

Deep Learning of Brain Lesion Patterns for Predicting Future Disease Activity in Patients with Early Symptoms of Multiple Sclerosis

Youngjin Yoo^{1,2,5(✉)}, Lisa W. Tang^{2,3,5}, Tom Brosch^{1,2,5}, David K.B. Li^{3,5},
Luanne Metz⁶, Anthony Traboulsee^{4,5}, and Roger Tam^{2,3,5}

¹ Department of Electrical and Computer Engineering,
University of British Columbia, Vancouver, BC, Canada
youngjin.yoo@alumni.ubc.ca

² Biomedical Engineering Program, University of British Columbia,
Vancouver, BC, Canada

³ Department of Radiology, University of British Columbia, Vancouver, BC, Canada

⁴ Division of Neurology, University of British Columbia, Vancouver, BC, Canada

⁵ MS/MRI Research Group, University of British Columbia, Vancouver, BC, Canada

⁶ Division of Neurology, University of Calgary, Calgary, AB, Canada

Abstract. Multiple sclerosis (MS) is a neurological disease with an early course that is characterized by attacks of clinical worsening, separated by variable periods of remission. The ability to predict the risk of attacks in a given time frame can be used to identify patients who are likely to benefit from more proactive treatment. In this paper, we aim to determine whether deep learning can extract, from segmented lesion masks, latent features that can predict short-term disease activity in patients with early MS symptoms more accurately than lesion volume, which is a very commonly used MS imaging biomarker. More specifically, we use convolutional neural networks to extract latent MS lesion patterns that are associated with early disease activity using lesion masks computed from baseline MR images. The main challenges are that lesion masks are generally sparse and the number of training samples is small relative to the dimensionality of the images. To cope with sparse voxel data, we propose utilizing the Euclidean distance transform (EDT) for increasing information density by populating each voxel with a distance value. To reduce the risk of overfitting resulting from high image dimensionality, we use a synergistic combination of downsampling, unsupervised pretraining, and regularization during training. A detailed analysis of the impact of EDT and unsupervised pretraining is presented. Using the MRIs from 140 subjects in a 7-fold cross-validation procedure, we demonstrate that our prediction model can achieve an accuracy rate of 72.9 % (SD = 10.3 %) over 2 years using baseline MR images only, which is significantly higher than the 65.0 % (SD = 14.6 %) that is attained with the traditional MRI biomarker of lesion load.

Keywords: Deep learning · Convolutional neural network · Clinical prediction · Multiple sclerosis lesion · Machine learning · MRI

1 Introduction

Multiple sclerosis (MS) is an immune mediated disorder characterized by inflammation, demyelination, and degeneration in the central nervous system. There is increasing evidence that early detection and intervention can improve long-term prognosis. However, the disease course of MS is highly variable, especially in its early stages, and it is difficult to predict which patients would progress more quickly and therefore benefit from more aggressive treatment. The McDonald criteria [1, 2], which are a combination of clinical and magnetic resonance imaging (MRI) indicators of disease activity, facilitate the diagnosis of MS in patients who present early symptoms suggestive of MS.

However, predicting which patients will meet a given set of criteria for disease activity within a certain time frame remains a challenge. MRI is invaluable for monitoring and understanding the pathology of MS *in vivo* from the earliest stages of the disease, but the commonly computed MRI biomarkers such as brain and lesion volume are not strongly predictive of future disease activity [3], especially when only baseline measures are available, which is often the case when a patient first presents. Researchers have attempted to define more sophisticated MRI features that are more predictive. Recently, Wottschel *et al.* employed a support vector machine trained on user-defined features to predict the conversion of clinically isolated syndrome (CIS), a prodromal stage of MS, to clinically definite MS [4]. The features included demographic information and clinical measurements at baseline, and also MRI-derived features such as lesion load (also known as burden of disease, BOD) and lesion distance measurements from the center of the brain.

User-defined features typically require expert domain knowledge and a significant amount of trial-and-error, and are subject to user bias. An alternate approach is to automatically learn patterns and extract latent features using machine learning. In recent years, deep learning [5] has received much attention due to its use of automated feature extraction to achieve breakthrough success in many applications, in some cases from high-dimensional data with complex content such as neuroimaging data. For example, deep learning of neuroimaging data has been used to perform various tasks such as the classification between mild cognitive impairment and Alzheimer’s disease (*e.g.*, [6]) and to model pathological variability in MS [7].

