Contents lists available at ScienceDirect

Journal of the Neurological Sciences

journal homepage: www.elsevier.com/locate/jns

Machine learning for filtering out false positive grey matter atrophies in single subject voxel based morphometry: A simulation based study

Hernán C. Külsgaard [a,*], José I. Orlando [a], Mariana Bendersky [b,d], Juan P. Princich [b], Luis S. R. Manzanera [c], Alberto Vargas [c], Silvia Kochen [b], Ignacio Larrabide [a]

[a] *Pladema Institute - UNICEN/CONICET, Tandil, Buenos Aires, Argentina*
[b] *ENyS - UNAJ/CONICET, Florencio Varela, Buenos Aires, Argentina*
[c] *Hospital Clinic, Barcelona, Spain*
[d] *III Normal Anatomy Department, School of Medicine, University of Buenos Aires, Buenos Aires, Argentina*

ARTICLE INFO

Keywords:
Voxel based morphometry
Magnetic resonance imaging
Grey matter atrophies
Machine learning

ABSTRACT

Single subject VBM (SS-VBM), has been used as an alternative tool to standard VBM for single case studies. However, it has the disadvantage of producing an excessively large number of false positive detections. In this study we propose a machine learning technique widely used for automated data classification, namely Support Vector Machine (SVM), to refine the findings produced by SS-VBM. A controlled set of experiments was conducted to evaluate the proposed approach using three-dimensional T1 MRI scans from control subjects collected from the publicly available IXI dataset. The scans were artificially atrophied at different locations and with different sizes to mimic the behavior of neurological disorders. Results empirically demonstrated that the proposed method is able to significantly reduce the amount of false positive clusters ($p < 0.05$), with no statistical differences in the true positive findings ($p > 0.05$). This evidence was observed to be consistent for different atrophied areas and sizes of atrophies. This approach could be potentially be applied to alleviate the intensive manual analysis that radiologists and clinicians have to perform to filter out miss-detections of SS-VBM, increasing its usability for image reading.

1. Introduction

Alzheimer's disease (AD) [1], Frontotemporal Lobar Degeneration (FLD) [2] and Medial Temporal Lobe Epilepsy (MTLE) [3] are brain disorders associated with grey matter (GM) reduction in the temporal, parietal and/or frontal lobes. Voxel Based Morphometry (VBM) [4,5] is a brain image analysis methodology that has been widely utilized in the context of medical research on these diseases for the last two decades. Based on Magnetic Resonance Imaging (MRI), VBM plays a key role to understand the brain atrophy patterns that are relevant for these brain disorders [1–3,6–10].

VBM statistically compares voxels between two groups of MRI scans to determine if there exists differences in brain tissues densities between two groups. The output of VBM is a Statistical Parametric Map (SPM), where each voxel corresponds to the result of its associated test and therefore to the difference in densities of both groups for that specific voxel (Fig. 1(A)). Combining adjacent groups of voxels highlighted as statistically different by VBM allows to pose hypothesis and draw conclusions regarding the underlying anatomy of the groups. By definition, VBM relies on two comparison groups and can only be applied to assess their differences in tissue density. As a result, it has no direct application in daily clinical practice, e.g. to study differences between an individual and a comparison group.

Single subject VBM (SS-VBM) [11], also referred to as single case VBM [12,13] or individual VBM [14], has been presented as an alternative to standard VBM for single case studies, where a single subject is compared to a control group (Fig. 1(B)). Several articles have been published in recent years applying SS-VBM to identify regions of grey matter reduction [13,15,16]. Hedderich et al. [15] demonstrated the improvement in diagnostic accuracy and inter-rater agreement between experts using a combination of SS-VBM and total intracranial volume. Suzuki et al. [13] used SS-VBM to find characteristic patterns of volume loss to differenciate patients with chorea-acanthocytosis and Huntington's disease. SS-VBM has also been used in the study of functional MRI. In particular, Roswandowitz et al. [16] used it to asses behavioural differences in functional MRI on patients with apperceptive and

[*] Corresponding author.
E-mail address: hkulsgaard@pladema.exa.unicen.edu.ar (H.C. Külsgaard).

https://doi.org/10.1016/j.jns.2020.117220
Received 25 May 2020; Received in revised form 6 October 2020; Accepted 2 November 2020
Available online 6 November 2020

associative phonagnosia.

Since its introduction in 1999 [17], several authors investigated methodological alternatives to improve SS-VBM performance. Salmond et al. [18] assessed the relation between the false-positive rate and the smoothing applied to the images. More recently, Muhlau et al. [11] described the usage of one- and two-sample *t*-tests to preserve statistical validity.

The main disadvantage of SS-VBM is that, being restrained to comparing a single individual to a group, it cannot distinguish between pathological grey matter reduction and normal/anatomical inter-subject variability. This issue has been previously studied, for instance, by applying SS-VBM on healthy subjects, where a high number of significantly different densities are detected by the method [12,19]. These “false detections” do not follow a single, easy to recognise pattern, such as a predominant region of occurrence or a specific size. This makes difficult to identify and discard such false detections when analyzing the results. Some preliminary solutions to this problem have been studied in the past. Scarpazza et al. [20], for instance, found that the voxel-wise GM density might not follow a normal distribution, and therefore a non-parametric statistical test might be more suited than a classic two-sample *t*-test. Chen et al. [21] developed a different approach to single subject, combining classic VBM results with machine learning for detecting increased and reduced GM volume.

Machine learning (ML) [22] refers to a family of artificial intelligence methodologies that allows to automatise human tasks by training a computerised model using a set of training samples. Support vector machine (SVM) is a supervised learning technique widely used for automated data classification [23]. Given a set of training samples, the model automatically learns the best way to separate them according to their known, manually annotated, classes and each samples’ features. These features correspond to numerical representations of predefined characteristics for each example. During training, the hyperplane that better separates them in the multidimensional feature space is found. At test time, this hyperplane is used to identify the class associated with new given input samples whose classes are unknown.

In this paper we propose to use a SVM model to classify clusters resulting from SS-VBM into atrophies (**AT**) and not-atrophies (**NAT**). Our aim is to enhance the applicability of SS-VBM, e.g. for radiological reading, by automatically identifying false positive responses and preserving the true detections of the SS-VBM method. To ensure a proper and controlled evaluation of the proposed approach, our study is conducted using real brain MRI scans of control subjects on which grey matter atrophies were artificially introduced using simulation [24]. We then generated synthetic atrophies at specific locations of the brain to model the effect of different relevant disorders on GM density. This allowed us to know in advance the exact location of GM density alterations, which provided a ground truth for later assessment.

2. Method

2.1. Subject data

One hundred control subjects without known pathologies were evaluated in our study, all of them collected from the Hammersmith Hospital group of the IXI data set (http://brain-development.org/ixidataset/) (46 females / 54 males, age 35 ±9 years, ranging from 20 to 54 years). Their associated 256×256×128 MRI volumes correspond to T1 weighted images acquired using an Philips Intera 3 T (Philips Medical Systems, Best, The Netherlands) scanner at a voxel resolution of $0.9375\times0.9375\times1.2\,\text{mm}^3$, with a repetition time of 9.6 msec, echo time of 4.6 msec, 208 phase encoding steps, echo train length of 208, reconstruction diameter of 240.0 mm, and a flip angle of 8°.

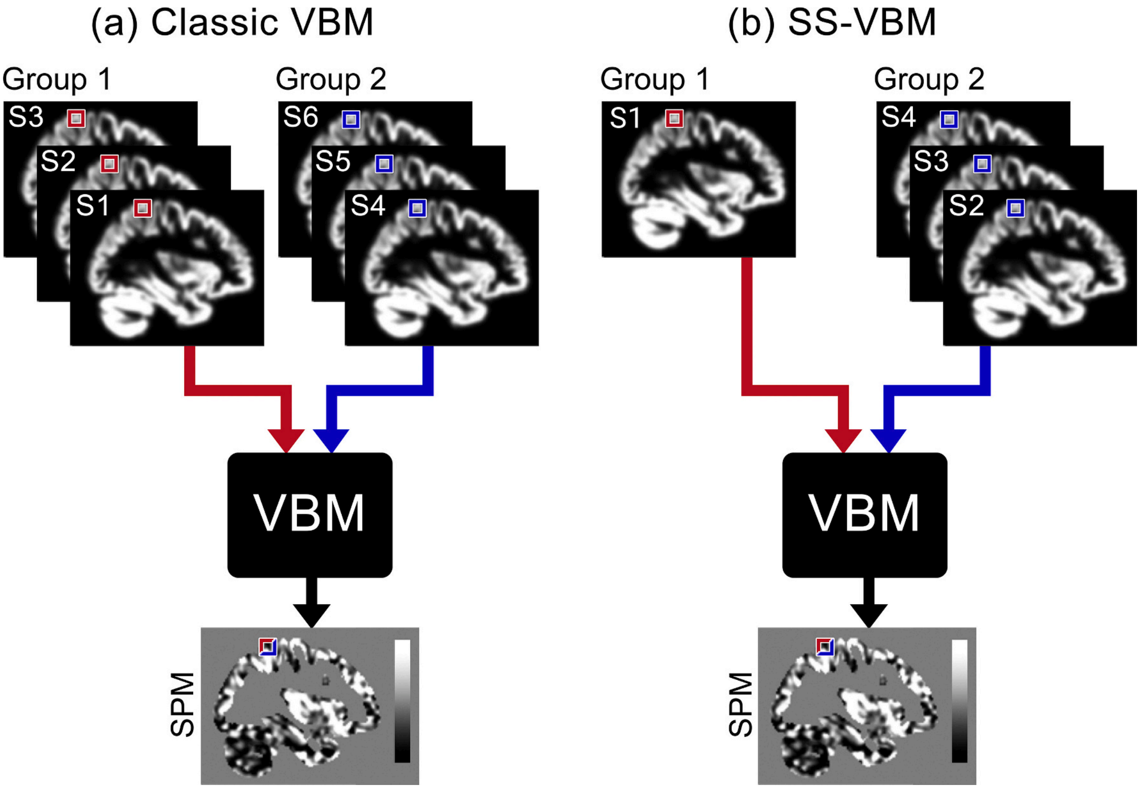

Fig. 1. VBM is a methodology for brain MRI analysis that statistically compares two groups of images at a voxel level. (A) The standard VBM approach compares two groups of subjects and the output is a Statistical Parametric Map (SPM) indicating the statistical significance of the differences between groups at each voxel. (B) Alternatively, SS-VBM statistically compares a single subject to a group, providing a SPM that reports for each voxels in the subject’s MRI if there is a significantly different tissue density with respect to the group. *VBM* voxel based morphometry, *SS-VBM* single subject voxel based morphometry, *SPM* statistical parametric map.

