Prediction of Disease Progression in Multiple Sclerosis Patients using Deep Learning Analysis of MRI Data

Adrian Tousignant[1,3] ADRIAN.TOUSIGNANTDURAN@MAIL.MCGILL.CA
Paul Lemaître[3] PLEMAITR@CIM.MCGILL.CA
Doina Precup[1] DPRECUP@CS.MCGILL.CA
Douglas L. Arnold[2] DOUGLAS.ARNOLD@MCGILL.CA
Tal Arbel[3] ARBEL@CIM.MCGILL.CA
[1] *School of Computer Science, McGill University, Montreal, Canada*
[2] *Montreal Neurological Institute, McGill University, Montreal, Canada*
[3] *Center for Intelligent Machines, McGill University, Montreal, Canada*

Abstract

We present the first automatic end-to-end deep learning framework for the prediction of future patient disability progression (one year from baseline) based on multi-modal brain Magnetic Resonance Images (MRI) of patients with Multiple Sclerosis (MS). The model uses parallel convolutional pathways, an idea introduced by the popular Inception net (Szegedy et al., 2015) and is trained and tested on two large proprietary, multi-scanner, multi-center, clinical trial datasets of patients with Relapsing-Remitting Multiple Sclerosis (RRMS). Experiments on 465 patients on the placebo arms of the trials indicate that the model can accurately predict future disease progression, measured by a sustained increase in the extended disability status scale (EDSS) score over time. Using only the multi-modal MRI provided at baseline, the model achieves an AUC of 0.66 ± 0.055. However, when supplemental lesion label masks are provided as inputs as well, the AUC increases to 0.701 ± 0.027. Furthermore, we demonstrate that uncertainty estimates based on Monte Carlo dropout sample variance correlate with errors made by the model. Clinicians provided with the predictions computed by the model can therefore use the associated uncertainty estimates to assess which scans require further examination.

Keywords: Deep learning, Multiple Sclerosis, MRI, multi-modal, disease progression

1. Introduction

Neurological diseases, such as Alzheimer disease (AD) and Multiple Sclerosis (MS), present major burdens on patients, their families, and on society as a whole (Naci et al., 2010; Brookmeyer et al., 2007). As an example of the financial burden alone, the mean lifetime cost per MS patient has been estimated to be $2.5 million (Patwardhan et al., 2005). Correctly identifying patients whose condition is at risk of worsening by predicting their future disease course would be enormously beneficial to patients as well as the health care system. More aggressive treatments could be offered to patients for whom the prognosis is poor, setting the stage for personalized medicine. Furthermore, clinical trials would be faster and cheaper if one could accurately predict future disease progression in the early phases of the trial. In this paper, we focus on the particular context of MS, an inflammatory, demyelinating, and degenerative disease of the central nervous system resulting in a range of physical and cognitive disabilities (MS Society of Canada, last accessed on 06/30/18). To date, treatments

have been successfully developed to suppress the acute inflammatory demyelinating lesions that are the hallmark of this disease, and the associated relapses, but these treatments have little of no effect on the relentless disability progression that develops in some patients.

MRIs are an integral part of the clinical management of MS due to their exquisite sensitivity to the development of new MS lesions. MRIs allow clinicians to monitor the evolution of lesions and the effectiveness of drugs being used to suppress them. New lesion formation has also been extremely valuable as a biomarker of treatment efficacy in early stage clinical trials (Barkhof et al., 1997; Frisoni et al., 2010). There have been several successful imaging studies which used MRI biomarkers to determine disability level and progression for a variety of diseases (Jack et al., 2004; Losseff et al., 1996; Ulla et al., 2013). In the context of MS, MRIs have been shown to help predict future lesion activity, defined as the presence of new or enlarged *T2-lesions* in future images, using a *Bag-of-Lesions* brain representation, and to identify potential treatment responders based on carefully designed image features (Doyle et al., 2017). There has also been some success in predicting the conversion to Clinically Definite MS (CDMS) using a SVM on radiomic lesion features (Wottschel et al., 2015) and using a CNN on lesion masks (Yoo et al., 2016). However, although handcrafted features have been successfully adapted to these contexts, there are currently no known imaging biomarkers predictive of future MS disability progression, and it has been shown that clinically-derived lesion features alone are not sufficiently predictive (Zhao et al., 2017). Given the need for data-driven models, deep learning provides an attractive approach to building predictors of disease progression based on MRI, particularly given the recent success of this methodology in a wide variety of tasks in computer vision (Krizhevsky et al., 2012; Deng et al., 2014), and in medical image analysis (Menze et al., 2015; Carass et al., 2017). The capacity of deep learning models to extract predictive features from data is particularly desirable, given that there are currently no clear biomarkers for MS progression.

