Computer-Assisted Segmentation of White Matter Lesions in 3D MR Images Using Support Vector Machine[1]

Zhiqiang Lao, Dinggang Shen, Dengfeng Liu, Abbas F. Jawad, Elias R. Melhem, Lenore J. Launer
R. Nick Bryan, Christos Davatzikos

Rationale and Objectives. Brain lesions, especially white matter lesions (WMLs), are associated with cardiac and vascular disease, but also with normal aging. Quantitative analysis of WML in large clinical trials is becoming more and more important.

Materials and Methods. In this article, we present a computer-assisted WML segmentation method, based on local features extracted from multiparametric magnetic resonance imaging (MRI) sequences (ie, T1-weighted, T2-weighted, proton density-weighted, and fluid attenuation inversion recovery MRI scans). A support vector machine classifier is first trained on expert-defined WMLs, and is then used to classify new scans.

Results. Postprocessing analysis further reduces false positives by using anatomic knowledge and measures of distance from the training set.

Conclusions. Cross-validation on a population of 35 patients from three different imaging sites with WMLs of varying sizes, shapes, and locations tests the robustness and accuracy of the proposed segmentation method, compared with the manual segmentation results from two experienced neuroradiologists.

Key Words. White matter lesion segmentation; support vector machine; machine learning

Cerebrovascular disease (CVD) in elderly individuals is very important. In particular, CVD increases the likelihood of clinical dementia (1–4) even in the absence of clinical stroke (5), albeit the literature is somewhat inconclusive as to whether CVD has simply an additive role to Alzheimer's disease (AD) or there are interactions between the two. Approximately one third of patients that meet clinical and pathologic diagnostic criteria for AD have some degree of vascular pathology (6,7). The impact of CVD on mild cognitive impairment—in which the etiology of the cognitive deficit is generally less clear—is likely to be even greater. Therefore, to identify biologic markers specific to the AD process, it is critical to also identify the extent of concurrent CVD related brain injury that is often clinically silent (8–11), because, at the very least, CVD increases the likelihood of clinical presentation of dementia, for the same level of AD-related pathology.

Population studies, such as the Cardiovascular Health Study or the Rotterdam Scan Study, have shown that brain lesions, especially white matter lesions (WMLs), are

Acad Radiol 2008; 15:300–313

[1] Department of Radiology, 3600 Market Street, Suite 380, University of Pennsylvania, Philadelphia, PA 19104 (Z.L., D.S., A.F.J., E.R.M., R.N.B., C.D.); Department of Biostatistics, Children's Hospital of Philadelphia, Philadelphia, PA (A.F.J.); Laboratory of Epidemiology, Demography, and Biometry, National Institute on Aging, Bethesda, MD (L.J.L.); Lister Hill National Center for Biomedical Communications, National Library of Medicine/National Institute of Health, Bethesda, MD (D.L.). Received August 23, 2007; accepted October 1, 2007. Supported (in part) by the Intramural Research Program of the NIH, National Institute of Aging contract N01-HC-95178. Image analysis was supported in part by R01-AG-1497. **Address correspondence to:** Z.L. e-mail: Zhiqiang.Lao@uphs.upenn.edu

doi:10.1016/j.acra.2007.10.012

associated with age, clinically silent stroke, higher systolic blood pressure, lower forced expiratory volume in 1 second, hypertension, atrial fibrillation, carotid and peripheral arterioscleroses, impaired cognition, and depression (12–14). Furthermore, it has been shown that stroke patients with a large WML load have an increased risk of hemorrhagic transformation, higher preoperative risk of a disabling or fatal stroke during endarterectomy, or intercerebral hemorrhage during anticoagulation therapy (15). The increased interest in brain lesion research may improve diagnosis and prognosis possibilities for patients with cardiovascular symptoms.

The relationship between diabetes mellitus and cognitive impairment, as well as with increased risk for dementia, has been documented by several clinical studies (16–18). This relationship is mediated by brain pathology, including cerebral infarcts, leukoaraiosis, and tissue atrophy (19–24). Precise measurement of such pathology from magnetic resonance imaging (MRI), and more importantly measurement of evolution of pathology over time, is very important for disease monitoring and evaluation of treatments for diabetes mellitus, such as controlling blood pressure and glycemia. All of these previous studies have employed subjective evaluation of brain lesions, such as the scale of de Groot et al (12), which examined the relationship between periventricular and subcortical WMLs and cognitive functioning in 1,077 elderly subjects randomly sampled from the general population, and are hampered by variations in the anatomical definition of brain abnormalities. Therefore such methods of evaluation of brain abnormality in diabetes mellitus are not easily reproducible, qualitative, and nearly impossible to use without paired reading and high level of quality control in large multisite studies and in longitudinal evaluations that might span several years (25). There is an increasing need for development of highly automated, validated, and reproducible computer-based image analysis tools, especially in large-scale longitudinal studies evaluating brain pathology in diabetes mellitus.

Because brain lesion patterns are very heterogeneous, ranging from punctuate lesions in the deep white matter to large confluent periventricular lesions, the scoring of such lesions is complicated. Moreover, it has been shown that different visual rating scales lead to inconsistencies among studies (26). Commonly used ordinal brain lesion scoring methods, such as used in the Cardiovascular Health Study (27) or the Rotterdam Scan Stud (12,28), offer semiquantitative information on the prevalence of such lesions. Exact spatial information is useful because it has been suggested that specific lesion patterns are associated with specific symptoms (29,30). Moreover, for longitudinal studies aiming to capture relatively small changes in brain lesion patterns, accurate information of lesion volume and location is essential. Expert-based delineation of brain lesions is known to be difficult to reproduce across raters, or even within the same rater, which makes it problematic and that combination of readings from independent reader may be necessary in a longitudinal study.

The use of an automated segmentation method that detects brain lesions with a high sensitivity and specificity could be advantageous. Most of the successful methods in the literature have been developed for the detection of multiple sclerosis (MS) lesions (31–46). In early approaches when multi-modality images are not easily available, features describing normal tissue statistics (either intensity property alone or both intensity and spatial properties) are usually extracted from available modality and then combined with various classifiers, such as: minimum distance classifier, Bayesian classifier, decision tree, for MS lesion segmentation purpose. In (31), Kamber et al built a voxel-wise probability normal tissue (GM [gray matter], WM [white matter], VN [ventricle]) distribution model in Talairach space and then use a decision tree to discriminate MS lesion tissue from normal tissue based on entropy minimization. A similar approach was pursued elsewhere (46), in which spatial statistics of normal brain tissues were first determined from a training set, and deviations from normal variation were flagged as lesions. In Udupa et al (33), major brain tissues (WM, GM, and CSF [cerebrospinal fluid]) were modeled as fuzzy connected regions; potential MS lesions are regarded as isolated islands and were further refined by human judgment.

Most current imaging studies offer the potential to combine multiparametric MRIs (ie, images obtained via different MRI protocols). The advantage of integrating information from multiple sequences is that it can reduce the uncertainty and increase the accuracy of the segmentation. One can categorize most state-of-the-art lesion segmentation algorithms in two main categories: supervised voxel-wise classification and unsupervised clustering. Leemput et al (37) proposed an unsupervised WML segmentation model via setting up a multivariate Gaussian model to describe normal tissue signal distribution, and using it to detect MS lesions as outliers. In supervised methods, a set of images in which the desired segmentation is known (expert manual delineation) is used as a training set to build and fine-tune the segmentation algorithm (35,44). Based on the well-known medical image processing system INSECT, Zijdenbos et al (41) proposed a