In this work, using the baseline MRIs of patients with early symptoms of MS but not yet meeting the McDonald 2005 criteria for MS diagnosis, we aim to predict which patients worsened to meet the conversion criteria within two years. MS exhibits a complex pathology that is still not well understood, but it is known that change in spatial lesion distribution may be an indicator of disease activity [8]. Our clinical motivation is to discover white matter lesion patterns that may indicate a faster rate of worsening, so that patients who exhibit such patterns can be selected for more personalized treatment. We investigate whether latent MRI lesion patterns extracted by deep learning can predict disease status conversion to meet the McDonald 2005 criteria with greater accuracy than BOD. The main idea is to employ convolutional neural networks (CNNs) to identify

latent lesion pattern features whose variability can maximally distinguish those patients at risk of short-term disease activity from those who will remain relatively stable.

2 Materials and Preprocessing

The baseline T2-weighted (T2w) and proton density-weighted (PDw) MR images of 140 subjects were used to predict each patient’s disease status at two years. The dataset consists of 60 non-converters and 80 converters. The image dimensions are $256 \times 256 \times 60$ with a voxel size of $0.937 \times 0.937 \times 3.000$ mm. Preprocessing consisted of skull stripping and linear intensity normalization. The T2w and PDw scans were segmented via a semi-automated multimodal method to produce lesion masks. The mask images were then downsampled to $128 \times 128 \times 30$ with Gaussian pre-filtering as a first dimensionality reduction step.

3 Methods

Prior to feature extraction, all images were spatially normalized to a standard template (MNI152) [9] using affine registration. Our CNN architecture is a 9-layer model (Fig. 1), consisting of three 3D convolutional layers interleaved with three max-pooling layers, followed by two fully connected layers, and finally a logistic regression output layer.

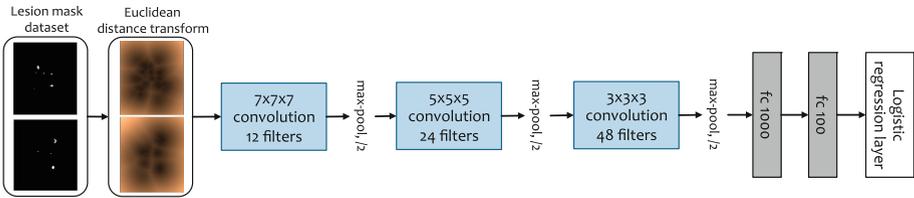


Fig. 1. The proposed convolutional neural network architecture (fc = fully connected layer) for predicting future disease activity in patients with early symptoms of MS. The Euclidean distance transform is used for increasing information density from sparse lesion masks.

3.1 Euclidean Distance Transform of Lesion Masks

MS lesions typically occupy a very small percentage of a brain image, and as a result the binary lesion masks contain mostly zeros. From our preliminary experiments, we observed that the CNN model learns mostly noisy patterns from the binary lesion masks, which is likely due to the fact that sparse lesion voxels can be ignored or deformed into noise spikes by various stages of convolution and pooling operations during training. As described in Sect. 4, the training and

test results show that the binary lesion masks are not appropriate as the input to the CNN model. We could have also used raw MR images as the input, but the lesion voxels would almost certainly be lost in the learning process due to their sparsity. To overcome this problem, we propose increasing the density of information in the lesion masks by the Euclidean distance transform (EDT) [10], which measures the Euclidean distance between each voxel and the closest lesion. The EDTs of the binary lesion masks form the input to our CNN model. From Fig. 1, we can see examples of how the spatial distribution of the lesions is densely captured and better amplified than those in the original binary masks. The impact of the transform on training a deep learning network will be presented in Sect. 4. We used the ITK-SNAP’s Convert3D tool [11] for applying the EDT.

3.2 CNN Training

It has been shown that pretraining can improve the optimization performance of supervised deep networks when training sets are limited, which often happens in the medical imaging domain [12], but the gains are dependent on data properties. We investigated the impact of using a 3D convolutional deep belief network (DBN) for pretraining to initialize our CNN model. Our convolutional DBN has the same network architecture as the convolutional and pooling layers of our CNN. For our DBN and CNN, we used the leaky rectified non-linearity [13] (negative slope $\alpha = 0.3$), which is designed to prevent the problem associated with non-leaky units failing to reactivate after encountering certain conditions due to large gradient flow. Our convolutional DBN was initialized using a robust method [14] that particularly considers the rectified non-linearity and has been shown to allow successful training of very deep networks on natural images, and trained using contrastive divergence [15]. To analyze the influence of EDT and pretraining on supervised training, we trained our CNN under four conditions: no EDT and no pretraining, no EDT with pretraining, with EDT and no pretraining, with both EDT and pretraining. For all four experiments, we used negative log-likelihood maximization with AdaDelta [16] (conditioning constant $\epsilon = 1e-12$ and decay rate $\rho = 0.95$) and a batch size of 20 for training. Since there are more converters than non-converters in the dataset, the class weights in the cost function (cross entropy) for supervised training were automatically adjusted in each fold to be inversely proportional to the class frequencies observed in the training set. We used Theano [17] and cuDNN [18] for implementing the CNN models.

