selected region to characterize their textures. We have used the Haar family of wavelets, which have already been successfully used in other medical image classification applications [26]. The DWT extracts the low and high frequency components of a signal, so that they can be analysed separately.

When the transform is applied to an image, four matrices of coefficients are obtained: namely LL_i, LH_i, HL_i and HH_i where i stands for the level of decomposition, which represent the approximations and details in the vertical, horizontal and diagonal directions respectively. They can be seen in the example that Figure 4 illustrates.

The first level of decomposition is applied on the original image, while the next levels' i are applied to the matrix of approximations of level i-1 as Figure 5 shows.

© 2017 The Author(s). This is an open access article published by Portland Press Limited on behalf of the Biochemical Society and distributed under the Creative Commons Attribution Licence 4.0 (CC BY).

Clinical Science (2017) **131** 1465–1481
DOI: 10.1042/CS20170051

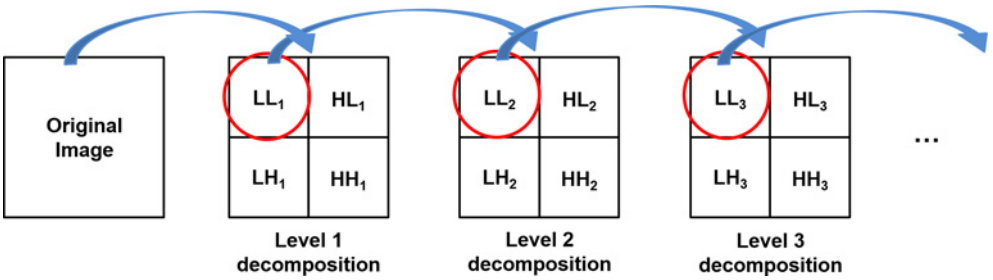

Figure 5. Example of the names of the coefficient matrices after a three-level DWT decomposition.

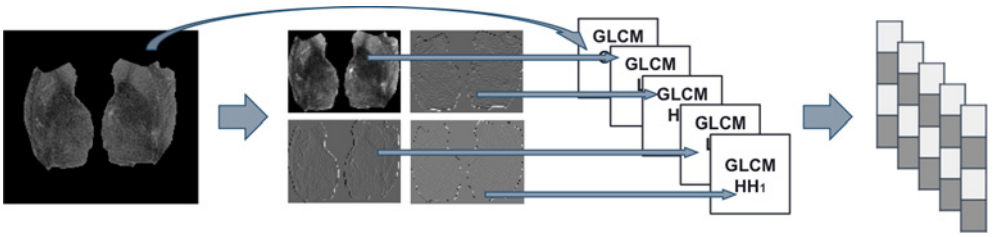

Figure 6. Diagram showing how the wavelet co-occurrence feature (WCF) descriptors are built.

One of the descriptors we used is based on the DWT, and it is built using the mean and S.D. of the histograms of the original image and each one of the matrices of coefficients yielded after three DWT levels (i.e. LL_1, LH_1, HL_1, HH_1, LL_2, LH_2, HL_2, HH_2, LL_3, LH_3, HL_3 and HH_3). Hence, we represent each region by a vector of 26 features. This descriptor is known as Wavelet statistical features (WSF) [26,32].

The other descriptor based on the DWT is built using the features proposed by [33] derived from the grey level co-occurrence matrix (GLCM) of the original image and each of the the coefficient matrices obtained after the first DWT level (i.e. LL_1, LH_1, HL_1 and HH_1). The features extracted from each GLCM are concatenated to form the final descriptor. A diagram depicting this process is shown in Figure 6.

To achieve some invariance to rotation, we averaged the features extracted from GLCMs computed with orientations 0, 45, 90 and 135°. These descriptors are called WCF [26,32]. In this work, we assessed two variants of the WCF descriptors, WCF_4 and WCF_{13}, depending on whether we extracted 4 or 13 features from the GLCMs respectively. WCF_4 is built using the Haralick features *Contrast*, *Correlation*, *Energy* and *Homogeneity*, and WCF_{13} is formed using all features proposed by [33] except the *Maximal Correlation Coefficient*. These two descriptors showed good performance in [34].

Local binary patterns

Local binary patterns (LBP) were introduced by [27]. In the original, version they worked with a 3 × 3 pixel block, but LBPs were later generalized, so size of the neighbourhood and number of sampling points were parameters of the method. Given a pixel c with co-ordinates (x_c,y_c), a pattern code is calculated by comparing it with the value of its P neighbours separated by a distance R, which in our case is 1, as per (eqn 1).

$$LBP_{R,P} = \sum_{p=0}^{P-1} s\left(g_p - g_c\right) 2^p \tag{1}$$

where g_c and g_p are the grey-level values of pixel c and its p-th neighbour, and function $s\,(g_p - g_c)$ is defined as:

$$s\left(g_p - g_c\right) = \begin{cases} 1 \text{ if } g_p - g_c \geqslant 0 \\ 0 \text{ if } g_p - g_c < 0 \end{cases}$$

Finally, the whole image is described by means of a histogram of the LBP values of all pixels, given by (eqn 1). As the position of the *first* neighbour (i.e. $p = 0$) is fixed, it being the pixel on the right hand side of c, the $LBP_{R,P}$ operator is not invariant to rotation. We remove such effect of rotation using the rotation invariant local binary pattern, $LBP^{ri}_{R,P}$, defined in [27].

© 2017 The Author(s). This is an open access article published by Portland Press Limited on behalf of the Biochemical Society and distributed under the Creative Commons Attribution Licence 4.0 (CC BY).

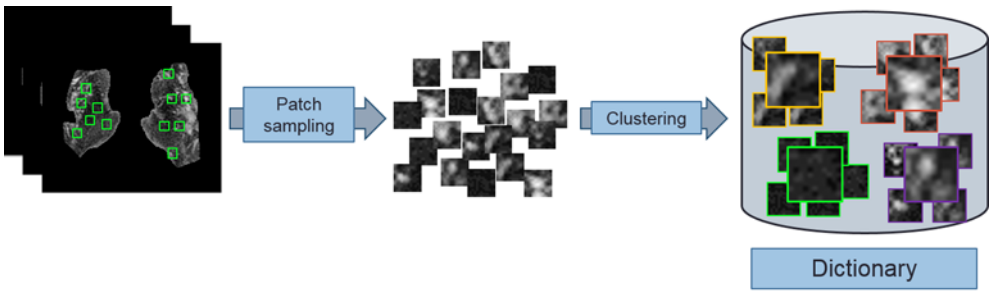

Figure 7. How the dictionary is created.

Small square patches are extracted from each basal ganglia region, which are characterised by means of the descriptors and grouped to form the "dictionary".