supervised MS lesion segmentation method using multispectral (T1-weighted [T1-w], T2-weighted [T2-w], proton density-weighted [PD]) intensity signal and spatial prior as features and artificial neural network for classification purpose. The method was developed in the context of a Phase III clinical trial and results were evaluated on 29 subjects, revealing that the obtained lesion measurements are statistically equivalent to those obtained by trained human observers. Wu et al (40) described a method to quantitatively measure volumes of three subtypes of MS lesions (T1 hyperintense enhancing, T1 hypointense, and CSF-like "black hole" lesions), as well as segmentation of GM, WM, and CSF simultaneously. This method used a expectation-maximization approach that iteratively integrated a statistical intensity feature–based (k-nearest-neighbor) classifier and three-channel TDS$^+$ (32,42,43), which makes use of a deformable digital brain atlas to eliminate the confounding misclassification, based on the assumption that WMLs are only within WM regions. Somewhat related is also the method (35) that used an artificial neural network trained on three multiparametric MRI and on probability distribution of GM, WM, and CSF. Anbeek et al (44) proposed a supervised and multispectral WML segmentation method, multispectral intensity (T1-w, T2-w, PD, fluid attenuation inversion recovery [FLAIR], and inversion [IR]) as well as spatial features are defined as features to discriminate WML tissue from normal tissue, KNN was then used for voxel-wise evaluation of the possibility to be WML tissue. This method was then extended to segment WM, GM, ventricle, CSF, and WML (47). Admiraal-Behloul (45) proposed a multimodality white matter hyperintensities segmentation approach that also employs multispectral intensity (T2-w, PD, and FLAIR) as well as tissue spatial distribution probability map as features. Signals in T2-w and FLAIR were used to define feature, and PD was just used for skull-stripping purpose. Fuzzy inference system was then used to further infer whether a voxel is WMH or normal tissue. Some other methods have combined space and time into the lesion characterization process (48,49), albeit these approaches focused primarily on quantifying the temporal variations of MS lesions, important in differentiating active from chronic lesions.

However, relatively less attention has been given to brain lesion segmentation in elderly individuals or AD or diabetic patients. Because MS lesions present different characteristics from lesions in elderly or diabetic individuals, those methods are not directly applicable to our studies, although they have formed the foundation for our development. Because of the decreased contrast between WM and GM in MRI in the elderly, techniques that require the segmentation of WM and GM for the extraction of the WMLs perform moderately well when applied to geriatric patients, especially when they were originally designed and trained to extract lesions in MS patients. A histogram-based segmentation method was used for identification of leukoaraiosis in elderly individuals (50). Although the results of that study were promising, we have found in large-scale studies that such relatively simple segmentation methods are not sufficiently robust. A method using a supervised classification method with relatively good sensitivity but somewhat limited specificity to lesions has been presented (51).

In this article, we present a novel computer-assisted WML segmentation approach that has been designed to process MRI scans of elderly diabetes patients and used in a large clinical study: Action to Control Cardiovascular Risk in Diabetes Memory in Diabetes (ACCORD-MIND, http://www.accordtrial.org/), a Phase III large clinical trial that aims to investigate the relationship between diabetes, treatment intensity, and thinking and memory in older patients (25). Our method uses a combination of image analysis and support vector machines (SVM). Image intensities from multiple MRI acquisition protocols, after coregistration, are used to form a voxel-wise attribute vector (AV) that helps to discriminate lesion from various normal tissue image profiles during segmentation. In general, there are four steps in our approach, as summarized in Fig 1. First, a preprocessing step includes co-registration of different MRI modalities of the same patient to skull-stripping, intensity normalization, and inhomogeneity correction. Second, a set of training samples is manually delineated by expert readers and then used to build a classification model via SVM and AdaBoost; this step is applied only once, during training. Third, the SVM model is used to perform the voxel-wise segmentation. Finally, false-positive voxels are further eliminated via postprocessing techniques described later, thereby producing final WML segmentation results. This methodology is described and then validated against expert human readings.

METHODS

Patients and MR Imaging

Images used in this study were offered by the ACCORD-MIND MRI trial, which is a prospective randomized four-site trial on conventional versus aggressive

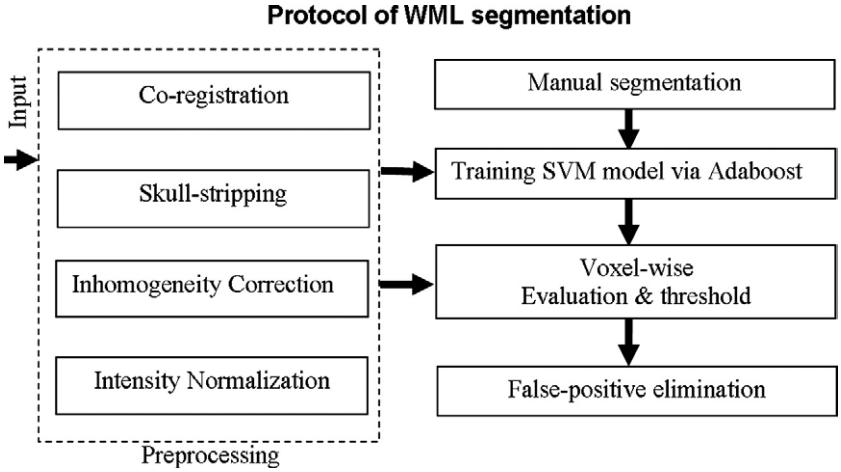

Figure 1. Summary of our computer-assisted white matter lesion segmentation protocol.

treatment of diabetes (http://www.accordtrial.org/). Mean age of these subjects was 62 (mean ± SD 62.2 ± 5.9, range 54–77, median 61). Twenty-seven were female and 18 were male. MRIs were performed during the baseline period on enrollment into the study. All 45 participants' exams consisted of transaxial T1-w, T2-w, PD, and FLAIR scans. All scans except T1-w were performed with a 3-mm slice thickness, no slice gap, a 240 × 240 mm field of view, and a 256 × 256 scan matrix. T1-w scans were performed with a 1.5-mm slice thickness, same slice gap, field of view, and scan matrix.

Preprocessing

The multiple images acquired from the same individual are coregistered to compensate for possible motion between scans. Mutual information–based affine registration (52), implemented in FMRIB Software Library (53), is employed for coregistration of multimodality images. The FLAIR image of each subject is used as a reference space, to which all other modality images are transformed. After coregistration, a deformable model based skull-stripping algorithm called Brain Extraction Tool (BET) (54), implemented in FMRIB Software Library (53), is used to generate an initial brain tissue mask from the coregistered T1-w image, and then this brain tissue mask is used to extract the brain region from all other modality images. Finally, for each image volume, inhomogeneities are corrected by N3 (55), and intensity normalization within and across different subjects is minimized by a global histogram matching method. To this end, for two three-dimensional images (or two-dimensional slices) $\boldsymbol{I}_1$ and $\boldsymbol{I}_2$ with histograms $H_{I_1}(i)$ and $H_{I_2}(i)$, respectively, the transformation, $\boldsymbol{T}$ ($T(i) = s \times i + t$, where i is the intensity value before and after transformation $\boldsymbol{T}$, s, and t are scaling and translation parameters of T, respectively), is found so that $\sum_{i=1}^{i=\max} [H_{I_1}(T(i)) - H_{I_2}(i)]^2$ is minimal, where *max* is the maximum intensity value.

Training

Manual segmentation.——Ten training sets with highly variable lesion loads were selected from these 45 ACCORD-MIND participants. WMLs in these subjects were manually segmented by two neuroradiologists (RNB, ERM). The segmentation of the first rater was regarded as the gold standard for training our classifier, whereas the segmentation of the second rater was used for evaluating interrater agreement and for comparing it against computer-rater agreement.

