



**HAL**  
open science

## Spatial Decision Forests for MS Lesion Segmentation in Multi-Channel Magnetic Resonance Images

Ezequiel Geremia, Olivier Clatz, Bjoern H. Menze, Ender Konukoglu, Antonio Criminisi, Nicholas Ayache

► **To cite this version:**

Ezequiel Geremia, Olivier Clatz, Bjoern H. Menze, Ender Konukoglu, Antonio Criminisi, et al.. Spatial Decision Forests for MS Lesion Segmentation in Multi-Channel Magnetic Resonance Images. *NeuroImage*, 2011, 57 (2), pp.378-390. 10.1016/j.neuroimage.2011.03.080 . inria-00616194

**HAL Id: inria-00616194**

**<https://inria.hal.science/inria-00616194>**

Submitted on 26 Jul 2013

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Spatial Decision Forests for MS Lesion Segmentation in Multi-Channel Magnetic Resonance Images

Ezequiel Geremia<sup>a,c</sup>, Olivier Clatz<sup>a</sup>, Bjoern H. Menze<sup>a,b</sup>, Ender Konukoglu<sup>c</sup>, Antonio Criminisi<sup>c</sup>, Nicholas Ayache<sup>a</sup>

<sup>a</sup>*Asclepios Research Project, INRIA Sophia-Antipolis, France*

<sup>b</sup>*Computer Science and Artificial Intelligence Laboratory, MIT, USA*

<sup>c</sup>*Machine Learning and Perception Group, Microsoft Research Cambridge, UK*

---

## Abstract

A new algorithm is presented for the automatic segmentation of Multiple Sclerosis (MS) lesions in 3D Magnetic Resonance (MR) images. It builds on a discriminative random decision forest framework to provide a voxel-wise probabilistic classification of the volume. The method uses multi-channel MR intensities (T1, T2, FLAIR), knowledge on tissue classes and long-range spatial context to discriminate lesions from background. A symmetry feature is introduced accounting for the fact that some MS lesions tend to develop in an asymmetric way. Quantitative evaluation of the proposed methods is carried out on publicly available labeled cases from the MICCAI MS Lesion Segmentation Challenge 2008 dataset. When tested on the same data, the presented method compares favorably to all earlier methods. In an a posteriori analysis, we show how selected features during classification can be ranked according to their discriminative power and reveal the most important ones.

*Keywords:*

Multi-sequence MRI, Segmentation, Multiple Sclerosis, Random Forests, MICCAI Grand Challenge 2008

---

## 1. Introduction

Multiple Sclerosis (MS) is a chronic, inflammatory and demyelinating disease that primarily affects the white matter of the central nervous system. Automatic detection and segmentation of MS lesions can help diagnosis and patient follow-up. It offers an attractive alternative to manual segmentation which remains a time-consuming task and suffers from intra- and inter-expert variability. MS lesions, however, show a high variability in appearance and shape which makes automatic segmentation a challenging task. MS lesions lack common intensity and texture characteristics, their shapes are variable and their location within the white matter varies across patients.

A variety of methods have been proposed for the automatic segmentation of MS lesions. For instance, in (Anbeek et al., 2004) and (Admiraal-Behloul et al., 2005), the authors propose to segment white matter signal abnormalities by using an intensity-based  $k$ -nearest neighbors method with spatial prior and a fuzzy inference system, respectively. A similar classifier combined with a template-driven segmentation was proposed in (Wu et al., 2006) to segment MS lesions into three different subtypes (enhancing lesions, T1 black holes, T2 hyperintense lesions). A false positive reduction based on a rule-based method, a level set method and a support vector machine classifier is presented in (Yamamoto et al., 2010) along with a multiple-gray level thresholding technique. Many general purpose brain tissue and brain tumor segmentation approaches can be modified easily for MS lesion segmentation. In (Bricq et al., 2008b), for example, the authors present an unsupervised algorithm based on hidden Markov chains for brain tissue segmentation in

MR sequences. The method provides an estimation of the proportion of white matter (WM), grey matter (GM) and cerebro-spinal fluid (CSF) in each voxel. It can be extended for MS lesions segmentation by adding an outlier detector (Bricq et al., 2008a).

Generative methods were proposed consisting in a tissue classification by means of an expectation maximization (EM) algorithm. For instance, the method presented in (Datta et al., 2006) aims at segmenting and quantifying black holes among MS lesions. The EM algorithm can be modified to be robust against lesion affected regions, its outcome is then parsed in order to detect outliers which, in this case, coincide with MS lesions (Van Leemput et al., 2001). Another approach consists in adding to the EM a partial volume model between tissue classes and combining it with a Mahalanobis thresholding which highlights the lesions (Dugas-Phocion et al., 2004). Morphological postprocessing on resulting regions of interest was shown to improve the classification performance (Souplet et al., 2008). In (Freifeld et al., 2009), a constrained Gaussian mixture model is proposed, with no spatial prior, to capture the tissue spatial layout. MS lesions are detected as outliers and then grouped in an additional tissue class. Final delineation is performed using probability-based curve evolution. Multi-scale segmentation can be combined with discriminative classification to take into account regional properties (Akselrod-Ballin et al., 2006). Beyond the information introduced via the spatial prior atlases, these methods are limited in their ability to take advantage of long-range spatial context in the classification task.

To overcome this shortcoming, we propose the use of an ensemble of dis-

criminative classifiers. Our algorithm builds on the random decision forest framework which has multiple applications in bioinformatics (Menze et al., 2009), and, for example, more recently also in the image processing community (Andres et al., 2008; Yi et al., 2009; Criminisi et al., 2010). Adding spatial and multi-channel features to this classifier proved effective in object recognition (Shotton et al., 2009), brain tissue segmentation in MR images (Yi et al., 2009), myocardium delineation in 3D echocardiography (Lempitsky et al., 2009) and organ localization in CT volumes (Criminisi et al., 2010).

Applying multi-channel and context-rich random forest classification to the MS lesion segmentation problem is novel, to our knowledge. The presented classifier also exploits a specific discriminative symmetry feature which stems from the assumption that the healthy brain is approximately symmetric with respect to the mid-sagittal plane and that MS lesions tend to develop in asymmetric ways. We then show how the forest combines the most discriminative channels for the task of MS lesion segmentation.

## 2. Materials

This section describes the data, algorithms and notations which are referred to in the rest of the article.

### *2.1. MICCAI Grand Challenge 2008 dataset*

The results in this article rely on a strong evaluation effort. This section presents the MICCAI<sup>1</sup> Grand Challenge 2008 datasets, which is the largest

---

<sup>1</sup>MICCAI is the annual international conference on Medical Image Computing and Computer Assisted Intervention.

dataset publicly available, and explains the way our method is compared against the winner of the challenge (Souplet et al., 2008). In the rest of the article, the MICCAI Grand Challenge 2008 on MS Lesions Segmentations will be referred as MSGC.

### *2.1.1. Presentation*

The MSGC (Styner et al., 2008a) aims at evaluating and comparing algorithms in an independent and standardized way for the task of MS lesion segmentation. The organizers make publicly available two datasets through their website. A dataset of labeled MR images which can be used to train a segmentation algorithm, and an unlabeled dataset on which the algorithm should be tested. The website offers to quantitatively evaluate the segmentation results on the unlabeled dataset using the associated private ground truth database, and to publish the resulting scores. This project is an original initiative to provide an unbiased comparison between MS lesions segmentation algorithms. In the rest of the article, the dataset for which labels are publicly available will be referred to as *public* dataset, whereas the dataset for which data is not available will be referred to as *private* dataset.

### *2.1.2. Data*

The public dataset contains 20 cases, 10 from the Children’s Hospital in Boston (CHB) and 10 from the University of North Carolina (UNC), which are labeled by a CHB expert rater. The private dataset contains 25 cases, 15 from CHB and 10 from UNC. The private dataset was annotated by a single expert rater at CHB and jointly by 2 expert raters UNC. For each case, the centers provided 3 MR volumes: a T1-weighted image, a T2-weighted

image and a FLAIR image. These were co-registered and sampled to fit the isotropic  $0.5 \times 0.5 \times 0.5 \text{ mm}^3$  resolution.

Both private and public datasets gather anatomical images from two different centers, CHB and UNC, and shows high variability in intensity contrast (cf. Section 5.2), image noise and bias field. Both public and private datasets contain highly heterogeneous cases and could thus be considered as realistic test cases.

### *2.1.3. Evaluation*

Quantitative evaluation is carried out on the private dataset using a set of known metrics defined in (Styner et al., 2008a) and summed up in Table 1. The two full sets of expert segmentations were used as reference for method comparison.

### *2.1.4. Top-ranked methods*

The challenge results highlight four top-ranked methods each reflecting a different approach to the task of MS lesion segmentation. A  $k$ -nearest neighbor classification of brain tissue relying on spatial location and intensity value was proposed in (Anbeek et al., 2008). This method provides a voxel-wise probabilistic classification of MS lesions. In (Bricq et al., 2008a), the authors present an unsupervised segmentation algorithm based on a hidden Markov chain model. The method takes into account neighborhood information, MR sequences and probabilistic priors in order to delineate MS lesions. Alternatively, the iterative method proposed in (Shiee et al., 2008, 2010) jointly performs brain tissue classification and MS lesion segmentation by combining statistical and topological atlases. Finally, in (Souplet et al., 2008), the

Name	Definition	Unit	Best	Worse
$TNR$	$\frac{TN}{FP+TN}$	%	100	0
$TPR$	$\frac{TP}{TP+FN}$	%	100	0
$FPR$	$\frac{FP}{FP+TN}$	%	0	100
$PPV$	$\frac{TP}{TP+FP}$	%	100	0
$VO$	$\frac{Vol(Seg \cap GT)}{Vol(Seg \cup GT)}$	%	0	100
$VD$	$\frac{Vol(Seg) - Vol(GT)}{Vol(GT)}$	%	0	$< \infty$
$SD$	$\frac{\sum_{v \in \partial(Seg)} \min_{u \in \partial(GT)} d(u, v) + \sum_{v \in \partial(GT)} \min_{u \in \partial(Seg)} d(u, v)}{card(Seg \cup GT)}$	$mm$	0	$< \infty$

Table 1: **The evaluation metrics** true negative rate ( $TNR$ ), true positive rate ( $TPR$ ), false positive rate ( $FPR$ ) and positive predictive value ( $PPV$ ) are defined using the following notations: true positives ( $TP$ ), true negatives ( $TN$ ), false positives ( $FP$ ) and false negatives ( $FN$ ). The volume overlap ( $VO$ ) and the relative absolute volume difference ( $VD$ ) evaluates the differences between the segmentation ( $Seg$ ) and the ground truth ( $GT$ ) by computing their volume ( $Vol$ ). The average symmetric surface distance ( $SD$ ) measures how close the segmentation and the ground truth are from each other using the Euclidean distance  $d$  on the set of boundary voxels noted  $\partial$ . The *best*, respectively *worse*, column contains the metric score of the perfect segmentation, respectively of a completely-off segmentation.

authors show that a global threshold on the FLAIR MR sequence, inferred using an EM brain tissue classification, suffices to detect most MS lesions. The final segmentation is then constrained to appear in the white matter by applying morphological operations.