As certain local binary patterns represent fundamental properties of texture, providing the vast majority of patterns present in textures [27], while others are known to be less descriptive of the texture, Ojala et al. [27] introduced a measure of 'uniformity' U ($LBP_{R,P}$), which counts the number of spatial transitions (i.e. bitwise 0/1 changes) in a binary pattern $\mathrm{LBP}_{R,P}$ for $\mathrm{LBP}_{R,P}$ less than 2 (i.e. $\mathrm{LBP}^{riu2}_{R,P}$) as expressed in (eqn 2).

$$LBP^{riu2}_{R,P} = \begin{cases} \sum_{p=0}^{P-1} s(g_p - g_c) \text{ if } U(LBP_{R,P}) \leqslant 2 \\ P+1 \qquad\qquad \text{otherwise} \end{cases}, \tag{2}$$

As the BG regions and the PVS are not very big, we tried to keep the texture analysis as local as possible, so in this work, we have used the values $R = 1$ and $P = 8$. The final descriptors we use are the histograms of the accumulated output of $\mathrm{LBP}_{1,8}$, $\mathrm{LBP}^{ri}_{1,8}$ and $\mathrm{LBP}^{riu2}_{1,8}$ operating in each BG region.

BoW

The BoW model [35] represents each image as a function of the frequency of appearance of certain visual elements, called visual words. The set of visual words is called the *dictionary* or *codebook*.

To build the dictionary, a set of keypoints from each image are sampled. Around each keypoint, a small square region (i.e. patch) is extracted and characterized by means of descriptors that retrieve information about the distribution of its pixels intensities. After that, the descriptors of the patches are clustered into K groups, each one having a prototype feature vector which is called *visual word*. This process is depicted in Figure 7.

In this work, we use a dense grid for sampling the keypoints and the k-means clustering method [36] for forming the visual words. The process of creating the dictionary is performed in each iteration of the cross-validation using the subsets of images used for training. We assessed different numbers of visual words to evaluate their impact on the classification.

Once the dictionary is built, each image of the dataset is described by means of a process called *image representation*. This consists of repeating, for each image, the same process of keypoint selection and characterization used in the creation of the dictionary, using also the same methods. Then, for each 'new' patch, we find the visual word of the dictionary that is most similar to it by means of calculating the Euclidean distance between their descriptors.

The histogram of the visual words representative of all patches in an image is used as its final descriptor. The image representation process is illustrated in Figure 8.

In this work, the patches are described using the SIFT [35]. Basically, SIFT descriptors are based on histograms of oriented gradients computed from the intensities of the regions that result from dividing a 16×16 pixel squared patch around each keypoint into 16 subregions of 4×4 pixels each. More details about SIFT can be found in [35]. Despite these consisting of two different parts, keypoint detector and patch descriptor, we only use the patch descriptor as we are sampling the keypoints in a dense grid.

Classification

In this work, we use a support vector machine (SVM) classifier, which is a supervised machine-learning approach that adjusts internal 'weights' by means of a training process (i.e. an optimization phase), minimizing the error between its calculated response and a 'ground truth' provided by an expert. This type of classifier has attracted attention in the last few years for analysing MR images [37-39]. SVM tries to find the optimal hyperplane that maximizes the distances

© 2017 The Author(s). This is an open access article published by Portland Press Limited on behalf of the Biochemical Society and distributed under the Creative Commons Attribution Licence 4.0 (CC BY).

Clinical Science (2017) **131** 1465–1481
DOI: 10.1042/CS20170051

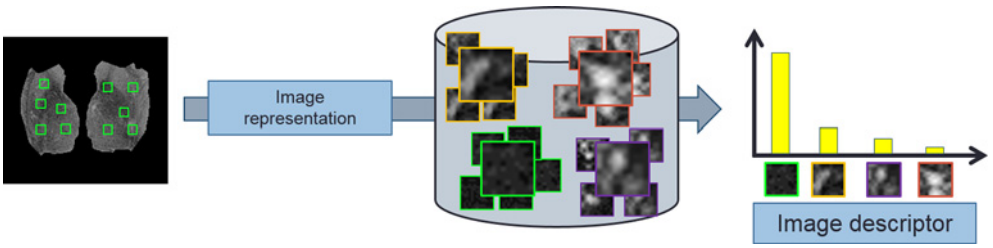

Figure 8. How the image representation is carried out.

Once the "dictionary" is formed, the histogram of the visual words representative of all patches is used as the final descriptor.

(i.e. margins) to the instances of the positive and negative classes in the training dataset. One of the parameters of SVM is the cost parameter *C*, which controls the trade-off between classes allowing training errors and forcing rigid margins.

SVM is a linear classifier: it tries to separate the data using a linear hyperplane. There are cases where the data are not linearly separable. In those cases, SVM may use the *kernel trick*: a kernel function K ($\mathbf{x}'$,$\mathbf{x}$) may transform the data into a higher dimensional space where it is possible to separate it linearly. After evaluating different kernels (i.e. linear, radial basis function, sigmoid), the best results were achieved with the radial basis function (RBF) kernel:

$$K\left(\mathbf{x}^{'}, \mathbf{x}\right) = \exp\left(-\gamma \left\|\mathbf{x}^{'} - \mathbf{x}\right\|^{2}\right) \quad (3)$$

We refer the reader interested in more details about SVM to [40].

We use several combinations of the regularization parameter *C* (i.e. 1, 5, 10, 50, 100, 250 and 500) and γ (i.e. 10^{-5}, 10^{-4}, 10^{-3}, 0.01 and 1), assessed with all descriptors, to find the optimal configuration. We use the implementation provided in the libSVM library[2] [41].

Validation of the classifier

We validated the classification with a stratified five-fold cross-validation as follows. The whole set, represented by the descriptors explained in 'Descriptors based on the Wavelet transform', 'Local binary patterns' and 'BoW' sections, was randomly partitioned into five equal-sized subsets with the same distribution as the original set. Of the five subsets, four were used to train the classifier and the remaining one was used as the test set. This process was repeated five times using a different subset each time as test set. The five results from the five folds were averaged to provide the final results.

This cross-validation process was repeated ten times, and the ten results were averaged to avoid possible bias due to a random separation of the folds. Data were normalized so that they had mean 0 and S.D. 1.

The overall results were validated in terms of accuracy, sensitivity and specificity, using the dichotomized ratings of Observer 1 as ground truth.

Statistical analyses

The descriptors that achieved the best performance would be used in a real automatic visual rating application. Therefore, we analysed the agreement of the visual ratings between the automatic classifier based on those descriptors and between each observer. We also analysed the association between the outcome of each PVS rating (i.e. from each observer and from the automatic classifier) and clinical parameters known to be related to PVS burden in the patients that comprise this sample (see 'Subjects and MRI protocol' section).

Interobserver agreement

We determined the weighted κ coefficient of the PVS ratings in the BG region (scale 0–4) between observers as per http://vassarstats.net/kappa.html (Copyright Richard Lowry 2001–2015). We also performed marginal homogeneity tests of the BG PVS visual ratings (scale 0–4) using the software application mh.exe ver. 1.2 (2016-03-01) (by John Uebersax).

After dichotomizing the BG PVS visual ratings produced by both the observers, we determined the κ coefficient between the observers and between the automatic classifier and each observer, using the κ function in MATLAB R2015a

[2]http://www.csie.ntu.edu.tw/∼cjlin/libsvm.

