CNN Prediction of Future Disease Activity for Multiple Sclerosis Patients from Baseline MRI and Lesion Labels

Nazanin Mohammadi Sepahvand[1], Tal Hassner[2], Douglas L. Arnold[3,4], and Tal Arbel[1]

[1] Centre for Intelligent Machines, McGill University, Montreal, Canada
nazsepah@cim.mcgill.ca
[2] The Open University of Israel, Israel
[3] Montreal Neurological Institute, McGill University, Montral, Canada
[4] NeuroRx Research, Montral, Canada

Abstract. New T2w and gadolineum-enhancing lesions in Magnetic Resonance Images (MRI) are indicators of new disease activity in Multiple Sclerosis (MS) patients. Predicting future disease activity could help predict the progression of the disease as well as efficacy of treatment. We introduce a convolutional neural network (CNN) framework for future MRI disease activity prediction in relapsing-remitting MS (RRMS) patients from multi-modal MR images at baseline and illustrate how the inclusion of T2w lesion labels at baseline can significantly improve prediction accuracy by drawing the attention of the network to the location of lesions. Next, we develop a segmentation network to automatically infer lesion labels when semi-manual expert lesion labels are unavailable. Both prediction and segmentation networks are trained and tested on a large, proprietary, multi-center, multi-modal, clinical trial dataset consisting of 1068 patients. Testing based on a dataset of 95 patients shows that our framework reaches very high performance levels (sensitivities of 80.11% and specificities of 79.16%) when semi-manual expert labels are included as input at baseline in addition to multi-modal MRI. Even with inferred lesion labels replacing semi-manual labels, the method significantly outperforms an identical end-to-end CNN which only includes baseline multi-modal MRI.

Keywords: Multiple Sclerosis· Magnetic Resonance Imaging · Disease Activity · Deep Learning

1 Introduction

Multiple sclerosis is traditionally known as a chronic inflammatory demyelinating disease of the central nervous system [7]. The presence of lesions in MRI is one of the hallmarks of MS. As a result, MRI has been used for diagnosis and to monitor disease progression and treatment response. The number of new or enlarging T2w lesions as well as gadolinium-enhancing lesions have been used as markers of disease activity [14, 16, 20] which in turn is used as a clinical outcome to monitor the progression of disease and also the efficacy of new treatments in clinical trials for RRMS [10, 21]. Hence, developing an automatic method to predict future disease activity from MRI could lead to better understanding of disease progression and help identify patients that can benefit from treatment. However, given the variability of lesion distribution in MRI, complexity of the evolution of lesions over time and the heterogeneity of the disease across the population in terms of clinical disease course, there are currently no established MR biomarkers that reliably predict the future disease activity.

We define future MRI disease activity as the *presence of any new/enlarging* $T2w$ *lesions or gadolinium lesions at any period within two years of the trial.* Examples of active and inactive patients are shown in Fig.1. Traditional biomarkers such as lesion counts and lesion volumes at baseline are not reliable predictors of future MRI disease activity due to the complexity of the disease. Furthermore, the presence of lesions at baseline does not necessary guarantee future MRI activity. Figure 2 illustrates how the absence of lesions at baseline does not guarantee the lack of future MRI disease activity, and a high lesion load at baseline does not guarantee the appearance of new/enlarging lesions two years ahead. As such, more sophisticated biomarkers are needed to reliably predict future MRI disease activity.

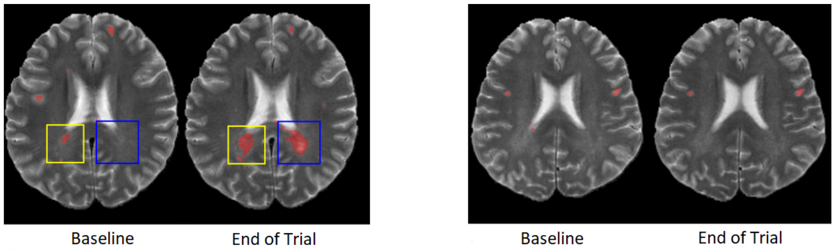


Fig. 1. Examples of active (left) and inactive patients (right). Lesions are highlighted in red. In each panel, the left image depicts T2w MRI at baseline and the right image is the same patient at the end of the second year. The patient in the left panel shows one enlarging (yellow bounding box) and one new lesion (blue bounding box) near the ventricles at year two. In the right panel, no new or enlarging lesions are present at year two.

