Deep Neural Net Forecasting of Multiple Sclerosis Disease Severity

Benjamin Insley,[1] Syed Rizvi,[2] Jonathon Cahill,[2] Joshua Stone,[3,4] and Carsten Eickhoff[1]

[1] *Center for Biomedical Informatics, Brown University, Providence, Rhode Island 02912, USA*
[2] *The Neurology Foundation, Brown University, Providence, Rhode Island 02912, USA*
[3] *Rhode Island Hospital, Providence, Rhode Island 02905, USA*
[4] *Brown University, Providence, Rhode Island 02912, USA*

Several machine learning methods were applied to multiple sclerosis (MS) patient data to predict 2-year disease course outcomes on two fronts; regression models used to estimate Expanded Disability Status Scale (EDSS) scores and binary classifier models that labeled patients as 'worsening' or 'non-worsening'. An ensemble of three regression models (two recurrent neural networks and a random forest regressor) accomplished a mean absolute error (MAE) of 0.51 and an accuracy of 50.0% on predicting 2-year EDSS scores on test data. A second ensemble of three binary classifier models (a dense neural network, a logistic regressor, and a linear support vector machine) achieved 67.6% accuracy on test data, with 32.4% precision and 35.9% recall for the 'worsening' class. A final ensemble combined both regression and binary classification by using a one-dimensional convolutional neural network as a preprocessing step. It produced a prediction on the 2-year EDSS score, and fed that prediction to a two-branched, binary classification neural network. This ensemble achieved 63.0% accuracy on test data, with 33.3% precision and 23.3% recall for the 'worsening' class.

INTRODUCTION

Multiple Sclerosis (MS) is a demyelinating disease of the central nervous system (CNS) that disrupts the body's ability to deliver signals across neurons [1–4]. MS is characterized by deteriorating motor and sensory function, but includes a host of other symptoms ranging from depression and anxiety to autonomic disfunction[5–8]. There are several physician-administered ordinal-scale instruments used to assess a patient's severity based on the intensity of certain physical and neurological impairments. The most widely-used scale is Kurtzke's Expanded Disability Status Scale (EDSS), which rates MS progression from 0 to 10 in 0.5 increments (excluding 0.5).

There are three main types of MS: relapsing-remitting (RRMS), secondary-progressive (SPMS), and primary progressive (PPMS). Each disease course is unique, and each show erratic, seemingly-unpredictable behavior. Extreme variance amongst different patients has made it difficult to establish standard progression rates, despite years of exploration into various MS models. Previous attempts include a multilayer regression model[9], a support vector machine (SVM)[6], logistic regression[4], Markov models[10], and a fully-connected perceptron deep neural net[7]. The most successful models had access to MRI data, which can be costly and time-consuming.

The surge of DNNs in recent years led to a type of model known as the recurrent neural network (RNN) specifically suited for reading time series data and forecasting the future. This advanced deep learning method, paired with other established machine learning techniques, may allow for a model to interpret a multitude of clinical signals from years of standard medical care to produce predictions on a patient's 2-year status. Here we explore the efficacy of these intricate models at predicting accurate disease outcomes in a data-impoverished setting.

Presented at the 71st Annual Meeting of The American Academy of Neurology (AAN), Philadelphia, PA.

MATERIALS AND METHODS

Rhode Island Hospital provided 7,631 records of individual MS patient appointments that included 1,164 unique patients. After filtering out subjects that had no diagnosis date and less than three years of data, the cohort was left with 3,848 visits and 705 unique patients. Table I shows the demographic breakdown of the dataset.

The data was first organized by patient into time series of medical appointments. Though appointments occurred at uneven time intervals, the records were bucketed by year and each patient time series was padded to be 5 years long. Tables II and III show the features available. The categorical signals were one-hot encoded (prepared as binary vectors where each index represents a category) and the numerical signals were normalized by removing the mean and dividing by the standard deviation.