In this work, we present the *first* automatic, end-to-end deep learning framework for the prediction of future patient disability progression based on multi-modal brain MRI of patients with MS. We use a deep 3D CNN with parallel convolutional pathways, an architecture inspired by the popular Inception net (Szegedy et al., 2015). The model is trained on two large proprietary, multi-scanner, multi-center, clinical trial datasets of patients with Relapsing-Remitting Multiple Sclerosis (RRMS). The model successfully predicts if a significant increase in future clinical disability will occur within a year, with an AUC of 0.66 ± 0.055. We show that supplementing the MRI with *T2* and *Gadolinium-enhanced* lesion labels (provided by experts) at baseline, if such labels are available, further increases prediction accuracy compared to using MRI alone, resulting in an AUC of 0.701 ± 0.027. Both proposed models outperform a VGG-like 3D CNN (Simonyan and Zisserman, 2014).

A key ingredient to the successful integration of machine learning methods into clinical workflows is the clinician's trust in the results provided by the learning system. Garnering this trust involves two important steps: (a) providing a quantitative measurement of confidence in the prediction results, and (b) establishing that when the system is confident in its predictions, these are indeed correct. The confidence measurements can then be communicated to the clinicians along with the predictions, so they can determine when further human review or particular attention is required. Hence, our proposed model also provides uncertainty estimates in addition to the predicted progression. This is accomplished by training the deep learning network with dropout, and taking repeated Monte Carlo (MC) samples of the prediction using dropout at test time (Leibig et al., 2016; Nair et al., 2018; Gal and Ghahramani, 2016; Tanno et al., 2017). The sample variance obtained through

this procedure provides a measure of uncertainty in the output. Our results show that, by applying a threshold on the uncertainty level of the output and only considering the patients for which the model is most confident, the prediction performance increases. This suggests a strong correlation between high model uncertainty and incorrect predictions.

2. Methodology

We propose a deep learning framework for the prediction of disability progression of MS patients based on MRI, within one year from the baseline scan. The model uses multi-modal MRI volumes as input. The volumes for a patient at baseline are concatenated together and the different modalities are used as input channels. The model consists of three consecutive 3D convolutional blocks, followed by 2 fully connected layers. A dropout layer is used after each convolutional block and fully connected layer. The purpose of these dropout layers is twofold. First, they serve to limit overfitting, which is crucial when working with a small dataset. Secondly, using dropout at test time allows us to take Monte Carlo (MC) samples of predictions, which can be used to quantify the uncertainty of the network in its output. Rectifier linear units are used at all layers except the output, which is a sigmoid.

The standard cross-entropy loss between the progression label in the training data and the model's prediction is then used to drive the learning process. Note that progression labels are binary and assigned after one year of follow-up visits, in accordance with clinical practice (as described in detail in Sec.3.1). Figure 1 shows a high-level overview of the model.

2.1. Convolutional Block

When working with MRI of MS patients, an important problem is the significant discrepancy in the location, shape and size of important lesions, which is in fact part of the difficulty of finding biomarkers based on brain MRI. An important goal in our work was to endow our model with the capability to incorporate information from features of different dimensions in an independent manner. To this end, the deep network is modularized into blocks with parallel convolutional pathways of different focal resolution sizes. Each convolutional block consist of 4 parallel pathways. The first pathway is a max-pooling layer with stride 2, whose purpose is to help propagate of information in a manner similar to residual connections. The 3 other pathways have resolution windows of 3, 5, and 7 respectively. Instead of naively using the expensive $5 \times 5 \times 5$ and $7 \times 7 \times 7$ convolution layer, we stack consecutive $3 \times 3 \times 3$ convolution layers, saving both in memory and in computation. Each convolutional layer has 64 kernels and a stride of 1, except for the last layer of each pathway, which has a stride of 2. The pathways are then concatenated and fed as input to the next block.

