Ensemble Learning for Predicting Multiple Sclerosis Disease Course

Y. Zhao[1] **, T. Chitnis**[2]**, and T. Doan**[1]
[1] Computer and Information Science Department, Fordham University, New York, NY, USA
[2]Department of Neurology, Harvard Medical School, Boston, MA, USA

Abstract— *Multiple Sclerosis (MS) is a long-lasting disease affecting nearly 1 million adults over the age of 18 in the US. Patients suffer from chronic progressive disabilities and the rate of disability accumulation varies across patients. The identification of patients who are more likely to accrue disability would allow clinicians to institute more rigorous monitoring procedures, and potentially initiate more potent therapies in the early stages of the disease. In our research, we apply machine learning techniques to predict a "worsening" or "non-worsening" MS disease progression at the five-year mark using the first two years of patient longitudinal data. Besides exploiting standalone models, we investigate the efficacy of ensemble learning methods including the latest novel XGBoost and LightGBM algorithms. We further analyze the top predictive indicators associated with disability accumulation. Our work is based on a real-world dataset consisting of 724 patients enrolled in the Comprehensive Longitudinal Investigation in MS at Brigham and Women's Hospital (CLIMB study). Our experimental results demonstrate that machine learning models can achieve approximately 80% and 70% predictive accuracy for the "worsening" and "non-worsening" cases respectively shortly after disease onset. Furthermore, ensemble learning methods are more effective and robust compared to standalone algorithms. Risk factors consistently revealed by our models could help clinicians monitor the MS patients more effectively.*

1. Introduction

Multiple sclerosis (MS) is an autoimmune disease of the central nervous system in which the immune system attacks the myelin sheath (a fatty layer of substance protecting the nerves), resulting in loss/blockage of signals from the brain. The majority of cases are affected by relapses involving neurological deficits such as vision blurring or loss, weakness, numbness, and imbalance or cognitive deficits. In the early stages of the disease, relapses generally improve or remit. However, there may be residual deficits due to relapses later on. In addition, there is a superimposed process of progressive disability that results in permanent deficits. Cumulative disability is typically measured using a 0-10 Expanded Disability Status Scale (EDSS) [1], in which 0 is normal and 6 corresponds to walking with a cane. There is considerable variability in MS disease development with some patients demonstrating a benign disease progression, while others progress to an EDSS of 6 within 5 years. All currently approved MS therapies primarily target relapses, and have limited effects on overall disability progression. However, there is increasing evidence suggesting that early and more aggressive treatment targeting relapses may delay or prevent the long-term accumulation of disability, but this effect must be balanced with the potential increase in side effects associated with more potent therapies. The identification of patients who are more likely to accrue disability would allow clinicians to institute more rigorous monitoring procedures and potentially initiate more potent therapies early in the course of the disease.

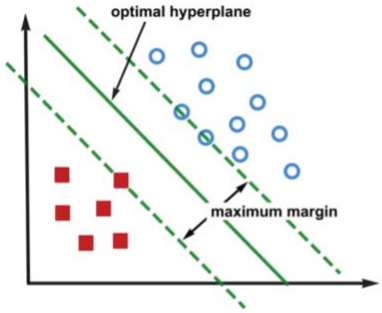
In our research, we work closely with doctors from Harvard Medical School and Brigham and Women's Hospital (BWH) in Boston, Massachusetts to predict the disability level of MS patients at the five-year mark using the first two years of their longitudinal data. Specifically, we predict which patients will accumulate disability ("worsening") and which are likely to remain without disability accumulation ("non-worsening") in their disease course. We define "worsening" as an increase of 1.5 or more from baseline EDSS to five-year mark EDSS, and "non-worsening" as all other cases, based on the fact that an EDSS increase of 1.0 or 1.5 is clinically significant and generally sustained, and is used as a primary or secondary endpoint in clinical trials of MS therapies.

In this paper, we present our findings by applying ensemble techniques to integrate information from multiple machine learning classifiers. Ensemble learning has been proven to produce better and more robust predictive performance compared to a single model. The technique has placed first in many prestigious machine learning competitions, such as the Netflix Competition, KDD and Kaggle. In our experiment, we created a heterogeneous meta-learner L from three established machine learning algorithms as our base classifiers: *Support Vector Machines* (SVM), *Logistic Regression* and *Random Forest*. We further investigated the efficacy of two homogeneous ensemble learners, *XGBoost* and *LightGBM*, which have gained much attention in recent years due to their superior performance.