Although several automatic prediction methods predict the conversion of patients with preliminary symptoms to MS [1, 3, 23], only recently has the first machine learning approach been proposed for the prediction of future MS disease activity, in terms of future new/enlarging T2w lesions, based on baseline MRI [5]. This approach—a random forest classifier based on a *Bag-of-Lesion* representation—led to promising results on a proprietary dataset (sensitivity at 68.0% and specificity at 57.0%).

In recent years, deep learning has provided a wide range of powerful alternative frameworks with impressive results in both computer vision [2, 12] and

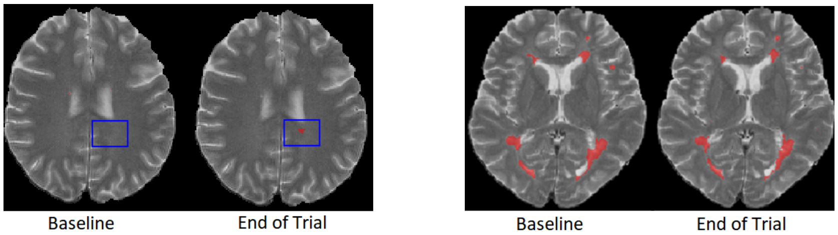


Fig. 2. Examples of patients whose baseline lesion loads are not good indicators of future disease activity: (left) active; (right) inactive. Lesions are highlighted in red. In each panel, the left image represents T2w MRI at baseline and the right image represents the same patient at the end of the second year. The patient in the left panel shows no baseline lesions, yet develops a new lesion (blue bounding box) at the end of the second year. The patient on the right, however, shows no new or enlarging lesions at the end of second year, despite having high lesion load at baseline.

medical imaging fields [4,13]. In this work, we present the *first automatic, deep learning framework, a 3D CNN, for predicting future disease activity* from baseline MRI of patients with RRMS. We show how activity prediction accuracy is further enhanced through the inclusion of binary T2w lesion labels at baseline as inputs. These lesion labels help the network focus on areas of the brains with lesions, thereby aiding in network training. As expert lesion labels are expensive, time consuming, and hard to obtain, we further evaluate the performance of our prediction network in settings where semi-manual expert T2w lesion labels are not available. To this end, we develop a 3D Unet segmentation network [17] for automatic inference of T2w lesion labels. These labels are then used as an input to the prediction network.

Both prediction and segmentation networks are trained on a proprietary, multi-scanner, multi-center, clinical trial dataset of 1068 patients with RRMS. The performance of both networks is examined at operating point of interest on Receiver Operating Characteristic (ROC) curves. Our results indicate that using only baseline MRI and semi-manual expert lesion labels at baseline leads to very high prediction accuracies for future MRI disease activity over the next two years (accuracies of 80.21% and precision of 91.82%). This leads to the possibility of future development of precision medicine in RRMS. We further show that inclusion of baseline T2w lesion labels inferred from a Unet segmentation network provides good results, while still significantly outperforming a CNN based only on baseline MRI as inputs.

2 Proposed Framework

We develop a 3D CNN to predict future disease activity from baseline MR sequences. The prediction network takes three baseline MRI sequences (T1w, T2w and FLAIR) as well as T2w lesion labels as input for each patient and produces future MRI disease activity as binary labels (*active*/*inactive*). Should semi-manual lesion labels not be available at baseline, these labels are estimated using a proposed segmentation network. Inferred lesions are then fed into our prediction network along with baseline MR modalities (see Fig. 4(a)). Our segmentation network is a modified Unet, trained to segment lesions from baseline MR sequences. This network takes as input four acquired MR sequences (T1w, T2w, FLAIR, and Proton Density) and generates a lesion label with the same dimensions as the input brain volume (Fig. 4(b)).