3. Experiments and Results

We evaluate the ability of the model learned by our approach to predict clinical progression in MS patients through a series of experiments. We first explore the ability of the network to predict future disease progression based on image features extracted from MRI at baseline. To this end, we train our model using 5 different MRI channels per patient, *T1c*, *T1p*, *T2w*, *Pdw*, and *FLAIR*.

The second experiment combines the 5 input MRI modalities with lesion labels masks, including both *T2w* lesion maps and *Gadolinium-enhanced* lesion maps, in order to explore whether this additional information, if available at baseline, improves the model's accuracy. We compare the

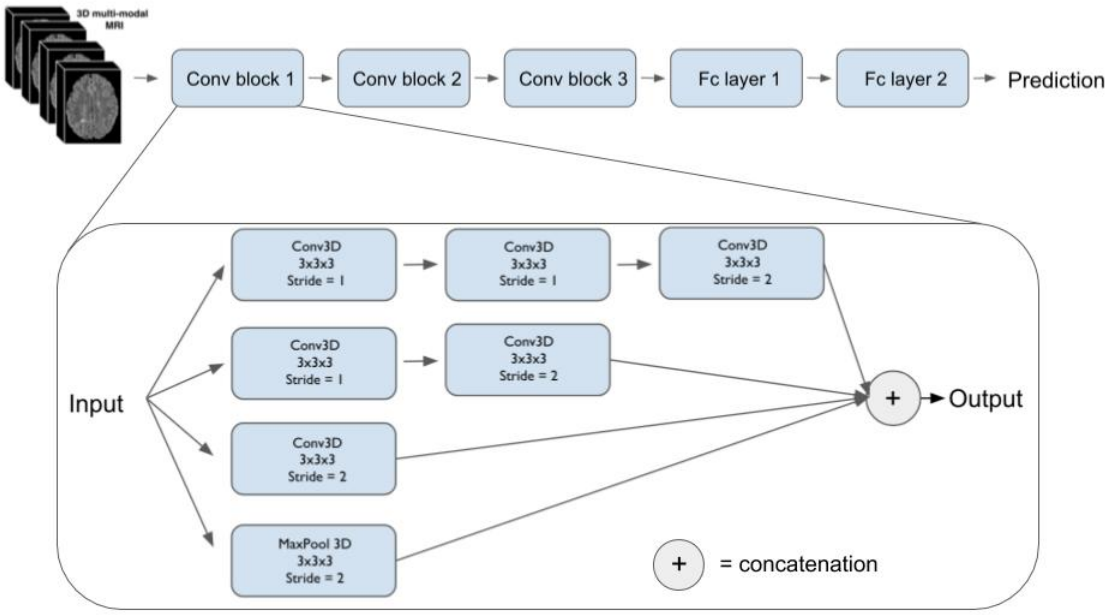


Figure 1: Proposed 3D CNN architecture. MRI volumes are provided as inputs. Information is propagated through three repeated 3D convolution blocks. The last layer is flattened and fed to two fully connected layers. We use ReLUs as the activation function and dropout with a drop probability of 0.5. A zoom-in of the convolutional block unit is provided. Each convolutional block is composed of 4 parallel pathways. The first three consist of different numbers of $3 \times 3 \times 3$ convolutional layers, while the last is a 3D MaxPooling layer to help with information propagation. There is a stride of 1 for every layer, except for the last layer of each pathway, where the stride is 2.

results of the proposed network to a 3D version of the popular VGG model (Simonyan and Zisserman, 2014) with a similar number of parameters. Our comparison model uses three series of conv-conv layers with kernel size of $3 \times 3 \times 3$ and stride of 1 followed by a max pooling layer of size $2 \times 2 \times 2$ and stride of 2. The number of kernels for each series of conv-conv layers is $64, 256$, and 512 respectively. These numbers were chosen to keep the total number of parameters as close as possible to those of our proposed model.