Attribute vector.——In general, the amount of intensity overlap between WMLs and normal tissue varies greatly across different modalities, as shown in Fig 2. In T1-w images, WMLs have intensities similar to GM, and in T2-w and PD images, WMLs look very similar to CSF. Although the FLAIR image has the least intensity overlap between WMLs and normal tissues, it has been suggested in the literature that FLAIR is less sensitive in the posterior fossa (56), may lead to "overestimation" of lesion load, and has a higher intervendor variability (57,58). Furthermore, FLAIR may present hyperintensity artifacts (59,60) that might lead to false positives, thereby rendering it difficult to use only the FLAIR images to segment WMLs. Therefore, it is important to integrate information from different modalities to minimize the ambiguity in

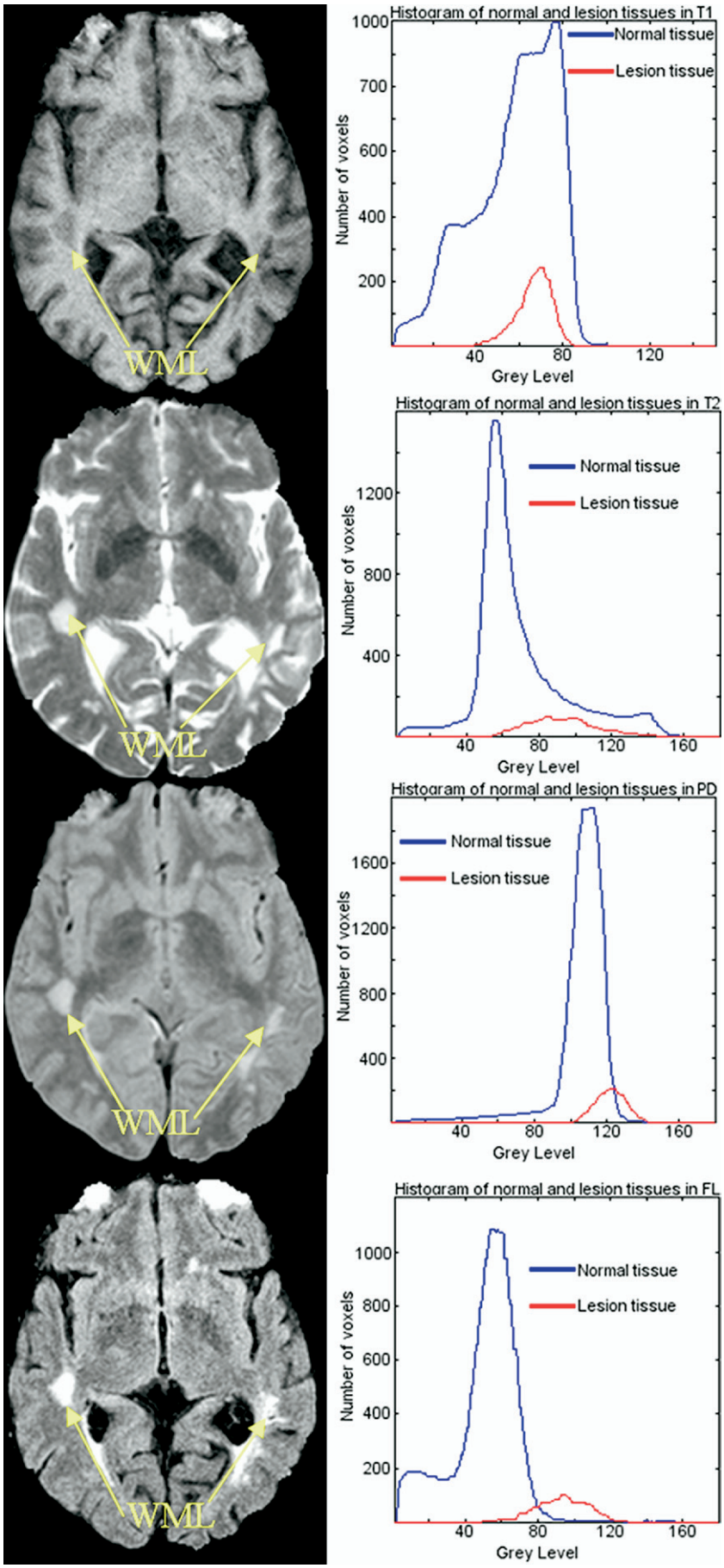

Figure 2. Intensity overlaps between white matter lesions tissue and normal tissue in T1, T2, proton density-weighted, and fluid attenuation inversion recovery scans, respectively (histograms for normal tissue have been scaled by 0.1 for visualization purpose).

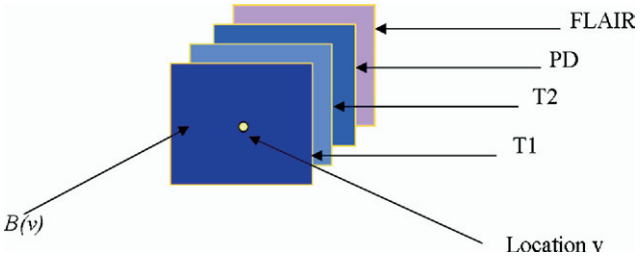

Figure 3. Image intensities from all modalities and all voxels in the spatial neighborhood of a voxel form an attribute vector that serves as an "imaging signature" of each voxel.

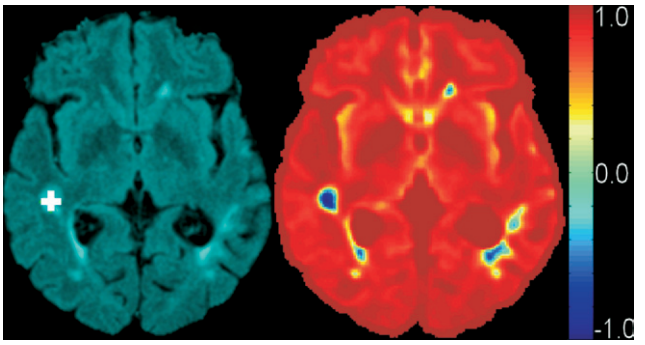

Figure 4. Discrimination ability of attribute vectors (AV). Left: Fluid attenuation inversion recovery scans image with selected lesion voxel marked as white cross. Right: Distance distribution in Hilbert space from all other voxels to this selected voxel. AVs of other lesion voxels are similar (having small distance in the attribute space) to the selected voxel, indicating that this imaging signature is characteristic of lesions.

identifying WMLs from normal tissue using only a single modality image.

An attribute vector (AV) is computed for each non-background voxel in a three-dimensional reference space for each subject, FLAIR image space serves as the reference space to which all other acquisitions are coregistered. To include spatial information from the vicinity of each voxel and make AV robust, each AV includes not only the four image intensities of that voxel, but also intensities of neighboring voxels, as shown in Fig 3. Moreover, to make AV robust to noise, each modality image of the same subject is smoothed by a Gaussian filter with a very small kernel (0.5 mm). Mathematically, for a voxel v in domain $\mathbf{\Omega}$, its AV is defined as $F(v) = \{I^m(\mu)\}$: $m \in \{T_1,T_2,PD,FLAIR\}$, and $\mu \in \boldsymbol{B}(v)$, where $\boldsymbol{B}(v)$ is a small neighborhood of v in Ω. In other words, the four image intensities of all voxels in the neighborhood of v are concatenated to an AV. The neighborhood size is selected based on the discrimination ability of AV, in our implementation it is 5 mm $\times$ 5 mm $\times$ 5 mm. Figure 4 shows the discrimination ability of this AV with respect to WML. In the figure on the left side, a voxel specified by white cross is selected; in the figure on the right side, the distance in Hilbert space (to be defined later in the article) between the AV of the marked voxel and the AVs of all other voxels is shown color-coded. It can be seen from the figure that lesion tissue shows high similarity to the selected voxel.

SVMs

SVMs are a relatively new machine learning tool and have emerged as a powerful technique for learning from data and, in particular, for solving binary classification problems. SVMs originate from Vapnik's statistical learning theory (61) and they formulate the learning problem as a quadratic optimization problem whose error surface is free of local minima and has global optimum. In a binary classification task such as the one in our study (normal tissue/lesion tissue), the aim is to find an optimal separating hyperplane (OSH) between the two datasets. Figure 5 illustrates a two-class problem with a hyperplane separating the two groups. SVMs find the OSH by maximizing the margin (minimum distance) between the classes. The main concepts of SVM are to first transform input data into a higher dimensional space (Hilbert space) by means of a kernel function and then construct an OSH between the two classes in the transformed space (Hilbert space). Those data vectors nearest to the constructed line in the transformed space are called the support vectors (Fig 5); they contain valuable information regarding the OSH. SVMs are an approximate implementation of the method of "structural risk minimization" aiming to attain low probability of generalization error. The theory of SVM can be referenced elsewhere (62).

The kernel function used in our application is Gaussian radial basis function kernel, defined as

$$K(x, y) = \exp\left(\frac{\|x - y\|^2}{2\alpha^2}\right)$$

where x and y are two feature vectors, and α controls the size of the Gaussian kernel.