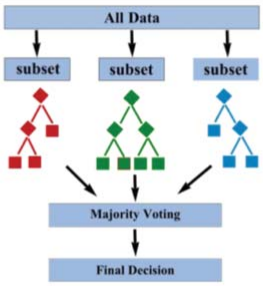
An additional motivation for our research is to study various risk factors affecting MS patients' disease progressions. To this end, we ranked the top predictors in our models and identified the most predictive factors. Detailed findings and discussions are presented in Section 5.

ISBN: 1-60132-502-9, CSREA Press ©

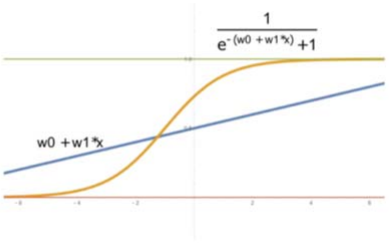
Fig. 1: Baseline Models



(a) Support Vector Machine



(b) Random Forest



(c) Logistic Regression

2. Data Acquisition and Preprocessing

Our data consist of 724 patients enrolled in the Comprehensive Longitudinal Investigation in MS at Brigham and Women's Hospital (CLIMB study) [2]. The features collected for this study consist of three categories: demographic, clinical and MRI related. The total number of variables in each category is 6, 17 and 7 respectively. In addition, patients in the CLIMB study are required to visit every six months where clinical test results such neurological function data are collected. MRI procedures are performed on these patients on an annual basis. To reflect the change of a patient's disease progression, we added a lagged variable (i.e., difference between the current value to the value in the previous period) for each clinical attribute. As a result, we have a total of 198 features in a two-year observation window.

Although we are more interested in predicting the "worsening" patients, they form the minority class in our training data. Indeed, we have only 165 progressive cases out of a total of 724 training samples. To address the class imbalance issue, we assigned a higher weight to minority class instances in all of our experimental models and selected the best weight for each model based on its performance on a validation set. [1]

3. Methods

In this section, we describe the methods we used to conduct our study. We present a performance comparison of these models and an analysis of the results in Section 4.

3.1 Baseline Models

We selected the following three established and popular machine learning algorithms as our baseline learners. These methods are illustrated in Figure 1. We describe three additional ensemble learning models in Section 3.2.

[1]Other class imbalance correction techniques (e.g., undersampling and SMOTE) were explored and resulted in worse performance.

3.1.1 Support Vector Machine (SVM)

SVM [3] performs classification tasks by constructing a decision boundary (i.e., hyperplanes) in a multidimensional space that separates instances of different class labels. As illustrated in Figure 1 (a), SVM strives to find a hyperplane that has the maximum margin, i.e the maximum distance between the hyperplane and the data points of both classes. Maximizing the margin distance provides reinforcement that future data points can be classified with more confidence.

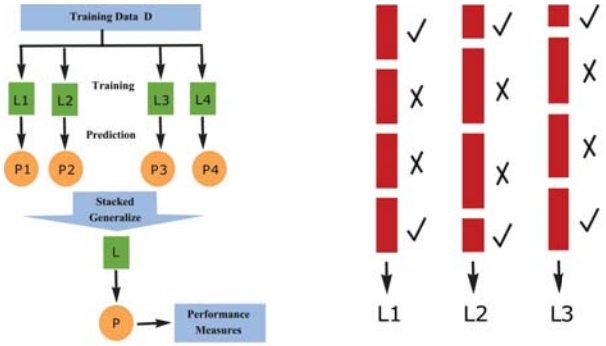
SVM is capable of transforming the data into a higher dimensional space using various kernel functions to enhance data separability. In our study, we adhered to the linear SVM to facilitate risk factor analysis.

3.1.2 Random Forest

A random forest [4] is a collection of *Decision Trees* (DTs). A DT model uses a tree structure to model the data in which each leaf node corresponds to a class label and attributes are represented as the internal nodes of the tree. Each branch represents a potential value of its parent node (i.e., an attribute). The major challenge in building a DT model is choosing the attribute for each node at each level. *Information Gain* and *Gini Index* are the two popular metrics used for attribute selection.