2.2. Atrophy simulation

The healthy subjects used in our study were artificially atrophied using the simulation software described in [24,25]. The simulator takes an input brain MRI scan and recreates an atrophy by shrinking the GM tissue in a specific location. As input it requires both, the 3D MRI image where the atrophy will be simulated, and a segmentation mask of the GM, white matter (WM) and cerebrospinal fluid (CSF). A spherical area is defined based on a given three-dimensional point and a given radius. A deformation force indicating the shrinking strength is applied on the area to simulate the atrophy. In our study we set this shrinking force parameter to 0.7, as suggested in the software documentation.

Three different locations were chosen to simulate atrophies on each subject, namely the hippocampus (HP), the parietal lobule (PL) and the superior frontal gyrus (SFG). These regions were selected due to their relevance to specific neurodegenerative diseases, more specifically Alzheimer's disease, Medial Temporal Lobe Epilepsy and Fronto-temporal Lobar Degeneration, respectively. For each subject, a total of 9 images with synthetic atrophies were derived by combining the three anatomical locations and the three atrophy sizes (20, 30 and 40 mm, Fig. 2).

2.3. Single subject voxel based morphometry (SS-VBM)

2.3.1. Preprocesing workflow

SS-VBM requires to preprocess the MRI datasets before performing statistical testing. First, each image was segmented into CSF, WM and GM, using the default segmentation functionality provided by SPM12 (Wellcome Trust Centre for Neuroimaging; http://www.fil.ion.ucl.ac.uk/spm/software/spm12/) running on Matlab (version 9.5.0, The MathWorks, Inc., Natick, USA). Only the GM tissue mask was used afterwards. A DARTEL template was created based on the original healthy GM segments to standardise all the subjects to a common reference space [26,27]. Flow fields were then generated for each subject, based on the created template, to register all of the GM segments. Finally, the resulting images were smoothed with a Gaussian kernel (σ=4 mm).

2.3.2. Statistical testing

We used two-sample *t*-tests in accordance to previous literature in the field [11]. In SS-VBM a single subject, playing the role of group 1 in standard VBM, is compared to several subjects, represented by group 2. A statistical test comparing each patient with simulated atrophies with the control group was performed, obtaining a SPM per individual. In each test, the analysed subject was removed from the control group to avoid comparing with its own healthy version. Additionally, age and sex were used as covariates for the test. The GM values extracted from the segmentation, in the range [0,1], were masked using an absolute threshold of 0.2. A statistical threshold of $p < 0.05$ was used, corrected with False Discovery Rate (FDR) from the implementation of CAT12 toolbox (Structural Brain Mapping Group; www.neuro.uni-jena.de/cat/).

The process described above was applied to each synthetically atrophied image, and compared to the healthy subjects (controls) group. The same process was performed for the healthy counterparts to establish a baseline for normal inter-subject anatomical variations.

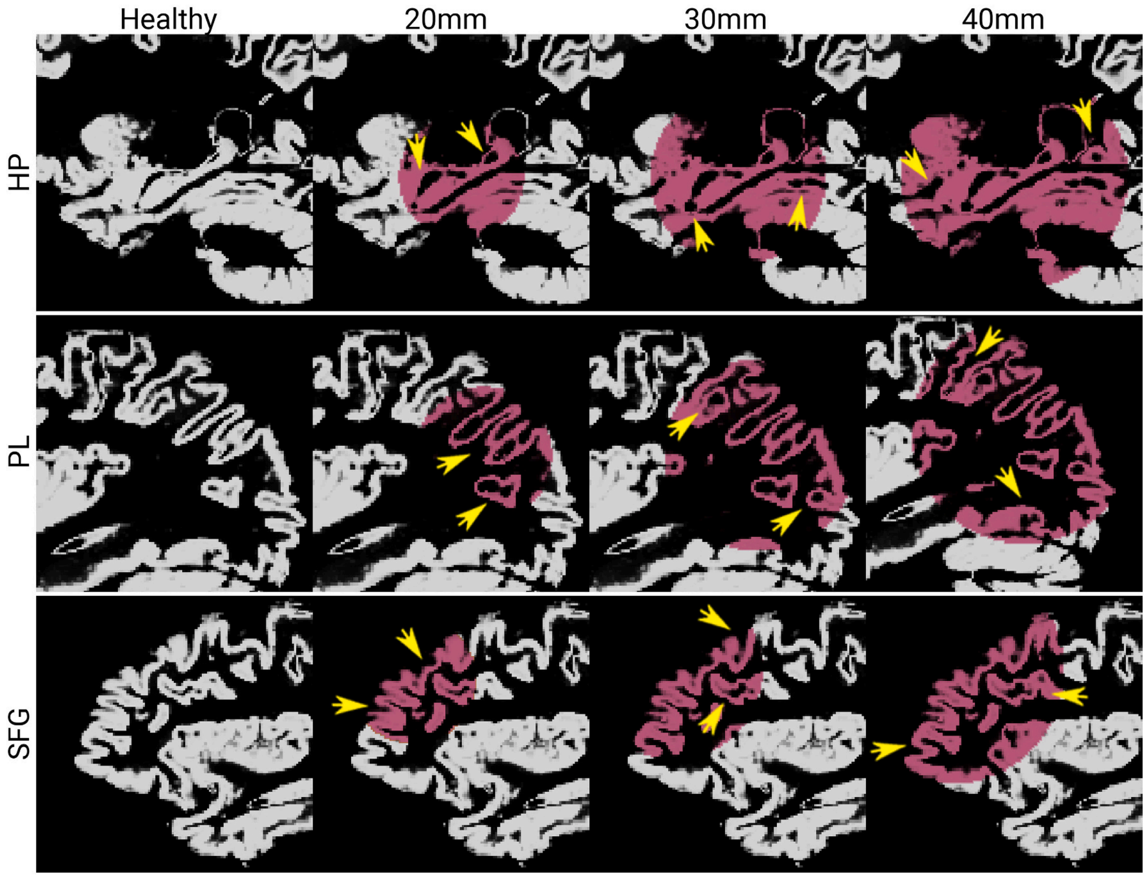

Fig. 2. Qualitative examples of simulated atrophies (highlighted in pink). From top to bottom: sagittal planes of the segmented GM from the same subject, zoomed on the HP (temporal lobe), the PL (parietal lobe) and the SFG (frontal lobe), respectively. From left to right: original healthy subject and simulated atrophies of sizes 20, 30 and 40 mm, respectively. Yellow arrows indicate the regions in which the atrophies are more evident. *GM* grey matter, *HP* hippocampus, *PL* parietal lobule, *SPG* superior frontal gyrus. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

2.4. Machine learning based cluster classification

Machine learning classification allows to automatically discriminate samples into different categories by discovering patterns in their features' distribution. Support Vector Machine (SVM) is a supervised learning technique widely used in computer science for data classification. Data samples are represented with feature vectors that determine their characteristics. Such features define a multidimensional space where each sample can be placed. Given a set of training samples and their known labels, a SVM model optimises a series of parameters to learn the hyperplane that better separates the categories in the feature space. This is done by minimising an objective function that penalises miss-classifications on the training set. In test time, the model classifies each new sample based on its location in the feature space with respect to the learned hyperplane. Furthermore, the signed Euclidean distance of the sample to the hyperplane can be interpreted as a score indicating the likelihood of belonging to the positive class. Scores above a certain threshold (generally 0) are labelled as belonging to the positive class, while those below the threshold are classified as negative samples.

We use this model to identify clusters that were miss-interpreted by SS-VBM as potential atrophies when being actually normal deviations in anatomy. Hence, our classification samples are the clusters obtained from SS-VBM tests, and we train a SVM model to classify them as actual atrophies (**AT**, positive class) or not-atrophies (**NAT**, negative class).

2.4.1. Data labelling

As mentioned in Section 2.3.2, two SS-VBM analysis were performed on each subject: one testing the differences between the original healthy image with the control group (Fig. 3(A)), and a second one between the image with synthetic atrophies and the control group (Fig. 3(B)). The second test was done to retrieve the set of clusters to be classified, while the former one was performed to determine which clusters were actually **NAT**.

Ground truth labels for **AT** and **NAT** clusters were assigned following a two-tier criterion (Fig. 3). To this end, both the original healthy MRI (Fig. 3(A)) and its synthetically atrophied counterpart (Fig. 3(B)) were compared using SS-VBM with respect to the control group. Two sets of clusters were identified, namely H (clusters detected in the healthy subject) and A (clusters detected in the atrophied counterpart). A cluster $a \in A$ was labelled as **AT** if and only if: (1) it lied within the region of the original MRI synthetically altered by the simulation procedure; and (2) it showed no coincidences with any other cluster $h \in H$. For the latter, two clusters $h \in H$ and $a \in A$ were assumed to be coincidences if their local maxima were too close to one another (less than 6 voxels apart, as measured by the Euclidean distance), and if they shared voxels (specificaly, more that 30% of voxels in the cluster).