The fitness of a hyperplane in feature space is usually measured by the distance between the hyperplane and those training points lying closest to it (the support vectors). A consequence of this is that we can completely specify our decision surface in terms of these support

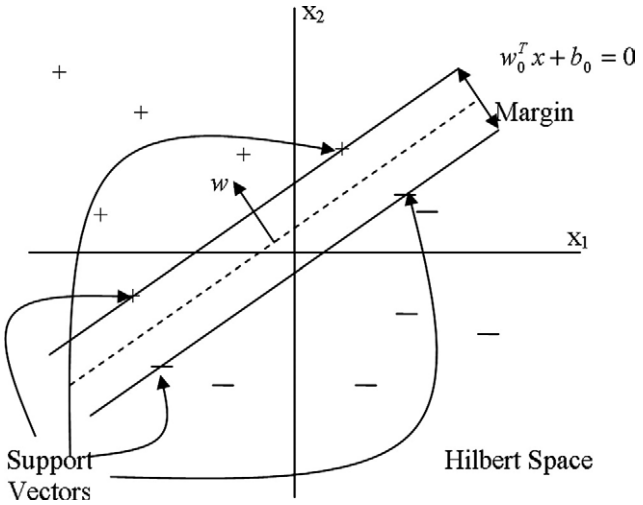

Figure 5. An example of two-class (+ and −) problem showing optimal separating hyperplane (dotted line) that support vector machines use to divide two groups' data, and the associated support vectors. Data shown by + and − represent binary class +1 and −1, respectively.

vectors. An overview of SVM pattern recognition techniques may be found elsewhere (63).

Training SVM via AdaBoost

After an AV is defined for each location in each training scan, a nonlinear pattern classifier is constructed from the entire training set (ie, by using all lesion voxels of all training scans as examples of imaging profiles to be recognized in new scans, along with a large number of normal tissue voxels). These example AVs are provided to SVM (64,65). Because the number of normal tissue voxels is far higher than the number of lesion voxels, it is essential to select only a representative set of normal tissue voxels comparable to the number of lesion voxels. This selection is not random, but it is rather guided by the classification results themselves, using the AdaBoost algorithm (66). This approach is based on a sequence of classifiers that rely increasingly on misclassified voxels, because those are presumably the voxels on which the classifier must focus. During this adaptive boosting procedure, each sample receives a weight that determines its probability of being selected in a training set for the next iteration. If a training sample is accurately classified, then its likelihood of being used again in subsequent iterations is reduced; conversely, if a training sample is inaccurately classified, then its likelihood of being used again is increased.

Segmenting a New Image

Voxel-wise segmentation of WML by SVM.——In the testing stage, T1-w, T2-w, and PD images of a new (not in the training set) scan are first coregistered to FLAIR space of the subject using mutual information registration method mentioned before, and then the pseudo-likelihood of each voxel being WML is measured by the generated SVM classifier, as described earlier. The output of SVM is a scalar measure of abnormality (as shown in Fig 6, left), which is further binarized by an optimal threshold to produce the labels for WMLs (as shown in Fig 6, right). These labels are called initial WML labels, because false-positive labels will be screened out by the methods proposed next.

Elimination of false-positive labels.——Misregistration between the four MRIs usually results in a number of false positives around the cortex because of the convoluted nature of the cortex, which amplifies the adverse effect of slight registration inaccuracies. By analyzing the spatial distribution of AVs from different samples, we found that AVs of false-positive voxels actually form a third class associated with the SVM training samples, which is far away from both classes of lesion and non–lesion training samples. In other words, these voxels don't match either lesion or normal tissue, according to the training set. Thus these false-positive voxels can be eliminated to a large extent by computing the distance of their AVs to the training samples in the Hilbert space that the SVM training model was built on, as described in the SVM section.

The distance measure in Hilbert space between two vectors v_1 and v_2 can be calculated as

$$D_h^2(v_1, v_2) = K(v_1, v_1) + K(v_2, v_2) - 2K(v_1, v_2),$$

where K is the Gaussian kernel function used by the SVM.

Suppose $L = \{v_i^\ell, 1 \le i \le m\}$ is the set of lesion AVs in training samples, m is the total number of AVs in L; $N = \{v_i^n, 1 \le i \le p\}$ is the set of normal tissue AVs in training samples, p is the total number of AVs in N; $F = \{v_i^f, 1 \le i \le q\}$ is the set of AVs of false positives, q is the total number of AVs in F. Figure 7 illustrates the distribution of L, N, and F in Hilbert space. Thus we measure the distance of each AV to a certain set of AVs in the following way. For each v_i^ℓ, its distance to L in Hilbert space is defined as $d_{v_i^\ell} = \min_{j=1}^{m} D_h^2(v_i^\ell, v_j^\ell)$, where $j \ne i$; similarly, for each v_i^n, its distance to N in Hilbert space is defined as $d_{v_i^n} = \min_{j=1}^{p} D_h^2(v_i^n, v_j^n)$, where $j \ne i$; for each v_i^f,

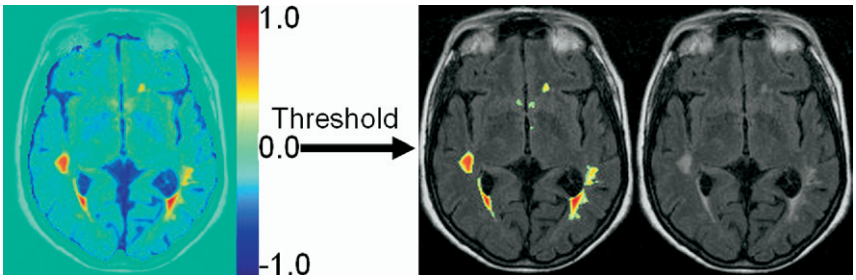

Figure 6. Illustration of voxel-wise segmentation by support vector machines (SVM). Left: The result of voxel-wise evaluation map showing different lesion rating for each voxel, based on generated SVM model (1: lesion; −1: normal). Right: White matter lesion segmentation results after thresholding the map on the left superimposed on fluid attenuation inversion recovery image. Threshold actually corresponds to SVM classification boundary as illustrated in Fig 7, with two classes labeled as −1 and 1 respectively, 0.0 is selected as a threshold.

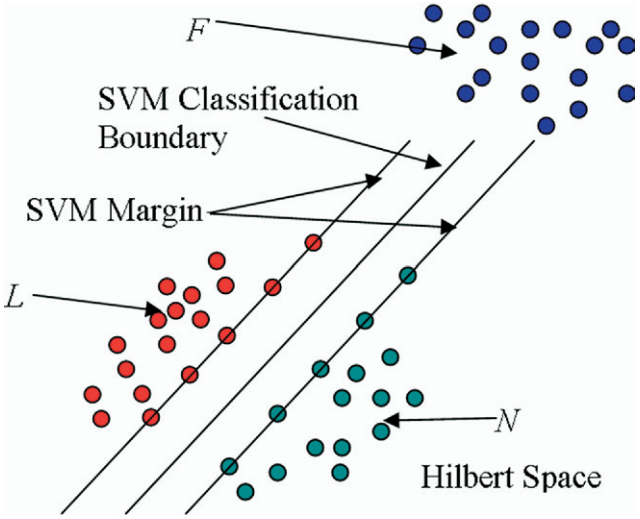

Figure 7. Illustration of *L*, *N*, and *F* distribution in Hilbert space. Green and red represent attribute vectors (AVs) of healthy and lesion tissue, respectively, whereas blue represents AVs of voxels that are misclassified mostly because minor registration errors between the four different acquisitions (T1, T2, PD, and fluid attenuation inversion recovery scans) causes them to have imaging profiles that are drastically different from the training set, and hence prone to misclassification.

its distance to L in Hilbert space is defined as $d^L_{v_i^f} = \min_{j=1}^{m} D_h^2(v_i^f, v_j^\ell)$; similarly its distance to N in Hilbert space is defined as $d^N_{v_i^f} = \min_{j=1}^{p} D_h^2(v_i^f, v_j^n)$. Figure 8 shows the distributions of these distances, which indicates that we can simply use this minimal distance measure to eliminate the false-positive samples, by selecting a suitable threshold.

Extra-axial hyperintense regions, like fat in the orbits, cannot always be completely removed by the skull-stripping algorithm used in preprocessing stage. Imaging profiles belonging to these regions are more similar to WMLs than that of normal tissue and therefore they are eliminated from the segmentation mask after SVM classification. This is done by morphologic operations combined by adaptive thresholding in skull-stripped FLAIR image. Figure 9 demonstrates one sample result from the algorithm.