Finally, we explore the effectiveness of using Monte Carlo (MC) sample variance as an estimate of model uncertainty. The objective is to evaluate whether filtering out the most uncertain examples on the ROC plot leads to higher AUC on the ROC curve constructed from the predictions of the remaining examples. If this is indeed the case, then the system is generally correct when certain, so uncertainty measurements are useful to flag problematic cases.

3.1. Dataset

The data used for our experiments are drawn from two large proprietary, multi-scanner, multi-center, clinical trial MRI volumes acquired from patients with Relapsing-Remitting MS (RRMS). The image contrast was standardized across sites based on dummy run scans and, once approved, scanners and sequences remained consistent at each site. The MRI volumes were then pre-processed with brain extraction (Smith, 2002), N3 bias field inhomogeneity correction (Sled et al., 1998), Nyul image intensity normalization (Nyúl and Udupa, 1999), and registration to the MNI-space. Only scans of patients from the placebo arms of both trials were used in the experiments, in order to eliminate the drug effects on our analysis of the natural disease course. In addition, only scans of patients who completed the trial were used. This amounted to scans from a total of 465 patients from two different clinical trials. One trial consisted of 1330 RRMS patients, of which 312 patients were on placebo and completed the trial. Each patient had two MRI scans, taken one year apart, resulting in 624 scans with a non-progression/progression split of 582/42. The second dataset consisted of 543 RRMS patients, of which 153 patients were in the placebo arm and also finished the trial. Patients were scanned at intervals of 24 weeks, resulting in 459 scans with a non-progression/progression split of 398/61. T1-weighted pre- and post- contrast (T1p, T1c), T2-weighted (T2w), Proton Density-weighted (Pdw) and Fluid-attenuated inversion (FLAIR) were used as input modalities for the network. In addition, two lesion masks (T2-weighted and Gadolinium enhanced) were provided were provided with the dataset. The T2-weighted masks were obtained through a semi-manual process in which an automated segmentation algorithm was corrected by a trained expert reader. The Gad-enhancing masks were obtained through consensus between two human experts. In addition to the scheduled scans, patients from both trials had trimestrial clinical follow-up visits, at which the expanded disability status scale (EDSS) score was assessed by a clinician. Each patient was followed by the same clinician over the course of the trial, who was unaware of the treatment assignment in order to prevent bias in the EDSS rating. Binary progression labels were assigned to a scan, in accordance with the clinical definition: change in the expanded disability status scale (EDSS) score was measured over time, and a patient was deemed to progress within a year if the clinical progrssion critrion was met at last once during the year (see Table 1). A vast majority of patients did not progress during the trial, resulting in a significant class imbalance, with a non-progression to progression ratio close to 9:1. To mitigate the effect of this imbalance, we over-sampled the minority class during training.

Table 1: Requirement for progression given a baseline EDSS score

Definition of Progression	
Baseline EDSS	Criteria
0	An increase of 1.5 or more in EDSS score sustained for 12 weeks or more
0.5 to 5.5	An increase of 1 or more in EDSS score sustained for 12 weeks or more
6.0 and up	An increase of 0.5 or more in EDSS score sustained for 12 weeks or more

3.2. Training the network

We trained our network using the RMSProp optimizer with the cross-entropy loss objective (Hinton, 2016). The network was trained for 100 epochs, with a learning rate of $1e^{-5}$ and a dropout rate of 0.5. Training took approximately 10 hours on a V100 GPU with 16GB of memory. We were forced to use a batch size of 2 only, as we faced memory issues due to the size of the input MRI volumes and the number of model parameters.