2.4.2. Algorithm setup

Being SVM a supervised learning model, a training set is required to learn the model parameters. To this end, each dataset corresponded to a specific atrophy size and atrophy region and were randomly divided in a training and test sets, comprising 70 and 30 images, respectively. During the training step, 10-fold cross-validation was used on the training set for model selection and calibration, randomly splitting every fold into 70% for training and 30% to validate the performance of each specific configuration.

Accordingly, for each atrophy size and region, a SVM model was trained. We refer to these as single size models (SS models). Furthermore, three additional models were trained, one per atrophied region, combining all corresponding atrophy sizes. We refer to them as multiple size models (MS models).

Each cluster was represented by 11 different features obtained either from the SS-VBM output or from the prepocessed GM segmentation, as described in Table 1. Features were standardised to zero mean and unit variance using their own mean and standard deviation, as estimated from the training set.

We used the SVM implementation provided by the Statistical and Machine Learning Toolbox from Matlab. Hyperparameters of the SVM model, namely the scale of the radial basis function of the kernel and the box constraint, were fixed using a Bayesian Optimisation approach provided in that implementation.

A large number of **NAT** clusters were observed in our training set. As

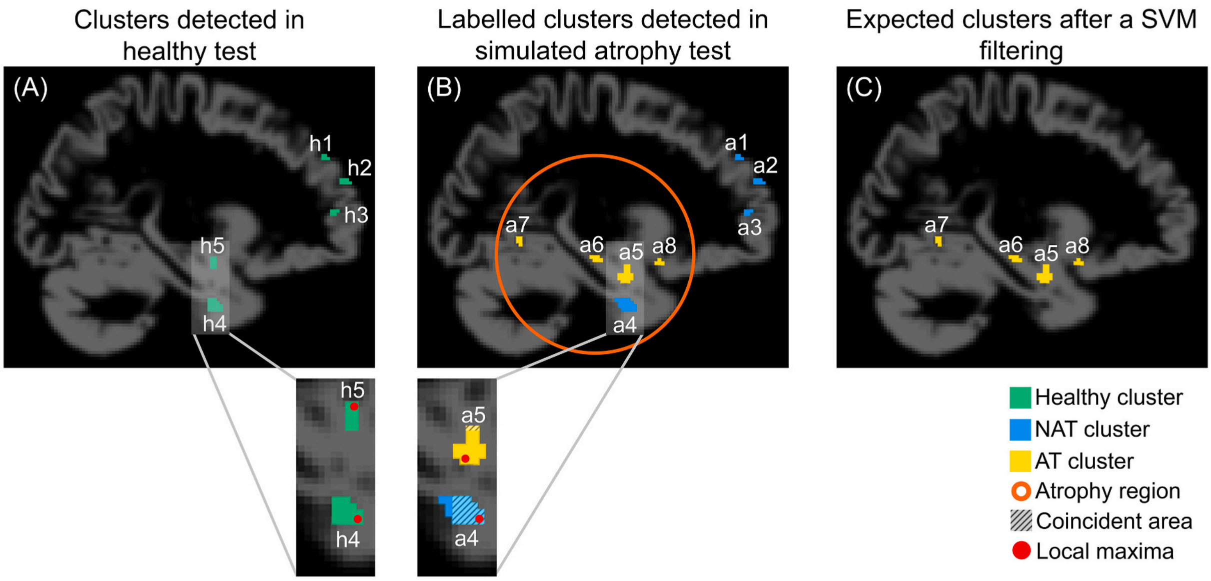

Fig. 3. Ground truth **AT** and **NAT** label assignment for SS-VBM detected clusters. (A) Clusters identified by the SS-VBM test on the original healthy image, corresponding to the subject 1 vs. the rest of the control group. These clusters were used to identify cluster coincidences for that individual. (B) Clusters identified by the SS-VBM test on the simulated atrophy image of subject 1 vs. the control group. Each of these clusters was checked for coincidence with the clusters on healthy subjects. In this example, cluster *a5* shared less than 30% of its volume with cluster *h5* and their local maxima were more than 6 voxels apart, so *a5* was labelled as **AT**. Cluster *a4* shared more than 30% of its volume with cluster *h4* and its local maxima were less than 6 voxels apart, then *a4* was labelled as **NAT**. Clusters *a6*, *a7* and *a8* were within the atrophy region and had no coincidences, so they were labelled as **AT**. Clusters *a1*, *a2* and *a3* were outside the atrophy region and were labelled as **NAT**. (C) Representation of an ideal filtering result for subject 1, where only **AT** were preserved and **NAT** were filtered out.

Table 1
List of features extracted for each cluster.

Feature No.	Data source	Description
1	SS-VBM output	Total number of voxels inside the cluster
2	SS-VBM output	Maximum value inside the cluster
3	SS-VBM output	Minimum value inside the cluster
4	SS-VBM output	Mean value inside the cluster
5	SS-VBM output	Median value inside the cluster
6	SS-VBM output	Standard deviation inside the cluster
7	Pre-processed GM	Maximum value of GM inside the cluster
8	Pre-processed GM	Minimum value of GM inside the cluster
9	Pre-processed GM	Mean value of GM inside the cluster
10	Pre-processed GM	Median value of GM inside the cluster
11	Pre-processed GM	Standard deviation value of GM inside the cluster

SS-VBM single subject voxel based morphometry, *GM* grey matter.

SVM models trained with imbalanced data sets are prone to be biased towards the majority class [28], the Different Error Costs technique was used to alleviate this effect. Hence, the cost of miss-classifying the minority class was set to 1 and the cost of the majority one fixed to the class ratio on the training set.

2.4.3. Evaluation metrics

The aim of this work is the reduction of **NAT** clusters in SS-VBM. First, we analysed the number of **NAT** clusters miss-detected per subject by SS-VBM and by the SVM models. Those were compared using 2-sample Kolmogorov-Smirnov tests, as none of the assumptions of the *t*-test were hold. Normality was checked using a Shapiro–Wilks test, and homoscedasticity was verified using Fisher test, with both of them resulting negative ($p > 0.05$). The same verification process was replicated for the distribution of the number of atrophy clusters correctly detected.

The metrics used to evaluate the classification performance of the SVM models were the receiver operating characteristic (ROC) curve and the area under the ROC curve (AUC-ROC). We used Star [29] to compare the ROC curves between the SS and MS models. Additionally, standard binary classification metrics such as sensitivity, specificity, precision, balanced accuracy and F1 score were reported, in accordance to the following equations:

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{1}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{2}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3}$$

$$\text{Balanced Accuracy} = \frac{\text{Sensitivity} + \text{Specificity}}{2} \tag{4}$$

$$\text{F1 Score} = \frac{2 \times \text{Sensitivity} \times \text{Precision}}{\text{Sensitivity} + \text{Precision}} \tag{5}$$

where TP, TN, FP and FN correspond to the number of true positive, true negative, false positive and false negative classifications, respectively. Sensitivity and specificity measure the ratio of detected **AT** and **NAT**, respectively, while precision refers to the ratio of correctly detected **AT**. Balanced accuracy indicates the percentage of correctly classified clusters. F1 score is defined as the harmonic mean of the sensitivity and precision and summarizes the overall **AT** detection performance.

3. Results

3.1. Cluster filtering

Fig. 4 depicts the number of **NAT** clusters per subject in the test sets, as obtained by SS-VBM before (grey) and after filtering the **AT** clusters using the proposed SVM approach (coloured by model). Results are grouped by anatomical region and by atrophy size (Fig. 4(A), (B) and (C)). Notice that the results for the SS and SM models were both included.

Similarly, Fig. 5 presents the number of correctly detected **AT** clusters per subject as obtained by SS-VBM before (grey) and after filtering using the SVM models (also coloured by model). In all cases, the clusters were retrieved taking into consideration the training/validation/test partition at subject level. Hence, no overlap exists between subjects on the training, validation and test sets.

The amount of **NAT** clusters after filtering with SVM models was significantly smaller than before ($p < 0.05$), regardless of the location of the simulated atrophies and their sizes. The differences between SS-VBM before and after SVM processing were not significant in the superior frontal gyrus for the MS model on the 30 mm atrophies test set ($p = 0.1$). On the other hand, the differences in the number of **AT** clusters detected before and after SVM filtering, were not significant in all the cases ($p > 0.05$) regardless the region or the size of the atrophy.

3.2. Classification performance of the SVM models

Table 2 shows the classification performance in the test sets, as obtained by all the SVM models. Results are grouped by atrophy region and size. To assess the performance of different SVM models in relation with the atrophy size used during training, each test set was classified using the SS model of the corresponding atrophy size and the MS model, both trained on atrophies at the same region, and compared (e.g. the 20 mm test set was classified by the 20 mm SS model and by the MS model, and so on). Additionally, the test set with multiple atrophy sizes was classified using all the SS and MS models. The filtering results in all the anatomical regions were also evaluated using ROC curves, as depicted in Figs. 6, 7 and 8.

In general, better results identifying **AT** and **NAT** were obtained for HP and PL compared to SFG. Although the sensitivity in SFG was higher than in HP and PL, specificity was remarkable lower, near to 0.5. In terms of sensitivity, specificity, precision and balanced accuracy, there was no predominant model. In the case of MS test sets, SS models obtained better results for all metrics. Observing the F1 scores for the 20 mm test set, values were lower than the rest of the test sets. For the AUC-ROC, SS models achieved higher values for the majority of the cases, but with a small margin.