© 2017 The Author(s). This is an open access article published by Portland Press Limited on behalf of the Biochemical Society and distributed under the Creative Commons Attribution Licence 4.0 (CC BY).

Table 1 Distribution of the visual ratings in the sample

PVS rating	0	1	2	3	4	Total
Number of images (%)	5 (1.89%)	128 (48.48%)	68 (25.76%)	44 (16.67%)	19 (7.20%)	264

(Copyright© 2007, Giuseppe Cardillo, updated 23 Dec 2009, http://uk.mathworks.com/matlabcentral/fileexchange/15365-cohen-s-kappa/content/kappa.m). We also conducted the McNemar's test between the ratings produced by the expert (i.e. Observer 1) and the automatic classifier to investigate whether the marginal frequencies between both were or not equal.

Clinical validation

The following clinical and demographic parameters were available for each study participant: age, hypertensive (or not) classification, stroke subtype (lacunar or cortical) classification and scores of WMH, atrophy and SVD burden. WMH were coded using Fazekas scores, for periventricular (PV) and deep lesions separately in the left and right hemispheres and a combined score for both hemispheres was recorded [42]. Brain atrophy was coded using a validated age-relevant template [43], with superficial and deep atrophy coded separately ranging from none to severe on a scale of 1–6 according to the centiles into which the template is divided, being 1($<25th$), 2 (25–50th), 3 (50–75th), 4 (75–95th), 5 ($>95th$) and if $>>5$, 6 is used. Total atrophy was calculated as the average of deep and superficial atrophy scores. SVD was coded as per [44] (0–4), which confers a point for each of the following conditions: if 1 or more cavitated, old lacunar lesions are present, if Fazekas PV score $\geqslant 3$ and/or Fazekas Deep score $\geqslant 2$, if BG PVS score is $\geqslant 2$ as per Potter et al. [16] (i.e. moderate-to-extensive), and if more than 1, brain microbleed is present.

We calculated the non-parametric bootstrapped correlations between BG PVS scores (before and after dichotomization, from observers and from the automatic classifier) and each clinical variable. We also performed binomial multivariable logistic regression to evaluate the clinical usefulness of our machine-learning scheme as per [16] and its sensitivity in various models. The latter was evaluated by comparison of correlated receiver operating characteristic (ROC) curves obtained from three models that have outcome variable as the dichotomized PVS rating from (A) the automatic classifier, (B) Observer 1 and (C) Observer 2. The first model (i.e. model 1) had the following predictors: age, total atrophy, hypertension, Fazekas score, whether the patient had a previous lacunar infarct or not, index stroke subtype and SVD score. The second model (i.e. model 2, implemented in [16]) had the same predictors as model 1 with the exception of SVD score. The third model (i.e. model 3) also had the predictors of model 1 with the exception of Fazekas score and whether the patient had a previous lacunar infarct or not, as these two parameters are contemplated within the SVD score. These analyses were done using MATLAB R2015a. Of note, the PVS outcome variable is also a contributor to the SVD score.

Analysis of the robustness against imaging confounds

All scans of the primary study that provided data for this analysis underwent quality checks. None of the T2-weighted sequences were corrupted by visible movement artefacts that could affect the automatic PVS rating procedure presented. However, there are other confounds that could have influenced the results. We calculated the number of scans misclassified on each of the ten iterations that contributed to the final result, on the absence and presence of the following imaging confounds visually identified by Observer 2 in the BG region blind to the neuroradiological reports: WMH found either bilaterally and scattered throughout the region or as a single cluster possibly indicative of a recent or old subcortical infarct, lacunes (symptomatic or asymptomatic), recent or old cortical strokes that partially affect the region, globus pallidus partially or totally hyperintense, partial volume effects of the CSF, and a combination of two or more of these factors.

We also counted the number of scans misclassified on each iteration for those people who had a lacunar infarct neuroradiologically determined, regardless of whether it was visible on T2-weighted in the BG region or not. This analysis would allow us to discuss whether the occurence of a recent lacunar infarct influenced the descriptors used by the classifier.

Results

The PVS ratings made by the experienced neuroradiologist (Observer 1), used to train the classifier, were distributed across the sample as Table 1 shows. The dichotomization of these ratings into none to mild compared with moderate to severe resulted in 133 and 131 datasets for each class respectively.

© 2017 The Author(s). This is an open access article published by Portland Press Limited on behalf of the Biochemical Society and distributed under the Creative Commons Attribution Licence 4.0 (CC BY).

Clinical Science (2017) **131** 1465–1481
DOI: 10.1042/CS20170051

Table 2 Average accuracy, sensitivity and specificity, as well as their respective standard deviations (std, std_{sens} and std_{spec}) of the SVM five-fold classification along the ten iterations

	C	γ	Acc. (%)	Sens. (%)	Spec. (%)	std	std_{sens}	std_{spec}
WSF	500	10^{-4}	73.47	78.71	68.14	0.90	1.13	1.26
WCF_4	50	10^{-4}	73.66	77.15	70.12	1.25	1.38	1.68
WCF_{13}	250	10^{-4}	75.95	77.86	73.96	0.87	1.63	1.04
$LBP_{1,8}$	50	10^{-3}	68.34	70.49	66.21	2.54	2.18	3.51
$LBP^{ri}_{1,8}$	50	10^{-4}	70.02	75.95	64.01	1.02	1.22	1.49
$LBP^{riu2}_{1,8}$	10	0.01	74.22	81.97	66.37	0.70	1.39	1.57
$WCF_4 + LBP^{ri}_{1,8}$	250	10^{-4}	78.84	79.84	77.80	1.12	1.60	1.25
$WCF_{13} + LBP^{ri}_{1,8}$	100	10^{-4}	78.13	78.62	77.58	1.16	2.07	1.55
BoW	5	10^{-4}	81.16	79.31	82.97	1.72	2.20	2.57

The cost parameter for the SVM scheme (*C*) and the regularisation parameter of the radial basis function kernel (γ) for which the best results with each set of descriptors (i.e. first column on the left) were obtained, are provided. Abbreviations: acc., average accuracy; sens., sensitivity; spec., specificity.

Table 3 κ coefficient, S.E.M. and 95% CI, given the observed marginal frequencies between each observer and the automatic classification method

	κ	S.E.M.	95% CI
Obs. 1 compared with classifier	0.6228	0.0481	(0.5286–0.7170)
Obs. 2 compared with classifier	0.6743	0.0455	(0.5851–0.7635)

Results of the SVM classification

Table 2 shows the best results using the descriptors based on the Wavelet transform (i.e. WSF, WCF_4 and WCF_{13}), the descriptors based on local binary patterns with $R = 1$ and $P = 8$ (i.e. $LBP_{1,8}$, $LBP^{ri}_{1,8}$ and $LBP^{riu2}_{1,8}$), the fusions of the descriptors WCF_4 and WCF_{13} with $LBP^{riu2}_{1,8}$ and the descriptors based on the BoW model.