Decision Trees tend to have high variance since they are likely to overfit the training data. A *Random Forest* model, illustrated in Figure 1 (b), creates a forest of DTs where each DT is trained with a subset of training instances and a subset of attributes. By pooling predictions from multiple DTs, a *Random Forest* reduces the variance of each individual DT and achieves a more robust and superior performance. In our study, we used a random forest of 50 decision trees where each tree was built with ten randomly selected attributes. The rest of the model parameters were assigned the default values in Python's *scikit-learn* package.

Fig. 2: Ensemble Learning Methods



(a) Ensemble Learning (b) Adaptive Boosting

3.1.3 Logistic Regression

Logistic regression [5] is a type of generalized linear model (GLM) which studies the association between a categorical response variable Y and a set of independent (explanatory) variables $X = \{X_1, X_2, \ldots, X_n\}$. As illustrated in Figure 1 (c), the Y variable is first modeled as a linear function of X with coefficients $W_0, W_1, W_2, \ldots, W_n$, and then the predictions (y_i's) are transformed into probability scores using a sigmoid function $f(y) = \frac{1}{1+e^{-y}}$. In a binary classification task, the scores indicate a corresponding instance's likelihood of belonging to the positive class. Thus, a cutoff (usually 0.5) can be established as a decision boundary to further categorize the instances into the more likely class. The "training" process involves adjusting the coefficients to maximize the cross-entropy of the model outputs and the true class labels.

3.2 Ensemble Methods

In addition to individual machine learning algorithms, we explored ensemble techniques [6] to integrate information from the four base classifiers described in Section 3.1. Ensemble learning is a family of algorithms that seek to create a "strong" classifier based on a group of "weak" classifiers. In this context, "strong" and "weak" refer to how accurately the classifiers can predict the target variable. Ensemble learning has been proven to have improved and more robust performance than a single model.

3.2.1 Meta-learner L

Figure 2 (a) illustrates the principle of ensemble learning. Specifically, multiple base classifiers, $L_1, L_2, \ldots L_n$, are built for the original classification task with training data D. A meta-learner L is constructed by combining the outcomes from the base classifiers to improve predictive accuracy. Our meta-learner L is an example of a $heterogeneous$ ensemble because its base learners are obtained from different machine learning algorithms. Our next model, *XGBoost* [7], explores the efficacy of a $homogeneous$ ensemble where the base classifiers are obtained using a single machine learning algorithm.

As illustrated in Figure 2 (a), for the task of combining the outcomes from the base learners, we applied stacked generalization [8] in which an additional linear regression model was trained to predict the target variable in D based on the individual predictions from the four base classifiers. Stacking typically yields better performance than a straightforward majority voting approach [8].

3.2.2 $XGBoost$

We investigated the performance of *XGBoost* [7], an algorithm that has gained much popularity and attention since its inception in 2016 and was the winning algorithm for a number of machine learning competitions [9]. *XGBoost* belongs to the family of *homogeneous* ensemble methods in which the base learners, $L_1, L_2, \ldots, L_n$, are created using a single machine learning algorithm exploiting the concept of "adaptive boosting". Figure 2 (b) illustrates the concept of "adaptive boosting". Specifically, a sequence of classifiers is generated with each new model aiming to correct the errors of the previous model. This correction is typically achieved by boosting the weights of the misclassified instances in the previous model so that the new model will have a higher likelihood of correctly classifying them. New models are added sequentially until no further improvements can be achieved. In *XGBoost*, instead of boosting the weights, the algorithm fits the new model to residuals of the previous model and then minimizes the loss when adding the latest model. The process is equivalent to updating your model with a gradient descent towards a locally optimal solution.

3.2.3 $LightGBM$

LightGBM [10] is a gradient boosting tree-based framework which implements two new techniques: *Gradient-based One-Side Sampling* (GOSS) and *Exclusive Feature Bundling* (EFB). In particular, with GOSS, the algorithm keeps all large gradient instances and only samples from the population of small gradient instances. Thus, GOSS focuses on large gradient instances as they are considered undertrained. With EFB, the algorithm bundles mutually exclusive features (i.e., they rarely take nonzero values simultaneously) to reduce the number of features. Compared to other tree-based algorithms, *LightGBM* produces much more complex trees by following a leaf-wise split rather than a level-wise split, which is the main factor contributing to *LightGBM*'s superior performance.