No statistically significant differences were observed between the ROC curves of SS and MS models in the single atrophy size test sets, regardless of the atrophied area. For the MS test sets, the 40 mm SS model performed similar to the MS model in all the anatomical regions ($p > 0.05$). The 30 mm SS model achieved similar results, with the exception of the PL region ($p < 0.05$). Conversely, the ROC curves for the 20 mm SS models were significantly different to the MS models in all the anatomical regions ($p < 0.05$).

3.3. Qualitative results

Fig. 9 presents a qualitative analysis of the outputs of SS-VBM before and after **AT** filtering using the MS model. The GM density map corresponds to an atrophy simulated image, generated with a radius of 40 mm in the HP.

First, a ground truth label was generated for each cluster using the criteria described in Section 2.4.1. Five clusters were labelled as **AT** inside the simulated atrophy region (Fig. 9(A)). After the SS-VBM testing, the clusters in Fig. 9(B) were obtained. Notice that there were

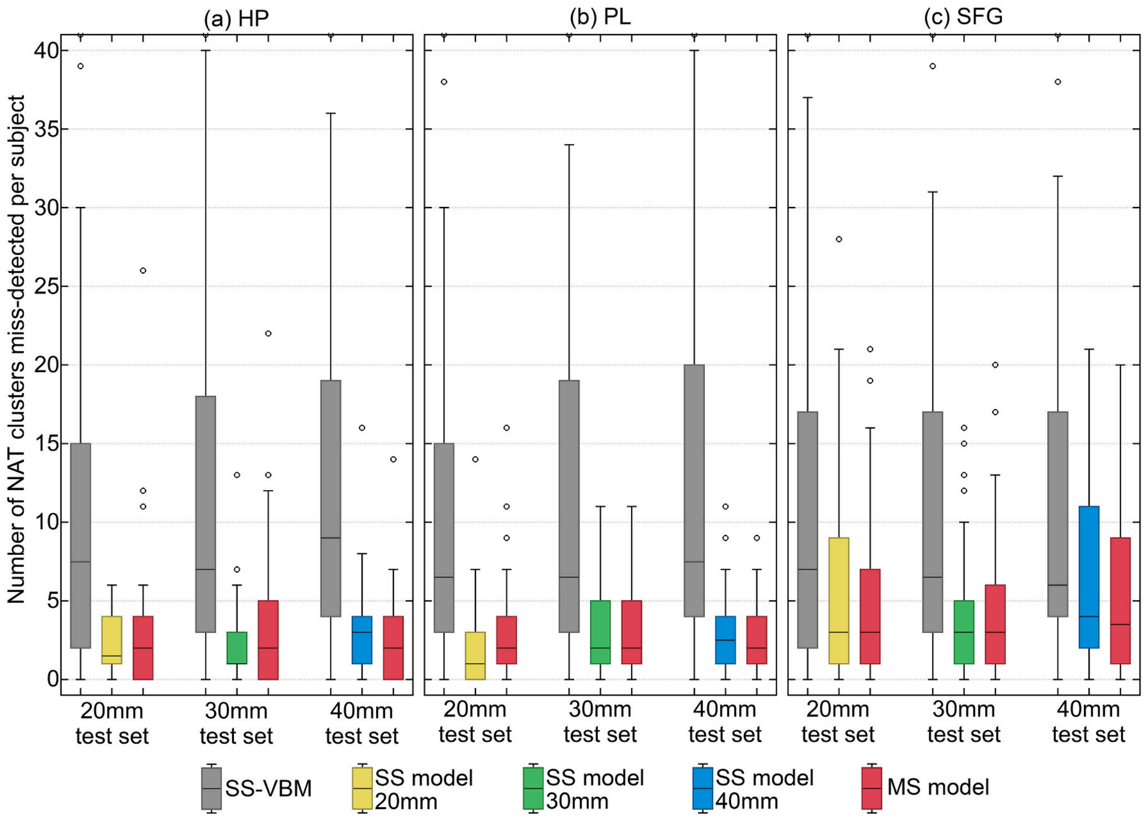

Fig. 4. Number of **NAT** clusters miss-detected per subject in the test sets, as obtained by SS-VBM before (grey box), after SVM filtering with SS models (yellow, green and blue boxes) and with the MS model (red box). From left to right: results in the data sets with atrophies simulated in (**a**) the hippocampus (HP), (**b**) the parietal lobule (PL) and (**c**) the superior frontal gyrus (SFG), with simulated atrophy sizes of 20 mm, 30 mm and 40 mm. *HP* hippocampus, *PL* parietal lobule, *SPG* superior frontal gyrus, *NAT* not-atrophy, *SS* single size, *MS* multiple size, *SS-VBM* single subject voxel based morphometry, *SVM* support vector machine. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

clusters detected outside the atrophy region (white arrows) that were not in the ground truth image.

Classifying the results with the MS model, we observed that the clusters outside the atrophy area were rated with the lowest scores while those inside were labelled with higher scores (Fig. 9(C)). After score thresholding, only the clusters within the atrophy region were classified as **AT**, except for one (see arrow in Fig. 9(D)). It can be seen that, by using a better adjustment of the score threshold, this cluster might be preserved.

4. Discussion

4.1. Filtering out NAT detection using SVMs

As previously discussed in the literature, the main limitation of SS-VBM is that it produces outputs with high false positive rates [12]. This is partially due to the data requirements of classic VBM that are relaxed in SS-VBM: more concretely, two groups are compared to each other instead of a single individual to a group. Although alternative approaches, such as the one in Scarpazza et al. [20], proposed changing the statistical tests as a proxy to reduce this effect, these modifications have not been adopted as a standard [13,15,16]. In this study, we introduce an alternative approach to improve SS-VBM performance by filtering out the **NAT** clusters using a SVM classifier. We empirically observed that using this machine learning approach allows to significantly reduce the number of **NAT** clusters by SS-VBM, without statistically affecting the number of detected **AT** clusters (Section 3.1). This observation is supported by the results presented in Figs. 4 and 5, and the corresponding statistical analysis. Therefore, the proposed SVM approach is expected not to affect the original performance of SS-VBM in terms of true positive detections, but most importantly to reduce the detected false positives. Notice, however, that the number of **NAT** is not reduced to zero in any case. This is not necessarily a problem or a limitation. SS-VBM is meant as a complementary tool for clinicians and/or radiologists to perform brain image analysis, i.e., an experienced clinician is expected to analyze SS-VBM results. In this sense, the proposed, novel filtering process should alleviate the reading task by drastically reducing the number of clusters to study.

The reasons to use simulated atrophies in this study is twofold: first, to ensure having gold standard labels for each cluster, and second, to control which anatomical regions were affected. In particular, atrophies were simulated at the hippocampus (temporal lobe), the parietal lobule (parietal lobe) and the superior frontal gyrus (frontal lobe). These are the most affected regions in clinically relevant disorders such as Alzheimer's disease, Frontotemporal Lobal Degeneration and Medial Temporal Lobe Epilepsy. It was experimentally observed that the filtering process using SVMs achieves similar performance, regardless of the anatomical region of analysis (Section 3.1). Notice, however, that each SVM model was trained using clusters produced by SS-VBM on images that were altered always on the same anatomical area. During model design, we observed a significant drop in performance when using cross-region classifiers (e.g. when evaluating a SVM model trained with atrophies detected in the HP area for detecting atrophies in the SFG, see Tables S1, S2 and S3 in the Online Supplementary Material).

Hence, we recommend to train a different SVM model for specific anatomical regions or brain disorders, and then select the corresponding one depending on the target application. Reciprocally, a radiologist or clinical reader might integrate the responses of multiple classifiers and compare their outputs and scores to further refine the output.