2.1 Activity prediction network

The prediction network is a 3D CNN network with five convolutional layers followed by two dense layers. The architecture of the prediction network is illustrated in Fig. 3. As evident from the figure, each convolutional layer consists of two consecutive 3 × 3 × 3 convolutions, each followed by a rectified linear unit (ReLu). In each convolutional layer, barring the first, a 2 × 2 × 2 max-pooling with strides of two follows the two convolution units. The initial number of filters (feature maps) is set to four and this number is doubled after each max-pooling. Two fully-connected layers—with 16 and one neuron(s), respectively—are appended to the output of last convolutional layer. At the end of each convolutional layer, batch normalization [9] is applied. The network is trained with a dropout probability of 0.5 applied to the layers before both dense layers.

2.2 Segmentation network

Should manual labels not be available, we wish to infer the lesion labels through an automatic framework. To this end, we develop a segmentation network which

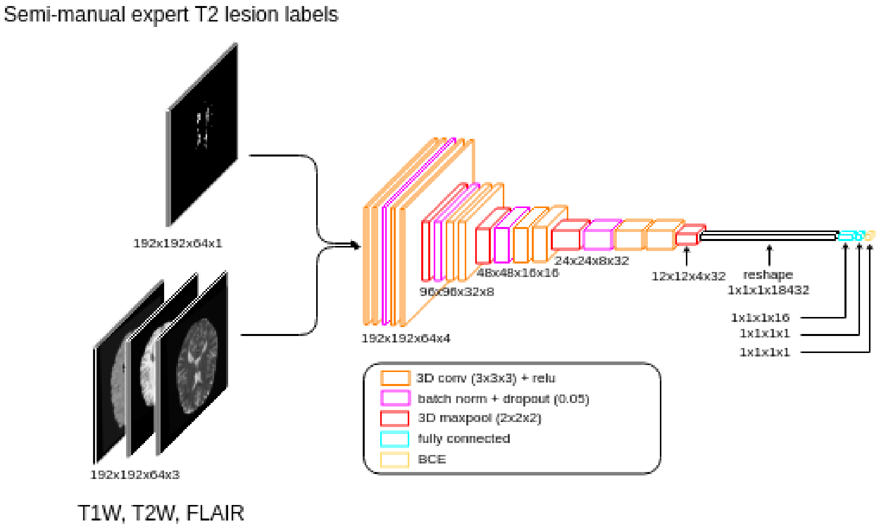


Fig. 3. Architecture of the activity prediction network. The network takes as input three MR sequences as well as T2w lesion labels at baseline and predicts future disease activity. All operations including convolution, max-pooling, and up-sampling are applied to 3D volumes.

is a modified Unet (see Fig. 4(b)). Similar to the standard Unet, our network consists of contracting (encoding) layers followed by expanding (decoding) layers. The structure of the encoding path is very similar to the prediction network. Five convolutional layers each containing two $3 \times 3 \times 3$ convolutions where each convolution unit is followed by a ReLu. Every layer, except the first, is followed by a $2 \times 2 \times 2$ max-pooling layer with strides of two for down-sampling.

In the decoding path, each layer consists of a deconvolution of $3 \times 3 \times 3$ with strides of two for upsampling, a concatenation with the correspondingly feature map from the contracting path, and two $3 \times 3 \times 3$ convolutions each followed by a ReLu. The number of feature maps is halved after each upsampling. Similar to the prediction network, batch normalization is applied to the output of each convolutional layer in the encoding path. In the decoding path, batch normalization is applied to the output of deconvolution unit in each layer.

2.3 Network training

Training of both segmentation and prediction networks is performed by the Adam optimizer [11] using standard cross entropy loss [8]. To deal with the class imbalance in the segmentation task, the two classes (*lesion*/*non-lesion*) are weighted in a manner which is inversely proportional to their frequencies. Specifically, the weight for each class (w_{c_i}) is defined as total number of voxels (vox_{tot}) divided by the total number of class voxels ($n_{vox_{c_i}}$):