The best descriptor in terms of overall accuracy was the descriptor based on the BoW model (81.15%) using a dictionary with 175 visual words, followed by the fusion of WCF_4 and $LBP^{riu2}_{1,8}$ (78.84%). Moreover, the former reached a sensitivity just slightly worse than the latter. The highest sensitivity is achieved by $LBP^{riu2}_{1,8}$, but its specificity is much worse than the BoW-based descriptor. It is also remarkable that, whereas WCF_4 does not get a good accuracy on its own, its accuracy improves by 7% when it is fused with the $LBP^{riu2}_{1,8}$ descriptor.

The automatic classifier used in the following sections will be the SVM based on the descriptors that achieved the best overall accuracy (i.e. the dense-SIFT based BoW model, with the SVM parameters $C = 5$ and $\gamma = 10^{-4}$ using the dictionary of 175 visual words). Once the visual dictionary is created and the classifier is trained, this method takes 0.0477 s to describe and classify each image.

Interobserver variability

The agreement of the BG PVS ratings (scale 0–4) between Observers 1 and 2 was κ =0.8269, S.E.M.: 0.0398, 95% CI: (0.749–0.9048). The maximum possible linear-weighted κ, given the observed marginal frequencies was 0.8729. McNemar's tests for each rating (0–4), and McNemar's tests of equal thresholds were significant in rating 1 ($P<0.003$).

The agreement of the dichotomized BG PVS ratings between Observers 1 and 2 was κ =0.6822, S.E.M.: 0.0369 and 95% CI: (0.6099–0.7545). The maximum possible linear-weighted κ, given the observed marginal frequencies was 0.8486.

Table 3 shows the agreement (i.e. κ coefficient, S.E.M., 95% CI and maximum possible linear-weighted κ, given the observed marginal frequencies) between each observer and the ratings assigned by the SVM classifier that yield the best accuracy (see Table 2). Since the classification experiment was repeated ten times, the reported agreements are the average of the corresponding ten agreements. The marginal proportions between the ratings from the expert (i.e. Observer 1) and the automatic classifier were nonsignificantly different from each other (McNemar's test, P=0.1086). See the 2 × 2 frequency in Table 4.

© 2017 The Author(s). This is an open access article published by Portland Press Limited on behalf of the Biochemical Society and distributed under the Creative Commons Attribution Licence 4.0 (CC BY).

Table 4 Two-by-two table between the ratings done by the expert (i.e. Observer 1), the predictions of the classifier and ratings from Observer 2

Ratings	Auto.	Classifier	Observer 2	
	0	1	0	1
Observer 1, rating 0	104	29	102	31
Observer 1, rating 1	18	113	11	120

Table 5 Non-parametric bootstrapped cross-correlation matrix for PVS ratings in the BG region

Parameter	Observer 1 scale 0–4	Observer 1 dichotomized	Observer 2 scale 0–4	Observer 2 dichotomized	Automatic classifier
Observer 1 (0–4)	1	0.9317	0.8130	0.6828	0.6588
Observer 1 (0–1)		1	0.7341	0.6901	0.6464
Observer 2 (0–4)			1	0.9057	0.7127
Observer 2 (0–1)				1	0.7030

All correlations shown were significant with $P<0.0001$.

Table 6 Coefficient estimates and significance (B (*P*-value)) of the associations for each predictor (i.e. clinical parameter) for model 2

Predictor	Automatic classifier	Observer 1 dichotomized	Observer 2 dichotomized
Age (years)	0.0569 (0.0044)*	0.0462 (0.0074)*	0.0613 (0.0009)*
Atrophy (scale 1–6)	0.0075 (0.9313)	–0.0411 (0.5726)	–0.0472 (0.5545)
Hypertension (0–1)	0.2930 (0.4649)	–0.2131 (0.5534)	0.3675 (0.3120)
Fazekas deep (0–3)	0.5394 (0.0874)	0.1231 (0.6570)	–0.0801 (0.7854)
Fazekas PV (0–3)	1.4615 (<0.0001)**	1.1553 (0.00012)*	1.3928 (<0.0001)**
Old lacunar infarcts (0–1)	0.9625 (0.0245)*	1.0139 (0.0063)*	1.1987 (0.0035)*
Index stroke lacunar (0–1)	0.2953 (0.4355)	0.4294 (0.1881)	0.1853 (0.5868)

The outcome variable is the dichotomized PVS score. * $P<0.05$, ** $P<0.001$

Clinical validation

Bootstrapped correlations between the PVS ratings and with the clinical parameters

Visual ratings done by Observer 1 (dichotomized and not dichotomized), Observer 2 (dichotomized and not dichotomized) and the automatic classifier were equally, significantly and positively correlated with age, PVS ratings in CS (dichotomized and not), atrophy (deep and superficial), Fazekas (deep and PV), hypertension, old lacunar infarcts and SVD score. None of the BG PVS ratings correlated with index stroke subtype (lacunar or cortical), and all were highly and significantly correlated with each other as Table 5 shows.

Applicability in clinical research

Table 6 shows the results of the binomial multivariable logistic regression. Age, Fazekas PV scores and the presence of old lacunar infarcts were significant and negatively associated with all BG PVS scores (i.e. those done by both observers and by the automatic classifier), as in [16]. The coefficient estimates tabulated (B) express the effects of each predictor variable on the log odds of being in one class (i.e. 1 or 0) compared with the reference class (i.e. 1 or 0 as per Observer 1).

Sensitivity analysis

Figure 9 shows the predicted probabilities of the outcome variables for each model. The distribution of the predicted '0's and '1's to be 0 and 1 respectively for the classifier and Observer 2 were similar across all the models. All outcomes (i.e. PVS ratings from the classifier, Observers 1 and 2) were consistently poorer for model 2, which does not include SVD scores as predictor, than for the other two models. The PVS ratings from Observer 1 were particularly sensitive to the presence and absence of the SVD scores as predictor in the model, being exceptionally high when more components of the SVD score (including it) were included (i.e. model 1).

Figure 10 shows the correlated ROC curves for each outcome variable (i.e. automatic classifier, Observers 1 and

© 2017 The Author(s). This is an open access article published by Portland Press Limited on behalf of the Biochemical Society and distributed under the Creative Commons Attribution Licence 4.0 (CC BY).