The method proposed in (Souplet et al., 2008) won the MICCAI MS Segmentation Challenge 2008. For this specific method, the segmentation results on public and private datasets were made available by the authors



and will be used as reference. In the rest of the article, methods will be identified by their reference.

### 2.2. Data preprocessing

We sub-sample and crop the images so that they all have the same size,  $159 \times 207 \times 79$  voxels, and the same resolution,  $1 \times 1 \times 2 \text{ mm}^3$ . Sub-sampling and cropping intend to reduce the time spent learning the classifier. The preprocessing procedure corrects for RF acquisition field inhomogeneities (Prima et al., 2001) and performs inter-subject intensity calibration (Rey, 2002). Spatial normalization is also performed by aligning the mid-sagittal plane with the center of the images (Prima et al., 2002).

Spatial prior is added by registering the MNI atlas (Evans et al., 1993) to the anatomical images, each voxel of the atlas providing the probability of belonging to WM, GM and CSF (cf. Figure 1). The Image Fusion module of MedINRIA (INRIA, 2010) is used to perform the affine registration of the MNI atlas onto every single case.

### 2.3. Notations

The multi-channel aspect of the method presented in this article requires to carefully define and name each channel. MR images from the MSGC dataset will be noted  $I_s$  where the index  $s \in \{T1, T2, FLAIR\}$  stands for an MR sequence. Registered spatial priors will be noted  $P_t$  where the index  $t \in \{WM, GM, CSF\}$  stands for a brain tissue class.

Although having different semantics, anatomical images and spatial priors will be treated under the unified term *signal channel* and denoted  $C \in \{I_{T1}, I_{T2}, I_{FLAIR}, P_{WM}, P_{GM}, P_{CSF}\}$ .

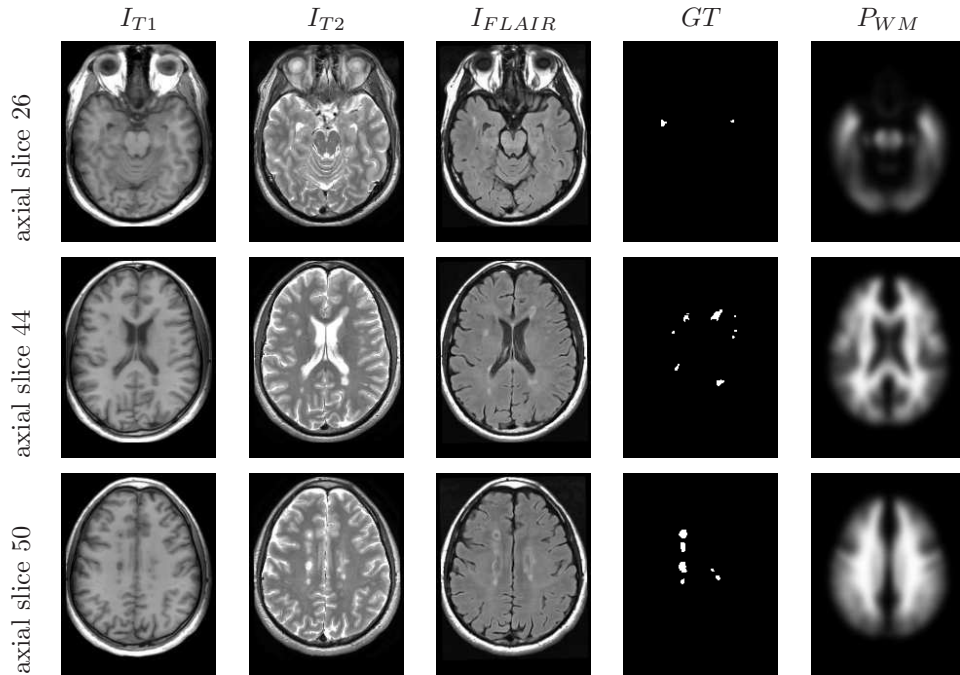


Figure 1: **Case CHB07 from the public MSGC dataset.** From top to bottom: three axial slices of the same patient. From left to right: preprocessed T1-weighted ( $I_{T1}$ ), T2-weighted ( $I_{T2}$ ) and FLAIR MR images ( $I_{FLAIR}$ ), the associated ground truth  $GT$  and the registered white matter atlas ( $P_{WM}$ ).

The data consists of a collection of voxel samples described by their spatial position  $\mathbf{x} = (x, y, z)$ . Voxels can be evaluated in all available signal channels. The value of the voxel  $\mathbf{x}$  in channel  $C$  is denoted  $C(\mathbf{x})$ .

### 3. Methods

This section describes our adaptation of the random decision forests to the segmentation of MS lesions and illustrates the visual features employed.

### 3.1. Context-rich decision forest

Our detection and segmentation problem can be formalized as a binary classification of voxel samples into either background or lesions. This classification problem is addressed by a supervised method: discriminative random decision forest, an ensemble classifier using decision trees as base classifiers. Decision trees are discriminative classifiers which are known to suffer from over-fitting (Breiman et al., 1984). A random decision forest (Amit and Geman, 1997) achieves better generalization by growing an ensemble of many independent decision trees on a random subset of the training data and by randomizing the features made available to each node during training (Breiman, 2001).

#### 3.1.1. Forest training

The training data consists of a set of labeled voxels  $\mathcal{T} = \{\mathbf{x}_k, Y(\mathbf{x}_k)\}$  where the label  $Y(\mathbf{x}_k)$  is given by an expert. When asked to classify a new image, the classifier aims to assign every voxel  $\mathbf{x}$  in the volume a label  $y(\mathbf{x})$ . In our case,  $y(\mathbf{x}) \in \{0, 1\}$ , 1 for lesion and 0 for background.

The forest has  $T$  components with  $t$  indexing each tree. During training, all observations (voxels)  $\mathbf{x}_k$  are pushed through each of the trees. Each internal node applies a binary test (Shotton et al., 2009; Yi et al., 2009; Lempitsky et al., 2009; Criminisi et al., 2010) as follows:

$$t^{\tau_{low}, \tau_{up}, \theta}(\mathbf{x}_k) = \begin{cases} true, & \text{if } \tau_{low} \leq \theta(\mathbf{x}_k) < \tau_{up} \\ false, & \text{otherwise} \end{cases} \quad (1)$$

where  $\theta$  is a function identifying the visual feature extracted at position  $\mathbf{x}_k$ . There are several ways of defining  $\theta$ , either as a local intensity-based

average, local spatial prior or context-rich cue. These are investigated in more detail in the next section. The value of the extracted visual feature is thresholded by  $\tau_{low}$  and  $\tau_{up}$ . The voxel  $\mathbf{x}_k$  is then sent to one of the two child nodes based on the outcome of this test.

During training, each node  $p$  is optimized using the partition of the training data  $\mathcal{T}_p$  it receives as input. At the end of the training process, each node  $p$  is assigned to the optimal binary test  $t^{\lambda_p^*}$ , where  $\lambda_p^* = (\tau_{low}^*, \tau_{up}^*, \theta^*)_p$ . The optimality criterion is the information gain, denoted  $IG$ , as defined in (Quinlan, 1993)

$$IG(\lambda, \mathcal{T}_p) = H(\mathcal{T}_p) - H(\mathcal{T}_p | (t^\lambda(\mathbf{x}_k))_k) \quad (2)$$

where  $\mathcal{T}_p \subset \mathcal{T}$  and where  $H$  denotes the entropy. More precisely, the term  $H(\mathcal{T}_p | (t^\lambda(\mathbf{x}_k))_k)$  measures the error made when approximating the expert labeling  $Y$  by the binary test  $t^\lambda$ . The optimal parameter  $\lambda_p^*$  maximizes the information gain

$$\lambda_p^* = \arg \max_{\lambda} IG(\lambda, \mathcal{T}_p) \quad (3)$$

for node  $p$ . As a result, the optimal binary test is the test discriminating lesion from background voxels such as maximizing the information gain.

Only a randomly sampled subset  $\Theta$  of the feature space is available at each node for optimization, while the threshold space is uniformly discretized. The optimal  $\lambda^* = (\tau_{low}^*, \tau_{up}^*, \theta^*)$  is found by exhaustive search jointly over the feature and threshold space. Random sampling of the features leads to increased inter-node and inter-tree variability which improves generalization (Breiman, 2001).

Trees are grown to a maximum depth  $D$ . At the node level, a leaf node is generated when the information gain is below a minimal value  $IG_{min}$ .

As a result of the training process, each leaf node  $l_t$  of every tree  $t$  receives a partition  $\mathcal{T}_{l_t}$  of the training data. The partition  $\mathcal{T}_{l_t}$  can be divided into two sets respectively containing background and lesion voxels and defined as

$$\mathcal{T}_{l_t}^b = \{(\mathbf{x}, Y(\mathbf{x})) \in \mathcal{T}_{l_t} | Y(\mathbf{x}) = b\} \quad (4)$$

where  $b \in \{0, 1\}$  stands for the background and lesion class, respectively. Subsequently, the following empirical posterior probability is defined

$$p_{l_t}(Y(\mathbf{x}) = b) = \frac{|\mathcal{T}_{l_t}^b|}{|\mathcal{T}_{l_t}|} \quad (5)$$

and stored at the leaf node.

Figure 2 illustrates how the decision trees partition the data in the feature space and how resulting probabilities are stored in leaf nodes.

### 3.1.2. Prediction

When applied to a new test volume  $\mathcal{T}_{test} = \{\mathbf{x}_k\}$ , each voxel  $\mathbf{x}_k$  is propagated through all the trees by successive application of the relevant binary tests. When reaching the leaf node  $l_t$  in all trees  $t \in [1..T]$ , posteriors  $p_{l_t}(Y(\mathbf{x}) = b)$  are gathered in order to compute the final posterior probability defined as follows:

$$p(y(\mathbf{x}) = b) = \frac{1}{T} \sum_{t=1}^T p_{l_t}(Y(\mathbf{x}) = b) \quad (6)$$

which is a mean over all the trees in the forest. This probability may be thresholded at a fixed value  $\tau_{posterior}$  if a binary segmentation is required.