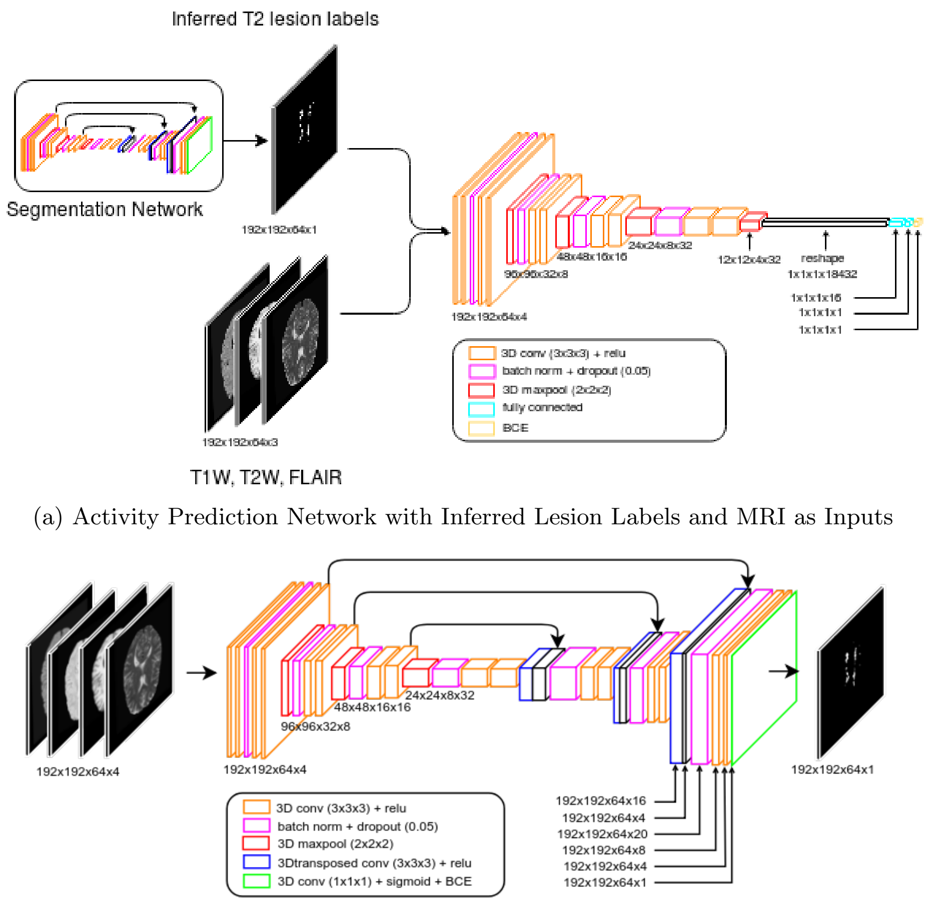
$$w_{c_i} = n_{vox_{tot}}/n_{vox_{c_i}} \quad c_i = 0(\text{inactive}), 1(\text{active}). \tag{1}$$

This produces a rough ratio of 1 (lesion) to 800 (non-lesion). The prediction task suffers from a similar imbalance problem: 75% of the patients in our trial are labeled as *active*. We address this imbalance by oversampling from the minority (inactive) class so that each batch has equal number of active and inactive samples. Since the ratio of active to inactive patients is 3 to 1, oversampling can be achieved by replicating inactive samples three times. This way, the total number of active and inactive samples are equal and therefore each batch can contain equal number of samples from both classes.

3 Experiments and results

3.1 Data sets

To validate our framework for predicting future activity from baseline MRI, we conducted experiments using a proprietary dataset consisting of 1068 MS patient brain images, acquired during a large, multi-center, multi-scanner clinical trial. The trial was two years long, and MR scans were obtained at the beginning of the trial (baseline) and at the end of the first and second years. While samples from all time points were used for training, validation, and testing of the segmentation network, only baseline samples were used for the prediction network. T1-weighted (T1w), T2-weighted (T2w), Proton Density-weighted (PDw) and



(a) Activity Prediction Network with Inferred Lesion Labels and MRI as Inputs

(b) Lesion Segmentation Network

Fig. 4. (a) Activity prediction network with automatically inferred labels as inputs, (b) lesion segmentation network. (a) The prediction network takes as input three MR sequences as well as T2w lesion labels generated by the segmentation network depicted in (b) and predicts future MRI disease activity. (b) The segmentation network takes as input four MR sequences and estimates T2w lesion labels at the baseline which are then fed into the prediction network demonstrated in (a). All operations including convolution, max-pooling and up-sampling are applied to 3D volumes.