Once organized as three-dimensional tensors with shape [patients, time steps, features], the data was processed by two different ensembles of machine learning models. The first ensemble is meant for binary classification, where each patient was labeled as 'worsening' or 'non-worsening' depending on their EDSS score 2 years ahead of their 5-year input sequence. Two neural nets comprised this ensemble; a convolutional net trained on

regression predicted 2-year EDSS scores, then fed its prediction as an additional feature to a two-branch neural net that produced the final binary output. With 705 total samples, 405 were used as a training set, 200 as a validation set, and 100 as a test set. The preprocessing network used mean squared error (MSE) as its loss function. The output network used binary cross-entropy as loss. Both used Adam as their optimizer.

The second ensemble was trained to predict EDSS scores two years into the future. It was composed of a convolutional-recurrent neural net, a time-distributed dense net, and a random forest regressor. Final predictions were determined by taking a weighted average of the individual predictions. The weights were 0.4, 0.1, and 0.5, respectively. Each model used MSE as its loss function. Both neural nets used stochastic gradient descent (SGD) as their optimizer. Since the random forest regressor has no means of validation during training, the 705 patients were split into a training set of 405 and a test set of 300.

Finally, the data was reorganized by appointment, such that there were no longer time series of signals, but rather single time slices used to predict 2-year outcomes (worsening or non-worsening). Another ensemble of three models, a fully-connected neural network with one skip connection, a logistic regressor, and a linear support vector machine (SVM), was trained on 2,356 samples (631 worsening) and tested on 1,492 samples (348 worsening). Due to the severe class imbalance, each model used its own method of data augmentation to get a more even split between 'worsening' and 'non-worsening'. The neural network adjusted the minority class weight to 10:1, such that it would be 10 times as costly to get a ''worsening' sample incorrect. In addition, Gaussian interpolation was used to upsample the minority class. This is when the 'worsening' samples are copied, Gaussian noise is added to their numerical features, and then new data points are created by taking two old samples and interpolating a new sample in between them. This artificially doubles the size of the minority class without directly copying those samples. The test cohort is not affected. The logistic regressor used a 3:1 class imbalance, and the linear SVM used interpolation upsampling without Gaussian noise. See Appendix for more information on model architectures.

Demographics	N
Number of subjects	705
Number of individual appointments	3,848
Number of females (%)	72.3
Average age at start of study	48
Average age at onset of diagnosis	37
Average EDSS score	2.62
Number of RRMS patients	486
Number of SPMS patients	152
Number of PPMS patients	45
Number of CIS patients	17
Number of RIS patients	10
Number of unspecified patients	1

TABLE I: Demographic data.

Provided Features	Data Type
Study ID*	Numerical
EDSS Score	Numerical
Type of MS	Categorical
Current Treatment (%)	Categorical
Past Treatment	Categorical
Gender	Categorical
Age	Numerical
Marital Status*	Categorical
Employment Status*	Categorical
ICD Code*	Categorical
Final Appointment Date*	Datetime
Current Appointment Date*	Datetime
Year Diagnosis*	Datetime

TABLE II: Features provided by Rhode Island hospital. Those features marked with an asterisk (*) are features not fed into the model. Some of the unused features were used to calculate other features.

RESULTS

CONCLUSION

In cases where MRI data is unavailable, this research shows the strength of a powerful neural network at providing insight into an MS patient's future. Recurrent architectures bolster traditional ML models by finding different representations of patient data that can better pick out progressive cases. However, achieving statistical power requires more data outside of clinical assessment. Future work will involve building a complimentary

Extrapolated Timeseries Features	Data Type
Disease Duration	Numerical
Age at Onset	Numerical
EDSS Subcategory	Categorical
Year-to-Date Visits (%)	Numerical
Initial Rate of Change†	Numerical
Initial Net Change†	Numerical
Average Rate of Change†	Numerical
Total Net Change†	Numerical
Previous EDSS Score‡	Numerical
Score at Onset	Numerical

TABLE III: Features calculated from those provided by Rhode Island Hospital. Those marked with a dagger (†) are features used only when a model had access to full time series of appointments for each patient. A double dagger (‡) indicates a feature used when a model was only looking at individual appointments. 'Initial' means calculated over the first three years, while *Average Rate of Change* and *Total Net Change* were calculated over the full sequence. *EDSS Subcategory* was used to denote what range a patient's EDSS score fell into: $1-3.5$, $4-5.5$, $6-8$, or >8.