A posterior map  $P_b$  is obtained by applying the same prediction procedure to all voxels. Thus for every voxel  $\mathbf{x}$ ,  $P_b(\mathbf{x}) = p(y(\mathbf{x}) = b)$  is the posterior

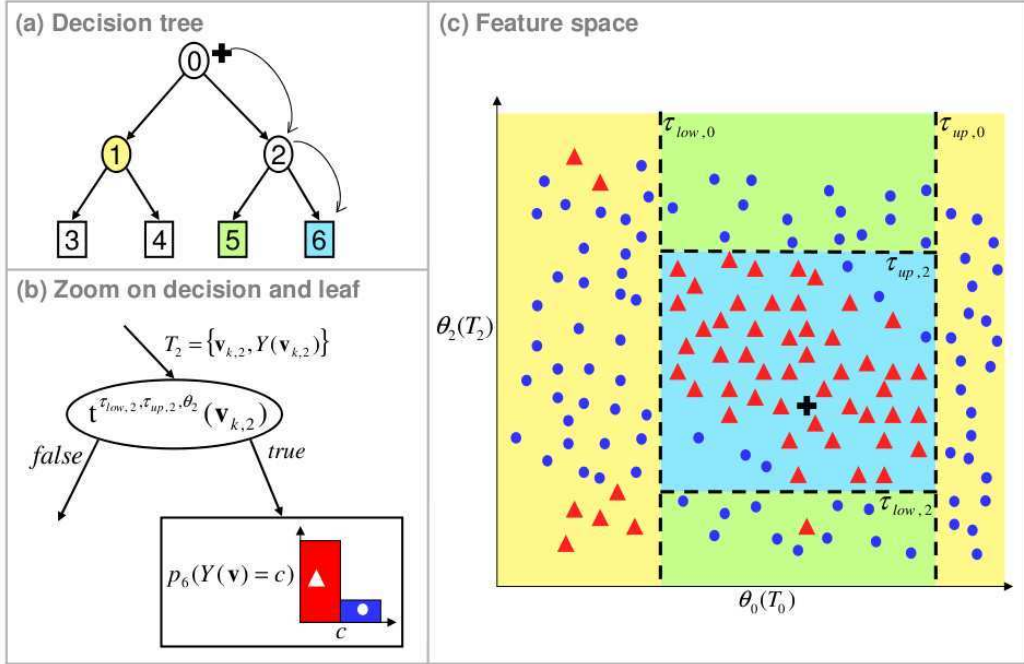


Figure 2: **Decision trees encode feature space partitions.** (a) A decision tree of depth  $D = 2$  is considered in this example. Decision node 1 and leaf nodes 5 and 6 are colored to track the partitions of the training data in the feature space. The black cross stands for an unseen sample (voxel) which is classified while propagated down the tree. (b) A zoom on node 2 shows that its binary test, denoted by  $t^{\tau_{low,2}, \tau_{up,2}, \theta_2}$ , is optimized over a partition of the training data, denoted by  $T_2 = \{\mathbf{v}_{k,2}, Y(\mathbf{v}_{k,2})\}$ . The leaf node 6 encloses the class distribution of the set of voxels reaching it during training. The classes are background in blue and lesion in red. (c) The dots stand for the training voxels and are colored according to their class. The black cross denotes a voxel from an unseen volume considered for prediction. Every decision node in the forest applies an axis-aligned feature test. Here we focus on decision nodes 0 and 2 using features  $\theta_0$  and  $\theta_2$ , respectively.

probability of belonging to the class  $b$ . This probability map can be thresholded at a fixed value to obtain a segmentation. Choosing  $\tau_{posterior} = 0.5$  is equivalent to looking for  $b^* = \arg \max_b P_b(\mathbf{x})$ .

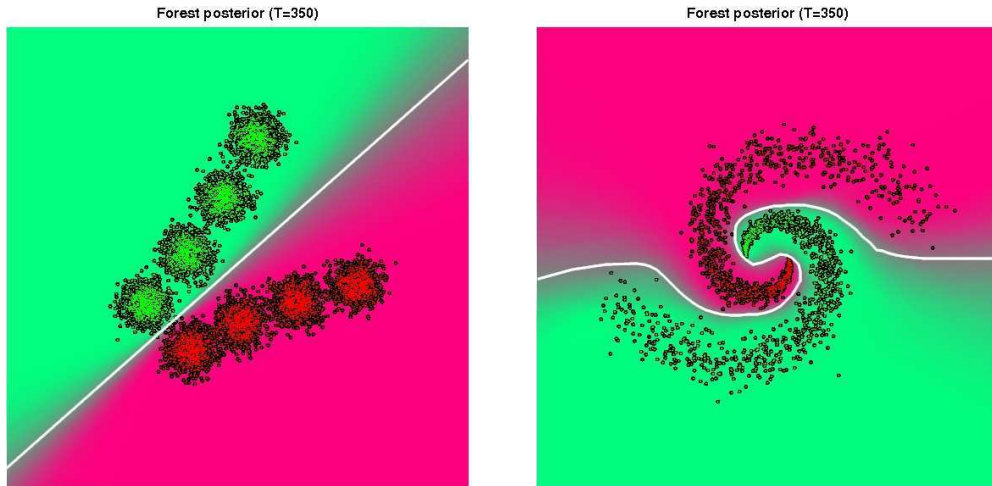


Figure 3: **Posterior maps learned from two distinct synthetic training sets.** In both cases, the training data consists of two classes, green and red, and is used to learn a large forest, here  $T = 350$ . The posterior map is obtained by classifying a dense grid in the feature space and is then overlaid with the associated training data (shown as points). Larger opacities indicate larger probability of a pixel belonging to a class while uncertain regions are indicated by less saturated colours. The white line plots the locus of points for which  $p(y(\mathbf{x}) = \textit{green}) = p(y(\mathbf{x}) = \textit{red})$ . We observe that 1) forest posteriors mimic the maximum-margin behavior, 2) uncertainty increases when moving away from the training data.

### 3.1.3. Advantages

The probabilistic random forest framework presented in this section shows considerable advantages over other classifiers, e.g. support vector machines (SVMs). Indeed, it combines efficient probabilistic classification and transparent feature selection as detailed below.

When a new case is presented to the classifier, every voxel goes through a sequence of decisions on different channels. As a result, the posterior probability affected to each voxel measures the confidence of this voxel being an MS lesions in the *multi-channel* space inferred from training.

Trees from the same forest are all independent from one another. The posterior map can thus be computed in parallel: each tree computes its own posterior map, they are then combined to form the final result (cf. Section 3.1.2). The training of the random forest can also be parallelized in a trivial manner by learning each tree independently from the others.

Moreover, unlike more “black-box” supervised methods such as SVMs or neural networks, the random forest framework enables us to enter the learned trees and identify the most discriminative features. In Section 5.4, we take advantage of this property to draw a detailed analysis of the most discriminative visual features for the task of MS segmentation.

The following properties motivate the use of random forests framework for the task of MS lesions segmentation: 1) when applying the random forest the binary trees can be evaluated extremely fast in prediction; 2) thanks to parallelism training on MR volumes is fast; 3) as a result of the training process, an optimal sequence of decisions, including most informative channels and visual features, is effortlessly available (cf. Section 5.4.2).



In many applications, random forests have been found to generalize better than SVM or boosting (Yin et al., 2010). The generalization power increases monotonically with increasing the forest size. Unlike SVM and boosting, random forests also estimate the confidence of the prediction as a by-product of the training process. Figure 3 shows the feature space of an exemplary segmentation problem. It illustrates that the random forest also generalizes well in regions of the feature space with sparse data support, which is beneficial. The uncertainty increases in those areas in feature space which have little data support and on the boundary between classes. This is an important expected behavior, similar to that of Gaussian Processes (Bishop, 2006).

Another interesting property of random forests lies in the way they separate the feature space. In Figure 3, probability maps show the highest uncertainty values on voxels equidistant of the two classes. This is a feature the random forest classifier shares with maximum-margin classifier such as SVMs (Bernhard Scholkopf and Smola, 1999).

### 3.2. Visual features

This section aims at presenting the visual features and at arguing the underlying motivation.

Two kinds of visual features are computed:

1. local features:

$$\theta_C^{loc}(\mathbf{x}) = C(\mathbf{x}) \quad (7)$$

where  $C$  is an intensity or a prior channel, and  $C(\mathbf{x})$  is the value of channel  $C$  at position  $\mathbf{x}$ ;

2. context-rich features comparing the voxel of interest with distant regions.

The first context-rich feature compares the local voxel value in channel  $C_1$  with the mean value in channel  $C_2$  over two 3D boxes  $R_1$  and  $R_2$  within an extended neighborhood:

$$\theta_{C_1, C_2, R_1, R_2}^{cont}(\mathbf{x}) = C_1(\mathbf{x}) - \frac{1}{Vol(R_1)} \sum_{\mathbf{x}' \in R_1} C_2(\mathbf{x}') - \frac{1}{Vol(R_2)} \sum_{\mathbf{x}' \in R_2} C_2(\mathbf{x}') \quad (8)$$

where  $C_1$  and  $C_2$  are both intensity or prior channels. The regions  $R_1$  and  $R_2$  are sampled randomly in a large neighborhood of the voxel  $\mathbf{x}$  (cf. Figure 4). The sum over these regions is efficiently computed using integral volume processing (Shotton et al., 2009). The random sampling of the features is part of the random forest framework and were used in previous work (Criminisi et al., 2009, 2010; Yi et al., 2009).

The second context-rich feature compares the voxel of interest at  $\mathbf{x}$  with its symmetric counterpart with respect to the mid-sagittal plane, noted  $S(\mathbf{x})$ :

$$\theta_C^{sym}(\mathbf{x}) = C(\mathbf{x}) - C(S(\mathbf{x})) \quad (9)$$

where  $C$  is an intensity channel.

A new version of the symmetry feature loosens the hard symmetric constraint as the size of the neighborhood increases, in order to take into account the fact that the brain is not perfectly symmetric. Instead of comparing with the exact symmetric  $S(\mathbf{x})$  of the voxel, the minimal difference value within a close neighborhood of  $S(\mathbf{x})$ , denoted  $\mathcal{S}$ , is chosen. Three different sizes are considered for this neighborhood: 6, 26 and 32 neighbors respectively (cf.

Figure 4). We obtain a softer version of the symmetric feature which reads:

$$\theta_{C,S}^{sym}(\mathbf{x}) = \min_{\mathbf{x}' \in \mathcal{S}} \{C(\mathbf{x}) - C(\mathbf{x}')\} \quad (10)$$

To summarize, three visual features are introduced: local, neighborhood and symmetry features. They can be thought as meta-features to be used *on top of* standard image filters, such as local moments, gradients or textures, rather than replacing them. In our case, visual features are evaluated on raw images and spatial priors, but any other channel could be added, e.g. intensity gradient as described in (Yi et al., 2009).

The presented features can be applied individually to every voxel, unlike e.g. geometric moments. This is an essential property which enables *voxel-wise* image classification. The presented method not only provides a voxel-wise classification, but also integrates neighborhood information in the classification process thanks to context-rich features. The use of 3D boxes to integrate neighborhood information is motivated by the fact that they are extremely efficient through integral volume processing (Shotton et al., 2009).

#### 4. Experiments and results

Results presented in this section aim at evaluating the segmentation results and comparing the context-rich random forest approach to methods presented during the challenge (Souplet et al., 2008; Anbeek et al., 2008; Bricq et al., 2008a; Shiee et al., 2008). Experiments described here are discussed in Section 5.

