RimNet: A deep 3D multimodal MRI architecture for paramagnetic rim lesions detection in Multiple Sclerosis

Abstract
In multiple sclerosis (MS), the presence of a paramagnetic rim at the edge of non-gadolinium-enhancing lesions indicates chronic focal neuroinflammation, is a negative prognostic imaging biomarker, and represents a potential outcome measure in MRI-based clinical trials. Here, we present a novel convolutional neural network (CNN) architecture (RimNet) for the automated detection of paramagnetic rim lesions employing multiple MR imaging contrasts. RimNet is based on using 3D patches centered on candidate lesions as input to several CNN branches which are interconnected between them, at both the first network blocks and the last fully connected layers. In this work, we evaluate different MRI contrast combinations for the detection of lesion rims and we further propose an enhanced MRI contrast. Imaging data were acquired at 3 tesla on 4 different scanners from 2 different vendors, totaling 90 patients (with 297 rim lesions and around 2000 lesions without rim). Our work is quantitatively evaluated (5-fold-nested cross-validation) against current clinical evaluation and compared to other baseline architectures. Both lesion-wise (false positive and true positive detection rates of 0.05 and 0.8, respectively) and patient-level results (sensitivity of 0.87 and specificity of 0.97) are comparable to the performance of experts, thus supporting usage of RimNet for speeding up and standardizing rim analysis in MS.

Keywords: Paramagnetic rim, multiple sclerosis, magnetic resonance imaging, multimodal, image classification, convolutional neural networks.

1. Introduction
Multiple sclerosis (MS) is an immune-mediated disorder characterized by focal inflammatory demyelinated lesions in the brain and spinal cord. Lesion detection on magnetic resonance imaging (MRI) represents the basis for MS diagnosis and routine follow-up (Thompson et al., 2018). A wide number of computer-assisted tools and machine learning algorithms have been developed for automated detection and segmentation of these lesions. Recently, other lesion-related features have gained attention as imaging biomarkers due to their pathological specificity and their clinical value (Maggi et al., 2018, 2019; Absinta et al., 2019). One of these imaging biomarkers is the presence of a paramagnetic rim around brain white matter (WM) lesions that is visible in susceptibility-based MRI (see Fig. 1A). The presence of such rims at the edges of ”older” non-gadolinium-enhancing lesions is the imaging hallmark of chronic active lesions, which are associated with plaque-edge smoldering inflammation and a worse prognosis (Absinta et al., 2016). On a patient level, the presence of chronic active paramagnetic rim lesions is associated with a more aggressive disease course (Absinta et al.,
Figure 1: Example of MS lesions on T2*\textsubscript{w} phase (left), magnitude (center) and FLAIR* (right) modalities. (A) and (D) are clear examples of rim+ and rim- lesions, respectively. In (C) there are two more subtle rim+ lesions. (B) is an example of a rim- lesion that has a rim-like intensity artefact.

These lesions have a potential prognostic clinical value and deserve to be explored as an outcome measure in MRI-based clinical trials (Absinta et al., 2013, 2016, 2018, 2019). Today, the presence/absence of paramagnetic rims in MS lesions is determined through visual inspection by experts, which is tedious, time consuming and prone to intra- and inter-rater variability. An automated method to distinguish rim-positive from rim-negative lesions (respectively, rim+ and rim-) could support physicians in their evaluation and ease the application of this promising MRI biomarker in daily clinical practice. From a machine learning perspective, the rim+ vs. rim- classification faces two major challenges: 1) the intensity features of the rim are not necessarily discernible from the internal lesion parenchyma and/or surrounding white matter tissue (Fig. 1C) or, on the contrary, some rim-like intensity artefacts may appear (Fig. 1B); and 2) the relatively low number of labeled cases for training due to the lowest frequency of rim+ vs. rim- lesions in MS patients (unbalanced dataset with large majority of rim- lesions).