RESULTS

Two representative results are shown in Fig 10. "Gold standard" (manual) and computer-assisted segmentation results are superimposed on the FLAIR images, respectively.

ROC Analysis

We have computed the receiver operating characteristic (ROC) curve for our computer-assisted lesion segmentation algorithm. The ROC curve is a graphical plot of the sensitivity versus (1 - specificity) for a binary classifier system as its discrimination threshold is varied (67). Figure 11 shows a zoomed version of ROC curve showing detail in the region of interest. Different symbols on the ROC curve show different thresholds we used. Additionally, "*" shows second rater's manual segmentation result compared with our gold standard.

By investigating the balance between sensitivity and specificity, we determined the optimal threshold to be 0.05. The specificity of a ROC is defined as $\frac{TN}{TN + FP}$, where TN represents the volume of true negatives and FP represents the volume of false positives. In WML segmentation, TN is always a very large number as compared with FP, which makes specificity of ROC insensitive to FP change.

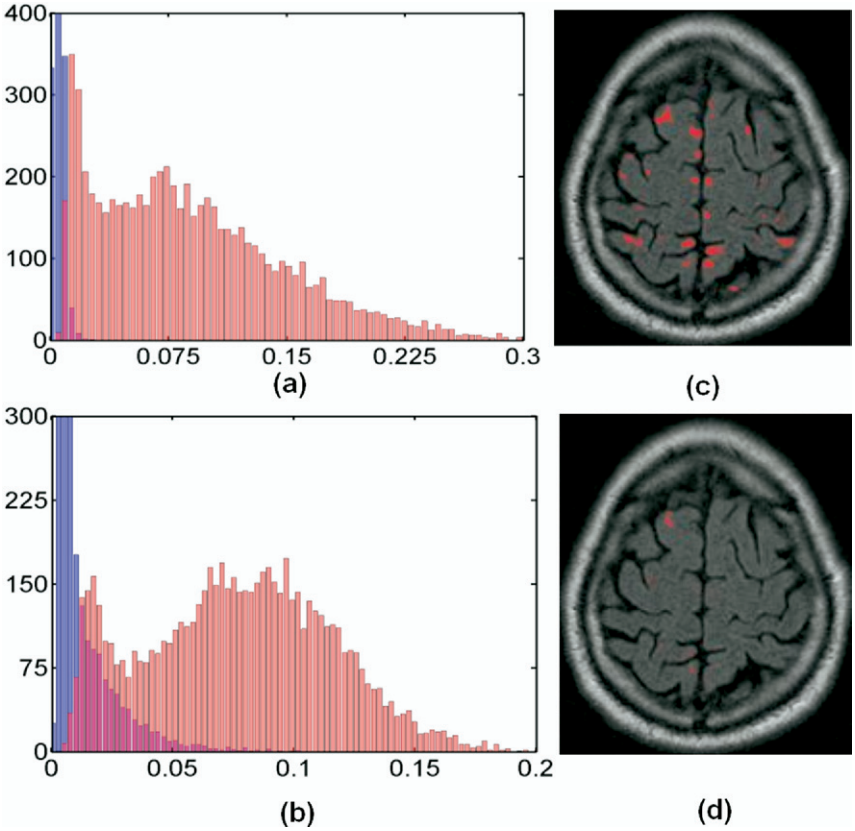

Figure 8. Demonstration of false positive elimination via attribute vector distance in Hilbert space. **(a)** Distance distribution of $\{d_{v_i^G}\}$ (blue, true positives), $\{d_{v_i^L}^L\}$ (red, false positives), and the overlap between $\{d_{v_i^G}\}$ and $\{d_{v_i^L}^L\}$ (violet). **(b)** Distance distribution of $\{d_{v_i^{G'}}\}$(blue, true negatives), $\{d_{v_i^N}^N\}$ (red, false positives), and the overlap between $\{d_{v_i^{G'}}\}$ and $\{d_{v_i^N}^N\}$ (violet). White matter lesion segmentation results **(c)** before false-positive elimination and **(d)** after false-positive elimination via thresholding the distance map.

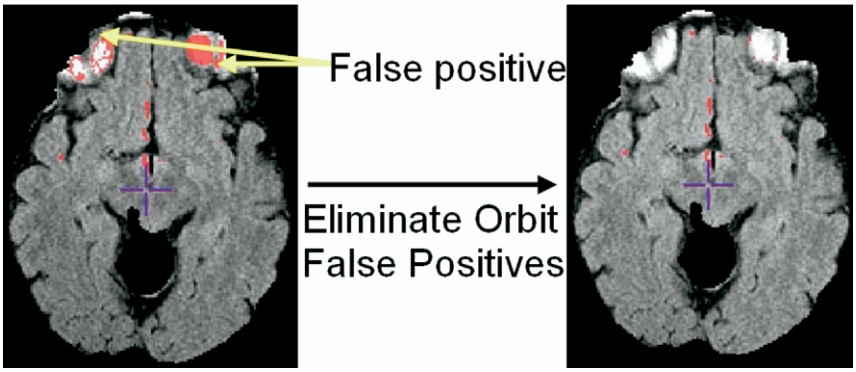

Figure 9. Demonstration of orbital false positive elimination. Left: Orbital false positives (red) overlaid on fluid attenuation inversion recovery scans before false-positive elimination. Right: After orbital false-positive elimination.

Statistical Analysis

We trained the classifier on data from 10 participants, and tested it on the remaining 35 cases. We have performed statistical comparisons between the lesion volume obtained by manual and computer-assisted segmentation (with threshold = 0.05) of these 35 subjects.

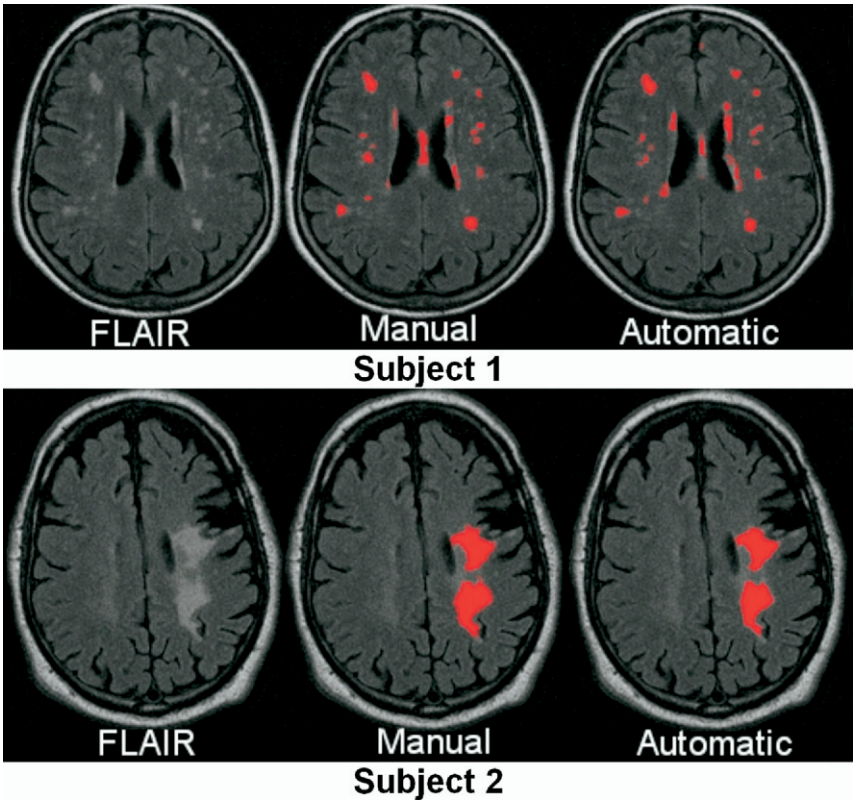

Figure 10. Comparison of white matter lesion segmentation results between gold standard and computer-assisted segmentation for two subjects. In subject 1, gold standard and computer-assisted lesion measurements are 11,714.9 mm^3 and 12,397.9 mm^3, respectively; in subject 2, gold standard and computer-assisted lesion measures are 15,978.5 mm^3 and 17,884.9 mm^3, respectively.