Overall Accuracy	67.0%
Worsening Precision	39.6%
Worsening Recall	33.9%
Non-Worsening Precision	75.7%
Non-Worsening Recall	79.9%
Worsening Relative Frequency	28.0%
Non-Worsening Relative Frequency	72.0%

TABLE IV: Binary classification results with access to full patient history (validation cohort).

Overall Accuracy	63.0%
Worsening Precision	33.3%
Worsening Recall	23.3%
Non-Worsening Precision	70.9%
Non-Worsening Recall	80.0%
Worsening Relative Frequency	30.0%
Non-Worsening Relative Frequency	70.0%

TABLE V: Binary classification results with access to full patient history (test cohort).

Model	Mean Absolute Error	Accuracy(%)
Convolutional-Recurrent	0.54	30%
Time-Distributed Dense	0.54	40%
Random Forest	0.56	45%
Full Ensemble	0.51	50%

TABLE VI: Regression results with access to full patient history (test cohort).

Overall Accuracy	67.6%
Worsening Precision	32.4%
Worsening Recall	35.9%
Non-Worsening Precision	79.8%
Non-Worsening Recall	77.2%
Worsening Relative Frequency	23.3%
Non-Worsening Relative Frequency	76.7%

TABLE VII: Binary classification results with access to single appointment (test cohort).

network that processes raw MRI data alongside clinical signals.

[1] M. Boggild, The BMJ , 339 (2009).
[2] E. Kingwell, Journal of Neurology, Neurosurgery, and Psychiatry , 61 (2012).
[3] J. Palace, The BMJ (2014).
[4] V. B. Libertje, Multiple Sclerosis Journal (2011).
[5] A. Langer-Gould, JAMA Neurology , 1686 (2006).
[6] Y. Zhao, PLOS One (2017).
[7] B. Bejarano, BMC Neurology , 67 (2011).
[8] A. C. J. Janssens, Journal of Clinical Epidemiology , 180 (2004).
[9] K. Tilling, Health Technology Assessment (Winchester, England) , 1 (2016).
[10] S. Gauthier, Neurology , 2059 (2007).
[11] H. Tremlett, Neurology , 172 (2006).
[12] A. Shirani, Multiple sclerosis (Houndmills, Basingstoke, England) , 442 (2012).
[13] R. Bergamaschi, Journal of Neurology, Neurosurgery, and Psychiatry (2007).
[14] L. Bottaci, The Lancet , 469 (1997).
[15] L. Leocani, Journal of Neurology, Neurosurgery, and Psychiatry (2006).
[16] S. Fiorini, arXiv (2016).
[17] E. Fisher, Annals of Neurology , 255 (2008).
[18] A. Signori, Multiple Sclerosis Journal (2017).
[19] V. G. Jokubaitis, Annals of Clinical and Translational Neurology (2015).
[20] S. Meyer-Moock, BMC Neurology , 58 (2014).
[21] M. Trust, Multiple Sclerosis Trust (2018).
[22] F. Chollet, *Deep Learning with Python* (Manning Publications Co., Shelter Island, NY, 2018).
[23] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, Journal of Machine Learning Research **12**, 2825 (2011).
[24] F. Chollet *et al.*, "Keras," `https://keras.io` (2015).
[25] J. D. Hunter, Computing In Science & Engineering **9**, 90 (2007).