3.3. Results

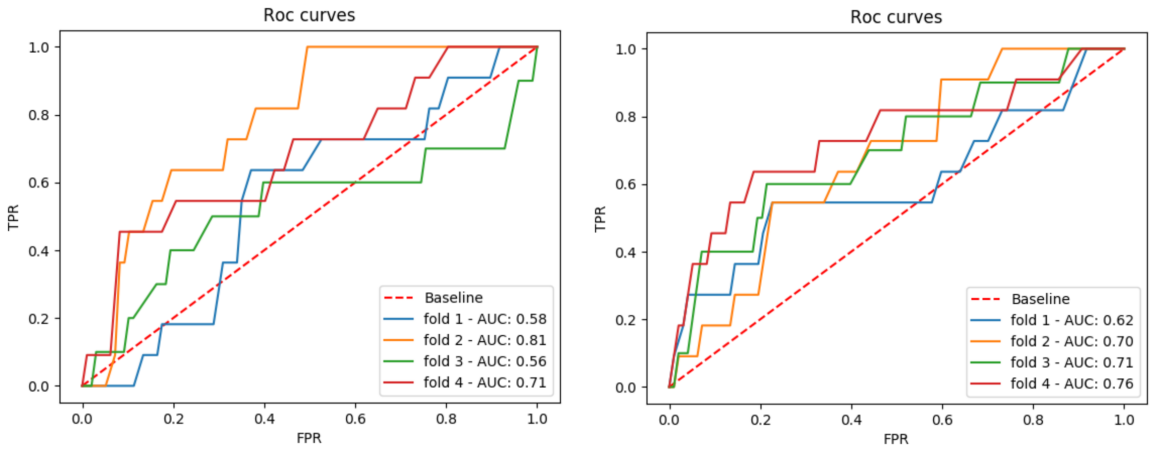
The results were evaluated based on the True Positive Rate ($TPR = \frac{TP}{TP+FN}$) and False Positive Rate ($FPR = \frac{FP}{FP+TN}$) receiver operating characteristic (ROC) curve. We obtained the TPR and FPR points by varying the threshold used to binarize the output of the model. These metrics were chosen to assess the performance instead of accuracy due to the large class imbalance. The dataset was split into training (75%), validation (15%) and test sets (10%). K-fold cross-validation was performed with $K = 4$; the choice of K was determined based on the size of the dataset. We ensured that data from all the follow-up visits for the same patient was always kept in the same fold, in order to not contaminate the validation and testing sets with training data. As recall and precision are both important for progression prediction, the F-score was used for early stopping.

When only the five MRI modalities were provided as input, the network attained an average area under the curve (AUC) of 0.66 ± 0.055. With the addition of T2 lesion and Gad-enhancing lesion masks as inputs, the progression prediction AUC improved substantially and the variability across folds was reduced. The AUC for this case was 0.701 ± 0.027. Using McNemar's test for comparison of classification learning algorithms, we can reject the null hypothesis that our model is similar to a model trained with scrambled labels with a p-value of $4.84e^{-6}$. Figure 2 shows the ROC curves for both experiments. Table 2 summarizes the quantitative results for both experiments against the baseline VGG-like 3D CNN which was included for comparison. Both experiments performed significantly better than the baseline, which had an AUC of 0.615 ± 0.053.

Table 2: Comparison of the model's performance

Networks	# of Parameters	AUC $\pm$ std
VGG-like 3D CNN	13.5M	0.615 ± 0.053
Proposed 3D CNN	14M	0.66 ± 0.055
Proposed 3D CNN with lesion masks	**14M**	$\mathbf{0.701 \pm 0.027}$

We quantified the uncertainty of our model's output by taking Monte Carlo dropout prediction samples at test time. We ran 20 forward passes of our model with a dropout rate of 0.5 and calculated the mean and sample variance of the generated values, representing the final prediction output and the associated model uncertainties, respectively. To assess the relationship between our network's confidence in its prediction and its performance, we generated the overall ROC curve (based on the entire set of results) and compared it against the ROC curves showing the results in which the top $n^{th}\%$ most uncertain predictions were removed. Figure 3 shows that excluding even a few of the most uncertain predictions can improve the model's overall AUC on the remaining predictions. Specifically, the reference curve has an AUC of 0.649. Removing the 2% most uncertain examples is enough to result in an increase in AUC to 0.659. The trend continues when removing the top



(*a*) 5 MRI modalities (*b*) 5 MRI modalities + 2 lesion labels

Figure 2: TPR vs FPR ROC curves: (a) Using the 5 MRI modalities only as inputs. AUC of 0.66 ± 0.055 (b) Supplementing the model with the 2 provided lesion label masks. AUC of 0.701 ± 0.027. Results for each fold of the cross-validation are shown.