Paired Spearman correlation (*SC*) measurements among first rater, second rater, and computer-assisted method shows high correlation among them ($P < .001$ and $\rho = 0.95$ between second rater and first rater; $P < .001$ and $\rho = 0.79$ between computer and first rater; $P < .001$ and $\rho = 0.74$ between computer and second rater). Although high in correlation measurement, mean ± standard deviation (median) of the lesion volumes obtained from first, second, and computer raters were 1,494 mm^3 ± 3,416 mm^3 (559 mm^3); 2,839 mm^3 ± 6,192 mm^3 (1461 mm^3); and 1,869 mm^3 ± 3416 mm^3 (393 mm^3) respectively, the mean volume of 2nd rater is approximate twice of 1st rater, which suggests that manual segmentation is subject to large inter-rater variability as shown in Fig 12.

To investigate the variation of the lesion load's distribution of the 35 evaluated subjects, the coefficient of variation (*CV*) was calculated. It is a statistical measurement of the dispersion of data around the mean and calculated as: $CV = \frac{\sigma}{\mu}$, where σ is the standard deviation and μ is the mean. The *CVs* for the three raters were 189%, 218%, and 182%, respectively, which shows that all three raters agree on the large variation in lesion load from this set of subjects. Because the distribution of lesion volume is skewed, log transformation (log_{10}) was performed and further statistical analysis was done using the log-transformed data. Comparisons between means of the log-transformed lesion load (volume in mm^3) among raters were performed using paired *t*-test. On average, the mean of first rater reading was 0.37 ($\log_{10}$ mm^3) lower than second rater ($P < .001$), and not significantly higher by 0.07 ($\log_{10}$ mm^3) than computer rater ($P < .001$). The second rater's reading was significantly higher ($P < .001$) than the computer rater by .38 ($\log_{10}$ cc). The agreement between the computer rater and the first rater is better than that between second rater and first rater. Computer-assisted segmentation is an extension of human rater's power to more accurately and precisely quantification lesion volume.

DISCUSSION

We have presented an approach to the problem of WML segmentation, based on integrating multiple MRI acquisi-

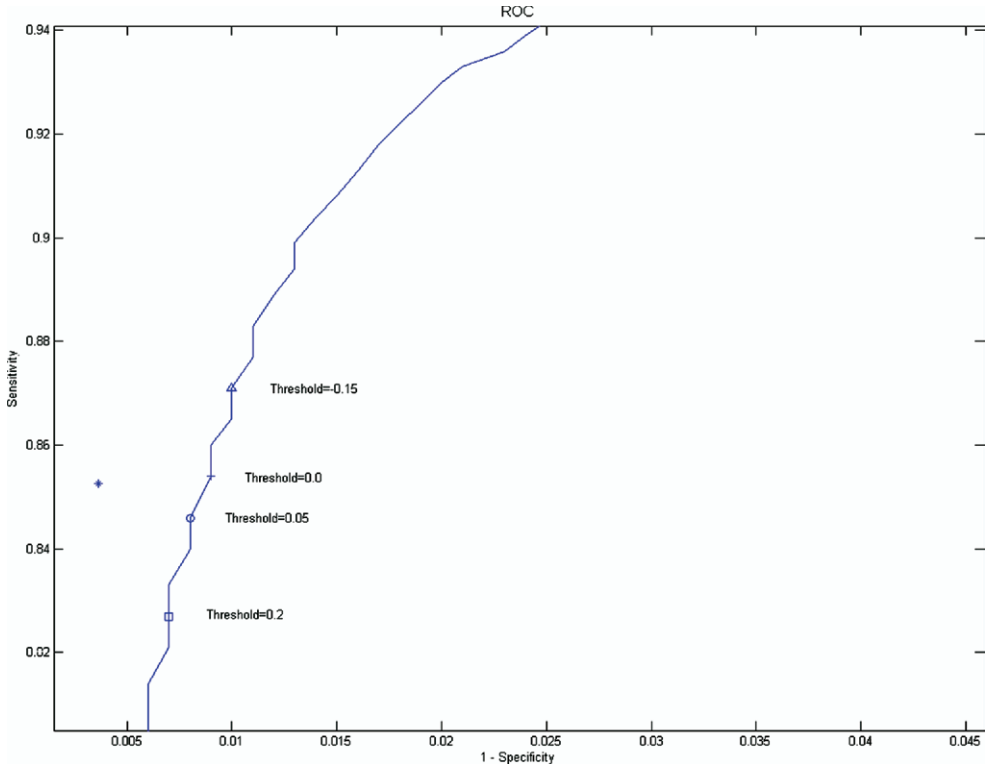

Figure 11. A zoomed part of receiver operating characteristic curve of our segmentation algorithm. ∗: Indicates the result of the second rater compared with gold standard (first rater). Other symbols on the curve denote different thresholds (ie, Δ threshold = −0.15, + threshold = 0.0, ○ threshold = 0.05, □ threshold = 0.2 (see Fig 6 for the definition of threshold).

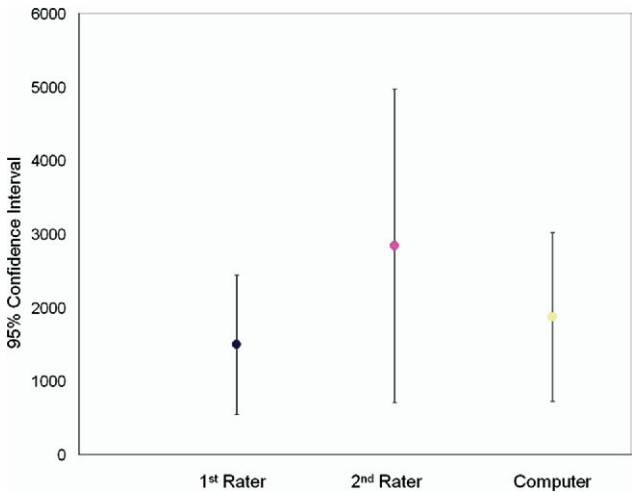

Figure 12. 95% CI (confidence intervals) for gold standard (first rater), second rater, and computer-assisted segmentation method (computer) over 35 subjects, respectively. Volume measurements are in mm^3.

tions and training a nonlinear pattern classification algorithm to recognize imaging profiles that are representative of a brain lesion. By combining four types of MRI acquisition protocols—namely FLAIR, T_2, PD, and T_1—a multivariant imaging signature is constructed for every image voxel and is subsequently evaluated by a nonlinear pattern classifier. Results that agree well with human experts were obtained.

The objective, quantitative, and reproducible evaluation of WML has been a challenge in many neuroimaging studies. Although qualitative readings have been employed by many studies, the relatively limited sensitivity and interrater agreement is an obstacle, particularly in longitudinal studies or in studies seeking to detect subtle effects. Our experiments confirm that, although human experts are relatively internally consistent in what they define as lesion, they can differ considerably between each other; thereby, their readings when combined increases the measurement mean and standard deviation and therefore decreases study power.

We did not rigorously evaluate the relative value of each acquisition protocol. Although FLAIR provides the best contrast between periventricular WMLs and ventricles, PD helps avoid potential "overestimation" of lesion load that has been observed with the FLAIR sequence, especially in the posterior fossa. PD may also help in eliminating false positives in regions in which FLAIR has

hyperintense artifacts. We experimented with different combinations of three modalities and observed that the quality of the segmentation deteriorated when omitting any of the four image types. Therefore it appears that all four protocols carry some discriminatory power, albeit of different degrees.

One of the challenges we faced during the development of the segmentation method was that the number of lesion training voxels that we had available were dramatically smaller than the number of training voxels for healthy tissue, because lesions constitute a very small percentage of the entire brain. Even smaller was the number of voxels that were misclassified. Therefore a natural bias toward healthy tissue was inevitable. Although one could randomly select an equal number of voxels for training, the bias toward the more frequent tissues would still persist. We overcame this problem by using adaptive boosting (ie, via an iterative procedure that progressively emphasized voxels being misclassified). Voxels that were incorrectly classified had more likelihood to influence subsequent iterations. Therefore, in the end, the classifier was constructed mainly from "difficult-to-classify" voxels.