3.3 Data Augmentation and Regularization

Due to the high dimensionality of the input images relative to the number of samples in the dataset, even after downsampling, the proposed network can suffer from overfitting. Data augmentation is one of the most popular approaches to reduce the risk of overfitting by artificially creating training samples to increase the dataset size. To generate more training samples, we performed data augmentation by applying random rotations (± 3 degrees), translations (± 2 mm), and scaling (± 2 percent) to the mask images, which increased the number of

training images by fourfold. To regularize training, we applied dropout [19] with $p = 0.5$, weight decay (L_2 -norm regularization) with penalty coefficient $2e-3$ and L_1 -norm regularization with penalty coefficient $1e-6$. Finally, we applied early stopping, which also acts as a regularizer to improve the generalization ability [20], with a convergence target of negative log-likelihood of 0.6. The convergence target was used to stop training when the generalization loss (defined as the relative increase of the validation error over the minimum-so-far during training) started to increase, which was determined by cross-validation.

4 Results and Discussion

To see the impact of EDT on unsupervised pretraining, we computed the root mean squared (RMS) reconstruction error with and without EDT for each epoch during training of the convolutional DBN. The reconstruction error remaining after each epoch during pretraining of the first convolutional layer is shown in Fig. 2. We observed that pretraining with EDT converged faster and produced lower reconstruction error at convergence than pretraining without EDT.

To analyze the impact of EDT and pretraining on supervised training, we compared four different scenarios which were described in Sect. 3.2 and shown in Fig. 3. Without EDT, the CNN converged much faster with pretraining, but the prediction errors at convergence were similar between those obtained with and without pretraining. In both cases, the training made little progress on the prediction error on the training set, and no progress on the test error, which remained high. Using EDT, the optimization did not converge without pretraining even after 500 epochs, but did converge with pretraining. Without pretraining, the prediction errors fluctuated early for both the training and test datasets, but soon remained constant, and training made no further progress. In contrast, with both EDT and pretraining, the prediction errors on both training and test data decreased fairly steadily up to about 200 epochs.

Figure 4 shows visualizations of the manifolds produced by the CNN outputs, reduced to two dimensions using t-distributed stochastic neighbor embedding (t-SNE) [21]. When EDT and pretraining were not used, the two groups

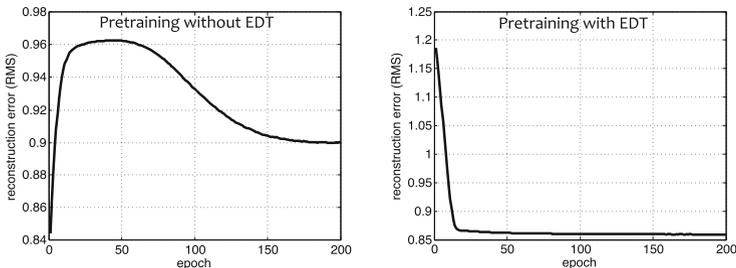


Fig. 2. The influence of EDT on unsupervised pretraining for a convolutional layer. Pretraining with EDT converged faster and produced lower reconstruction error after convergence.

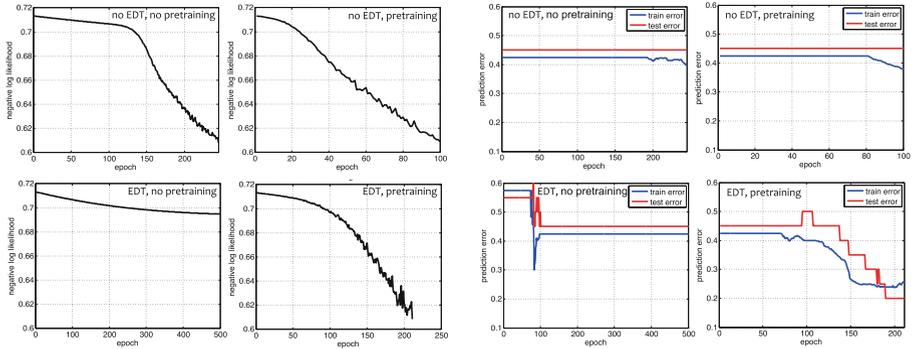


Fig. 3. The influence of EDT and pretraining on supervised training. The left 4 images show training costs and the right 4 images show prediction errors on both training and test datasets for each epoch during supervised training in a selected fold of cross-validation.