Fluid-attenuated inversion (FLAIR) are among MR sequences available for each subject and each timepoint. The dimensions of each volume are $192 \times 192 \times 64$, providing a resolution of $1mm \times 1mm \times 3mm$. Pre-processing included brain extraction [19], bias field inhomogeneity correction using N3 [18], Nyul image intensity normalization [15], and registration of all images to MNI-space.

In addition to four MR modalities, semi-manual expert T2w lesion labels are included in this dataset. These are comprised of 3D volumes with binary labels for lesions/non-lesions at each voxel. These labels are available for each patient at each time point and used for training the segmentation network.

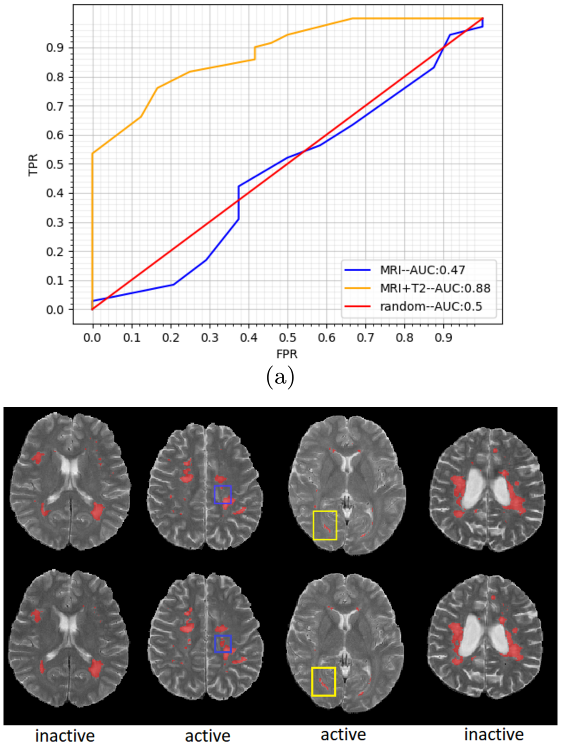
They were obtained through a semi-manual process where lesion labels were first generated by an in-house automated segmentation algorithm and then corrected by a trained expert reader.

Clinical data provided for this trial includes gadolinium and new/enlarging T2w lesion counts at the end of both first and second year. Due to missing clinical data for some patients, the total number of patient data available for the prediction task was 937. To calculate new/enlarging T2w lesion counts, lesion labels were generated through expert validation of an automatic longitudinal MS lesion segmentation framework [6]. Gadolinium counts were also provided for each patient in this dataset. These were estimated from post contrast T1w MRIs obtained after administration of contrast media (gadolinium). Gadolinium lesion segmentation was performed manually by trained experts. Binary MRI disease activity labels were defined based on the provided lesion counts. A patient was defined as being *active* if they had one or more new/enlarging T2 or gadolinium lesions.

3.2 Results

The dataset is divided into a training (80%), a validation (10%), and a test set (10%) in such a way that the ratio of active/inactive is the same (3/1) for all three splits. In this section, we first report the results of the prediction network trained to predict future disease activity from baseline multi-modal MRI as well as semi-manual expert lesion labels. This will be compared against using multi-modal MRI alone as inputs to the network. Next, the results of the lesion segmentation network will be shown. This leads to a quantitative analysis of the performance of the same network but with the semi-manual expert lesion labels replaced with the inferred lesions generated automatically by the segmentation network. Finally, a full comparison of the activity prediction results for all cases will be provided.