Although in the experiments reported here we mainly used a binary segmentation output, our approach actually derives a continuous spatial map that provides a pseudo-likelihood of each voxel corresponding to abnormal tissue. Such a continuous map can ultimately be a more appropriate way to characterize certain types of tissues, such as periventricular abnormalities, which might present a continuous scale of pathology.

Although an extensive experimental comparison between this approach and alternative methods in the literature is beyond the scope of this article, our algorithm has several features that render it novel and likely to be relatively more robust. In particular, we used the currently most robust machine learning method (ie, SVM), which is known to provide optimal generalization ability. Moreover, we used AdaBoost, which is a significant aspect of our approach, because it progressively learns from misclassified examples. Put simply, instead of weighting all voxels similarly, the "difficult" ones are identified by the algorithm, which places relatively more emphasis on them. Automated removal of false positives via examination of the distance in the Hilbert space is also an important novelty of our approach; this step is most often performed manually in a postprocessing step.

The framework of our method is somewhat similar to Abeeck's approach (44). Methodologically, both approaches start with a number of preprocessing steps (intrasubject coregistration, skull stripping, inhomogeneity correction, and intensity normalization) and use a supervised classifier (KNN in Anbeek's approach, SVM in this article) to separate lesions from normal brain tissue. The key difference is AV definition. In Anbeek's approach, both intensity and spatial information are used in AV definition. Because of the arbitrary occurrence nature of WML, it is difficult to form a "complete" training set that covers all occurrences in practice; our approach uses only intensity in AV definition, combined with AdaBoost-based training sample selection method, thus is easier in forming a "complete" empiric training set. KNN is known to be computational intensive with high-dimensional AV, that's why the AV definition in Anbeek's approach includes intensity and spatial features only on a single voxel, which may make such AV definition vulnerable to misregistration. SVM does not have such limitation and our AV definition includes not only multispectral signals on a certain voxel but also its small neighborhood, which makes it more robust to misregistration.

Several improvements and extensions to our basic methodology are possible. In particular, the current algorithm examines the data voxel by voxel, with the exception of using some signal information from a small neighborhood around each voxel. The anatomic context around each voxel could potentially help improve accuracy and reduce false positives. In our previous work (46), we have used a statistical atlas derived from deformable registration of many images of healthy individuals without any pathology, and we detected abnormalities as deviations from the normal spatial variation of healthy tissue. We anticipate that adding such statistically based anatomic information to the signal-based information examined here, is likely to improve segmentation accuracy. A second direction of work that might benefit the segmentation is toward coregistration among different modalities. The mutual information based coregistration method we are currently using provides a pairwise alignment between two modalities. A better way would align images of all modalities simultaneously to reach a consistent solution. Although the smoothing procedure in our WML segmentation protocol performed pretty well in dealing with coregistration error, this step can be improved by a neighborhood voting strategy (ie, for a certain voxel), measuring the correlation of selected AVs in the neighborhood across different modalities to AVs in the training set and selecting the one with the highest correla-

tion coefficient. This will lead to more robust AVs, and more accurate segmentation result.

In summary, by combining four different MRI acquisition protocols and using them to train a nonlinear pattern classification technique, we developed a relatively robust and fully automated segmentation method for white matter abnormalities. We are in the process of applying this method to data from over a dozen different centers in multisite studies and have obtained stable results, which further bolsters our confidence that this approach can facilitate large-scale neuroimaging studies seeking to quantify vascular disease.

ACKNOWLEDGMENTS

We would like to thank the committee of ACCORD-MIND project, which is funded by the NIA through an intra-agency agreement with NIHLBI (Y3-HC-3065), for providing the datasets, valuable comments and giving us permissions to publish this paper. We also like to thank Ms. Lisa Desiderio for assistance in coordinating this study. Finally, we would like to thank patients recruited by ACCORD-MIND project.

REFERENCES

1. Prins ND, van Dijk EJ, den Heijer T, et al. Cerebral white matter lesions and the risk of dementia. Arch Neurol 2004; 61:1531–1534.
2. Snowdon DA, Greiner LH, Mortimer JA, et al. Brain infarction and the clinical expression of Alzheimer's disease. The Nun study. JAMA 1997; 277:813–817.
3. Schneider JA, Wilson RS, Cochran EJ, et al. Relation of cerebral infarctions to dementia and cognitive function in older persons. Neurology 2003; 60:1082–1088.
4. Schneider JA, Wilson RS, Bienias JL, et al. Cerebral infarctions and the likelihood of dementia from Alzheimer disease pathology. Neurology 2004; 62:1148–1155.
5. Vermeer SE, Prins ND, den Heijer T, et al. Silent brain infarcts and the risk of dementia and cognitive decline. N Engl J Med 2003; 348:1215–1222.
6. Gearing M, Mirra SS, Hedreen JC, et al. The Consortium to Establish a Registry for Alzheimer's Disease (CERAD). Part X. Neuropathology confirmation of the clinical diagnosis of Alzheimer's disease. Neurology 1995; 45:461–466.
7. Ince P, Xuerb J, Mackenzie IR, et al. Neuropathology of a community sample of elderly demented and nondemented people. Brain Pathol 2000; 10:592–593.
8. Vermeer SE, Prins ND, den Heijer T, et al. Silent brain infarcts and the risk of dementia and cognitive decline. N Engl J Med 2003; 348:1215–1222.
9. Zekry D, Duyckaerts C, Moulias R, et al. Degenerative and vascular lesions of the brain have synergistic effects in dementia of the elderly. Acta Neuropathol 2002; 103:481–487.
10. Snowdon DA, Greiner LH, Mortimer JA, et al. Brain infarction and the clinical expression of Alzheimer disease. JAMA 1997; 277:813–817.
11. Schneider JA, Wilson RS, Bienias JL, et al. Cerebral infarctions and the likelihood of dementia from Alzheimer disease pathology. Neurology 2004; 13:1148–1155.
12. de Groot JC, de Leeuw FE, Oudkerk M, et al. Cerebral white matter lesions and cognitive function: the Rotterdam Scan Study. Ann Neurol 2000; 47:145–151.
13. de Groot JC, de Leeuw FE, Oudkerk M, et al. Cerebral white matter lesions and depressive symptoms in elderly adults. Arch Gen Psychiatry 2000; 57:1071–1076.
14. Longstreth WT Jr, Manolio TA, Arnold A, et al. Clinical correlates of white matter findings on cranial magnetic resonance imaging of 3301 elderly people. The Cardiovascular Health Study. Stroke 1996; 27:1274–1282.
15. Briley DP, Haroon S, Sergent S, et al. Does leukoaraiosis predict morbidity and mortality? Neurology 2000; 54:90–94.
16. Allen KV, Frier BM, Strachan MWJ. The relationship between type 2 diabetes and cognitive dysfunction: longitudinal studies and their methodological limitations. Eur J Pharmacol 2004; 490:169–175.
17. Arvanitakis Z, Wilson RS, Bienias JL, et al. Diabetes mellitus and risk of Alzheimer disease and decline in cognitive function. Arch Neurol 2004; 61:661–666.
18. Luchsinger JA, Tang MX, Stern Y, et al. Diabetes mellitus and risk of Alzheimer's disease and dementia with stroke in a multiethnic cohort. Am J Epidemiol 2001; 154:635–641.
19. Araki Y, et al. MRI of the brain in diabetes mellitus. Neuroradiology 1994; 36:101–103.
20. Eguchi K, Kario K, Shimada K. Greater impact of coexistence of hypertension and diabetes on silent cerebral infarcts. Stroke 2003; 34:2471–2474.
21. Heijer T, Vermeer SE, van Dijk, et al. Type 2 diabetes and atrophy of medial temporal lobe structures on brain MRI. Diabetologia 2003; 46: 1604–1610.
22. Inoue T, Fushimi H, Yamada Y, et al. Asymptomatic multiple lacunae in diabetics and non-diabetics detected by brain magnetic resonance imaging. Diabetes Res Clin Pract 1996; 31:81–86.
23. Kario K, Ishikawa J, Hoshide S, et al. Diabetic brain damage in hypertension: role of renin-angiotensin system. Hypertension 2005; 45:887–893.
24. Schmidt R, Launer LJ, Nilsson LG, et al. Magnetic resonance imaging of the brain in diabetes: the Cardiovascular Determinants of Dementia (CASCADE) study. Diabetes 2004; 53:687–692.
25. Williamson J, Miller ME, Bryan RN, et al. The Action to Control Cardiovascular Risk in Diabetes Memory in Diabetes Study (ACCORD-MIND): rationale, design, and methods. Am J Cardiol 2007; 99:112i–122i.
26. Mäntylä R, Erkinjuntti T, Salonen O, et al. Variable agreement between visual rating scales for white matter hyperintensities on MRI comparison of 13 rating scales in a poststroke cohort. Stroke 1997; 28:1614–1623.
27. Bryan RN, Manolio TA, Schertz LD, et al. A method for using MR to evaluate the effects of cardiovascular disease of the brain: the cardiovascular health study. Am J Neuroradiol 1994; 15:1625–1633.
28. De Groot JC, De Leeuw FE, Oudkerk M, et al. Periventricular cerebral white matter lesions predict rate of cognitive decline. Ann Neurol 2002; 52:335–341.
29. Benson RR, Guttmann CR, Wei X, et al. Older people with impaired mobility have specific loci of periventricular abnormality on MRI. Neurology 2002; 58:48–55.
30. Smith CD, Snowdon DA, Wang H, et al. White matter volumes and periventricular white matter hyperintensities in aging and dementia. Neurology 2000; 54:838–842.
31. Kamber M, Shingal R, Collins DL, et al. Model-based 3-D segmentation of multiple sclerosis lesions in magnetic resonance brain images. IEEE Trans Med Imaging 1995; 14:442–453.
32. Warfield S, Dengler J, Zaers J, et al. Automatic identification of grey matter structures from MRI to improve the segmentation of white matter lesions. J Image Guided Surg 1995; 1:326–338.
33. Udupa J, Wei L, Samarasekera S, et al. Multiple sclerosis lesion quantification using fuzzy-connectedness principles. IEEE Trans Med Imaging 1997; 16:598–609.
34. Welti D, Gerig G, Radu EW, et al. Spatio-temporal segmentation and characterization of active multiple sclerosis lesions in serial MRI data. Proc Information Proc Med Imaging 2001; 438–445.
35. Zijdenbos AP, Dawant BM, Margolin RA, et al. Morphometric analysis of white matter lesions in MR images: method and validation. IEEE Trans Med Imaging 1994; 13:716–724.