Clinical Science (2017) **131** 1465–1481
DOI: 10.1042/CS20170051

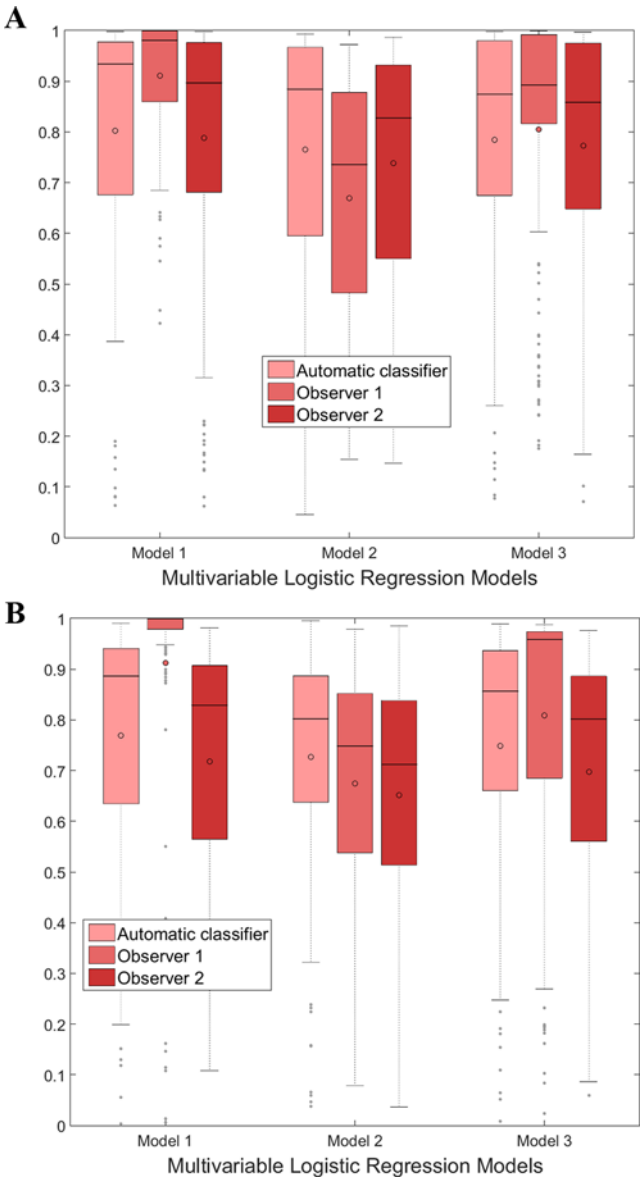

Figure 9. Illustration of the results of the multivariable logistic regression models.

Boxplots showing the distributions of the predicted probabilities of the outcome variable '1' (**A**) and '0' (**B**) (i.e. PVS ratings from the automatic classifier, from Observer 1 or Observer 2) for each logistic regression model.

2) also for each model. The area under the curve (AUC) from the automatic classifier experiences the least variation across the three curves: 0.93, 0.90 and 0.92 for models 1, 2 and 3 respectively (maximum variation 3%) indicating highest consistency in model accuracy, followed by Observer 2 (maximum variation 5%).

Performance on the presence/absence of imaging confounds

As Table 7 shows, only 9.6–16.6% of the scans that have a small T2-weighted hyperintense lesion such as lacunes, WMH or subcortical new or old infarcts in the BG region of size comparable with those of the PVS were misclassified, compared with 16% of the scans that have two or more of these confounds, and 13.6% of those who had none. These percentages were higher when the T2-weighted hyperintense covered a larger region (i.e. cortical stroke or globus pallidus hyperintense), but the number of scans that had these confounds were very small (7 and 5 respectively out of 264). The number of patients who had a recent lacunar infarct (neuroradiologically determined) and for which the PVS rating was miscalculated was the same as the number of patients that did not have any imaging confound and for which the PVS rating done by the classifier was wrong (compared with the ratings of the neuroradiologist).

© 2017 The Author(s). This is an open access article published by Portland Press Limited on behalf of the Biochemical Society and distributed under the Creative Commons Attribution Licence 4.0 (CC BY).

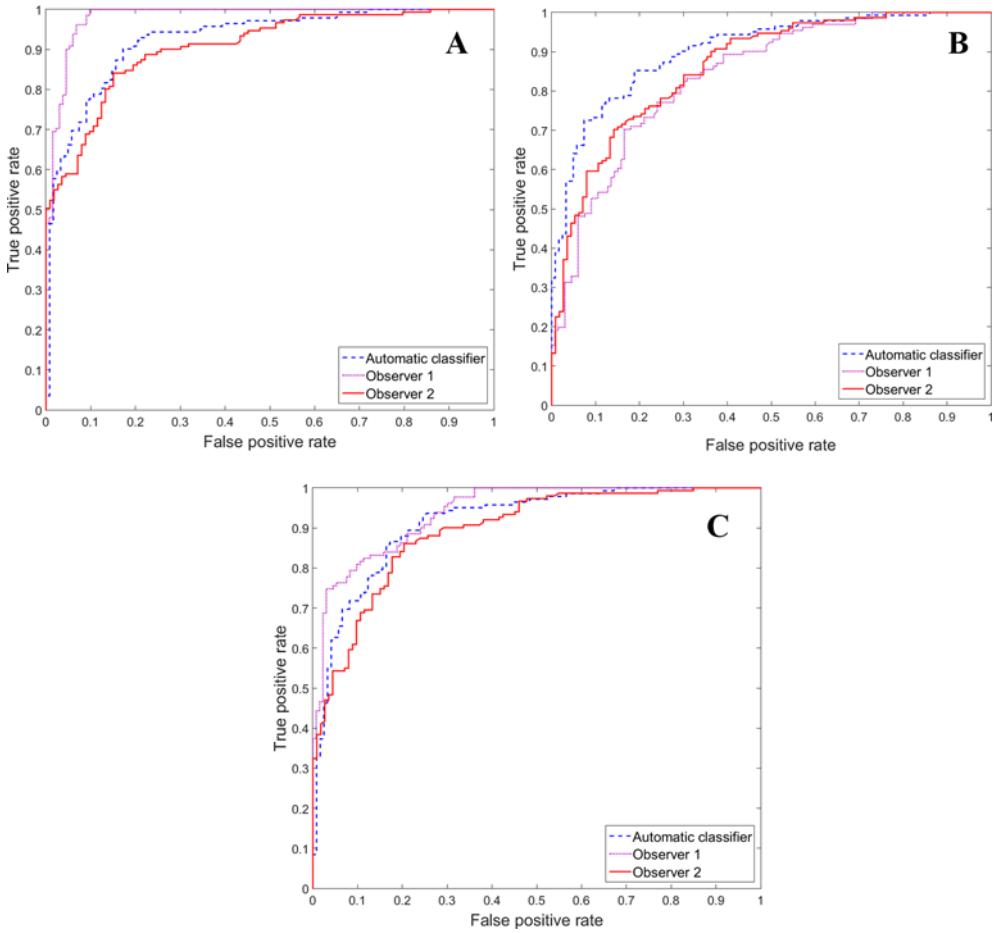

Figure 10. Receiver Operating Characteristics (ROC) for each classification.

ROC curves showing the performance for each outcome variable in the regression models 1, 2 and 3 ((**A**), (**B**) and (**C**) respectively). In model 1, (A) AUCs of the classifier, Observer 1 and Observer 2 were 0.9265, 0.9813 and 0.9074 respectively. In model 2, (B) AUCs of the classifier, Observer 1 and Observer 2 were 0.9041, 0.8395 and 0.8622 respectively. In model 3, (**C**) AUCs of the classifier, Observer 1 and Observer 2 were 0.9152, 0.9411 and 0.8934 respectively.