I) Prediction with semi-manual expert T2w lesion labels. A 3D CNN network that takes as input three MR sequences (T1w, T2w and FLAIR) as well as T2w lesion labels at baseline is trained on a proprietary dataset. Testing the performance of this network on a test set of 95 patients results in a sensitivity of 80.11% and a specificity of 79.16% (accuracies of 80.21%, precision of 91.82%) suggesting that our network is able to reliably predict the future MRI disease activity. To further evaluate the performance of our prediction network, an identical CNN network with only baseline MR sequences as input is trained. ROC curves, defined as True Positive Rate (TPR) vs False Positive Rate (FPR), for these two experiments are shown in Fig.5 (a). As is evident from the ROC curves, training an identical network with the same parameters and hyper-parameters values and using only baseline MR images leads to a performance barely better than random (with sensitivity of 8.45% and specificity of 79.16%). The low sensitivity in the baseline network is due to the fact that the network predicts the majority of the patients as inactive. This result shows that including lesion labels can significantly improve the performance of the prediction network over MRI



(a)

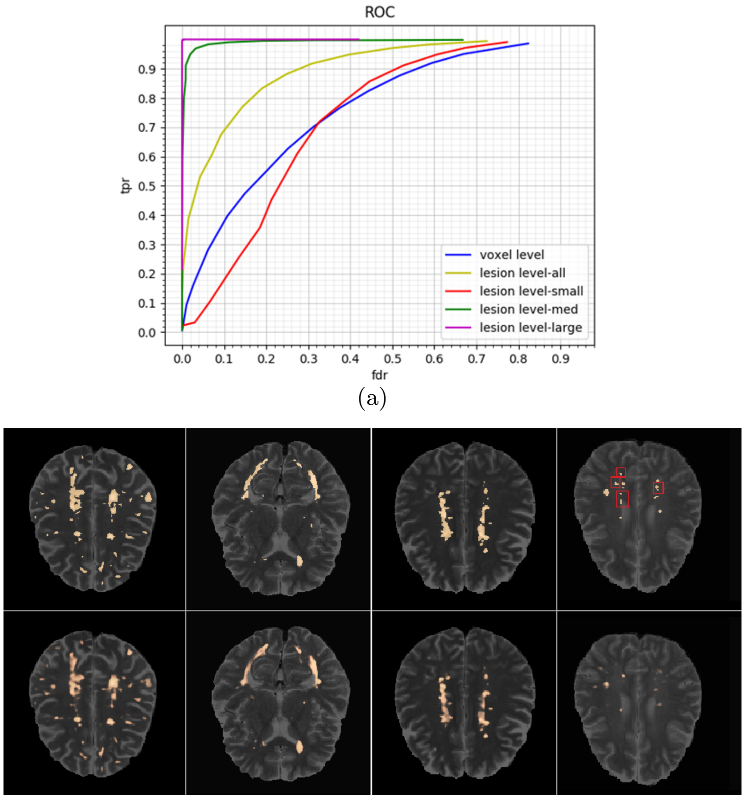
(b)

Fig. 5. Quantitative and qualitative results of the prediction network. (a) ROC curve with different inputs: Orange line: three MR sequences plus semi-manual, expert T2w lesion labels; blue: baseline MRI sequences alone; red: random performance. (b) Examples of brains for which the network prediction is successful: (top) Baseline MRI and lesion labels (red); (bottom) Year 2 MRI and lesion labels. Blue box depicts example of a new lesion. Yellow box depicts an enlarging lesion.

alone. Qualitative examples of images for which network prediction is successful are also depicted in Fig. 5(b).

II) T2w Lesion Segmentation and Detection. We examine the case where semi-manual lesion labels are not available, and are instead generated automatically. To this end, we train a Unet segmentation network with four MR sequences (T1w, T2w, FLAIR, and PDw) as input to estimate the lesion labels. Later, the inferred lesion labels will be fed, along with three MR sequences (T1w, T2w, FLAIR) at baseline, to the 3D CNN activity prediction network.

The segmentation network is evaluated by reporting ROC curves for the test data. ROC curves, defined as TPR vs. False Detection Rate (FDR), are reported in Fig. 6 for both voxel-level segmentation and lesion-level detection. To obtain



(a)

(b)

Fig. 6. Quantitative and qualitative results of the segmentation network. (a) ROC curves for both voxel level segmentation (blue) and lesion level detection split into three groups: small (red curve), medium (green) and large (magenta) lesions. (b) Examples of semi-manual expert segmentation (top) vs. output of segmentation network (bottom). While the segmentation network performs well for the three first brain images, a few small lesions are missed in the last case (inside red bounding boxes).