Another reason to justify the use of atrophy simulation is that it

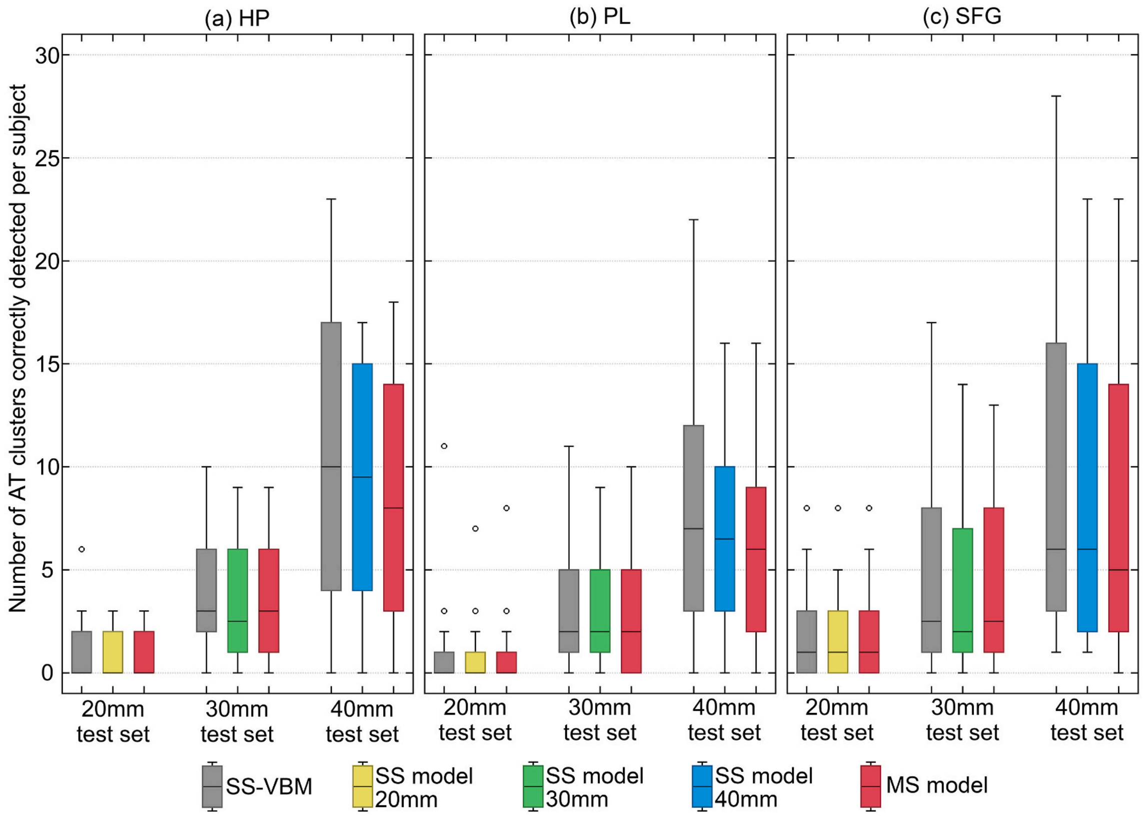

Fig. 5. Number of **AT** clusters correctly detected per subject in the test sets, as obtained by SS-VBM before (grey box), after SVM filtering with SS models (yellow, green and blue boxes) and with the MS model (red box). From left to right: results in the data sets with atrophies simulated in (**a**) the hippocampus (HP), (**b**) the parietal lobule (PL) and (**c**) the superior frontal gyrus (SFG), with simulated atrophy sizes 20 mm, 30 mm and 40 mm. *HP* hippocampus, *PL* parietal lobule, *SPG* superior frontal gyrus, *AT* atrophy, *SS* single size, *MS* multiple size, *SS-VBM* single subject voxel based morphometry, *SVM* support vector machine. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 2
Classification performance of the SVM models in terms of binary classification metrics.

Atrophy location	Test set	20 mm		30 mm		40 mm		MS			
	SVM Model	20 mm	MS	30 mm	MS	40 mm	MS	20 mm	30 mm	40 mm	MS
HP	Sensitivity	0.765	**0.794**	0.857	**0.902**	**0.862**	0.804	0.547	0.687	**0.887**	0.813
	Specificity	**0.810**	0.716	**0.834**	0.737	0.765	**0.803**	0.847	**0.852**	0.696	0.740
	Precision	**0.268**	0.203	**0.589**	0.488	0.728	**0.749**	0.577	**0.640**	0.528	0.545
	Bal. Acc.	**0.787**	0.755	**0.845**	0.819	**0.813**	0.803	0.697	0.769	**0.792**	0.777
	F1 Score	**0.397**	0.323	**0.698**	0.633	**0.789**	0.775	0.562	**0.662**	0.662	0.652
	AUC-ROC	0.854	**0.862**	**0.917**	0.906	**0.897**	0.893	0.798	0.868	0.876	**0.881**
PL	Sensitivity	0.750	**0.812**	**0.869**	0.860	**0.803**	0.762	0.611	0.746	**0.794**	0.778
	Specificity	**0.833**	0.748	0.766	**0.786**	0.786	**0.826**	**0.858**	0.761	0.741	0.784
	Precision	**0.282**	0.220	0.497	**0.517**	0.678	**0.711**	**0.579**	0.500	0.495	0.536
	Bal. Acc.	**0.791**	0.780	0.817	**0.823**	**0.795**	0.794	0.734	0.754	0.767	**0.781**
	F1 Score	**0.410**	0.347	0.633	**0.646**	**0.736**	0.735	0.595	0.599	0.610	**0.634**
	AUC-ROC	0.880	**0.886**	0.899	**0.914**	**0.882**	0.872	0.798	0.840	**0.872**	0.869
SFG	Sensitivity	0.941	**0.961**	0.864	**0.924**	**0.912**	0.866	0.772	0.765	**0.933**	0.866
	Specificity	0.576	**0.579**	**0.654**	0.589	0.538	**0.599**	0.604	**0.654**	0.529	0.596
	Precision	0.230	**0.234**	**0.460**	0.434	0.568	**0.590**	0.436	**0.467**	0.440	0.459
	Bal. Acc.	0.759	**0.770**	**0.759**	0.757	0.725	**0.733**	0.688	0.710	**0.731**	0.731
	F1 Score	0.369	**0.377**	**0.600**	0.591	0.700	**0.702**	0.557	0.580	0.598	**0.600**
	AUC-ROC	**0.864**	0.820	**0.841**	0.835	**0.814**	0.796	0.761	0.805	**0.813**	0.802

SVM support vector machine, *MS* multiple size, *AUC-ROC* area under the curve of the receiver operating characteristic, *HP* hippocampus, *PL* parietal lobule, *SFG* superior frontal gyrus, *Bal.Acc.* balanced accuracy.

allowed us to produce atrophies with different sizes and intensities. This setting is extremely difficult to achieve using real atrophy cases, requiring a considerably large set of volunteers plus the labelling effort, which would be difficult to standardise and quantify. By producing synthetic atrophies, we have an effect that is highly reproducible and easy to quantify on the resulting images, allowing us to draw more accurate conclusions. The sizes of the atrophies were chosen based on the corresponding lobes volume, from a small atrophy affecting only a few regions to a big atrophy covering the majority of the lobe.

In this sense, it is important to highlight that the proposed SVM filtering approach performed equally well regardless of the atrophy size (Table 2, Figs. 4 and 5). Furthermore, models trained with clusters from

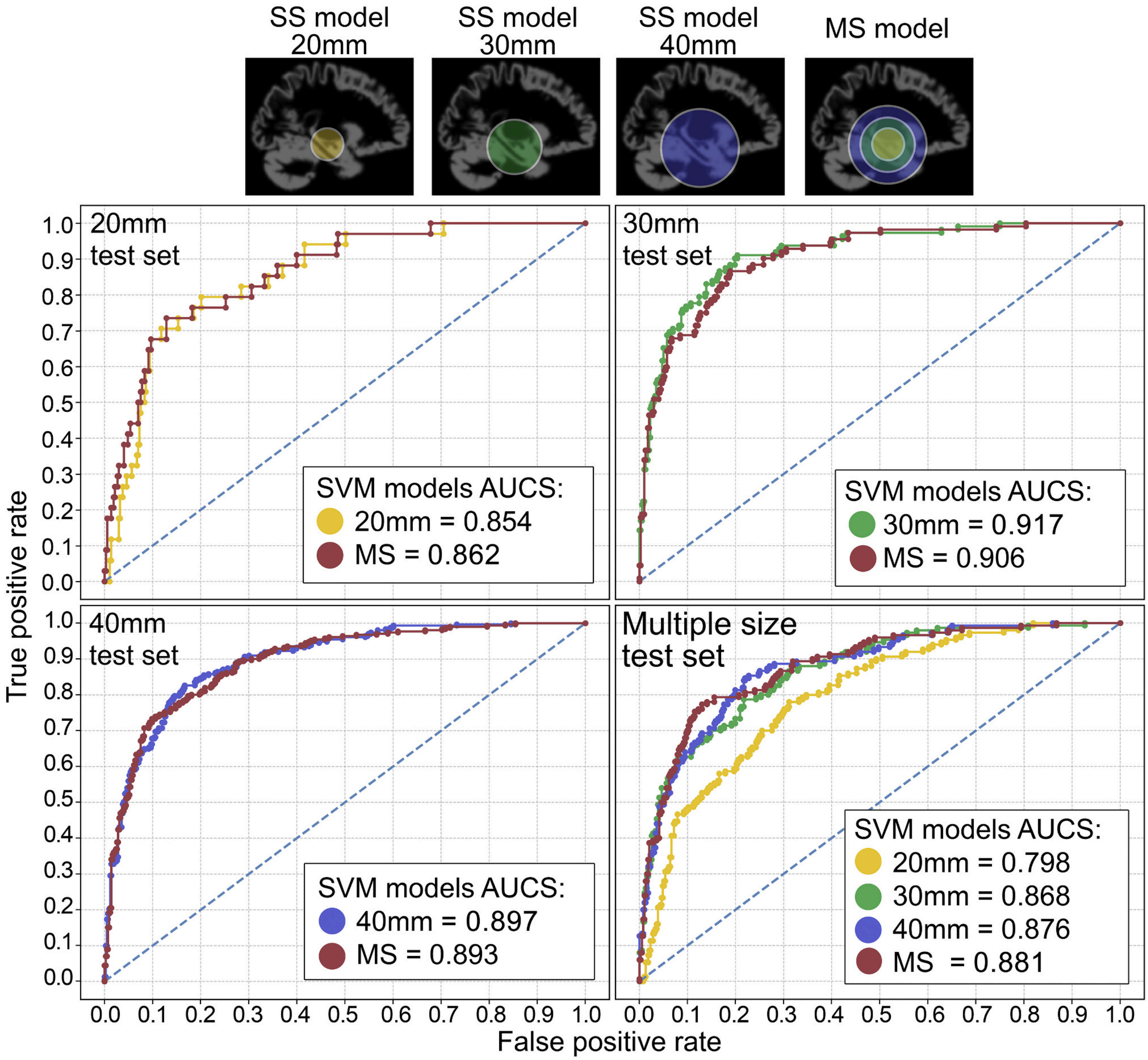

Fig. 6. Classification results at the HP atrophy in the test sets composed only for single atrophy sizes of 20, 30, and 40 mm and the MS one, as measured by ROC curves. Classification was performed using SS models (20, 30 and 40 mm) and the MS models. *SS* single size, *MS* multiple size, *HP* hippocampus, *PL* parietal lobule, *SPG* superior frontal gyrus, *ROC* receiver operating characteristic.

atrophies from multiple sizes (MS models), performed similarly to the individual ones (SS models), except for the combined test set (Figs. 6, 7 and 8), in which the MS model outperformed the others, or was the second best in its worst performance. This indicates that a MS model might be more suitable for a real clinical scenario where atrophies of different sizes are expected to happen.