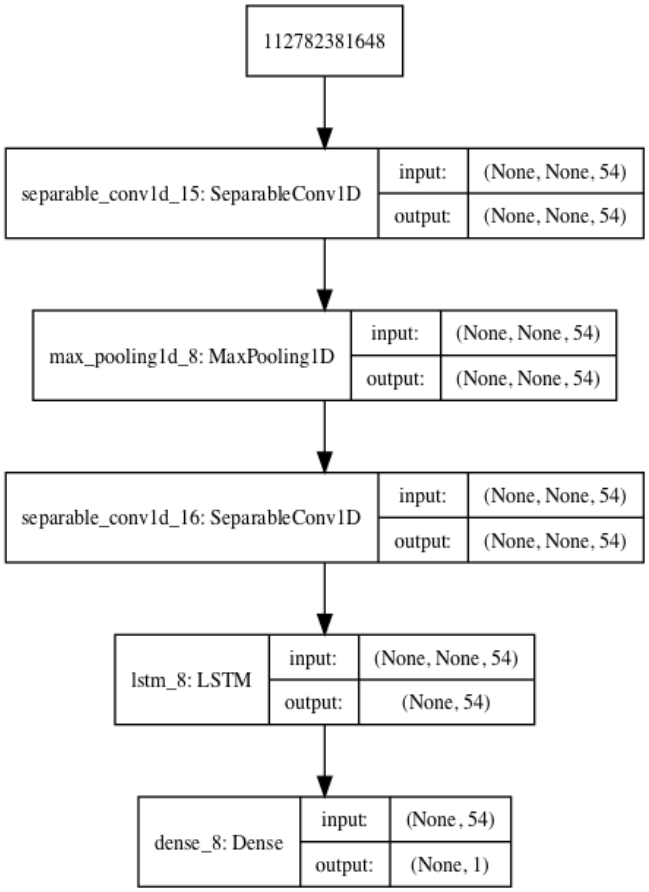

FIG. 1: Convolutional-recurrent model used in the temporal regression network.

Appendices

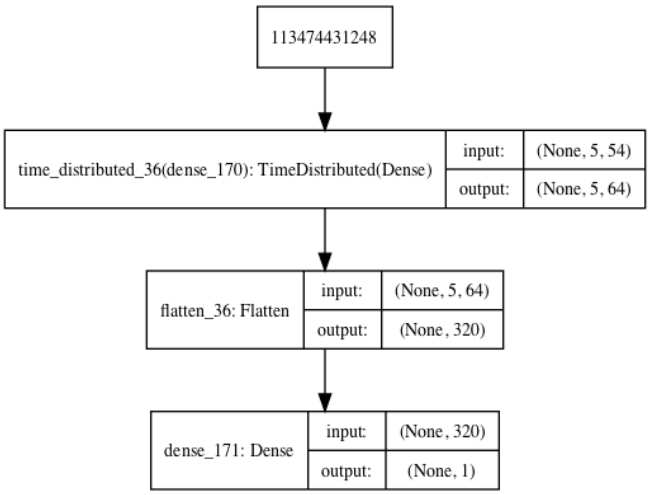

FIG. 2: Time-Distributed Dense model used in the temporal regression network.

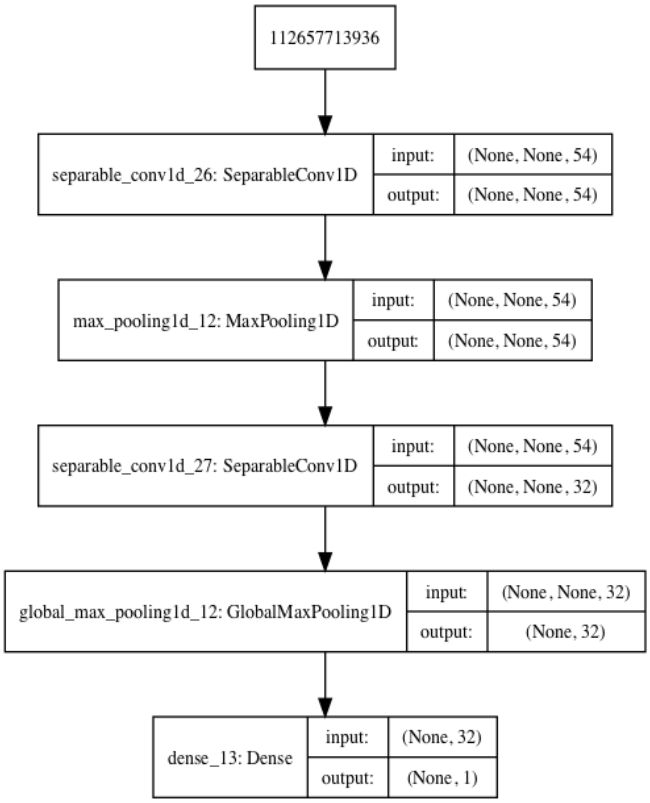

FIG. 3: Convolutional model used in the temporal binary classification network.