No solutions to automated rim+/rim- classification problem have been published to date. In this paper, we propose a tailored 3D patch-based CNN architecture to exploit different MR imaging contrasts. Previously proposed classification networks exploit multimodal inputs at different architecture levels. For instance, 2D approaches merge different feature representations before feeding them to the network ((Yang et al., 2018)). The caveat of this strategy is that it does not process distinct information coming from each representation but rather a joint combination in which individual contributions are weakened. Some approaches mix multimodal feature maps at all network layers (Zhang et al., 2018). However, this type of multimodal fusion dramatically increases the number of trainable parameters, particularly in the last fully-connected layers. This can pose additional training challenges. Other methods rather rely on connecting multiple CNN branches (one per modality) at the last layers only (with dense layers (Huang et al., 2019) or ensemble of shallow classifiers (Kumar et al., 2016)) thus keeping the feature extraction independent for each modality. Our proposed architecture is a compromise between the above approaches, mixing multimodal features at first and last blocks of the network. Moreover, we present
a quantitative evaluation (at both lesion and patient level) on a large cohort, with images from four different scanners, in comparison with manual annotations of two experts. RimNet performance is equivalent to that of the experts, thus supporting its potential for accelerating and standardizing the rim analysis in MS patients.

2. Dataset

The study received approval by the local ethics committee and all patients gave written informed consent prior to participation. A total of 90 patients were scanned at three hospitals (blinded for review): 55 at Hospital 1, 30 at Hospital 2 and 5 at Hospital 3. Age and Expanded Disability Status Scale (EDSS) (Kurtzke, 1983) is shown in Fig. 2A. Distribution of rim+/− lesions per patient is shown in Fig. 2B.

**MRI acquisition.** Brain images were acquired at 3 tesla on four different MRI scanners: MAGNETOM Skyra and MAGNETOM Prisma (Siemens Healthcare, Erlangen, Germany) in Hospital 1, MAGNETOM Prisma (Siemens Healthcare, Erlangen, Germany) in Hospital 2 and Ingenia (Philips, Best, The Netherlands) in Hospital 3. In all centers, three-dimensional (3D) T2*-weighted segmented echo-planar imaging (EPI), giving high resolution T2*-weighted magnitude and phase images (referred as T2* and phase), and 3D T2-FLAIR images were acquired. A 3D T1-weighted MPRAGE sequence was available in Hospital 1 and a 3D MP2RAGE in Hospital 2. T2* images had a 0.65x0.65x0.65mm³ resolution in Hospital 1 and Hospital 2 and 0.55x0.55x0.55mm³ resolution in Hospital 3. 3D FLAIR, MPRAGE and MP2RAGE images were acquired with an isotropic resolution of 1mm in all centers. FLAIR* images were computed by upsampling FLAIR to T2* resolution and then co-registering and voxel-wise multiplying it by the T2* images (Sati et al., 2012).

**Expert annotations of rims.** Two neurologists with experience in MS imaging annotated rim+ lesions manually. For exploring the presence of rims, the unwrapped phase images were primarily used. This modality depicts the phase shift mainly produced by iron-laden macrophages and relative myelin content at the lesion edge. Moreover, FLAIR* was used in the annotation process to localize MS lesions and therefore discard potential false positives due to rim-shaped artefacts. After a first screening conducted individually
by each expert, 37% of the lesions underwent consensus review, which was done in a second joint screening by the two experts, yielding 297 rim+ lesions detected in consensus (which will be considered as the ground truth). Rim- lesions were annotated as follows. Automated lesion segmentation (La Rosa et al., 2018) was done either using MPRAGE (or MP2RAGE) and FLAIR (Hospital 1 and Hospital 2) or only FLAIR (Hospital 3). Rim- lesions were extracted by excluding the manually annotated rim+ lesions from those. In order to balance samples, we discarded all lesions smaller than 100 voxels and bigger than 10,000 voxels, which were 2028 rim- and 3 rim+ lesions. Overall our dataset of 90 patients contains 295 and 2022 rim+ and rim- lesions, respectively.

3. Proposed multimodal detection framework

Pre-processing and patch extraction. Following the experts’ workflow, the phase contrast was used for automated detection. We also explored the contribution of T2*, in which rims with high iron content can be clearly visualized. Both MRI contrasts did not undergo any pre-processing. Additionally, we also used the FLAIR* with the purpose of focusing on the lesion without need of relying on a binary segmentation mask. To this end, FLAIR* is enhanced by element-wise squaring it and then applying a box linear filter (kernel size of 6x6x6 voxels). The resulting image, eFLAIR*, encodes morphometric features of the lesions (see Fig. 3).