(converters and non-converters) showed poor linear separability in the learned manifold space. The two groups were more distinguishable in the manifold space learned from the CNN with EDT and pretraining.

Table 1. Performance comparison (%) between 5 different prediction models for predicting short-term (2 years) clinical status conversion in patients with early MS symptoms. The same training parameters were used for all the CNNs. We performed a 7-fold cross-validation on 80 converters and 60 non-converters and computed the average performance for each prediction model.

Prediction model	Accuracy	Sensitivity	Specificity	AUC
Logistic regression with BOD	65.0 ± 14.6	54.3	80.9	67.6 ± 14.9
CNN (no EDT, no pretraining)	57.9 ± 4.9	94.9	8.3	51.6 ± 4.4
CNN (no EDT, pretraining)	57.9 ± 5.9	98.7	3.6	51.1 ± 4.9
CNN (EDT, no pretraining)	54.3 ± 6.2	71.4	28.6	50.0 ± 0.0
CNN (EDT, pretraining)	72.9 ± 10.3	78.6	65.1	71.8 ± 10.2

For evaluating prediction performance, we used a 7-fold cross-validation procedure in which each fold contained 120 subjects for training and 20 subjects. Note that the number of training images for each fold was increased to 480 by data augmentation. For comparison to the established approach used in clinical studies, a logistic regression prediction model applied to the classic MRI biomarker of BOD was used. The results of the comparison are shown in Table 1. When EDT was not used, the CNN (with and without pretraining) produced lower prediction accuracy rates than those attained by the logistic regression model with BOD. In addition, these cases produced very high sensitivity but

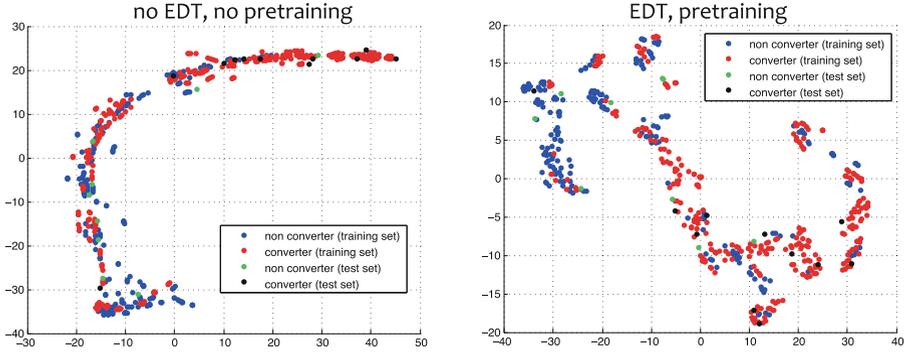


Fig. 4. Visualizations to show the influence of EDT and pretraining on the learned manifold space, reduced to two dimensions using t-SNE [21]. Each subject in the dataset is represented by a two-dimensional feature vector. The axes represent the feature element values of each two-dimensional feature vector in the learned low-dimensional map. The converter and non-converter groups are more linearly separable in the manifold space when using the EDT and pretraining.

low specificity, possibly due to overfitting on the sparse lesion image data. When EDT was used without pretraining, the CNN did not converge for every fold in the cross-validation and also produced lower prediction accuracy rates than lesion volume. The gap between sensitivity and specificity was reduced but still remained large. The CNN with EDT and pretraining improved the prediction performance by approximately 8% in accuracy and 4% in AUC when compared to the logistic regression model with BOD. In addition, the SDs for both accuracy and AUC decreased by approximately 4–5%, showing a more consistent performance across folds. This model also achieved the best balance between sensitivity and specificity.