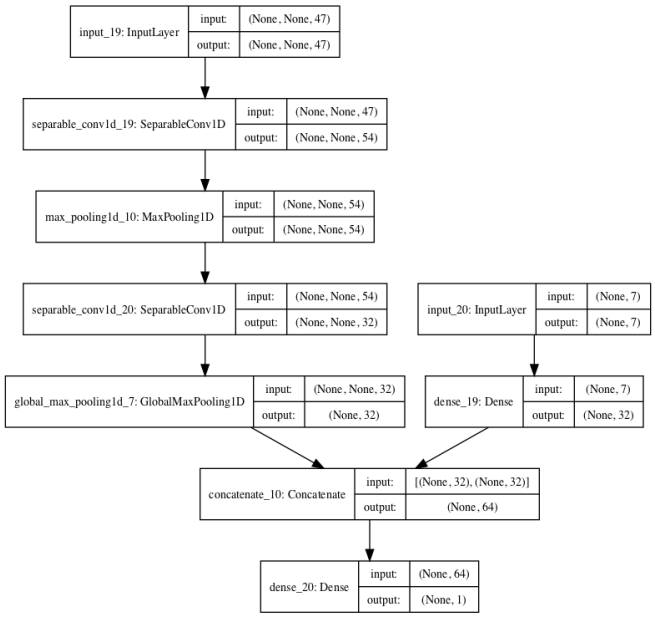

FIG. 4: Branched model used in the temporal binary classification network. The longer branch is purely convolutional, the shorter branch is fully-connected, and the outputs are concatenated then fed through another dense layer.

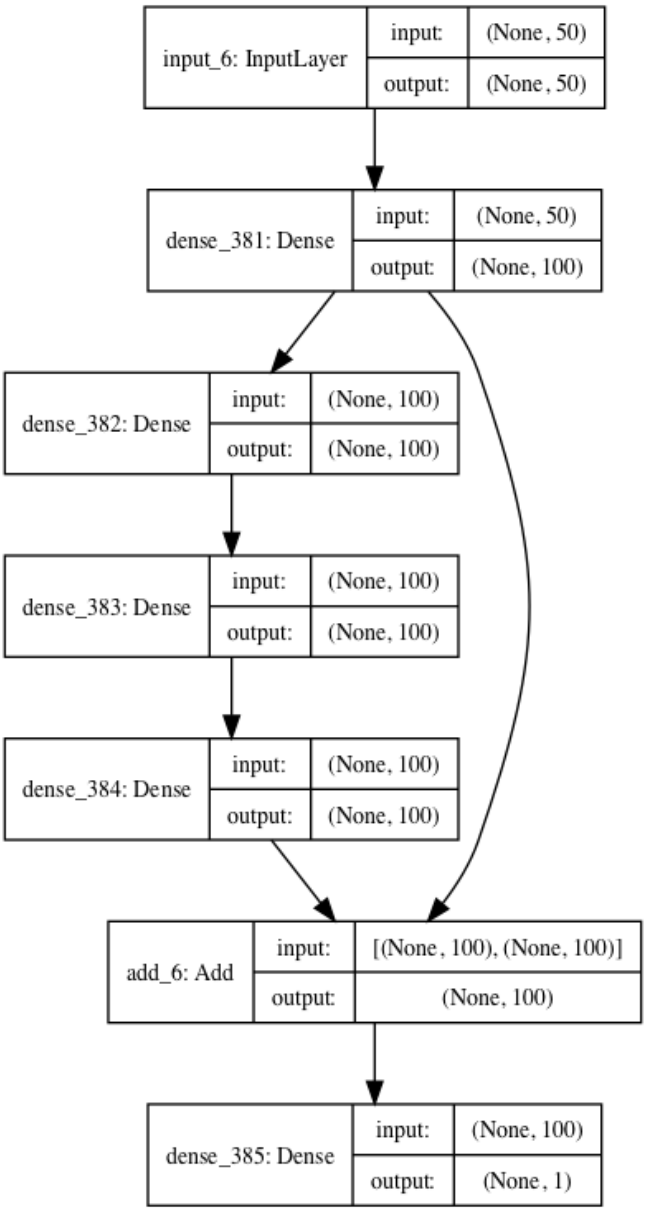

FIG. 5: Dense network used in the single-appointment binary classification network.