Complementary, all tests were performed using smoothing kernel widths of 2 mm and 6 mm. The results, evaluated in terms of the same metrics, were similar to those obtained using the 4 mm kernel (see Tables S4 and S5 in the Online Supplementary Material). The latter was chosen as it was the best in terms of the balance between the presented metrics. The 4 mm kernel width also guaranteed a larger number of patients with detections in the atrophy region.

4.2. *Potential clinical applications*

One potential clinical application for the proposed approach is to aid in MRI reading in population based studies. By means of SS-VBM, radiologists and physicians could obtain a first set of clusters indicating abnormal areas or potential atrophies. Subsequently, the proposed SVM approach would be used to assign each cluster a score and rank them according to the confidence of the model. Finally, clinicians would visually assess the data and decide if they should be ignored or further studied.

4.3. *Limitations*

The proposed approach is intended to filter out **NAT** clusters miss-detected by SS-VBM and, as such, it is not able to detect atrophies originally missed during the SS-VBM phase of the analysis. Hence, the sensitivity of the final model depends on the ability of SS-VBM to identify potential atrophy regions. Nevertheless, SS-VBM has proven to have high sensitivity with poor specificity so, in principle, it could be argued that SVM filtering will increase the specificity of the model by reducing the number of false positive detections.

When analyzing the classification results at a cluster level (Table 2),

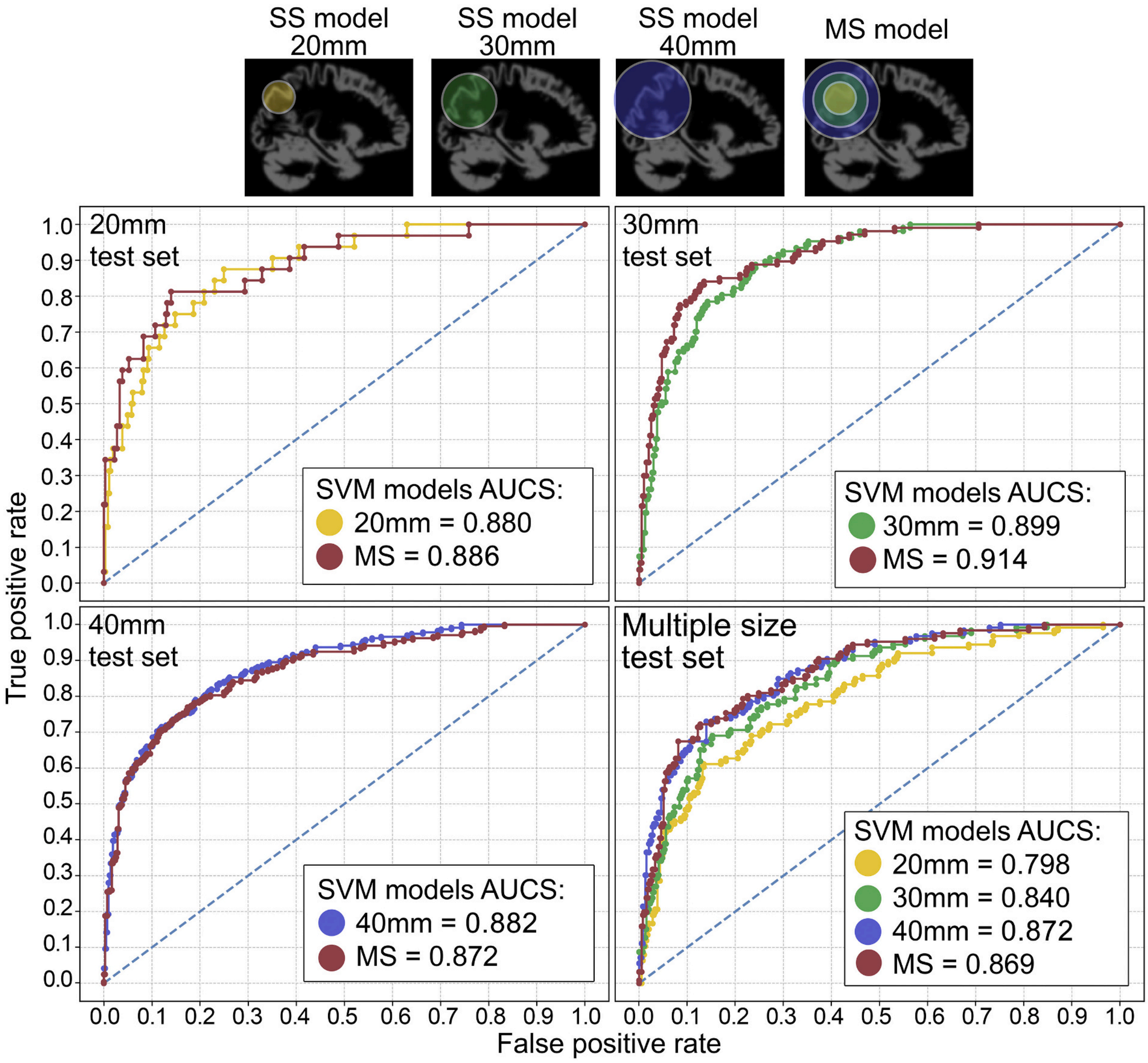

Fig. 7. Classification results at the PL atrophy in the test sets composed only for single atrophy sizes of 20, 30, and 40 mm and the MS one, as measured by ROC curves. Classification was performed using SS models (20, 30 and 40 mm) and the MS models. *SS* single size, *MS* multiple size, *HP* hippocampus, *PL* parietal lobule, *SPG* superior frontal gyrus, *ROC* receiver operating characteristic.

it was observed that, although the sensitivity was high, the precision was low, hence affecting the corresponding F1 scores. This indicates that there is still a large amount of **NAT** clusters that are missclassified, mostly in cases with small size atrophies (e.g. 20 mm). We believe this is mostly a consequence of outlier cases exhibiting a larger number of **NAT** clusters than the general trend (see Fig. 4). Incorporating other features targetting other aspects of the cluster such as their density in specific regions of the brain might aid to alleviate this limitations. Nevertheless, it is worth mentioning that, when analyzing the results in a per patient basis (Fig. 4), the distribution of **NAT** clusters after the filtering process is significantly reduced with respect to the original SS-VBM counterpart, without significantly affecting the amount of **AT** clusters.

Notice that one potential source of false positives could be the identification of normal variations in the anatomy. This is due to the fact that SS-VBM does not incorporate explicit constraints to limit the discoveries only to pathological differences. Every anatomical structure can vary to a certain extent from the usual presentation, which does not necessarily render it abnormal or pathological. The identification of cerebral sulci is not straightforward. Interruptions and branches complicate the identification of the same sulcus in different brains. This poses difficulties for parcellating the cortex with automated methods. Often, variations are discovered in the structure, origin, branching pattern of sulci or presence of additional or accessory gyri, resulting in high inter- and intrasubject variability in radiological readings [30–32]. The importance of such anatomical variations in the clinical setting is based on the fact that they represent a variant of the normal presentation. They can present diagnostic dilemmas, affect surgical procedures or, in this case, be interpretated as AT clusters in the SS-VBM analysis. The incorporation of our SVM filtering stage might alleviate this issue, although it is out of the scope of this study. Future analysis should evaluate if the proposed method is able to reduce the amount of false positive clusters associated with non-pathological anatomical variations.

From a machine learning point of view, it is worth mentioning that our proposed framework is general enough to be implemented using other classifiers and features. In this study we propose a proof-of-concept approach in which the combination of a relatively easy to compute set of features and a standard linear classification model is able to effectively reduce the number of false positive detections of SS-VBM. Future work could be focused on improving our results by applying ensemble based models such as Random Forests [33], using a larger set of image and non-image based features (e.g. computed from the original