Exhaustive segmentation results are available for both public and private datasets under the following url:

<ftp://ftp-sop.inria.fr/asclepios/Published-Material/Ezequiel.Geremia/>

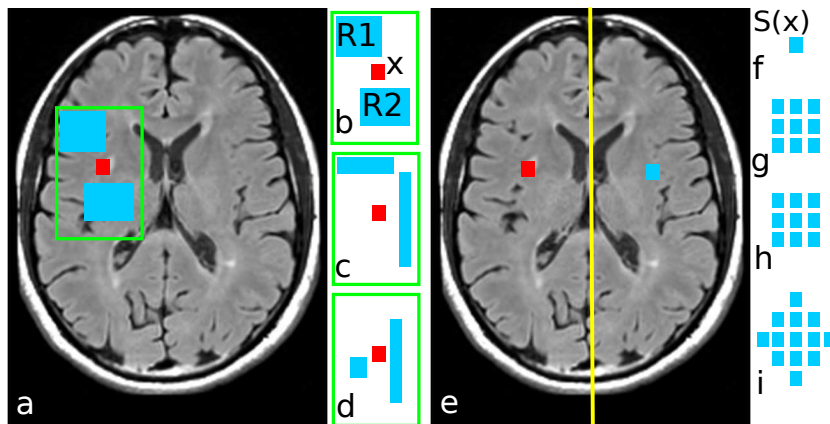


Figure 4: **2D view of context-rich features.** (a) A context-rich feature depicting two regions  $R_1$  and  $R_2$  with constant offset relative to  $\mathbf{x}$ . (b-d) Three examples of randomly sampled features in an extended neighborhood. (e) The symmetric feature with respect to the mid-sagittal plane. (f) The hard symmetric constraint. (g-i) The soft symmetry feature considering neighboring voxels in a sphere of increasing radius. See text for details.

#### 4.1. Results on the public MSGC dataset

For quantitative evaluation, the 20 available cases from the public dataset are classified and compared to other methods (Souplet et al., 2008; Anbeek et al., 2008; Bricq et al., 2008a; Shiee et al., 2008). A three-fold cross-validation is carried out on this dataset: the forest is trained on  $\frac{2}{3}$  of the cases and tested on the other  $\frac{1}{3}$ , this operation is repeated 3 times in order to collect test errors for each case.

The binary classification is evaluated using two measures, true positive rate ( $TPR$ ) and positive predictive value ( $PPV$ ), both equal 1 for perfect segmentation (cf. Table 1).

Forest parameters are fixed to the following values: number of random regions  $|\Theta| \simeq 950$ , number of trees  $T = 30$ , tree depth  $D = 20$ , lower bound

for the information gain  $IG_{min} = 10^{-5}$ , posterior threshold  $\tau_{posterior} = 0.5$ . Parameters  $T$  and  $D$  are set here to maximum values, Section 5.3 explains how these parameters can be optimized in order to improve segmentation results.

Tables in supplemental material report extensive results allowing comparison on every case of the MSGC public dataset. It shows that the learned context-rich random forest achieves better  $TPR$  in all cases (cf. top bar plot), and better  $PPV$  in 70% of the cases (cf. center bar plot). Computed  $p$ -values for the pair-sample  $t$ -test show that these improvements are significant for both  $TPR$  ( $p = 1.3 \cdot 10^{-7}$ ) and  $PPV$  ( $p = 0.0041$ ) scores.

Metric [%]	Souplet et al.	Context-rich RF	$RI$ [%]	$p$ -value
$TPR$	19.21 ± 13.68	<b>39.39 ± 18.40</b>	<b>105</b>	<b><math>1.3 \cdot 10^{-7}</math></b>
$PPV$	29.55 ± 16.26	<b>39.78 ± 20.19</b>	<b>35</b>	<b>0.0041</b>

Table 2: **Comparison of context-rich random forests with method presented in (Souplet et al., 2008) on the public dataset.** Relative improvement over (Souplet et al., 2008), defined as  $RI = (score_{RF} - score_{other})/score_{other}$ , are significant for both  $TPR$  ( $p = 1.3 \cdot 10^{-7}$ ) and  $PPV$  ( $p = 0.0041$ ) scores. Significant improvements over (Souplet et al., 2008) are highlighted in bold.

#### 4.2. Results on the private MSGC dataset

A context-rich random forest was learned on the whole public dataset from the MS Lesion Challenge, i.e. 20 labeled cases. Forest parameters are fixed to the following values: number of random regions  $|\Theta| \simeq 950$ , number of trees  $T = 30$ , tree depth  $D = 20$ , lower bound for the information gain  $IG_{min} = 10^{-5}$ , posterior threshold  $\tau_{posterior} = 0.5$ . Considerations that lead to these parameter values are detailed in Section 5.3.

The MSCG website carried out a complementary and independent evaluation of our algorithm on the previously unseen private dataset. The results, reported in Table 3, confirm a significant improvement over (Souplet et al., 2008). The presented spatial random forest achieves, on average, slightly larger true positive ( $TPR$ ), which is beneficial (cf. Table 1), and comparable false positive ( $FPR$ ) rates but lower volume difference ( $VD$ ) and surface distance ( $SD$ ) values (cf. Table 3). Pair-sample  $p$ -values were computed for the  $t$ -test on the private dataset. Results show significant improvement over the method presented in (Souplet et al., 2008) on  $SD$  ( $p = 4.2 \cdot 10^{-6}$ ) for the CHB rater, and on  $SD$  ( $p = 6.1 \cdot 10^{-3}$ ) for the UNC rater.

## 5. Discussion

### 5.1. Interpreting segmentation results

Quantitative evaluation of segmentation results, for both public (cf. Tables 2) and private (cf. Tables 3, 4, 5 and 6) datasets, show that the presented random forest framework compares favorably to top-ranked methods. More specifically, they show significant improvement on the algorithm presented in (Souplet et al., 2008). Exhaustive results are available in the supplemental material allow case-by-case comparison with other methods.

The MSGC website (Styner et al., 2008b) gathers the results of the methods presented during the MSGC in 2008 as well as more recent methods which results were submitted directly to the website. The resulting ranking affects a score to each method. The score relates the performance of the method against expected inter-expert variability which is known to be high for MS lesions segmentation. A score of 90 would equal the accuracy of a human

Rater	Metric [%]	Souplet et al.	Context-rich RF	$RI$ [%]	$p$ -value
CHB	$VD$	$86.48 \pm 104.9$	$52.94 \pm 28.63$	-38.7	0.094
	$SD$	<b><math>8.20 \pm 10.89</math></b>	<b><math>5.27 \pm 9.54</math></b>	<b>-35.7</b>	<b><math>4.2 \cdot 10^{-6}</math></b>
	$TPR$	$57.45 \pm 23.22$	$58.08 \pm 20.03$	+1.0	0.90
	$FPR$	$68.97 \pm 19.38$	$70.01 \pm 16.32$	+1.5	0.70
UNC	$VD$	$55.76 \pm 31.81$	$50.56 \pm 41.41$	-9.4	0.66
	$SD$	<b><math>7.4 \pm 8.28</math></b>	<b><math>5.6 \pm 6.67</math></b>	<b>-24.3</b>	<b><math>6.1 \cdot 10^{-3}</math></b>
	$TPR$	$49.34 \pm 15.77$	$51.35 \pm 19.98$	+3.9	0.54
	$FPR$	$76.18 \pm 17.07$	$76.81 \pm 11.70$	+0.1	0.83

Table 3: **Average results computed by the MSGC on the private dataset and compared to the method presented in (Souplet et al., 2008).** The relative mean improvement on the algorithm presented in (Souplet et al., 2008) on the private dataset is defined as follows  $RI = (score_{RF} - score_{Souplet})/score_{Souplet}$ . The  $RI$  and the  $p$ -values are reported on top for each metric of the MSGC, associated  $p$ -values are reported below. Independent quantitative evaluation confirms improvement on the algorithm presented in (Souplet et al., 2008). Boldface highlights significant improvements. The spatial random forest achieves, on average, slightly larger true positive ( $TPR$ ), which is beneficial and comparable false positive ( $FPR$ ) rates but lower volume difference ( $VD$ ) and surface distance ( $SD$ ) values.

rater. The method presented in (Bricq et al., 2008a) and our context-rich random forest approach are first and second respectively. They show very close scores, 82.1354 and 82.0755 respectively. A score of 82 places the accuracy of the method just below that of a human expert. In addition, we know that the reliability of automatic methods is generally higher than that of human experts. Approaching the performance of a human rater is thus even more interesting.

Rater	Metric [%]	Anbeek et al.	Context-rich RF	$RI$ [%]	$p$ -value
CHB	$VD$	46.93 $\pm$ 50.41	52.94 $\pm$ 28.63	+12.8	0.62
	$SD$	<b>7.85 <math>\pm</math> 11.00</b>	<b>5.27 <math>\pm</math> 9.54</b>	<b>-32.8</b>	<b><math>5.9 \cdot 10^{-3}</math></b>
	$TPR$	59.14 $\pm$ 21.79	58.08 $\pm$ 20.03	-1.80	0.84
	$FPR$	78.51 $\pm$ 20.56	70.01 $\pm$ 16.32	-10.8	$8.6 \cdot 10^{-2}$
UNC	$VD$	100.7 $\pm$ 132.3	50.56 $\pm$ 41.41	-49.8	$9.1 \cdot 10^{-2}$
	$SD$	<b>9.77 <math>\pm</math> 9.00</b>	<b>5.6 <math>\pm</math> 6.67</b>	<b>-42.7</b>	<b><math>3.6 \cdot 10^{-4}</math></b>
	$TPR$	48.86 $\pm$ 21.64	51.35 $\pm$ 19.98	+5.10	0.38
	$FPR$	83.19 $\pm$ 18.59	76.81 $\pm$ 11.70	-7.68	0.11

Table 4: **Average results computed by the MSGC on the private dataset and compared to the method presented in (Anbeek et al., 2008).** The relative mean improvement on the algorithm presented (Anbeek et al., 2008) on the private dataset is defined as follows  $RI = (score_{RF} - score_{Anbeek})/score_{Anbeek}$ . The  $RI$  and the  $p$ -values are reported on top for each metric of the MSGC, associated  $p$ -values are reported below. Independent quantitative evaluation confirms improvement on the algorithm presented in (Anbeek et al., 2008). Boldface highlights significant improvements. The spatial random forest achieves, on average, better results on UNC than on CHB labels: higher true positive ( $TPR$ ), which is beneficial and lower false positive ( $FPR$ ) rates, lower volume difference ( $VD$ ) and surface distance ( $SD$ ) values.

Although segmentation results include most MS lesions delineated by the expert (cf. Figures 5 and 6), we observe that some MS lesions are missing. Missed MS lesions are located in specific locations which are not represented in the training data, e.g. in the corpus callosum (cf. Figure 5, slice 38). This is a limitation of the supervised approach. In this very case, however, the posterior map highlights the missed lesion in the corpus callosum as belonging to the lesion class with high uncertainty. Low confidence (or high



Rater	Metric [%]	Bricq et al.	Context-rich RF	$RI$ [%]	$p$ -value
CHB	$VD$	73.03 $\pm$ 78.80	52.94 $\pm$ 28.63	-27.5	0.19
	$SD$	6.65 $\pm$ 6.55	5.27 $\pm$ 9.54	-20.8	0.20
	$TPR$	<b>46.70 <math>\pm</math> 19.94</b>	<b>58.08 <math>\pm</math> 20.03</b>	<b>+24.4</b>	<b><math>1.0 \cdot 10^{-3}</math></b>
	$FPR$	51.06 $\pm$ 25.23	70.01 $\pm$ 16.32	+37.1	$8.3 \cdot 10^{-5}$
UNC	$VD$	51.33 $\pm$ 27.00	50.56 $\pm$ 41.41	-1.50	0.92
	$SD$	6.61 $\pm$ 5.23	5.6 $\pm$ 6.67	-15.2	0.17
	$TPR$	<b>39.50 <math>\pm</math> 16.06</b>	<b>51.35 <math>\pm</math> 19.98</b>	<b>+30.0</b>	<b><math>1.2 \cdot 10^{-3}</math></b>
	$FPR$	60.80 $\pm$ 22.75	76.81 $\pm$ 11.70	+26.3	$8.6 \cdot 10^{-4}$

Table 5: **Average results computed by the MSGC on the private dataset and compared to the method presented in (Bricq et al., 2008a).** The relative mean improvement on the algorithm presented (Bricq et al., 2008a) on the private dataset is defined as follows  $RI = (score_{RF} - score_{Bricq}) / score_{Bricq}$ . The  $RI$  and the  $p$ -values are reported on top for each metric of the MSGC, associated  $p$ -values are reported below. Independent quantitative evaluation confirms improvement on the algorithm presented in (Bricq et al., 2008a). Boldface highlights significant improvements. The spatial random forest achieves, on average, slightly larger true positive ( $TPR$ ), which is beneficial, but also slightly larger false positive ( $FPR$ ) rates, and lower volume difference ( $VD$ ) and surface distance ( $SD$ ) values.

uncertainty) reflects the incorrect spatial prior inferred from an incomplete training set. Indeed, in the training set, there is no example of MS lesions appearing in the corpus callosum.