Image patches were extracted centered at the center of mass of automatically detected lesions (as suggested in La Rosa et al. (2018)). Our preliminary analysis concluded that a patch of 28x28x28 voxels was a good trade-off between a reasonable input size and a good coverage of lesions within a patch.

Data augmentation. We applied some preprocessed data augmentation to tackle the unbalanced class problem, as this experimentally showed better results than adding class weights in the loss function. This includes rotating each rim+ lesion by 90°, 180° and 270° in the X and Y axes, which leads to having a balanced dataset, with a 50.5% of rim+ lesions. Besides, three elastically deformed versions of each lesion were generated, effectively quadrupling the training data. We also performed online data augmentation during the training first by randomly rotating the patches by 0°, 90°, 180° or 270° in the Z axis, then flipping it along an axis (X, Y, Z or none) and finally translating the patch 2
Network architecture. The prototype RimNet, see Fig. 4, is built upon three parallel CNNs based on the VGGNet (Simonyan and Zisserman, 2014), which has been proven to outperform other state-of-the-art architectures in similar multimodality approaches (Le et al., 2017). In our architecture, each branch is trained with unique weights and receives a different modality as input, to which patch-based global normalization has been previously applied. Each single CNN contains three successive blocks of two convolutional layers and a max pooling layer. For the convolutional layers, Xavier initialization (Glorot and Bengio, 2010) is performed, a hyperbolic tangent (tanh) is used as activation function and batch normalization is applied. In order to exploit multimodal low-level features, the output of the first eFLAIR* block is concatenated to the outputs of the other modalities’ first blocks. Finally, the three four-dimensional output tensors of each CNN are concatenated before being fed to a final succession of fully connected layers.

Thanks to the straight-forward parallel flow of the data of the three modalities, the network benefits from the prediction capacities of each contrast. Multimodal capabilities are exploited through the last fully connected layers. More importantly, by mixing the first blocks’ outputs, the network is led to extract eFLAIR* low-level features which, combined with phase and T2* low-level features, demonstrate good prediction capabilities (see Section 4).
4. Results

**Evaluation procedure.** We compared RimNet with different baseline approaches. Mono-modal networks were considered simply as one CNN branch of the RimNet directly connected to the fully connected layers. As multimodal baseline network we relied on recent work (Huang et al., 2019), which concatenates the outputs of three monomodal branches of convolutions and pooling layers, one for each modality, and feeds them to a cascade of fully connected layers. In order to avoid over-fitting and to better reflect the performance in a real clinical scenario, results were conducted with a 5-fold nested cross-validation procedure, splitting our dataset into 4 training folds and 1 testing fold (80-20%). All architectures were trained using the ADAM optimizer (Kingma and Ba, 2014) with learning rate decay and early stopping. For each fold’s distribution, a 4-fold inner cross-validation determined the number of epochs trained with each learning rate ($10^{-4}$, $7.5 \cdot 10^{-5}$, $5.0 \cdot 10^{-5}$, $2.5 \cdot 10^{-5}$, $1.0 \cdot 10^{-5}$). The change of learning rates was triggered after 3 consecutive epochs of validation loss increase. Early stopping was applied when the last learning rate change was triggered.

**Single-modality exploration.** A prior exploration of the prediction capabilities of each modality shows that all modalities can, with different contributions, help to predict rim+ lesions (see Fig. 5). Undoubtedly, the phase was the modality that best detected rim+ lesions, which is consistent with the results presented in recent literature (Absinta et al., 2018). The fairly good performance of eFLAIR* in a single-modality approach suggests that morphometrical features such as size and shape play an important role in our classification problem.