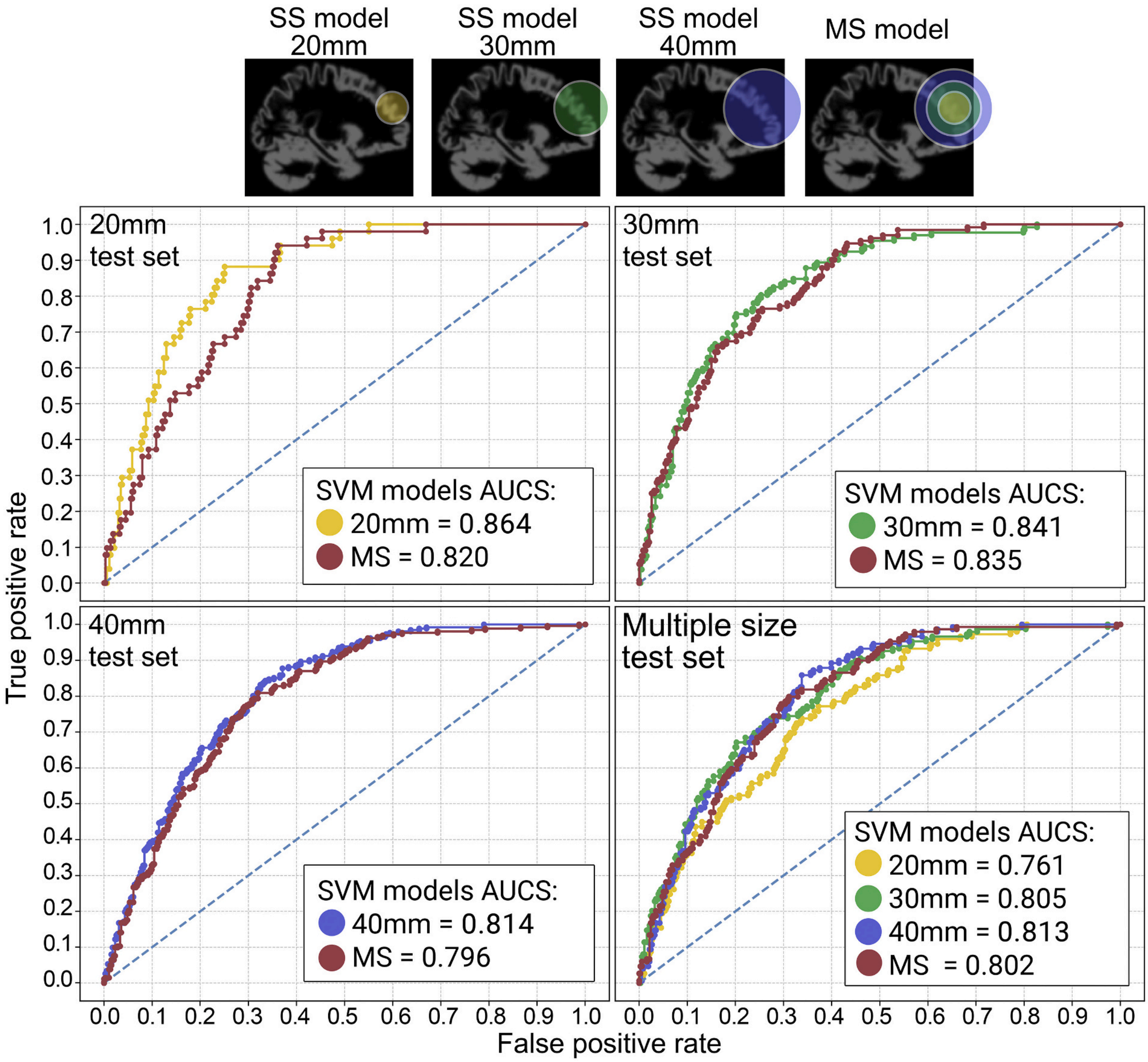

Fig. 8. Classification results at the SFG atrophy in the test sets composed only for single atrophy sizes of 20, 30, and 40 mm and the MS one, as measured by ROC curves. Classification was performed using SS models (20, 30 and 40 mm) and the MS models. *SS* single size, *MS* multiple size, *HP* hippocampus, *PL* parietal lobule, *SPG* superior frontal gyrus, *ROC* receiver operating characteristic.

input MRI scan or even retrieved from clinical records). Furthermore, other more complex artificial intelligence approaches such as convolutional neural networks [34] might be exploited to avoid manually engineering features by learning them automatically.

The results of this simulation based study supports this application scenario where the atrophy sizes used for training and testing were not random, like in clinical practice, but fixed values. Yet, a complementary evaluation with real cases is requiered to analyze if it is universally applicable. Notice, however, that one of the key difficulties of such study relies on the availability of gold standard labels indicating the areas with pathological GM anomalies. Previous studies have pointed out the difficulty of such analysis [35]. One way to overcome this limitation could be to appeal to the consensus of a board of experts to annotate the MRI scans, and take the majority voting of their responses as a ground truth label. Yet, this does not ensure a perfect gold standard annotation. As previously pointed out, the proposed evaluation based on synthetic atrophy allowed bypassing the need of these expensive annotations, ensuring an unbiased estimation of the model performance.

5. Conclusion

In this paper we introduced a machine learning based approach to remove false positive clusters from SS-VBM. Our careful evaluation on a series of artificially atrophied MRI datasets showed that the proposed method is able to significantly reduce the amount of false positive clusters identified by SS-VBM, while preserving the true positive findings. These results were consistent for every atrophy region and size evaluated. Although further evaluation of true diseased cases is still needed, we envision that this approach could be applied in the future to provide objective information to alleviate the intensive manual analysis that radiologists and clinicians perform, by filtering out miss-detections by SS-VBM.

Availability of data and material

This study was conducted using MRI data from the Hammersmith Hospital subset of the publicly available IXI dataset, which is accessible through the following URL address: https://brain-development.org/ixi-dataset/.

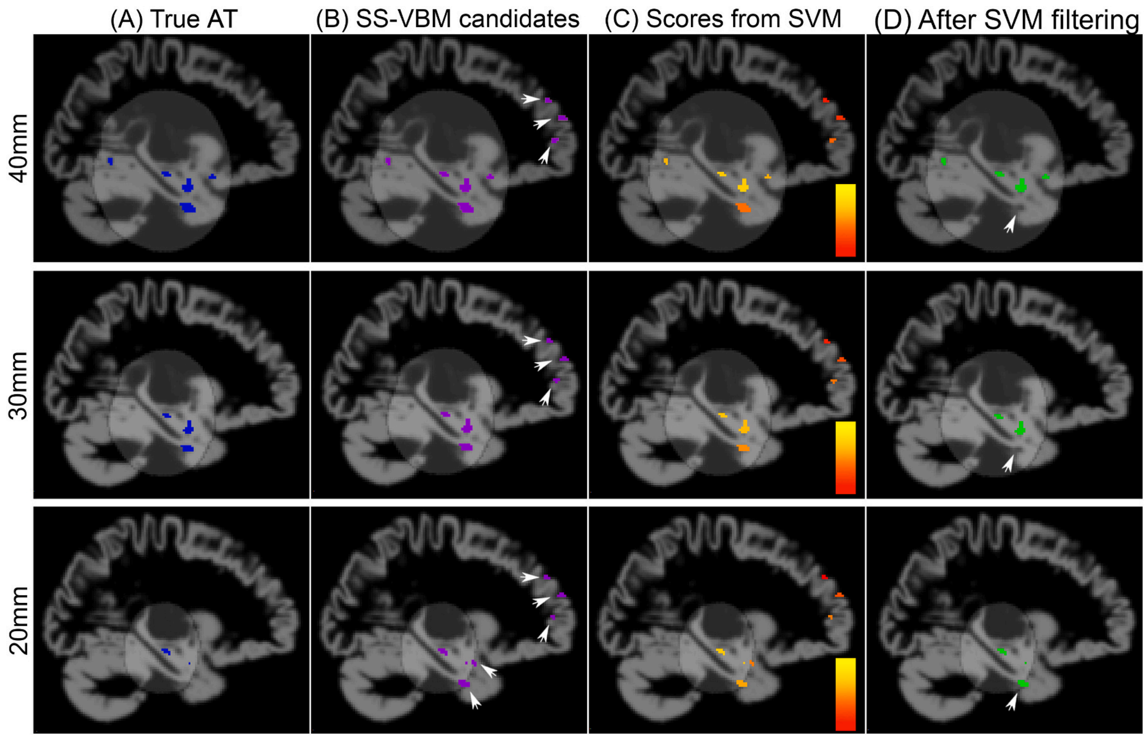

Fig. 9. Qualitative results of the same subject with simulated atrophies of 40 mm (first row), 30 mm (second row) and 20 mm (third row) in the HP region. (A) Ground truth clusters labelled as **AT** (highlighted in blue). (B) Clusters indicated as potential atrophies by SS-VBM analysis (highlighted in magenta). White arrows point to **NAT** (false positive) cluster detections. (C) Scores assigned by the MS SVM classifier to each of the clusters. Low scores indicate potential **NAT** clusters, while high scores correspond to potential **AT** clusters. Notice that depending on the threshold selected it would be possible to filter out the **NAT** clusters without significantly affecting the number of **AT** clusters. (D) Results after filtering using the MS SVM model (highlighted in green). Notice that the **NAT** clusters were removed, but also one **AT** cluster for the 40 and 30 mm atrophies (pointed with arrows). In the case of 20 mm, the arrow indicates a miss-detected **NAT** cluster. *HP* hippocampus, *AT* atrophy, *NAT* not-atrophy, *SS-VBM* single subject voxel based morphometry, *SVM* support vector machine, *MS* multiple size. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Authors' contributions

Conception and study design: H.C.K., J.I.O. and I.L.; Data processing and analysis: H.C.K., M.B., J.P.P., L.S.R.M., A.V. and S.K.; Statistical analysis: H.C.K., J.I.O. and I.L.; Interpretation of results: H.C.K., J.I.O., M.B., J.P.P., L.S.R.M., A.V., S.K. and I.L.; Drafting the manuscript work and revising it: All authors; Approval of final version to be published: All authors; Agreement to be accountable for the integrity and accuracy of all aspects of the work: All authors.

Declaration of Competing Interest

The authors declare that they have no conflict of interest.

Acknowledgments

This study was partially funded by Agencia Nacional de Promoción Científica y Tecnológica (ANPCyT), Argentina; PICT 2016-0116 and by an NVIDIA, United states hardware grant. H.C.K. PhD studies are funded by a CONICET Doctoral Research scholarship, Argentina.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jns.2020.117220.

References

[1] Catriona D. Good, Ingrid S. Johnsrude, John Ashburner, Richard N.A. Henson, Karl J. Friston, Richard S.J. Frackowiak, A voxel-based morphometric study of ageing in 465 normal adult human brains, Neuroimage 14 (1) (2001) 21–36, https://doi.org/10.1006/nimg.2001.0786.

[2] J.M.S. Pereira, G.B. Williams, J. Acosta-Cabronero, G. Pengas, M.G. Spillantini, J. H. Xuereb, J.R. Hodges, P.J. Nestor, Atrophy patterns in histologic vs clinical groupings of frontotemporal lobar degeneration, Neurology 72 (19) (2009) 1653–1660, https://doi.org/10.1212/WNL.0b013e3181a55fa2.

