A fully automated, hierarchical classification method for detecting white matter lesions in Multiple Sclerosis

Marco Battaglini[1], Nicola De Stefano[1], and Mark Jenkinson[2]
[1]*Neurological and behavioral sciences, University of Siena, Siena, Tuscany, Italy,* [2]*FMRIB Centre, Nuffield Department of Clinical Neurosciences, University of Oxford*

INTRODUCTION: Segmentation of white matter (WM) multiple sclerosis (MS) lesions is an important task. Development of fully automated methods (no user intervention) is a goal for which many attempts have been made. However, performance to date is not yet good enough for the use in clinical research studies, especially for different sets of images. One major problem with previous approaches is the large false positive rates. We have found this to be true for neural network classifiers [1,2], and so to improve this we propose using a **two-level, hierarchical classification** method: a first-level, voxel-wise classifier followed by a **novel second-level, cluster-wise** classifier. The second-level classifier works with the output of the first-level classifier and therefore cannot correct false negative errors (missed lesions), but is capable of reducing false positives (false lesions).

METHODS: ***Data:*** a small sample from a multisite clinical study dataset was taken (14 different sets of images), including 24 MS patients with T1w, T2w and PD images (0.97x0.97x3mm) plus manual lesion segmentations. Patients were split into a first-level training set of 10 subjects, a second-level training set of 5 subjects and a testing set of 9 subjects (assigned at random). ***Preprocessing:*** the T1w images were segmented into tissue-types white matter (WM), grey matter (GM) and cerebrospinal fluid (CSF). All images were linearly registered/transformed to the T2w space, and were intensity normalized by dividing by their 95th percentile. From these, an artificial **"PseudoFlair" (PsF)** image was calculated: PsF=(PD*T2w*T1w)/(PD+T2w); see Fig. 1b. ***Voxel-wise Features:*** (i) the PsF image intensity; (ii) its 3x3 neighbourhood; (iii) 3x3 and 9x9 multi-scale intensities (two 3x3 arrays of averages, using patches of size 3x3 and 9x9 voxels within the slice) [2]. ***Voxel-wise Classifier:*** a two-layer neural-network (perceptron) with 28 inputs (features), 100 hidden units and one sigmoidal output node. ***Second-level Classification:*** the voxel-wise output from the first-level classifier was thresholded at 0.5, clustered and used to calculate features. ***Cluster-wise Features:*** for each cluster, a set of ratios was calculated between statistics (mean, median, 75th/90th percentiles, or maximum intensity values) from two regions: (i) inside the cluster, *to* (ii) the exterior border of the cluster or all WM or all GM. These were done for each "sequence" separately (T1w, PD, PsF and "probability" output from the first-level classifier). A subset of 19 features was then selected, based on initial tests for discrimination power. ***Cluster-wise Classifier:*** a two-layer neural network with 19 input features, 100 hidden units and one sigmoidal output node. ***Performance Measures:*** Different thresholds (from 0.5 to 1.0) were applied to both the voxel-wise and cluster-wise neural network outputs when run on the testing set. In each case the True Positives (TP), False Positives (FP) and False Negatives (FN) were measured both for voxels and for lesion (cluster) counts, where a "true" cluster is defined by overlap with a manually segmented lesion. Voxel-based results are given in, terms of True Positive Rate TPR=TP/(TP+FN) and False Discovery Rate FDR=FP/(FP+TP), while lesion count results are given in terms of the ratio of the number of FP or TP clusters to the number of true lesions.

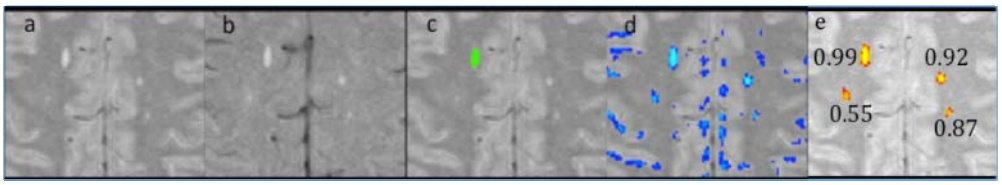

Fig. 1: (a) PD image; (b) PseudoFlair (PsF) image; (c) manual segmentation; (d) output from voxel-wise classifier, thresholded at 0.5; (e) output from the cluster-wise classification. Just 4 clusters were above the 0.5 threshold (second-level output values) and the numbers shown are the second-level output "probabilities". The colours shown in (e) are based on first-level output. Note that the yellow voxels (high probability) are more representative of the central lesion, with red voxels highlighting the border.

RESULTS: **Fig. 1** shows example images (inputs, manual lesion segmentation and the classifier outputs). A 0.5 threshold was applied in both cases and clearly shows many False Positives in the voxel-wise output (although the False Negative rate was less than 1%). The cluster-wise output is obviously much cleaner and, in this example, shows three "false positive" lesions, at lower "probability" values. **Fig. 2** shows results of the four quantitative performance measures. The most significant feature of these is that use of the **cluster-wise classifier** results in a **large reduction in False Positives** (left column) with only a **very minor change in True Positives** (right column). This is the case for both the lesion counting measures and the voxel-based measures. The largest change is seen in the False Positive lesion count, although this is grossly overestimated for low thresholds where fewer, but very large, clusters are formed, artificially lowering the cluster count. However, even for higher thresholds the reduction in false positives is substantial and offers clear improvements on what is possible with the first-level (voxel-wise) results, regardless of what threshold is applied.

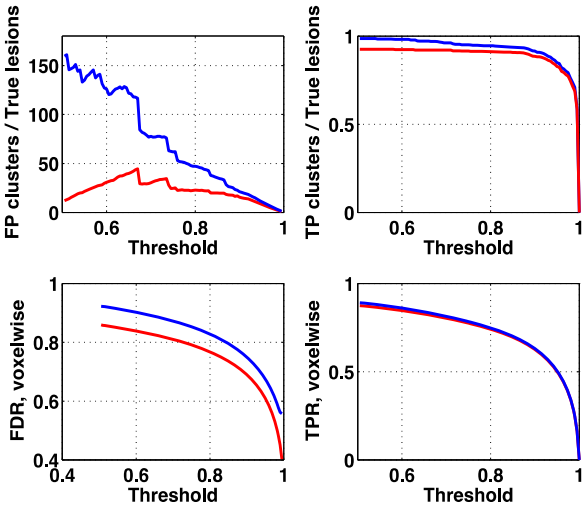

Fig. 2: Performance measures for lesion counts (top row) and voxels (bottom row). In each case the voxel-wise classifier results are shown in blue and the cluster-wise classifier results in red. Note that the cluster-wise classifier reduces False Positives greatly (for lesion count and for voxels) but only reduces the True Positives by a very small amount. [TP=True Positive; FP=False Positive; TPR=True Positive Rate; FDR=False Discovery Rate]

CONCLUSIONS: We have proposed a new, simple and user independent approach for detecting lesions using a two-level hierarchical classification scheme that employs both voxel-wise and cluster-wise classifiers (neural networks in this case, but not restricted to this at all). Results from different sets of images showed large reductions in False Positives while maintaining True Positives, demonstrating great potential for this approach. Future work will optimize the feature set and techniques for thresholding the classifier ("probability") outputs. We will also investigate what influence the errors in the manual segmentations have in this data.

REFERENCES [1] Zijdenbos et al, TMI 1994; [2] Battaglini, De Stefano & Jenkinson, ISMRM 2010