Table 7 Number of scans misclassified per number of iterations, on the presence/absence of imaging confounds

Confounds	Total number of scans	Number of scans misclassified		
		1–3 iter.	4–6 iter.	7–10 iter.
Unilateral WMH	31	4	1	3
Lacunes (symptomatic or not)	25	4	4	3
Bilateral WMH	30	5	0	5
Cortical stroke/CSF partial vol.	7	0	0	2
Globus pallidus hyperintense	5	1	1	1
Two or more of the above	56	4	3	9
None	110	12	8	15
Lacunar stroke	119	17	5	15

Discussion

We developed an automatic framework to classify T2-weighted MRI as having none or few PVS in the BG region compared with having many of them, in response to the need for such tool given the role of PVS in SVD and vascular dementia progression. Our framework uses a conventional SVM classifier based on the information from SIFT descriptors that operate on patches from the BG region using a dense grid following the 'BoW' model. These descriptors

© 2017 The Author(s). This is an open access article published by Portland Press Limited on behalf of the Biochemical Society and distributed under the Creative Commons Attribution Licence 4.0 (CC BY).

Clinical Science (2017) **131** 1465–1481
DOI: 10.1042/CS20170051

provided the highest classification accuracy (81.16%) from those evaluated. This accuracy is slightly lower than the one reported in [24] with the same descriptors (82.34%). The reason is the different validation of the classifier used in both works: in [24], the classification was carried out by randomly splitting the dataset into train (70%) and test sets (30%), whereas in this case we have used five-fold cross-validation. This classifier took an average of 0.0477 s to describe and classify each image. The framework proved to be useful in clinical settings and outperformed the visual classification done by a trained observer.

The image processing pipeline that preprocessed the data where the descriptors were extracted was designed following the visual rating guidelines for PVS from Potter et al. [16] (http://www.sbirc.ed.ac.uk/documents/epvs-rating-scale-user-guide.pdf), which are based on assessing the PVS from a region of interest on the axial MRI slice with the most visible PVS. All agreements between the automatic classifier, the dichotomized ratings of the experienced neuroradiologist (Observer 1) and those from the trained observer (Observer 2), as shown in 'Interobserver variability' section were above 0.6. However, the agreement between the dichotomized ratings from both observers ($\kappa = 0.6822$) was slightly higher than the agreement between the classifier and any of the observers (0.6228 with Observer 1 and 0.6743 with Observer 2). The fact that the classifier had better agreement with Observer 2 than with Observer 1 might be because Observer 2 followed the same guidelines used to design the pipeline for the automatic classifier, whereas Observer 1 may have also applied their individual experience and neuroradiological knowledge while rating the PVS. The cross-correlation between the classifier output and the dichotomized ratings of both observers, shown in Table 5, followed the same pattern: the correlation of the classifier with Observer 2 was higher than with Observer 1 (0.7030 and 0.6464 respectively). This cross-correlation between the output of the classifier and the dichotomized ratings of Observer 2 (0.7030) was comparable and even slightly higher than between the dichotomized ratings of both the observers (0.6901).

The statistical model built to evaluate the applicability of the automatic classifier to the clinical research showed excellent and similar goodness-of-fit irrespective of whether the outcome variable was the automatic classifier (AUC $=0.90$), Observer 1 (AUC $=0.84$) or Observer 2 (AUC $=0.86$). Also, age, the burden of PV WMH (i.e. Fazekas PV) and the presence of old lacunar infarcts were associated with the PVS burden irrespective of whether these were rated automatically or visually by any of the observers, proving the usefulness of the automatic framework proposed. A separate sensitivity analysis of this and similar correlated models showed that the automatic classifier was the least susceptible to be influenced by the overall burden of SVD shown in the MRI scan while the ratings from the neuroradiologist captured better the full flavour of the SVD features. The degree in which this result was favoured by the single-slice approach adopted by the classifier [12,16] is not known. Further evaluation on the whole extent of the three anatomical regions defined by [16], with added scrutiny to exclude lacunes is needed. Nevertheless, given that the accuracy of the classifier on the presence of imaging confounds was not different from it in the absence of them, and that the output was quite robust against the whole SVD burden, we do not foresee any problem for this automatic classification scheme to be applied to longitudinal or multicentre studies, as long as the training and testing datasets have similar acquisition protocols.

A possible limitation of the present work is the fact that the segmentation of the BG region is not always accurate (due to, for example, not finding the anatomical points described in 'Image preprocessing' section), causing a potential misclassification. As we wanted to assess the validity of a fully automatic method, we kept those suboptimal segmentations. Another limitation of the study may be the dichotomization of the visual ratings used in the automatic classification. Due to limitations in the sample size, we needed to simplify the classification, so we dichotomized the visual rating scale as it was done in recent studies [1]: a reliable five-class classification model is not possible to be trained with such few instances in some classes (e.g. out of 264 subjects there were only five with a rating 0 or 19 with rating 4). Further analyses using bigger samples and considering the full ratings (i.e. 0–4) need to be done.

Conclusion and future work

In the present paper, we have proposed an automatic framework based on image analysis and machine learning to predict the burden of enlarged PV spaces on the BG as 'none or few' or 'moderate to severe' based on the PVS visual rating scale [16]. We compared different descriptors computed from the BG region. The BoW-based descriptors achieved the best accuracy (81.16%) in the classification, carried out using a SVM trained using the visual ratings provided by an experienced neuroradiologist (i.e. Observer 1) as ground truth.

We also compared the predictions of the classifier with the visual ratings done by Observer 1 and also with those done by a trained image analyst (i.e. Observer 2). The interobserver agreement with the Observer 2 ($\kappa = 0.6743$) was higher than that with the Observer 1 ($\kappa = 0.6228$) and comparable with that between both observers ($\kappa = 0.6822$).

© 2017 The Author(s). This is an open access article published by Portland Press Limited on behalf of the Biochemical Society and distributed under the Creative Commons Attribution Licence 4.0 (CC BY).

The cross-correlation with the Observer 2 (0.7030) is also higher than that with the Observer 1 (0.6464), and slightly higher than that between both the observers (0.6901).

Finally, we built three correlated logistic regression models with some clinical variables as independent variables and the ratings predicted by the automatic method and both observers as outcome variables and demonstrated that although the automatic classifier does not capture the overall SVD severity, it can be used in clinical research as it consistently gives a meaningful output in relation to clinical parameters.

For future work, we will try to improve the classification performance by means of extracting the whole BG region and use the information from all slices where the extracted region appears (i.e. 3D analysis), as it may provide information that we are currently not taking into account. We will also try to use data from patients from other studies to increase our sample size and perform a five-class classification (i.e. ratings from 0 to 4). Supervised machine-learning schemes like the one presented here would require the ground truth PVS counts or segmentations from a large number of datasets done by an expert to be able to count and/or segment PVS. Such data are currently unavailable. However, the output from this classifier could be used as input to the fully automatic PVS unsupervised segmentation approach developed by [15], (mentioned in the 'Introduction' section) which needs the PVS ratings to tune its algorithm and make it fully automatic. Finally, the classifier presented here could be adapted to get the visual rating of the PVS in the CS.

Clinical perspectives

- In the brain, enlarged PVS are commonly assessed visually, for which several rating scales exist. However, for epidemiological and large clinical database analyses, there is a need for a fully automatic rating of the PVS load that overcomes interobserver and interscale differences, which the current study aims to address.