10% most uncertain results, which leads to an AUC of 0.681 on the remaining predictions, an improvement of almost 5%.

4. Conclusion and Future Work

In this paper, we develop an end-to-end 3D CNN with parallel convolutional layers for predicting future progression in MS patients using MRI images and clinical assessment of disability. Our results also indicate that supplementing the model with *T2* and *Gad-enhancing* lesion labels, should they be available, further improves prediction accuracy. Finally, our model includes an uncertainty estimate based on MC dropout sample variance. Filtering out very uncertain examples leads to improved results for the remaining predictions. While this work focused on MC sample variance as an uncertainty metric, exploring alternative ways of quantifying uncertainty would be of interest. Additionally, adapting our architecture to leverage longitudinal clinical information (e.g. age, disability stage) could help improve predictions. Finally, exploring ways to interpret the predictions of the model and identify which regions of the brain contributed to the final decisions could help uncover new MS biomarkers, guiding the way of future research and furthering our understanding of the disease.

Acknowledgments

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

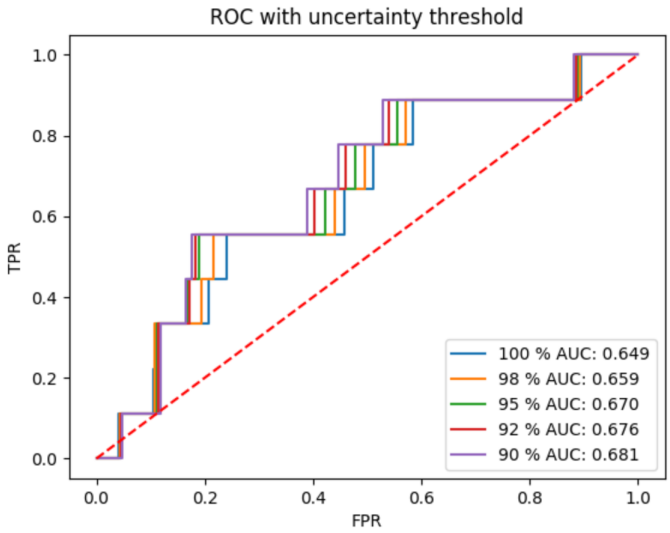


Figure 3: TPR vs. FPR of retained test time predictions when thresholding based on the MC sample variance. The percentage of predictions retained at each curve's uncertainty threshold is indicated in the color coded legend as well as the corresponding AUC. The reference curve (100%) performance, when no uncertainty thresholding is performed, is also shown (blue).

References

Frederik Barkhof, Massimo Filippi, David H. Miller, Philip Scheltens, Adriana Campi, Chris H. Polman, Giancarlo Comi, Herman J. Adèr, Nick Losseff, and Jacob Valk. Comparison of mri criteria at first presentation to predict conversion to clinically definite multiple sclerosis. *Brain*, 120(11):2059–2069, 11 1997.

Ron Brookmeyer, Elizabeth Johnson, Kathryn Ziegler-Graham, and H Michael Arrighi. Forecasting the global burden of alzheimer's disease. *Alzheimer's & dementia*, 3(3):186–191, 2007.

Aaron Carass, Snehashis Roy, Amod Jog, Jennifer L Cuzzocreo, Elizabeth Magrath, Adrian Gherman, Julia Button, James Nguyen, Ferran Prados, Carole H Sudre, et al. Longitudinal multiple sclerosis lesion segmentation: resource and challenge. *NeuroImage*, 148:77–102, 2017.

Li Deng, Dong Yu, et al. Deep learning: methods and applications. *Foundations and Trends® in Signal Processing*, 7(3–4):197–387, 2014.

Andrew Doyle, Doina Precup, Douglas L Arnold, and Tal Arbel. Predicting future disease activity and treatment responders for multiple sclerosis patients using a bag-of-lesions brain representation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 186–194. Springer, 2017.