36. Alfano B, Brunetti A, Larobina M, et al. Automated segmentation and measurement of global white matter lesion volume in patients with multiple sclerosis. J Magn Reson Imaging 2000; 12:799–807.
37. Van Leemput K, Maes F, Vandermeulen D, et al. Automated segmentation of multiple sclerosis lesions by model outlier detection. IEEE Trans Med Imaging 2001; 20:677–688.
38. Udupa JK, Nyúl LG, Ge Y, et al. Multiprotocol MR image segmentation in multiple sclerosis: experience with over 1,000 studies. Acad Radiol 2001; 8:1116–1126.
39. Van Leemput K, Maes F, Vandermeulen D, et al. Automated segmentation of multiple sclerosis lesions my model outlier detection. KU-Leuven, 2000.
40. Wu Y, Warfield SK, Tan IL, et al. Automated segmentation of multiple sclerosis lesion subtypes with multichannel MRI. NeuroImage 2006; 32:1205–1215.
41. Zijdenbos AP, Forghani R, Evans AC. Automatic "pipeline" analysis of 3-D MRI data for clinical trials: application to multiple sclerosis. IEEE Trans MedImaging 2002; 21:1280–1291.
42. Warfield SK, Kaus M, Jolesz FA, et al. Adaptive, template moderated, spatially varying statistical classification. Med Image Anal 2000; 4:43–55.
43. Wei X, Warfield SK, Zou KH, et al. Quantitative analysis of MRI signal abnormalities of brain white matter with high reproducibility and accuracy. J Magn Reson Imaging 2002; 15:203–209.
44. Anbeek P, Vincken KL, van Osch MJ, et al. Probabilistic segmentation of white matter lesions in MR imaging. NeuroImage 2004; 21:1037–1044.
45. Admiraal-Behloul F, van den Heuvel DM, Olofsen H, et al. Fully automatic segmentation of white matter hyperintensities in MR images of the elderly. NeuroImage 2005; 28:607–617.
46. Yu S, Pham DL, Shen D, et al. Automatic segmentation of white matter lesions in T1-weighted brain MR images. IEEE International Symposium on Biomedical Imaging; Macro to Nano. 2001; 253–256.
47. Anbeek P, Vincken KL, van Bochove GS, et al. Probabilistic segmentation of brain tissue in MR imaging. NeuroImage 2005; 27:795–804.
48. Gerig G, Welti D, Guttmann CR, et al. Exploring the discrimination power of the time domain for segmentation and characterization of lesions in serial MR data. Med Image Anal 2000; 4:31–42.
49. Meier DS, Guttmann CRG. MRI time series modeling of MS lesion development. NeuroImage 2006; 32:531–537.
50. Jack CRJ, O'Brien PC. FLAIR histogram segmentation for measurement of leukoaraiosis volume. J Magn Res Imaging 2001; 14:668–676.
51. Mohamed FB, Vinitski S, Gonzalez CF, et al. Increased differentiation of intracranial white matter lesions by multispectral 3D-tissue segmentation: preliminary results. Magn Reson Imaging 2001; 19:207–218.
52. Viola P, Wells WM III. Alignment by maximization of mutual information. Proc Int Conf Computer Vision 1995. Los Alamitos, CA.
53. Smith SM, Jenkinson M, Woolrich MW, et al. Advances in functional and structural MR image analysis and implementation as FSL. NeuroImage 2004; 23:208–219.
54. Smith SM. Fast robust automated brain extraction. Human Brain Mapping 2002; 17:143–155.
55. Sled J, Zijdenbos A, Evans A. A nonparametric method for automatic correction of intensity nonuniformity in MRI data. IEEE Trans Med Imaging 1998; 17:87–97.
56. Gawne-Cain ML, O'Riordan JI, Coles A, et al. MRI lesion volume measurement in multiple sclerosis and its correlation with disability: a comparison of fast fluid attenuated inversion recovery (fFLAIR) and spin echo sequences. J Neurol Neurosurg Psychiatry 1998; 64:197–203.
57. Bastianello S, Bozzao A, Paolillo A, et al. Fast spin-echo and fast fluid-attenuated inversion-recovery versus conventional spin-echo sequences for MR quantification of multiple sclerosis lesions. Am J Neuroradiol 1997; 18:699–704.
58. Rovaris M, Comi G, Rocca MA, et al. Relevance of hypointense lesions on fast fluid-attenuated inversion recovery MR images as a marker of disease severity in cases of multiple sclerosis. Am J Neuroradiol 1999; 20:813–820.
59. Bakshi R, Caruthers SD, Janardhan V, et al. Intraventricular CSF pulsation artifact on fast fluid-attenuated inversion-recovery MR images: analysis of 100 consecutive normal studies. Am J Neuroradiol 2000; 21:503–508.
60. Tanaka N, Abe T, Kojima K, et al. Applicability and advantages of flow artifact–insensitive fluid-attenuated inversion-recovery MR sequences for imaging the posterior fossa. Am J Neuroradiol 2000; 21:1095–1098.
61. Vapnik VN. Statistical learning theory. New York, NY: Wiley, 1998.
62. Vapnik VN. The nature of statistical learning theory (statistics for engineering and information science). 2nd ed. New York, NY: Springer-Verlag, 1999.
63. Burges CJC. A tutorial on support vector machines for pattern recognition. Data Mining Knowledge Discovery 1998; 2:121–167.
64. Vapnik VN. Statistical learning theory. New York, NY: Wiley, 1998.
65. Lao Z, Shen D, Xue Z, et al. Morphological classification of brains via high-dimensional shape transformations and machine learning methods. Neuroimage 2004; 21:46–57.
66. Duda RO, Hart PE, Stork DG. Pattern classification. New York, NY: John Wiley and Sons, Inc, 2001.
67. Zweig MH, Campbell G. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. Clin Chem 1993; 39:561–577.