On the contrary, the random forest is able to detect *suspicious regions* with high certainty. *Suspicious regions* are visually very similar to MS lesions and widely represented in the training data, but they are not delineated by the expert, e.g. the left frontal lobe lesion again in Figure 5, slice 38. The

Rater	Metric [%]	Shiee et al.	Context-rich RF	$RI$ [%]	$p$ -value
CHB	$VD$	$84.17 \pm 120.8$	$52.94 \pm 28.63$	$-37.1$	0.22
	$SD$	<b><math>7.95 \pm 16.65</math></b>	<b><math>5.27 \pm 9.54</math></b>	<b><math>-33.5</math></b>	<b><math>9.0 \cdot 10^{-3}</math></b>
	$TPR$	$55.40 \pm 23.60$	$58.08 \pm 20.03$	$+4.83$	0.64
	$FPR$	$68.85 \pm 23.75$	$70.01 \pm 16.32$	$+1.69$	0.76
UNC	$VD$	$69.63 \pm 115.2$	$50.56 \pm 41.41$	$-27.4$	0.42
	$SD$	$7.104 \pm 8.93$	$5.6 \pm 6.67$	$-21.1$	0.15
	$TPR$	$49.79 \pm 24.54$	$51.35 \pm 19.98$	$+3.14$	0.76
	$FPR$	$74.28 \pm 20.08$	$76.81 \pm 11.70$	$+3.41$	0.50

Table 6: **Average results computed by the MSGC on the private dataset and compared to the method presented in (Shiee et al., 2008).** The relative mean improvement on the algorithm presented (Shiee et al., 2008) on the private dataset is defined as follows  $RI = (score_{RF} - score_{Shiee})/score_{Shiee}$ . The  $RI$  and the  $p$ -values are reported on top for each metric of the MSGC, associated  $p$ -values are reported below. Independent quantitative evaluation confirms improvement on the algorithm presented in (Shiee et al., 2008). Boldface highlights significant improvements. The spatial random forest achieves, on average, slightly larger true positive ( $TPR$ ), which is beneficial but slightly larger false positive ( $FPR$ ) rates, and lower volume difference ( $VD$ ) and surface distance ( $SD$ ) values.

appearance model and spatial prior implicitly learned from the training data points out that hyper-intense regions in the FLAIR MR sequence which lay in the white matter (cf. Section 5.4) can be considered as MS lesions with high confidence.

Recent histopathological studies have shown that grey matter regions are also heavily affected by the MS disease (Geurts and Barkhof, 2008). In our case, the public dataset does not show any MS lesion in the grey matter of

the brain. Subsequently, the decision forest learns that MS lesions preferably appear in the white matter. Adding new cases showing grey matter MS lesions in the training set would allow the forest to automatically adapt the segmentation to include this kind of lesions. This observation stresses the necessity of gathering large and heterogeneous datasets for training purposes.

When focusing on quantitative measures, we observe that cases UNC01 and UNC06 from the public dataset show surprisingly low scores (cf. Table 2). The labels by the CHB expert for these two cases are abnormal: the ground truth is mirrored with respect to the anatomical images. This may be considered as a label error and explains the low scores for these two specific cases. The MSGC website confirmed this observation and subsequently corrected the online database. We also observe that learning on the whole public dataset and testing on the private dataset (cf. Table 3) produces better average results than the three-fold cross-validation carried out on the public dataset (cf. Table 2). Again this illustrates the benefit of learning the classifier on large enough datasets capturing better the variability of the data.

### *5.2. Influence of preprocessing*

Data normalization is critical in order to ensure that the feature are evaluated in a coherent way in all the images presented to the forest. The evaluation of context-rich features,  $\theta^{cont}$ , is sensitive to rotation: spatial normalization is performed using rigid registration (Prima et al., 2001). In the same way, the evaluation of intensity-based features requires inter-case intensity calibration.

Classification results for cases from the CHB (cf. Figures 5 and 6) and

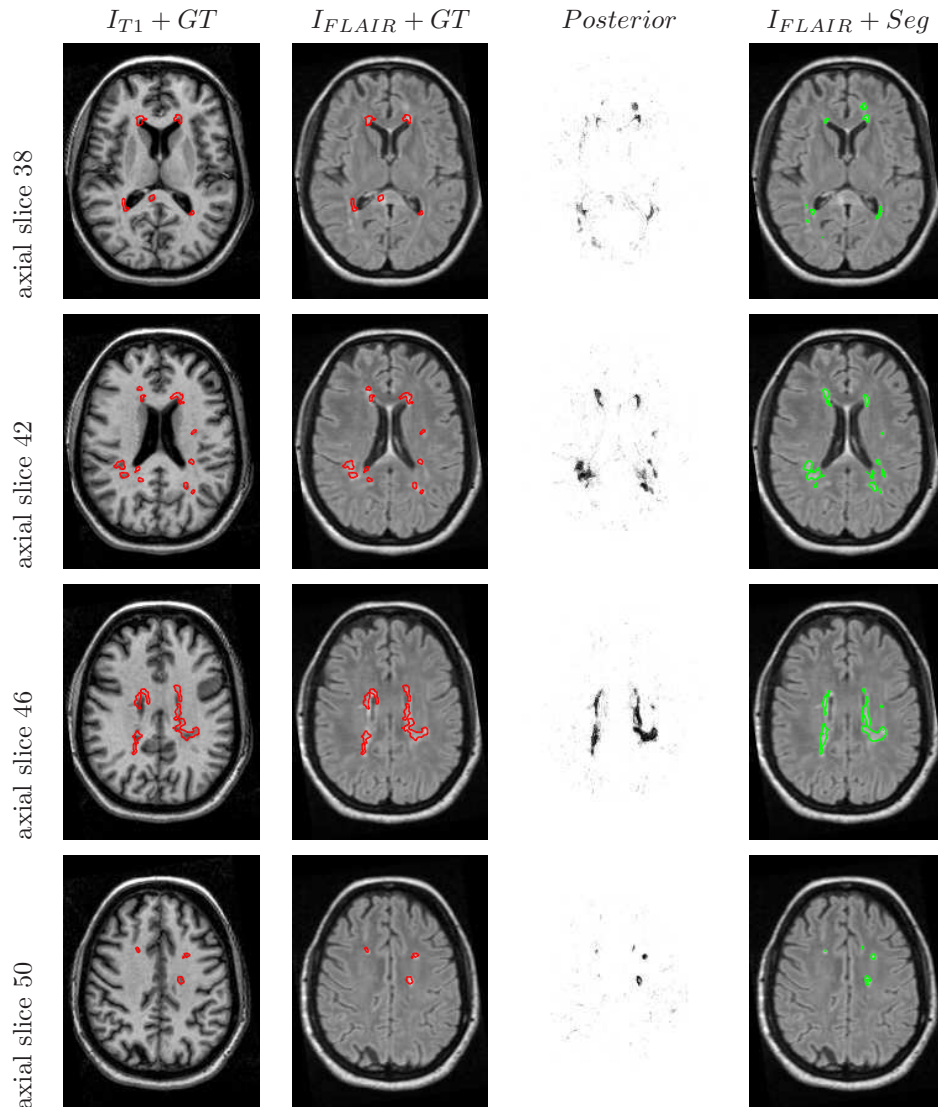


Figure 5: **Segmenting Case CHB05 from the public MSGC dataset.** From left to right: preprocessed T1-weighted ( $I_{T1}$ ), T2-weighted ( $I_{T2}$ ) and FLAIR MR images ( $I_{FLAIR}$ ) overlaid with the associated ground truth  $GT$ , the posterior map  $Posterior = (P_{lesion}(\mathbf{v}_k))_k$  displayed using an inverted grey scale and the FLAIR sequence overlaid with the segmentation ( $Seg = (Posterior \geq \tau_{posterior})$  with  $\tau_{posterior} = 0.5$ ). Segmentation results show that most of lesions are detected. Although some lesions are not detected, e.g. peri-ventricular lesion in slice 38, they appear enhanced in the posterior map. Moreover the segmentations of slices 38 and 42 show peri-ventricular regions, visually very similar to MS lesions, but not delineated in the ground truth.