- The performance of the three different types of descriptors were assessed, with the SVM classifier achieving the best accuracy with the BoW-based descriptors. These outputs were compared with visual ratings made by an experienced neuroradiologist (Observer 1) and by a trained image analyst (Observer 2), with agreements and cross-correlations between the classifier and Observer 1, the classifier and Observer 2, and comparable between both observers being $\kappa = 0.67$, $\kappa = 0.62$ and $V = 0.68$ respectively. Models using clinical variables as independent variable showed that the goodness-of-fit of the model for the classifier was good and slightly better than that of Observer 2.

- An automatic classifier can be used to assess PVS burden from brain MRI, and can provide clinically meaningful results. The automatic scheme proposed can be used for large epidemiological studies and clinical databases, as it has been shown to be unbiased to the coexisting pathologies and is consistent.

Acknowledgements

We thank the study participants, radiographers and staff at the Brain Research Imaging Centre Edinburgh, SINAPSE (Scottish Imaging Network A Platform for Scientific Excellence) Collaboration Centre. We also thank the Wellcome Trust for providing the data.

Funding

This work was supported by the Wellcome Trust [grant number 088134/Z/09]; and the Row Fogo Charitable Trust [grant number BRO-D.FID3668413].

Author contribution

M.dC.V.H. and V.G.C.: study design, image processing and analysis, statistical analysis, writing and approving the final version of the manuscript. P.A.A.: image acquisition protocol design, editing, revising and approving the final version of the manuscript. S.M.: patient recruitment and clinical examination, editing, revising and approving the final version of the manuscript. F.M.C.: statistical analysis, editing, revising and approving the final version of the manuscript. J.M.W.: primary study design, neuroradiological ratings, study funding, editing, revising and approving the final version of the manuscript.

© 2017 The Author(s). This is an open access article published by Portland Press Limited on behalf of the Biochemical Society and distributed under the Creative Commons Attribution Licence 4.0 (CC BY).

Clinical Science (2017) **131** 1465–1481
DOI: 10.1042/CS20170051

Competing interests

The authors declare that there are no competing interests associated with the manuscript.

Abbreviations

AUC, area under the curve; BG, basal ganglia; BoW, bag of visual words; CI, confidence interval; CS, centrum semiovale; CSF, cerebrospinal fluid; DWT, discrete Wavelet transform; FAST, FMRIB's Automated Segmentation Tool; FMRIB, Oxford Centre for Functional MRI of the Brain; FSL, FMRIB software library; GLCM, grey level co-occurrence matrix; LBP, local binary pattern; PV, periventricular; PVS, perivascular space; ROC, receiver operating characteristic; SIFT, scale-invariant feature transform; std, standard deviation; SVD, small vessel disease; T1w, T1-weighted; T2w, T2-weighted; WCF, wavelet co-occurrence feature; WMH, white matter hyperintensity; WSF, Wavelet statistical feature.

References

1 Potter, G.M., Doubal, F.N., Jackson, C.A., Chappell, F.M., Sudlow, C.L., Dennis, M.S. et al. (2015) Enlarged perivascular spaces and cerebral small vessel disease. *Int. J. Stroke* **10**, 376–381
2 Maclullich, A.M.J., Wardlaw, J.M., Ferguson, K.J., Starr, J.M., Seckl, J.R. and Deary, I.J. (2004) Enlarged perivascular spaces are associated with cognitive function in healthy elderly men. *J. Neurol. Neurosurg. Psychiatry* **75**, 1519–1523
3 Wuerfel, J., Haertle, M., Waiczies, H., Tysiak, E., Bechmann, I., Wernecke, K.D. et al. (2008) Perivascular spaces–MRI marker of inflammatory activity in the brain? *Brain* **131**, 2332–2340
4 Aribisala, B.S., Wiseman, S., Morris, Z., Valdes-Hernandez, M.C., Royle, N.A., Maniega, S.M. et al. (2014) Circulating inflammatory markers are associated with magnetic resonance imaging-visible perivascular spaces but not directly with white matter hyperintensities. *Stroke* **45**, 605–607
5 Patankar, T.F., Baldwin, R., Mitra, D., Jeffries, S., Sutcliffe, C., Burns, A. et al. (2007) Virchow–Robin space dilatation may predict resistance to antidepressant monotherapy in elderly patients with depression. *J. Affect. Disord.* **97**, 265–270
6 Laitinen, L.V., Chudy, D., Tengvar, M., Hariz, M.I. and Bergenheim, A.T. (2000) Dilated perivascular spaces in the putamen and pallidum in patients with Parkinson's disease scheduled for pallidotomy: a comparison between mri findings and clinical symptoms and signs. *Mov. Disord.* **15**, 1139–1144
7 Doubal, F.N., MacLullich, A.M.J., Ferguson, K.J., Dennis, M.S. and Wardlaw, J.M. (2010) Enlarged perivascular spaces on MRI are a feature of cerebral small vessel disease. *Stroke* **41**, 450–454
8 Pantoni, L. (2010) Cerebral small vessel disease: from pathogenesis and clinical characteristics to therapeutic challenges. *Lancet Neurol.* **9**, 689–701
9 Wardlaw, J.M., Smith, E.E., Biessels, G.J., Cordonnier, C., Fazekas, F. et al. (2013) Neuroimaging standards for research into small vessel disease and its contribution to ageing and neurodegeneration. *Lancet Neurol.* **12**, 822–838
10 Staals, J., Makin, S. D.J., Doubal, F.N., Dennis, M.S. and Wardlaw, J.M. (2014) Stroke subtype, vascular risk factors, and total MRI brain small-vessel disease burden. *Neurology* **83**, 1228–1234
11 Valdés Hernández, M.d.C., Piper, R.J., Wang, X., Deary, I.J. and Wardlaw, J.M. (2013) Towards the automatic computational assessment of enlarged perivascular spaces on brain magnetic resonance images: a systematic review. *J. Magn. Reson. Imaging* **38**, 774–785
12 Wang, X., Valdés Hernández, M.d.C., Doubal, F., Chappell, F.M., Piper, R.J., Deary, I.J. et al. (2016) Development and initial evaluation of a semi-automatic approach to assess perivascular spaces on conventional magnetic resonance images. *J. Neurosci. Methods* **257**, 34–44
13 Ramirez, J., Berezuk, C., McNeely, A.A., Scott, C.J.M., Gao, F. and Black, S.E. (2015) Visible Virchow-Robin spaces on magnetic resonance imaging of Alzheimer's disease patients and normal elderly from the Sunnybrook Dementia Study. *J. Alzheimers Dis.* **43**, 415–424
14 Cai, K., Tain, R., Das, S., Damen, F.C., Sui, Y., Valyi-Nagy, T. et al. (2015) The feasibility of quantitative MRI of perivascular spaces at 7T. *J. Neurosci. Methods* **256**, 151–156
15 Ballerini, L., Lovreglio, R., Valdés Hernández, M.d.C., Gonzalez-Castro, V., Maniega, S.M., Pellegrini, E. et al. (2016) Application of the ordered logit model to optimising frangi filter parameters for segmentation of perivascular spaces. *Procedia Comput. Sci.* **90**, 61–67
16 Potter, G.M., Chappell, F.M., Morris, Z. and Wardlaw, J.M. (2015) Cerebral perivascular spaces visible on magnetic resonance imaging: development of a qualitative rating scale and its observer reliability. *Cerebrovasc. Dis.* **39**, 224–231
17 Munsell, B.C., Wee, C.-Y., Keller, S.S., Weber, B., Elger, C., da Silva, L. A.T. et al. (2015) Evaluation of machine learning algorithms for treatment outcome prediction in patients with epilepsy based on structural connectome data. *Neuroimage* **118**, 219–230
18 Beheshti, I., Demirel, H. and Alzheimer's Disease Neuroimaging Initiative (2015) Probability distribution function-based classification of structural MRI for the detection of Alzheimer's disease. *Comput. Biol. Med.* **64**, 208–216
19 Ithapu, V., Singh, V., Lindner, C., Austin, B.P., Hinrichs, C., Carlsson, C.M. et al. (2014) Extracting and summarizing white matter hyperintensities using supervised segmentation methods in Alzheimer's disease risk and aging studies. *Hum. Brain Mapp.* **35**, 4219–4235
20 Roy, P.K., Bhuiyan, A., Janke, A., Desmond, P.M., Wong, T.Y., Abhayaratna, W.P. et al. (2015) Automatic white matter lesion segmentation using contrast enhanced FLAIR intensity and markov random field. *Comput. Med. Imaging Graph.* **45**, 102–111
21 de Brebisson, A. and Montana, G. (2015) Deep neural networks for anatomical brain segmentation. *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 20–28
22 Chen, L., Tong, T., Ho, C.P., Patel, R., Cohen, D., Dawson, A.C. et al. (2015) Identification of cerebral small vessel disease using multiple instance learning. *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015*, pp. 523–530, Springer
23 González-Castro, V., Valdés Hernández, M.d.C., Armitage, P.A. and Wardlaw, J.M. (2016) Texture-based classification for the automatic rating of the perivascular spaces in brain MRI. *Procedia Comput. Sci.* **90**, 9–14