Giovanni B. Frisoni, Nick C. Fox, Clifford R. Jack, Philip Scheltens, and Paul M. Thompson. The clinical use of structural mri in alzheimer disease. *Nature Reviews Neurology*, 6(2):67–77, 2010.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, pages 1050–1059. JMLR.org, 2016.

Geoffrey Hinton. Neural networks for machine learning. Coursera, 2016.

C. R. Jack, M. M. Shiung, J. L. Gunter, P. C. O'Brien, S. D. Weigand, D. S. Knopman, B. F. Boeve, R. J. Ivnik, G. E. Smith, R. H. Cha, E. G. Tangalos, and R. C. Petersen. Comparison of different mri brain atrophy rate measures with clinical disease progression in ad. *Neurology*, 62(4):591–600, 2004.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

Christian Leibig, Vaneeda Allken, Philipp Berens, and Siegfried Wahl. Leveraging uncertainty information from deep neural networks for disease detection. *bioRxiv*, 2016.

N. A. Losseff, S. L. Webb, J. I. O'Riordan, R. Page, L. Wang, G. J. Barker, P. S. Tofts, W. I. McDonald, D. H. Miller, and A. J. Thompson. Spinal cord atrophy and disability in multiple sclerosis: A new reproducible and sensitive MRI method with potential to monitor disease progression. *Brain*, 119(3):701–708, 1996.

Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34 (10):1993–2024, 2015.

MS Society of Canada. What is MS?, last accessed on 06/30/18. https://mssociety.ca/about-ms/what-is-ms.

Huseyin Naci, Rachael Fleurence, Julie Birt, and Amy Duhig. Economic burden of multiple sclerosis. *Pharmacoeconomics*, 28(5):363–379, 2010.

Tanya Nair, Doina Precup, Douglas L. Arnold, and Tal Arbel. Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation. In Alejandro F. Frangi, Julia A. Schnabel, Christos Davatzikos, Carlos Alberola-López, and Gabor Fichtinger, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, pages 655–663, Cham, 2018. Springer International Publishing.

László G Nyúl and Jayaram K Udupa. 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.

MB Patwardhan, DB Matchar, GP Samsa, DC McCrory, RG Williams, and TT Li. Cost of multiple sclerosis by level of disability: a review of literature. *Multiple Sclerosis Journal*, 11(2):232–239, 2005.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

John G Sled, Alex P Zijdenbos, and Alan C Evans. A nonparametric method for automatic correction of intensity nonuniformity in mri data. *IEEE transactions on medical imaging*, 17(1):87–97, 1998.

Stephen M Smith. Fast robust automated brain extraction. *Human brain mapping*, 17(3):143–155, 2002.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

Ryutaro Tanno, Daniel E Worrall, Aurobrata Ghosh, Enrico Kaden, Stamatios N Sotiropoulos, Antonio Criminisi, and Daniel C Alexander. Bayesian image quality transfer with cnns: exploring uncertainty in dmri super-resolution. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 611–619. Springer, 2017.

Miguel Ulla, Jean Marie Bonny, Lemlih Ouchchane, Isabelle Rieu, Beatrice Claise, and Franck Durif. Is r2* a new mri biomarker for the progression of parkinson's disease? a longitudinal follow-up. *PLoS One*, 8(3):e57904, 2013.

V Wottschel, DC Alexander, PP Kwok, DT Chard, ML Stromillo, N De Stefano, AJ Thompson, DH Miller, and O Ciccarelli. Predicting outcome in clinically isolated syndrome using machine learning. *NeuroImage: Clinical*, 7:281–287, 2015.

Youngjin Yoo, Lisa W Tang, Tom Brosch, David KB Li, Luanne Metz, Anthony Traboulsee, and Roger Tam. 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*, pages 86–94. Springer, 2016.

Yijun Zhao, Brian C Healy, Dalia Rotstein, Charles RG Guttmann, Rohit Bakshi, Howard L Weiner, Carla E Brodley, and Tanuja Chitnis. Exploration of machine learning techniques in predicting multiple sclerosis disease course. *PLoS One*, 12(4):e0174866, 2017.