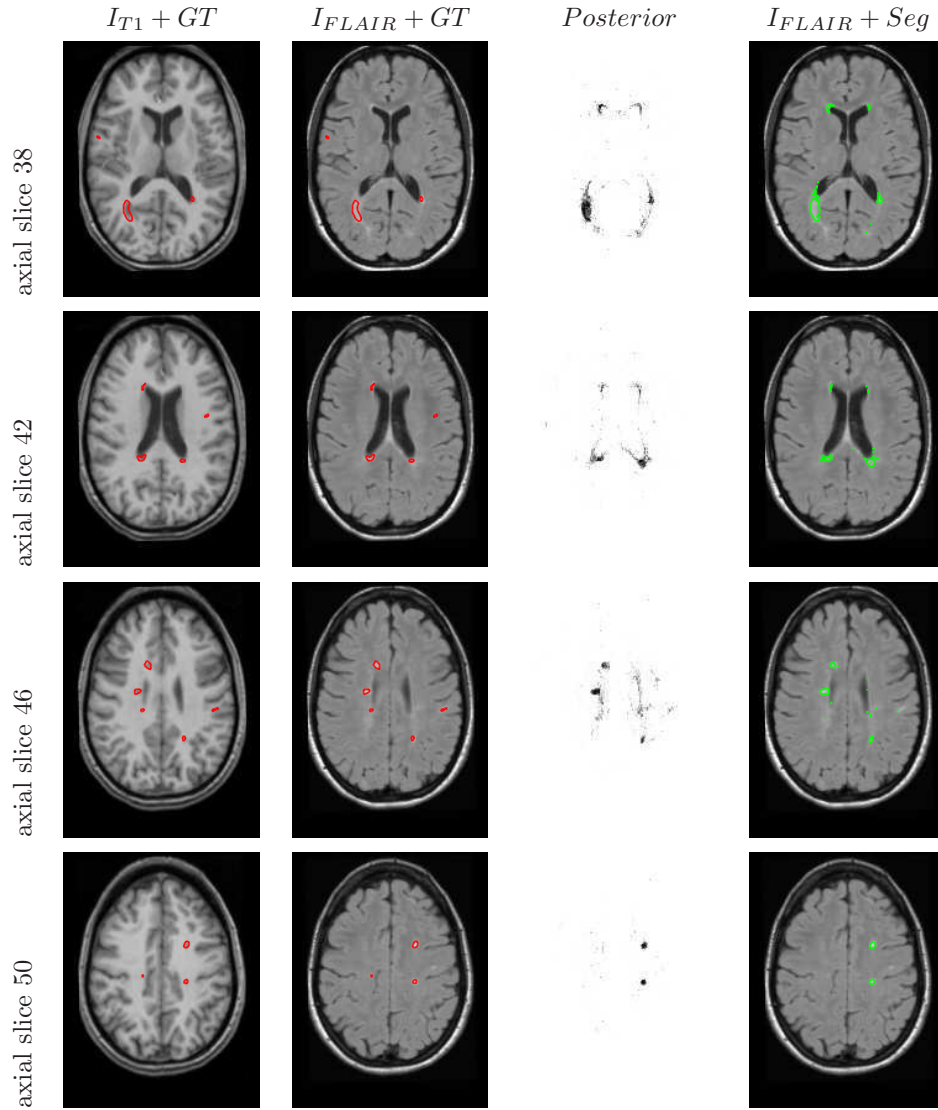


Figure 6: **Segmenting case UNC02 from the public MSGC dataset.** From left to right: preprocessed T1-weighted ( $I_{T1}$ ), T2-weighted ( $I_{T2}$ ) and FLAIR MR images ( $I_{FLAIR}$ ) overlaid with the associated ground truth  $GT$ , the posterior map  $Posterior = (P_{lesion}(\mathbf{v}_k))_k$  displayed using an inverted grey scale and the FLAIR sequence overlaid with the segmentation ( $Seg = (Posterior \geq \tau_{posterior})$  with  $\tau_{posterior} = 0.5$ ).

the UNC (cf. Figures 7 and 8) centers are obtained with the same forest. It is mandatory to use the same preprocessing as during training (cf. Section 2.2).

By doing so, the cases from different datasets, e.g. T1-weighted and FLAIR images in Figures 5 and 6, show very similar intensity values for a specific brain tissue and a given MR sequence. However, we observe that the contrast in the T1-weighted and FLAIR images is more marked in case CHB05 (cf. Figure 5) than in case UNC02 (cf. Figure 6). Despite contrast changes, classification results are coherent. This illustrates the stability of our method, the random forest framework together with its preprocessing step to slight inter-image contrast variations.

The trees are generated in parallel on 30 nodes and gathered to form the forest. Cropping and sub-sampling the training images aims at reducing, by a factor larger than 10, the execution time spent to learn a single tree. On IBM e325 dual-Opterons 246 at a maximum frequency of 2Ghz, learning a tree on 20 sub-sampled images and with parameters fixed in Section 4.2 on a single CPU takes, on average, 8 hours.

### *5.3. Influence of forest parameters*

The number of the trees and their depth, respectively denoted  $T$  and  $D$ , characterize the generalization power and the complexity of the non-parametric model learned by the forest. This section aims at understanding the contribution of each of these meta-parameters.

A 3-fold cross-validation on the public dataset is carried out for each parameter combination. Segmentation results are evaluated for each combination using two different metrics: the area under the receiver operating characteristic (ROC) curve and the area under the precision-recall curve.

The ROC curve plots  $TPR$  vs.  $FPR$  scores computed on the test data for every value of  $\tau_{posterior} \in [0, 1]$ . The precision-recall curve plots  $PPV$  vs.  $TPR$  scores computed on the test data for every value of  $\tau_{posterior} \in [0, 1]$ . Results are reported in Figure 7.

We observe that 1) for a fixed depth, increasing the number of trees leads to better generalization; 2) for a fixed number of trees, low depth values lead to underfitting while high depth values lead to overfitting; 3) overfitting is reduced by increasing the number of trees.

This analysis was carried out *a posteriori*. Tuning the meta-parameters of the forest on the training data is not a valid practice. Using out-of-bag samples for forest parametrization is indeed preferable. Due to the fact that little training data is available for the MS lesion class, available labeled data was exclusively used to train the forest. From this perspective, the forest parameters were set to arbitrary but high enough values to avoid under- and overfitting:  $T = 30$  and  $D = 20$ .

Forest parameters were indeed selected in a safety-area with respect to under- and overfitting. The safety-area corresponds to a sufficiently flat region in the evolution of the areas under the ROC and the precision-recall curve. As shown in Figure 8, increasing the number of trees tends to benefit the generalization power of the classifier. We also observe that the performance of the classifier stabilizes for large enough forests.

#### 5.4. Analysis of feature relevance

During training, features considered for node optimization form a large and heterogeneous set (cf. Section 3.2). Unlike other classifiers, random forests provide an elegant way of ranking these features according to their

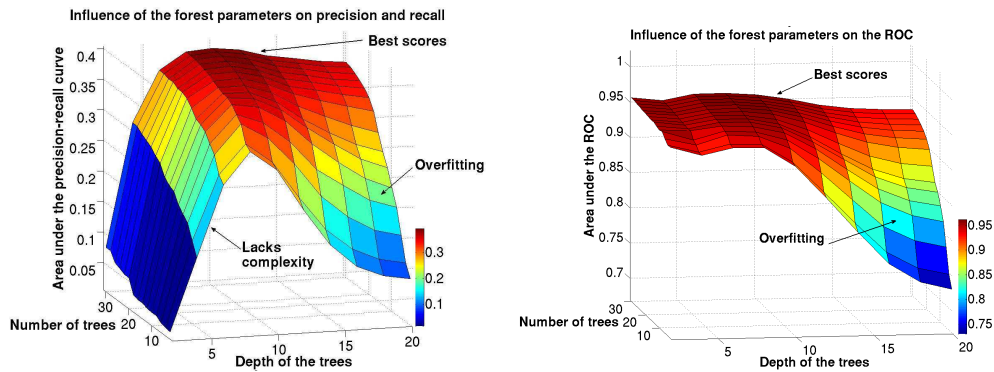


Figure 7: **Influence of forest parameters on segmentation results.** Both curves were plotted using mean results from a 3-fold cross validation on the public dataset. Left: the figure shows the influence of forest parameters on the area under the precision-recall curve. Right: the figure shows the influence of forest parameters on the area under the ROC curve. The ideal classifier would ensure area under the curve to be equal to 1 for both curves. We observe that 1) for a fixed depth, increasing the number of trees leads to better generalization; 2) for a fixed number of trees, low depth values lead to underfitting while high values lead to overfitting; 3) overfitting vanishes by increasing the number of trees.



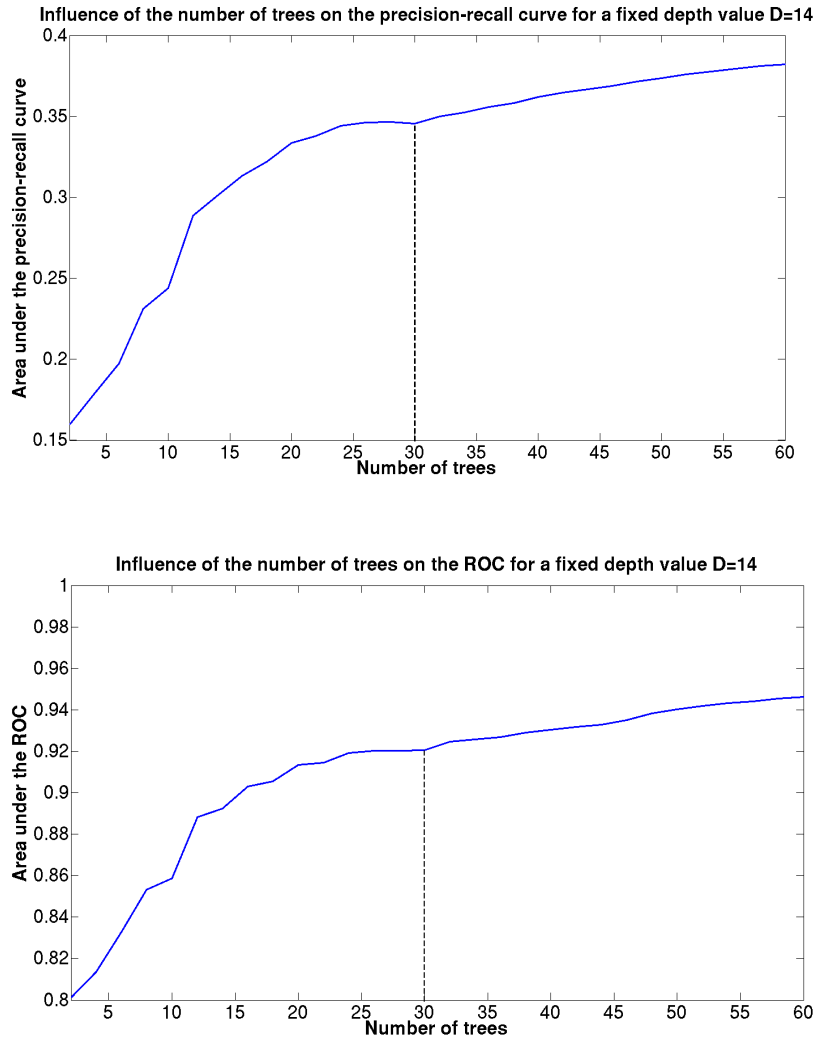


Figure 8: **Influence of the number of trees on segmentation results.** Both curves were plotted using mean results from a 3-fold cross validation on the public dataset. Top: the figure shows the influence of the number of trees on the area under the precision-recall curve. Bottom: the figure shows the influence of the number of trees on the area under the ROC curve. We observe that, for a fixed depth  $D = 14$ , increasing the number of trees improves generalization as stated in (Breiman, 2001). The increase in performance stabilizes around the value  $T = 30$ .

discriminative power. In this section, we aim at better understanding which are the most discriminative channels and visual cues (local, context-rich or symmetric) used in the classification process.

#### *5.4.1. Most discriminative visual features*

The first approach consists in counting the nodes in which a given feature was selected. We observe that local features were selected in 24% of the nodes, context-rich features were selected in 71% of the nodes whereas symmetry features were selected in 5% of the nodes (cf. Figure 9). In this case, no distinction is made as for the depth at which a given feature was selected.

Context-rich features exhibit high variability (900 of them are randomly sampled at every node). This variability combined with their ability to highlight regions which differ from their neighborhood explains why they were chosen. Together with local features, context-rich features learn a multi-channel appearance model conditioned by tissue spatial priors. Symmetry features are under-represented in the forest and thus prove to be the least discriminative ones. This is due to the fact that a large proportion of periventricular MS lesions tend to develop in a symmetric way. Nevertheless, symmetric features appear in top levels of the tree (up to third level) which indicates that they provide an alternative to local and context-rich features when these two fail.