5 Conclusion

We have presented a CNN architecture that learns latent lesion features useful for identifying patients with early MS symptoms who are at risk of future disease activity within two years, using baseline MRIs only. We presented methods to overcome the sparsity of lesion image data and the high dimensionality of the images relative to the number of training samples. In particular, we showed that the Euclidean distance transform and unsupervised pretraining are both key steps to successful optimization, when supported by a synergistic combination of data augmentation and regularization strategies. The final results were markedly better than those obtained by the clinical standard of lesion volume.

Acknowledgements. This work was supported by the MS/MRI Research Group at the University of British Columbia, the Natural Sciences and Engineering Research Council of Canada, the MS Society of Canada, and the Milan and Maureen Ilich Foundation.

References

1. Polman, C., Reingold, S., Edan, G., et al.: Diagnostic criteria for multiple sclerosis: revisions to the McDonald criteria. *Ann. Neurol.* **58**(2005), 840–846 (2005)
2. Polman, C.H., Reingold, S.C., Banwell, B., et al.: Diagnostic criteria for multiple sclerosis: revisions to the McDonald criteria. *Ann. Neurol.* **69**(2011), 292–302 (2010)
3. Odenthal, C., Coulthard, A.: The prognostic utility of MRI in clinically isolated syndrome: a literature review. *Am. J. Neuroradiol.* **36**, 425–431 (2015)
4. Wottschel, V., Alexander, D., Kwok, P., et al.: Predicting outcome in clinically isolated syndrome using machine learning. *NeuroImage: Clin.* **7**, 281–287 (2015)
5. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. *Nature* **521**, 436–444 (2015)
6. Suk, H., Lee, S., Shen, D., et al.: Hierarchical feature representation and multi-modal fusion with deep learning for AD/MCI diagnosis. *NeuroImage* **101**, 569–582 (2014)
7. Brosch, T., Yoo, Y., Li, D.K.B., Traboulsee, A., Tam, R.: Modeling the variability in brain morphology and lesion distribution in multiple sclerosis by deep learning. In: Golland, P., Hata, N., Barillot, C., Hornegger, J., Howe, R. (eds.) *MICCAI 2014, Part II. LNCS*, vol. 8674, pp. 462–469. Springer, Heidelberg (2014)
8. Giorgio, A., Battaglini, M., Rocca, M.A., et al.: Location of brain lesions predicts conversion of clinically isolated syndromes to multiple sclerosis. *Neurology* **80**, 234–241 (2013)
9. Mazziotta, J., Toga, A., Evans, A., et al.: A probabilistic atlas and reference system for the human brain: international consortium for brain mapping (ICBM). *Philos. Trans. Roy. Soc. B: Biol. Sci.* **356**, 1293–1322 (2001)
10. Maurer, C., Qi, R., Raghavan, V.: A linear time algorithm for computing exact Euclidean distance transforms of binary images in arbitrary dimensions. *IEEE Trans. Pattern Anal. Mach. Intell.* **25**, 265–270 (2003)
11. Yushkevich, P.A., Piven, J., Cody Hazlett, H., et al.: User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. *NeuroImage* **31**, 1116–1128 (2006)
12. Tajbakhsh, N., Shin, J.Y., Gurudu, S.R., et al.: Convolutional neural networks for medical image analysis: full training or fine tuning? *IEEE Trans. Med. Imaging* **35**, 1299–1312 (2016)
13. Maas, A., Hannun, A., Ng, A.: Rectifier nonlinearities improve neural network acoustic models. In: *International Conference on Machine Learning*, vol. 30 (2013)
14. He, K., Zhang, X., Ren, S., et al.: Delving deep into rectifiers: surpassing human-level performance on imagenet classification. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1026–1034 (2015)
15. Lee, H., Grosse, R., Ranganath, R., et al.: Unsupervised learning of hierarchical representations with convolutional deep belief networks. *Commun. ACM* **54**, 95–103 (2011)
16. Zeiler, M.: ADADELTA: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701* (2012)
17. Theano Development Team: Theano: A Python framework for fast computation of mathematical expressions. *arXiv e-prints abs/1605.02688*, May 2016
18. Chetlur, S., Woolley, C., Vandermersch, P., et al.: cuDNN: efficient primitives for deep learning. *arXiv preprint arXiv:1410.0759* (2014)

19. Srivastava, N., Hinton, G., Krizhevsky, A., et al.: Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **15**, 1929–1958 (2014)
20. Orr, G.B., Müller, K.R.: *Neural networks: tricks of the trade*. Springer, Heidelberg (2003)
21. Van Der Maaten, L.: Accelerating t-SNE using tree-based algorithms. *J. Mach. Learn. Res.* **15**, 3221–3245 (2014)