lesion-level detection statistics, TPR and FDR, from voxel segmentations, semi-manual, expert lesions smaller than three voxels are removed, in accordance with clinical protocol [22]. Candidate detected lesions are inferred using a simple connected component labeling method: a lesion is labeled as a true positive if the segmentation and its 18-connected neighborhood overlaps with at least three, or more than 50%, of the expert lesion voxels. Otherwise, it is labeled as a false positive. Insufficient overlap leads to a false negative. Lesion-level detection results are split into three groups of small (3-10 voxels), medium (11-50 voxels) and large (51+ voxels) according to their sizes and the ROC curves for each group is plotted separately in Fig. 6. The results indicate that, although the segmentation network performs very well for large and medium size lesions, it performs worse for small lesions. Qualitative examples of inferred lesion label slices shown against expert labels are also depicted in Fig. 6(b).

III) Comparison of Results. We now compare the results of three experiments, where the architecture of the prediction network is fixed, and network inputs are varied:

End-to-end CNN-MRI: We train an end-to-end CNN with all four MR sequences (T1w, T2w, FLAIR, and PDw) as inputs, with no lesion labels at baseline.

End-to-end CNN-MRI+T2: The PDw sequence is replaced with *semi-manual, expert* lesion labels at baseline. PDw sequence is selected as it contains information that is available in the remaining sequences and is thus the least informative modality among all available MR sequences.

End-to-end CNN-MRI+segm T2: The semi-manual, lesion labels in the previous experiment are replaced with inferred labels from the segmentation network.

ROC curves for these three experiments are provided in Fig. 7. In addition, accuracy, precision, sensitivity, and specificity at an operation point of interest (FPR=0.2) for each experiment are reported in Table 1. It is evident that the performance of the network with MRI alone as inputs is barely better than random, and that adding the semi-manual, expert lesion labels as input considerably improves the performance of the prediction network. One possible explanation for the poor performance with MRI alone is that, due to the limited size of the dataset, in order to avoid overfitting, restiction in the capacity of the model results in insufficient ability to learn from only baseline MR images. Our solution is to add lesion labels as an extra input modality to facilitate the training process by helping the network focus on the areas of interest.

The results also suggest that, although adding inferred lesion labels did not improve performance over MRI alone as much as adding the semi-manual expert

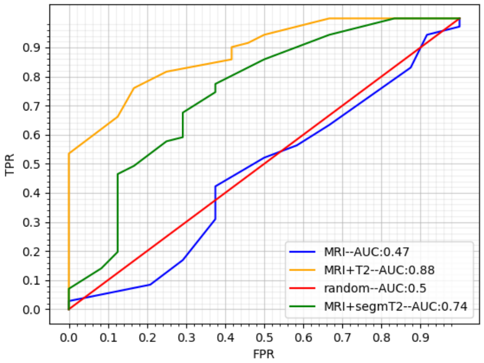


Fig. 7. ROC curves from three experiments on the prediction network with varying inputs. Orange line: three MR sequences plus semi-manual, expert T2w lesion labels, green line: three MR sequences plus inferred T2w lesion labels, blue: four MR sequences, red: random performance.

Table 1. Quantitative comparison of results of three experiments on the same test data, reported at FPR=0.2.

Method	Accuracy	Precision	Specificity	Sensitivity
End-to-end CNN-MRI	26.31%	50.45%	79.16%	8.45%
End-to-end-MRI+segmT2	58.95%	88.09%	79.16%	54.12%
End-to-end-MRI+T2	80.21%	91.82%	79.16%	80.11%

labels, the gain in performance as compared with the baseline MRI alone is significant. The degradation in prediction performance when using inferred lesion labels over semi-manual labels is mainly due to the segmentation's drop in performance for small lesions, which in this case, make up approximately 40% of all lesions. This suggests that improving the accuracy of the segmentation network (not the focus of this work) would lead to significant accuracy improvements in the automatic prediction of future MRI disease activity.