A finer estimation of the feature importance consists in weighting the counting process. For a given feature, instead of only counting the nodes in which it appears, we also take into account the proportion of lesion voxels it helps discriminating: “the larger the proportion of lesion voxels it helps to

discriminate, the larger the weight of the feature”. This leads us to define for a fixed depth value  $d$ , the importance of a given feature type, denoted ( $IFT$ ), as:

$$IFT(\alpha) = \frac{1}{|\mathcal{T}^1|} \sum_p |\mathcal{T}_p^1| \cdot \chi_\alpha(\theta_p) \quad \alpha \in \{loc, cont, sym\} \quad (11)$$

where  $\alpha$  denotes a feature type,  $p$  indices the nodes in layer  $d$ ,  $\mathcal{T}^1$  is the training set of lesion voxels which partition  $\mathcal{T}_p^1$  reached node  $p$ , and  $\chi$  is the indicator function such that:

$$\chi_\alpha(\theta_p) = \begin{cases} 1, & \text{if } \theta_p \text{ is of type } \alpha \\ 0, & \text{otherwise} \end{cases} \quad (12)$$

The feature importance evaluates to 21.1 % for local features, 76.6 % for context-rich features and 2.3 % for symmetry features. Results are comparable to those obtained only by counting the features in the forest, but the real advantage of this measure is to allow us to draw depth-by-depth feature importance analysis in a normalized way.

The feature importance as a function of the depth of the tree is reported in Figure 10. Presented results are averaged values over a forest containing  $T = 30$  trees. Again, we observe that context-rich features are predominantly selected as the most discriminative, which confirms the trend reported in Figure 9. However, as shown in Figure 10, the preponderance of context-rich features is not uniform throughout the tree. Indeed, local features are the most discriminative in layers 0 and 2. A careful analysis of selected channels helps understanding why local features are selected in the top layers of the tree (cf. Section 5.4.2).

The selected context-rich features show high variability. More specifically,

the long-range regions are distributed all over the neighborhood. Depth-by-depth analysis does not show any specific pattern in the position of the regions with respect to the origin voxel. In addition, the volume of the regions also show high variability. The observed heterogeneity of selected context-rich features aims at coping with the variability of MS lesions (shape, location and size).

The symmetry feature is under-represented in the forest. Its discriminative power is thus very low compared to local and context-rich features. This observation induces two complementary interpretations to explain why symmetry features are the least significant: 1) most of MS lesions appear in peri-ventricular regions and in a symmetric way, 2) most of MS lesions can be clearly identified by their signature across MR sequences and their relative position in the white matter of the brain. However, in deeper layers of the trees, the symmetry feature is more significant and tends to classify ambiguous asymmetrical regions. When looking into the selected features, we also notice that the hard symmetric constraint is preferred over the loose symmetric constrain (cf. Section 3.2). Indeed, the feature importance evaluates to 1.6% for the hard symmetric feature, and to 0.7% for the loose symmetric feature. Moreover, in the rare cases where the loose constrain is selected, the 6-neighbors version predominates (cf. Section 3.2). This observation supports the idea that considering brain hemispheres as symmetric is an accurate approximation in our specific setting (cf. Sections 2 and 3).

#### *5.4.2. Most discriminative channels*

The second approach focuses on the depth at which a given feature was selected. For every tree in the forest, the root node always applies a test on

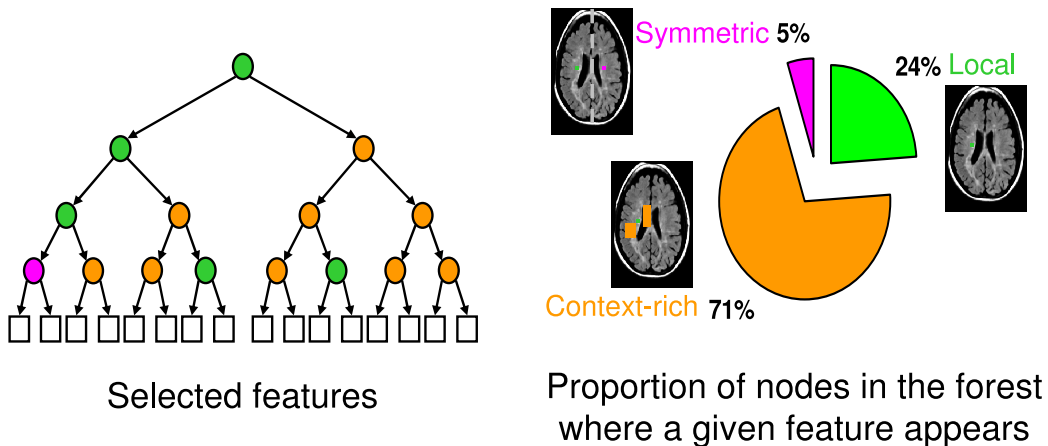


Figure 9: **Ranking features according to the proportion of nodes in which they appear.** Context-rich features are selected in 71% of the nodes, local features are selected in 24% of the nodes whereas symmetry features are selected in 5% of the nodes.

the FLAIR sequence ( $\theta_{FLAIR}^{loc}$ ). It means that out of all available features, containing local, context-rich and symmetry multi-channel features,  $\theta_{FLAIR}^{loc}$  was found to be the most discriminative. At the second level of the tree, a context-rich feature on spatial priors ( $\theta_{WM,GM}^{cont}$ ) appears to be the most discriminative over all trees in the forest. It aims at discarding all voxels which do not belong to the white matter.

The optimal decision sequence found while training the context-rich forest can thus be thought as a threshold on the FLAIR MR sequence followed by an intersection with the white matter mask (cf. Figure 11). Interestingly, this sequence matches the first and second steps of the pipeline proposed by the challenge winner method (Souplet et al., 2008). Note that in our case, it is automatically generated during the training process. Deeper layers in the trees, then, refine the segmentation of MS lesions by applying more accurate decisions.

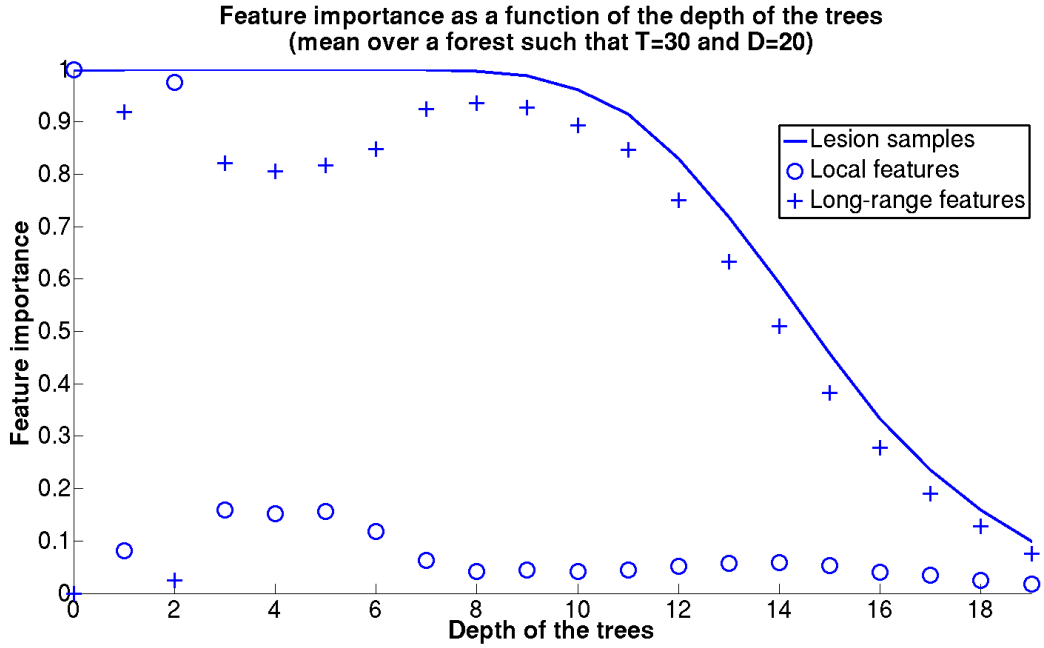


Figure 10: **Type of feature selected by layer of the tree.** For a fixed depth, the red circle stands for the importance of the context-rich feature ( $\theta^{cont}$ ), while the green circle stands for the importance of the local feature ( $\theta^{loc}$ ). For clarity, symmetry features ( $\theta^{sym}$ ) are omitted as they are under-represented in the forest. The blue line monitors the proportion of training samples of the lesion class which do not reside in leaf nodes, for each layer of the tree. We observe that context-rich features are predominantly selected as the most discriminative ones except in layers 0 and 2.

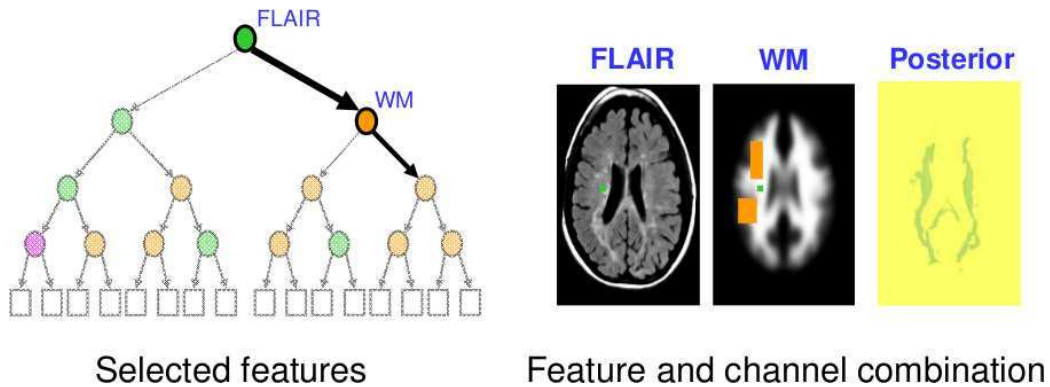


Figure 11: **Combination of features and channels learned by the forest to discriminate MS lesions.** The first layer of all trees in the forest performs a threshold on the FLAIR MR sequence. The second one discards all voxels which do not belong to the white matter. The posterior map is obtained by using a forest with trees of depth 2 and thus highlights hyper-intense FLAIR voxels which lie in peri-ventricular regions.

The feature importance (cf. Equation 11) can be extended in a straightforward way to be parametrized not only by the type of feature (local, context-rich, symmetric) but also by the channel. When globally looking at the selected channels (cf. Figure 12), we notice that their importance varies throughout the tree: first layers, as mentioned before, favor detection of bright spots in the white matter by successively testing the FLAIR MR sequence, spatial priors on WM and GM tissues and finally testing on the T2 MR sequence; deeper layers take into account other modalities to adjust the segmentation.