4. Experimental Results

All experiments were conducted by running 10-fold cross-validation. In addition to overall predictive accuracy, Sensitivity and Specificity were used to measure the performance

Table 1: Performance Comparison of Six Models

Threshold	Model	Sensitivity	Specificity	Overall
0.5	SVM	0.60	0.70	0.68
	Logistic Regression	0.70	0.71	0.71
	Random Forest	0.72	0.73	0.73
	XGBoost	0.50	0.87	0.79
	LightGBM	0.51	0.86	0.78
	Meta-L^*	0.61	0.84	0.79
0.45	SVM	0.75	0.64	0.67
	Logistic Regression	0.76	0.62	0.65
	Random Forest	0.83	0.51	0.58
	XGBoost	0.58	0.82	0.77
	LightGBM	0.52	0.85	0.77
	Meta-L^*	0.71	0.74	0.73
0.4	SVM	0.81	0.51	0.58
	Logistic Regression	0.81	0.57	0.62
	Random Forest	0.91	0.34	0.47
	XGBoost	0.68	0.76	0.74
	LightGBM	0.58	0.82	0.77
	Meta-L^*	**0.78**	**0.65**	**0.68**
0.35	SVM	0.92	0.34	0.47
	Logistic Regression	0.86	0.49	0.57
	Random Forest	0.98	0.11	0.31
	XGBoost	**0.79**	**0.69**	**0.71**
	LightGBM	0.70	0.76	0.75
	Meta-L^*	0.86	0.50	0.58
0.3	SVM	0.96	0.21	0.38
	Logistic Regression	0.91	0.41	0.52
	Random Forest	0.99	0.06	0.27
	XGBoost	**0.81**	**0.64**	**0.68**
	LightGBM	**0.78**	**0.68**	**0.70**
	Meta-L^*	0.93	0.35	0.48

Meta-L: ensemble of SVM, Logistic Regression and Random Forest.

in the positive and negative classes respectively. The hyper-parameters in this study were selected via a grid search based on the highest AUC (Area Under the ROC Curve) score.

Table 1 presents the main results of our experiment. Since we are most interested in predicting the "worsening" patients, we applied different thresholds in the ROC curve to increase a model's Sensitivity at the cost of lowering the Specificity. For each threshold displayed in column 1 of Table 3, we present a performance comparison of the six models described in Section 3 using overall sensitivity, specificity, and overall accuracies. Consequently, we can observe the trade-offs between an increase in sensitivity and a decrease in specificity for each model as we shift the threshold. A healthcare institution can select a desired threshold depending on its level of tolerance on the insufficient performance of the "non-worsening" class (i.e., low Specificity). In addition, we observe that:

- The highest accuracy on the "worsening" class which is of practical value is approximately 80%. This is because further improvement would lead to a less than 50% performance of the "non-worsening" class. Given 80% as the benchmark on the "worsening" class, $XGBoost$ and $LightGBM$ are the best candidates with each achieving close to 70% on the other class at the threshold of 0.35 and 0.3 respectively. Meta-learner L is the next runner-up with 65% accuracy on the "non-worsening" class.
- It is worth noting that some algorithms are more sensitive to the shift of threshold values. For example, *Random Forest* degenerated quickly as the threshold value moved below 0.4. On the other hand, $XGBoost$ and $LightGBM$ maintained a steady trade-off between the two classes as we varied the thresholds. We conclude that they are the models of our choice in our study due to their superior performance and robustness.