4 Conclusions

We present the first deep learning framework for predicting future MRI disease activity in RRMS patients from baseline MR sequences. We show that prediction accuracy based on MRI alone does not perform much better than chance but improves substantially when baseline lesion labels are provided as additional inputs to the network. These results suggest the possibility of early prediction of future disease activity for RRMS patients, paving the way for the possibility of precision medicine in RRMS. Including semi-manual expert T2w lesion labels offers a remarkable boost in prediction accuracy showing accuracies of 80.21%, precision of 91.82%, specificity of 79.16% and sensitivity of 80.11%. When these are not available, we offer a deep learning framework for the automatic estimation of lesion labels. Our tests demonstrate that future MRI disease prediction using machine-generated lesion labels is not as accurate as prediction using semi-manual expert labels, but offers significant improvement over MRI sequences alone. Future work on improving lesion segmentation, particularly for small lesions, should significantly increase the accuracy of future disease prediction.

Acknowledgements This work was supported by an award from the International Progressive MS Alliance (PA-1603-08175).

References

1. Barkhof, F., et al.: Comparison of MRI criteria at first presentation to predict conversion to clinically definite multiple sclerosis. Brain **120**(11), 2059–2069 (1997)
2. Bengio, Y., Courville, A., Vincent, P.: Representation learning: A review and new perspectives. TPAMI **35**(8), 1798–1828 (2013)
3. Brosch, T., et al.: Modeling the variability in brain morphology and lesion distribution in multiple sclerosis by deep learning. MICCAI pp. 462–469 (2014)

4. Carass, A., et al.: Longitudinal multiple sclerosis lesion segmentation: Resource and challenge. Neuroimage **148**, 77–102 (2017)
5. Doyle, A., et al.: Predicting future disease activity and treatment responders for multiple sclerosis patients using a bag-of-lesions brain representation. MICCAI pp. 186–194 (2017)
6. Elliott, C., et al.: Temporally consistent probabilistic detection of new multiple sclerosis lesions in brain MRI. IEEE TMI **32**(8), 1490–1503 (2013)
7. Gold, R., et al.: Placebo-controlled phase 3 study of oral BG-12 for relapsing multiple sclerosis. New England Journal of Medicine **367**(12), 1098–1107 (2012)
8. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press (2016)
9. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. ICML pp. 448–456 (2015)
10. Kaunzner, U., Gauthier, S.: MRI in the assessment and monitoring of multiple sclerosis: an update on best practice. Therapeutic advances in neurological disorders **10**(6), 247–261 (2017)
11. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. ICLR (2014)
12. Krizhevsky, A., Sutskever, I., Hinton, G.: Imagenet classification with deep convolutional neural networks. NIPS pp. 1097–1105 (2012)
13. Menze, B., et al.: The multimodal brain tumor image segmentation benchmark (brats). IEEE TMI **34**(10), 1993 (2015)
14. Moccia, M., de Stefano, N., Barkhof, F.: Imaging outcome measures for progressive multiple sclerosis trials. Multiple Sclerosis Journal **23**(12), 1614–1626 (2017)
15. Nyúl, L., Udupa, J.: On standardizing the mr image intensity scale. Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine **42**(6), 1072–1081 (1999)
16. Río, J., et al.: MR imaging in monitoring and predicting treatment response in multiple sclerosis. Neuroimaging Clinics **27**(2), 277–287 (2017)
17. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. MICCAI pp. 234–241 (2015)
18. Sled, J., Zijdenbos, A., Evans, A.: A nonparametric method for automatic correction of intensity nonuniformity in MRI data. IEEE TMI **17**(1), 87–97 (1998)
19. Smith, S.: Fast robust automated brain extraction. Human brain mapping **17**(3), 143–155 (2002)
20. Sormani, M.P., Bruzzi, P.: Mri lesions as a surrogate for relapses in multiple sclerosis: a meta-analysis of randomised trials. The Lancet Neurology **12**(7), 669–676 (2013)
21. Stangel, M., et al.: Towards the implementation of no evidence of disease activityin multiple sclerosis treatment: the multiple sclerosis decision model. Therapeutic advances in neurological disorders **8**(1), 3–13 (2015)
22. Windham, B., et al.: Small brain lesions and incident stroke and mortality: a cohort study. Annals of Internal Medicine **163**(1), 22–31 (2015)
23. Yoo, Y., et al.: Deep learning of brain lesion patterns for predicting future disease activity in patients with early symptoms of multiple sclerosis. In: Deep Learning and Data Labeling for Medical Applications, pp. 86–94. Springer (2016)