## 6. Conclusion

We demonstrated the power of the RF formulation applied to the difficult task of MS lesion segmentation in multi-channel MR images. We presented

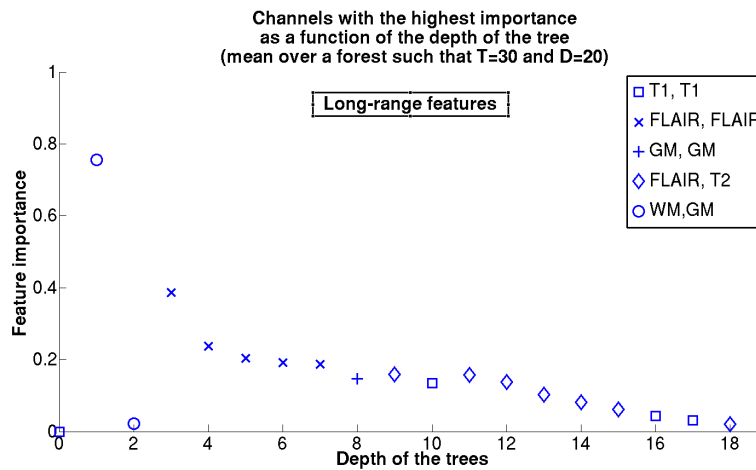
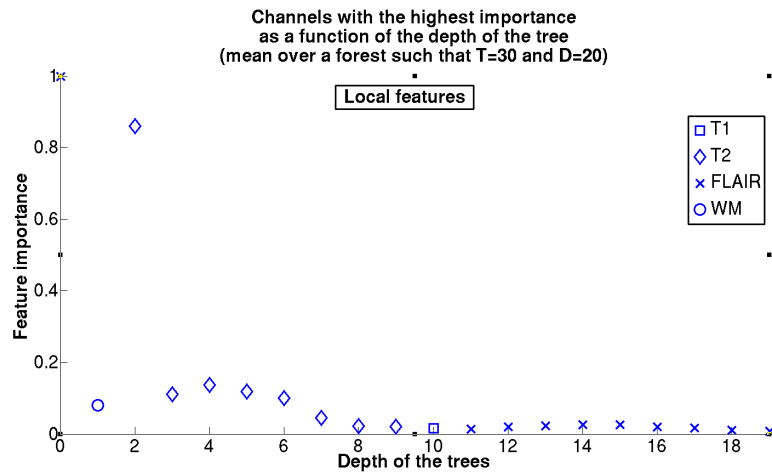


Figure 12: **Channel importance as a function of the depth of the tree.** Plots draw the channel importance drawn as a function of the depth of the tree for both local (top) and long-range features (bottom). For a fixed depth, only the most discriminative channel is depicted. Note how successive layers of the tree test complementary channels: the first layer performs a local test on the FLAIR MR sequence in order to detect bright spots, the second one discards all voxels which do not belong to the white matter by using context-rich information over the WM and GM channels. Note that a large spectrum of available channels is tested throughout the tree.



three kinds of 3D features based on multi-channel intensity, prior and context-rich information. Those features are part of a context-rich random decision forest classifier which demonstrated improved results on one of the state of the art algorithms on the public MS challenge dataset. In addition, the random decision forest framework provided a means to automatically select the most discriminative features to achieve the best possible segmentation. Future work could include the use of more sophisticated features to reduce even further the preprocessing requirements. The context-rich random forest framework presented in this article is generic which is an additional strength of the method. It can be applied as is to any other segmentation task, e.g. brain tumors segmentation in multi-sequence MR images of the brain. Finally, one could investigate an extension of the proposed approach to larger multi-class problems in order to try to simultaneously segment brain tissues (WM, GM, CSF) along with MS lesions.

## References

- Admiraal-Behloul, F., van den Heuvel, D., Olofsen, H., van Osch, M., van der Grond, J., van Buchem, M., Reiber, J., 2005. Fully automatic segmentation of white matter hyperintensities in mr images of the elderly. *NeuroImage* 28 (3), 607 – 617.
- Akselrod-Ballin, A., Galun, M., Basri, R., Brandt, A., Gomori, M. J., Filippi, M., Valsasina, P., 2006. An integrated segmentation and classification approach applied to multiple sclerosis analysis. In: *CVPR '06: IEEE*. pp. 1122–1129.
- Amit, Y., Geman, D., 1997. Shape quantization and recognition with randomized trees. *Neural Computation* 9 (7), 1545–1588.
- Anbeek, P., Vincken, K., van Osch, M., Bisschops, R., van der Grond, J., March 2004. Probabilistic Segmentation of White Matter Lesions in MR Imaging. *NeuroImage* 21 (3), 1037–1044.
- Anbeek, P., Vincken, K., Viergever, M., 2008. Automated MS-lesion segmentation by K-Nearest neighbor classification. In: *The MIDAS Journal - MS Lesion Segmentation (MICCAI 2008 Workshop)*.
- Andres, B., Köthe, U., Helmstaedter, M., Denk, W., Hamprecht, F. A., 2008. Segmentation of SBFSEM volume data of neural tissue by hierarchical classification. In: *DAGM-Symposium*. pp. 142–152.
- Bernhard Scholkopf, C. J. B., Smola, A. J., 1999. *Advances in Kernel Methods: Support Vector Learning*. MIT Press, Cambridge, MA.

- Bishop, C., 2006. Pattern Recognition and Machine Learning. Springer.
- Breiman, L., 2001. Random forests. *Machine Learning* 45 (1), 5–32.
- Breiman, L., Friedman, J. H., Olshen, R. A., J., S. C., 1984. Classification and Regression Trees. Wadsworth Press.
- Bricq, S., Collet, C., Armspach, J.-P., september 2008a. Ms lesion segmentation based on hidden markov chains. In: 11 th International conference on medical image computing and computer assisted intervention. Paper selected for "a grand challenge : 3D segmentation in the clinic". MICCAI, new-York September 6-10,.
- Bricq, S., Collet, C., Armspach, J.-P., december 2008b. Unifying framework for multimodal brain mri segmentation based on hidden markov chains. *Medical Image Analysis* 12 (6), 639–652.
- Criminisi, A., Shotton, J., , Bucciarelli, S., 2009. Decision forests with long-range spatial context for organ localization in CT volumes. In: MICCAI workshop on Probabilistic Models for Medical Image Analysis (MICCAI-PMMIA).
- Criminisi, A., Shotton, J., Robertson, D., Konukoglu, E., 2010. Regression forests for efficient anatomy detection and localization in CT studies. In: MICCAI workshop on Medical Computer Vision: Recognition Techniques and Applications in Medical Imaging (MICCAI-MCV).
- Datta, S., Sajja, B. R., He, R., Wolinsky, J. S., Gupta, R. K., Narayana, P. A., 2006. Segmentation and quantification of black holes in multiple sclerosis. *NeuroImage* 29 (2), 467 – 474.

- Dugas-Phocion, G., Ballester, M. Á. G., Malandain, G., Ayache, N., Lebrun, C., Chanalet, S., Bensa, C., 2004. Hierarchical segmentation of multiple sclerosis lesions in multi-sequence MRI. In: ISBI. IEEE, pp. 157–160.
- Evans, A. C., Collins, D. L., Mills, S. R., Brown, E. D., Kelly, R. L., Peters, T. M., 1993. 3D statistical neuroanatomical models from 305 MRI volumes. In: IEEE-Nuclear Science Symposium and Medical Imaging Conference. pp. 1813–1817.
- Freifeld, O., Greenspan, H., Goldberger, J., 2009. Multiple sclerosis lesion detection using constrained GMM and curve evolution. *J. of Biomed. Imaging* 2009, 1–13.
- Geurts, J. J., Barkhof, F., September 2008. Grey matter pathology in multiple sclerosis. *Lancet neurology* 7 (9), 841–851.
- INRIA, 2010. MedINRIA.  
URL [www-sop.inria.fr/asclepios/software/MedINRIA/index.php](http://www-sop.inria.fr/asclepios/software/MedINRIA/index.php)
- Lempitsky, V. S., Verhoek, M., Noble, J. A., Blake, A., 2009. Random forest classification for automatic delineation of myocardium in real-time 3D echocardiography. In: FIMH. LNCS 5528. Springer, pp. 447–456.
- Menze, B. H., Kelm, B. M., Masuch, R., Himmelreich, U., Petrich, W., Hamprecht, F. A., 2009. A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC Bioinformatics* 10, 213.
- Prima, S., Ayache, N., Barrick, T., Roberts, N., 2001. Maximum likelihood estimation of the bias field in MR brain images: Investigating different

- modelings of the imaging process. In: MICCAI. LNCS 2208. Springer, pp. 811–819.
- Prima, S., Ourselin, S., Ayache, N., 2002. Computation of the mid-sagittal plane in 3d brain images. *IEEE Trans. Med. Imaging* 21 (2), 122–138.
- Quinlan, J. R., 1993. C4.5: Programs for Machine Learning. Morgan Kaufmann.
- Rey, D., October 2002. Détection et quantification de processus évolutifs dans des images médicales tridimensionnelles : application à la sclérose en plaques. Thèse de sciences, Université de Nice Sophia-Antipolis, (in French).
- Shiee, N., Bazin, P., Pham, D., 2008. Multiple sclerosis lesion segmentation using statistical and topological atlases. In: *The MIDAS Journal - MS Lesion Segmentation (MICCAI 2008 Workshop)*.
- Shiee, N., Bazin, P.-L., Ozturk, A., Reich, D. S., Calabresi, P. A., Pham, D. L., 2010. A topology-preserving approach to the segmentation of brain images with multiple sclerosis lesions. *NeuroImage* 49 (2), 1524 – 1535.
- Shotton, J., Winn, J. M., Rother, C., Criminisi, A., 2009. Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. *Int. J. Comp. Vision* 81 (1), 2–23.
- Souplet, J.-C., Lebrun, C., Ayache, N., Malandain, G., 2008. An automatic segmentation of T2-FLAIR multiple sclerosis lesions. In: *The MIDAS Journal - MS Lesion Segmentation (MICCAI 2008 Workshop)*.

- Styner, M., Lee, J., Chin, B., Chin, M., Commowick, O., Tran, H., Markovic-Plese, S., Jewells, V., Warfield, S., Sep 2008a. 3D segmentation in the clinic: A grand challenge II: MS lesion segmentation. In: MIDAS Journal. pp. 1–5.
- Styner, M., Warfield, S., Niessen, W., van Walsum, T., Metz, C., Schaap, M., Deng, X., Heimann, T., van Ginneken, B., 2008b. MS lesion segmentation challenge 2008.  
URL <http://www.ia.unc.edu/MSseg/index.php>
- Van Leemput, K., Maes, F., Vandermeulen, D., Colchester, A. C. F., Suetens, P., 2001. Automated segmentation of multiple sclerosis lesions by model outlier detection. *IEEE Trans. Med. Imaging* 20 (8), 677–688.
- Wu, Y., Warfield, S., Tan, I., III, W. W., Meier, D., van Schijndel, R., Barkhof, F., Guttman, C., 09 2006. Automated segmentation of multiple sclerosis lesion subtypes with multichannel mri. *Neuroimage* 32 (3), 1205–1215.
- Yamamoto, D., Arimura, H., Kakeda, S., Magome, T., Yamashita, Y., Toyofuku, F., Ohki, M., Higashida, Y., Korogi, Y., 2010. Computer-aided detection of multiple sclerosis lesions in brain magnetic resonance images: False positive reduction scheme consisted of rule-based, level set method, and support vector machine. *Computerized Medical Imaging and Graphics* 34 (5), 404 – 413.
- Yi, Z., Criminisi, A., Shotton, J., Blake, A., 2009. Discriminative, semantic

segmentation of brain tissue in MR images. LNCS 5762. Springer, pp. 558–565.

Yin, P., Criminisi, A., Winn, J., Essa, I., 2010. Bilayer segmentation of webcam videos using tree-based classifiers. *Trans. Pattern Analysis and Machine Intelligence (PAMI)* 33.