**Multimodality approaches.** ROC and precision recall (PR) curves are shown, with the latter having special importance due to our highly imbalanced dataset (Saito and Rehmsmeier, 2015), in Fig. 5. RimNet clearly outperformed (AUC=0.966) the multimodal baseline (AUC=0.944) and all monomodal approaches (AUC=[0.885, 0.775, 0.754]). While ROC curves may not be very conclusive, differences in the true positive rates (TPR) when setting low false positive rates (FPR) are unambiguous. For example, with a FPR of 0.05,
Figure 6: Patient-wise analysis depending on the number of rim+ lesions set to consider a patient as chronic active. RimNet is evaluated choosing a FPR of 0.05. On the left, comparison with individual expert performance. On the right, confusion matrices showing the comparison between RimNet and experts assessment.

corresponding TPR values were 0.800 for the RimNet, 0.772 for the RimNet_NE (trained with FLAIR* instead of eFLAIR*), 0.695 for RimNet_S (RimNet simplified without the T2* branch) and 0.708 for the baseline. RimNet miss rate (20%) was as low as experts’, who each failed to identify 12.9% and 24% of rim+ lesions that were agreed upon at consensus review.

The outperformance of RimNet over the multimodal baseline architecture shows the contribution of fusion outputs after the first block. We hypothesize that this improves the extraction of low-level morphometrical features and uses them to enhance phase and T2* feature extractions. Our results also show a better performance with the proposed eFLAIR* (with a lesion highlighting and context softening) than with the FLAIR*. We think this is mainly due to the squaring function, as it cannot be easily learned by CNNs. However, further analysis exploring the features used by the network is needed to confirm this hypothesis.

**Patient-wise analysis.** At the patient-level, Absinta et al. (2019) showed that having four or more rim+ lesions was correlated to an earlier motor and cognitive disability, which characterizes a worse clinical prognosis. We thus evaluate the RimNet patient classification performance based on the number of rim+ lesions per patient. To do so, we set a lesion-wise accepted FPR of 0.05 and evaluated if patients were harboring or not chronic inflammation (respectively "chronic active" and "non-chronic active" patients) based on previously proposed thresholds of rim+ lesions per patient, ranging from 1 to 4. The first column in Fig. 6 shows that the performance of RimNet when attempting to differentiate chronic active from non-chronic active patients is comparable (particularly when considering 3 or
more rim+ lesions) to experts’ individual performance (computed by considering the post-consensus lesion labels as ground truth). With the previously used threshold of 4 rim+ lesions, RimNet misclassifies only 5 of the 90 patients, with sensitivity of 0.870 (0.826 and 0.739 for experts) and specificity of 0.970 (0.985 and 1.000 for experts), respectively.

5. Conclusion

We present the prototype RimNet, the first automated method for detection of paramagnetic rims in MS lesions. We rely on multimodal MRI input 3D patches, including T2*-phase and magnitude contrasts that depict different magnetic properties of the rim. Moreover, we show the ability of an enhanced FLAIR* sequence, which captures morphological features of the lesions, to help the detection. We concatenate the different modalities at both the first blocks and the last layers of the network. We extensively validated our contributions in a cohort of 90 patients, scanned in 4 different MR machines, and in comparison with other mono and multimodal baseline architectures. RimNet showed better performance (TPR=0.8) than the multimodal baseline (TPR=0.708), both with fixed FPR=0.05. Overall, RimNet excelled at distinguishing rim+ lesions (only fails at 20% of rim+ lesions, while experts fail at 12.9% and 24.1% of them) while keeping a low number of false positives (5%). Jointly with the very good patient-wise performance, which is also comparable to the experts’, RimNet appears as a promising automated method to support physicians in the rim analysis in MS patients, by providing a first accurate rim+ estimation in less than a second (300ms1) instead of 10 minutes per patient – the average time needed by an expert (without counting previous lesion detection). As future work, we aim at (i) including other MRI modalities (e.g. MPRAGE, MP2RAGE, FLAIR), (ii) using a pure testing set for further validation.

Acknowledgments

Acknowledgments withheld.

References


1. Experiment conducted on a desktop Intel(R) Core i7-4790 CPU machine at 3.60GHz and a GeForce GTX 1080Ti GPU.


Takaya Saito and Marc Rehmsmeier. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PloS one*, 10(3), 2015.