© 2017 The Author(s). This is an open access article published by Portland Press Limited on behalf of the Biochemical Society and distributed under the Creative Commons Attribution Licence 4.0 (CC BY).

24 González-Castro, V., Valdés Hernández, M.d.C., Armitage, P.A. and Wardlaw, J.M. (2016) Automatic rating of perivascular spaces in brain MRI using bag of visual words. *Image Analysis and Recognition: 13th International Conference, ICIAR 2016, Proceedings*, pp. 642–649, Springer International Publishing
25 Vapnik, V. (1995) *The Nature of Statistical Learning Theory*, 2nd Edition, Springer
26 Alegre, E., González-Castro, V., Alaiz-Rodríguez, R. and García-Ordás, M.T. (2012) Texture and moments-based classification of the acrosome integrity of boar spermatozoa images. *Comput. Methods Programs Biomed.* **108**, 873–881
27 Ojala, T., Pietikainen, M. and Maenpaa, T. (2002) Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Anal. Mach. Intell.* **24**, 971–987
28 Valdés Hernández, M.d.C., Armitage, P.A., Thrippleton, M.J., Chappell, F., Sandeman, E., Muñoz Maniega, S. et al. (2015) Rationale, design and methodology of the image analysis protocol for studies of patients with cerebral small vessel disease and mild stroke. *Brain Behav.* **5**, e00415
29 Lutkenhoff, E.S., Rosenberg, M., Chiang, J., Zhang, K., Pickard, J.D., Owen, A.M. et al. (2014) Optimized brain extraction for pathological brains (optiBET). *PLoS ONE* **9**, e115551
30 Zhang, Y., Brady, M. and Smith, S. (2001) Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *IEEE Trans. Med. Imaging* **20**, 45–57
31 Zuiderveld, K. (1994) Contrast limited adaptive histogram equalization. *Graphics Gems IV*, pp. 474–485, Academic Press Professional, Inc.
32 Arivazhagan, S. and Ganesan, L. (2003) Texture classification using wavelet transform. *Pattern Recognit. Lett.* **24**, 1513–1521
33 Haralick, R.M., Shanmugam, K. and Dinstein, I. (1973) Textural features for image classification. *IEEE Transactions on Systems, Man and Cybernetics SMC-3* 610–621
34 Alegre, E., González-Castro, V., Suárez, S. and Castejón, M. (2009) Comparison of supervised and unsupervised methods to classify boar acrosomes using texture descriptors. *ELMAR, 2009. International Symposium ELMAR*, pp. 65–70
35 Sivic, J. and Zisserman, A. (2003) Video google: a text retrieval approach to object matching in videos. *Proceedings of the Ninth IEEE International Conference on Computer Vision, 2003*, vol. 2, pp. 1470–1477
36 MacQueen, J. (1967) Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, pp. 281–297, University of California Press, Berkeley, California
37 Nam, K.W., Castellanos, N., Simmons, A., Froudist-Walsh, S., Allin, M.P., Walshe, M. et al. (2015) Alterations in cortical thickness development in preterm-born individuals: implications for high-order cognitive functions. *Neuroimage* **115**, 64–75
38 Tong, T., Wolz, R., Gao, Q., Guerrero, R., Hajnal, J.V. and Rueckert, D. (2014) Multiple instance learning for classification of dementia in brain MRI. *Med. Image Anal.* **18**, 808–818
39 Feis, D.-L., Brodersen, K.H., von Cramon, D.Y., Luders, E. and Tittgemeyer, M. (2013) Decoding gender dimorphism of the human brain using multimodal anatomical and diffusion MRI data. *Neuroimage* **70**, 250–257
40 Schölkopf, B. and Smola, A.J. (2001) *Learning with kernels: Support vector machines, regularization, optimization, and beyond*, MIT Press
41 Chang, C.-C. and Lin, C.-J. (2011) LIBSVM: A library for support vector machines. *ACM Trans. Intelligent Syst. Technol.* **2**, 27
42 Fazekas, F., Chawluk, J.B., Alavi, A., Hurtig, H.I. and Zimmerman, R.A. (1987) MR signal abnormalities at 1.5 T in Alzheimer's dementia and normal aging. *Am. J. Neuroradiology* **8**, 351–356
43 Farrell, C., Chappell, F., Armitage, P.A., Keston, P., MacLullich, A., Shenkin, S. et al. (2008) Development and initial testing of normal reference MR images for the brain at ages 65–70 and 75–80 years. *Eur. Radiol.* **19**, 177–183
44 Staals, J., Booth, T., Morris, Z., Bastin, M.E., Gow, A.J., Corley, J. et al. (2015) Total MRI load of cerebral small vessel disease and cognitive ability in older people. *Neurobiol. Aging* **36**, 2806–2811

© 2017 The Author(s). This is an open access article published by Portland Press Limited on behalf of the Biochemical Society and distributed under the Creative Commons Attribution Licence 4.0 (CC BY).