[3] Leonardo Bonilha, Chris Rorden, Gabriela Castellano, Fernando Cendes, Li M. Li, Voxel-based morphometry of the thalamus in patients with refractory medial temporal lobe epilepsy, Neuroimage 25 (3) (2005) 1016–1021, https://doi.org/10.1016/j.neuroimage.2004.11.050.

[4] John Ashburner, Karl J. Friston, Voxel-based morphometry—the methods, Neuroimage 11 (6) (2000) 805–821, https://doi.org/10.1006/nimg.2000.0582.

[5] John Ashburner, Karl J. Friston, Why voxel-based morphometry should be used, Neuroimage 14 (6) (2001) 1238–1243, https://doi.org/10.1006/nimg.2001.0961.

[6] Antonio Giorgio, Luca Santelli, Valentina Tomassini, Rose Bosnell, Steve Smith, Nicola De Stefano, Heidi Johansen-Berg, Age-related changes in grey and white matter structure throughout adulthood, Neuroimage 51 (3) (2010) 943–951, https://doi.org/10.1016/j.neuroimage.2010.03.004.

[7] Sean J. Colloby, John-Paul Taylor, et al., Patterns of cerebellar volume loss in dementia with lewy bodies and alzheimer's disease: a vbm-dartel study, Psychiatry Res. Neuroimaging 223 (3) (2014) 187–191, https://doi.org/10.1016/j.pscychresns.2014.06.006.

[8] Jonathan D. Rohrer, Gerard R. Ridgway, Marc Modat, Sebastien Ourselin, Simon Mead, Nick C. Fox, Martin N. Rossor, Jason D. Warren, Distinct profiles of brain atrophy in frontotemporal lobar degeneration caused by progranulin and tau mutations, Neuroimage 53 (3) (2010) 1070–1076, https://doi.org/10.1016/j.neuroimage.2009.12.088.

[9] Suzanne G. Mueller, Kenneth D. Laxer, Nathan Cashdollar, Shannon Buckley, Crystal Paul, Michael W. Weiner, Voxel-based optimized morphometry (vbm) of gray and white matter in temporal lobe epilepsy (tle) with and without mesial temporal sclerosis, Epilepsia 47 (5) (2006) 900–907, https://doi.org/10.1111/j.1528-1167.2006.00512.x.

[10] C.L. Yasuda, M.E. Morita, A. Alessio, A.R. Pereira, M.L.F. Balthazar, A.L.F. Costa, A. L.C. Costa, T.A. Cardoso, L.E. Betting, C.A.M. Guerreiro, et al., Relationship between environmental factors and gray matter atrophy in refractory mtle, Neurology 74 (13) (2010) 1062–1068, https://doi.org/10.1212/WNL.0b013e3181d76b72.

[11] M. Mühlau, A.M. Wohlschläger, C. Gaser, M. Valet, A. Weindl, S. Nunnemann, A. Peinemann, T. Etgen, R. Ilg, Voxel-based morphometry in individual patients: a

pilot study in early Huntington disease, Am. J. Neuroradiol. 30 (3) (2009) 539–543, https://doi.org/10.3174/ajnr.A1390.

[12] Cristina Scarpazza, Sartori Giuseppe, M.S. De Simone, Andrea Mechelli, When the single matters more than the group: very high false positive rates in single case voxel based morphometry, Neuroimage 70 (2013) 175–188, https://doi.org/10.1016/j.neuroimage.2012.12.045.

[13] Fumio Suzuki, Noriko Sato, Miho Ota, Atsuhiko Sugiyama, Yoko Shigemoto, Emiko Morimoto, Yukio Kimura, Noritaka Wakasugi, Yuji Takahashi, Akinori Futamura, et al., Discriminating chorea-acanthocytosis from huntington's disease with single-case voxel-based morphometry analysis, J. Neurol. Sci. 408 (2020) 116545, https://doi.org/10.1016/j.jns.2019.116545.

[14] Olivier Colliot, Neda Bernasconi, Najmeh Khalili, Samson B. Antel, Véronique Naessens, Andrea Bernasconi, Individual voxel-based analysis of gray matter in focal cortical dysplasia, Neuroimage 29 (1) (2006) 162–171, https://doi.org/10.1016/j.neuroimage.2005.07.021.

[15] Dennis M. Hedderich, Michael Dieckmeyer, Tiberiu Andrisan, Marion Ortner, Lioba Grundl, Simon Schön, Per Suppa, Tom Finck, Kornelia Kreiser, Claus Zimmer, et al., Normative brain volume reports may improve differential diagnosis of dementing neurodegenerative diseases in clinical practice, Eur. Radiol. (2020) 1–9, https://doi.org/10.1007/s00330-019-06602-0.

[16] Claudia Roswandowitz, Stefanie Schelinski, Katharina von Kriegstein, Developmental phonagnosia: linking neural mechanisms with the behavioural phenotype, NeuroImage 155 (2017) 97–112, https://doi.org/10.1016/j.neuroimage.2017.02.064.

[17] F.G. Woermann, S.L. Free, M.J. Koepp, S.M. Sisodiya, J.S. Duncan, Abnormal cerebral structure in juvenile myoclonic epilepsy demonstrated with voxel-based analysis of mri, Brain 122 (11) (1999) 2101–2108, https://doi.org/10.1093/brain/122.11.2101.

[18] C.H. Salmond, J. Ashburner, F. Vargha-Khadem, A. Connelly, D.G. Gadian, K. J. Friston, Distributional assumptions in voxel-based morphometry, Neuroimage 17 (2) (2002) 1027–1030, https://doi.org/10.1006/nimg.2002.1153.

[19] Hernan Claudio Kulsgaard, Delfina Braggio, Mariana Bendersky, Lucia Alba Ferrara, Ignacio Larrabide, A study of single subject vbm and dartel on healthy subjects, in: 14th International Symposium on Medical Information Processing and Analysis 10975, International Society for Optics and Photonics, 2018, https://doi.org/10.1117/12.2511457 page 109750Q.

[20] Cristina Scarpazza, Thomas E. Nichols, Donato Seramondi, Camille Maumet, Giuseppe Sartori, Andrea Mechelli, When the single matters more than the group (ii): addressing the problem of high false positive rates in single case voxel based morphometry using non-parametric statistics, Front. Neurosci. 10 (6) (2016), https://doi.org/10.3389/fnins.2016.00006.

[21] Shihui Chen, Jian Zhang, Xiaolei Ruan, Kan Deng, Jianing Zhang, Dongfang Zou, Xiaoming He, Feng Li, Bin Guo, Hongwu Zeng, et al., Voxel-based morphometry analysis and machine learning based classification in pediatric mesial temporal lobe epilepsy with hippocampal sclerosis, Brain Imag. Behav. (2019) 1–10, https://doi.org/10.1007/s11682-019-00138-z.

[22] Trevor Hastie, Robert Tibshirani, Jerome Friedman, The Elements of Statistical Learning. Springer Series in Statistics, Springer New York Inc., New York, NY, USA, 2001, https://doi.org/10.1007/978-0-387-84858-7.

[23] Corinna Cortes, Vladimir Vapnik, Support-vector networks, Mach. Learn. 20 (3) (1995) 273–297, https://doi.org/10.1023/A:1022627411411.

[24] Bilge Karaçali, Christos Davatzikos, Simulation of tissue atrophy using a topology preserving transformation model, IEEE Trans. Med. Imaging 25 (5) (2006) 649–652, https://doi.org/10.1109/TMI.2006.873221.

[25] Bilge Karaçali, Christos Davatzikos, Estimating topology preserving and smooth displacement fields, IEEE Trans. Med. Imaging 23 (7) (2004) 868–880, https://doi.org/10.1109/TMI.2004.827963.

[26] John Ashburner, A fast diffeomorphic image registration algorithm, Neuroimage 38 (1) (2007) 95–113, https://doi.org/10.1016/j.neuroimage.2007.07.007.

[27] John Ashburner, Karl J. Friston, Computing average shaped tissue probability templates, Neuroimage 45 (2) (2009) 333–341, https://doi.org/10.1016/j.neuroimage.2008.12.008.

[28] VASILE Palade, Class imbalance learning methods for support vector machines, in: Imbalanced Learning: Foundations, Algorithms, and Applications 83, 2013, https://doi.org/10.1002/9781118646106.ch5.

[29] Ismael A. Vergara, Tomás Norambuena, Evandro Ferrada, Alex W. Slater, Francisco Melo, Star: a simple tool for the statistical comparison of roc curves, BMC Bioinforma 9 (1) (2008) 265, https://doi.org/10.1186/1471-2105-9-265.

[30] J. Rademacher, V.S. Caviness Jr., H. Steinmetz, A.M. Galaburda, Topographical variation of the human primary cortices: implications for neuroimaging, brain mapping, and neurobiology, Cereb. Cortex 3 (4) (1993) 313–329, https://doi.org/10.1093/cercor/3.4.313.

[31] Michio Ono, Stefan Kubik, Chad D. Abernathey, Atlas of the Cerebral Sulci. Tps, 1990.

[32] Albert L. Rhoton Jr., The cerebrum, Neurosurgery 51 (suppl_4) (2002), https://doi.org/10.1097/00006123-200210001-00002. S1–1.

[33] Leo Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32, https://doi.org/10.1023/A:1010933404324.

[34] Yann LeCun, Yoshua Bengio, Geoffrey Hinton, Deep learning, Nature 521 (7553) (2015) 436–444, https://doi.org/10.1038/nature14539.

[35] Swati Sharma, Vincent Noblet, François Rousseau, Fabrice Heitz, Lucien Rumbach, J.-P. Armspach, Evaluation of brain atrophy estimation algorithms using simulated ground-truth data, Med. Image Anal. 14 (3) (2010) 373–389, https://doi.org/10.1016/j.media.2010.02.002.