Table 2: List of Top 10 Predictive Features

Rank	SVM	Logistic Regression	Random Forest
1	**ΔEDSS**	**ΔEDSS**	**ΔEDSS**
2	**PYRAMIDAL_FUNCTION**	**PYRAMIDAL_FUNCTION**	EDSS
3	ΔLESION_VOLUME	ΔAMBULATORY_INDEX	**PYRAMIDAL_FUNCTION**
4	**ΔDISEASE_CATEGORY**	MRI_STATUS	AMBULATORY_INDEX
5	ΔAMBULATORY_INDEX	**ΔDISEASE_CATEGORY**	DISEASE_ACTIVITY
6	AMBULATORY_INDEX	BOWEL_BLADDER_FUNCTION	DISEASE_STEP
7	BOWEL_BLADDER_FUNCTION	DISEASE_ACTIVITY	ΔAMBULATORY_INDEX
8	ΔTOTAL_GD	ΔTOTAL_GD	ΔSENSORY_FUNCTION
9	DISEASE_ACTIVITY	AMBULATORY_INDEX	DISEASE_CATEGORY
10	ΔWALKING_ABILITY	DISEASE_COURSE_SUBTYPE	ΔTDS_BPF
Rank	**XGBoost**	**LightGBM**	
1	**ΔEDSS**	EDSS	
2	EDSS	**ΔEDSS**	
3	**DISEASE_CATEGORY**	**DISEASE_CATEGORY**	
4	MRI_STATUS	MRI_STATUS	
5	**PYRAMIDAL_FUNCTION**	**PYRAMIDAL_FUNCTION**	
6	ATTACKPREV2Y	ΔAMBULATORY_INDEX	
7	FAMILY_MS	ATTACKPREV2Y	
8	AMBULATORY_INDEX	FAMILY_MS	
9	DISEASE_ACTIVITY	BOWEL_BLADDER_FUNCTION	
10	VISIT_AGE	DISEASE_ACTIVITY	

Δ:	change in the indicated variable
AMBULATORY_INDEX:	Ordinal scale of gait capacity
ATTACKPREV2Y:	number of attacks in the previous two years
BOWEL_BLADDER_FUNCTION:	measure of sensory function ranging from 0 (normal) to 6 (loss of bowel and bladder function)
DISEASE_ACTIVITY:	physician reported metric of curent inflammatory or progressive disease status
DISEASE_CATEGORY:	code indicating disease categories such as primary progressive, secondary progressive, etc.
DISEASE_STEP:	scale of disability
EDSS:	overall disability measure
FAMILY_MS:	code indicating family history of MS, including mother, father, sibling, cousin, etc.
LESION_VOLUME:	T2 lesion volume meastured on brain MRI
MRI_STATUS:	presence of new MrI lesions
PYRAMIDAL_FUNCTION:	measure of paramidal function ranging from 0 to 6
SENSORY_FUNCTION:	measure of bowel bladder function ranging from 0 (normal) to 6 (sensation lost below the head)
TDS_BPF:	1.5T TDS+ calculated brain parenchymal fraction
TOTAL_GD:	total number of Gad+ lesions
VISIT_AGE:	age of the subject

5. Risk Factor Analysis

We next examined the major factors that are predictive to MS progression. Five linear and tree-based algorithms, *SVM*, *Random Forest*, *Logistic Regression*, *XGBoost* and *LightGBM* were selected for the study. These models were chosen because their feature importance was well defined. In particular, for linear models, the importance is proportional to the magnitude of the coefficients. For tree-based models, the ranking follows the order of attributes that the algorithm selected to split the branches. Table 2 presents the top ten predictive features identified by each of the five models. Based on Table 2:

- Examining the top five risk factors associated with each model, we identified two consistent principal predictors (highlighted in bold) across all models. The first one, as expected, is the change of a patient's *EDSS* score (Δ*EDSS*). The second one is a patient's pyramidal function measure (*PYRAMIDAL_FUNCTION*).

In addition, a patient's MS disease category (*DISEASE_CATEGORY*) is another important variable that appeared in four out of the five models. Specifically, *SVM* and *Logistic Regression* are dependent on Δ*DISEASE_CATEGORY*, while *XGBoost* and *LightGBM* rely on the value of *DISEASE_CATEGORY* itself.

- Expanding our investigation to the top 10 risk factors associated with each model, we could identify two more common risk factors across all models, namely *DISEASE_ACTIVITY* and *AMBULATORY_INDEX* (or its related change ΔAMBULATORY_INDEX).
- In addition to the seven common risk factors revealed by all models, the measure of a patient's bowel and bladder function (*BOWEL_BLADDER_FUNCTION*) is the next important risk factor to watch out for because it appeared in three out of the five models.
- Furthermore, *SVM*, *Logistic Regression* and Random Forest rely on a patient's total number of Gad+ lesions (*TOTAL_GD*) and TDS+ calculated brain parenchymal fraction (*TDS_BPF*) in making their predictions, whereas *XGBoost* and *LightGBM* utilize a patient's genetic information , i.e., (*FAMILY_MS*) in their decisions.

6. Conclusions and Future Work

In this paper, we applied machine learning techniques to predict levels of disability accumulation for MS patients at the five-year mark based on two-year clinical observations. We built our model using 724 real-world patients enrolled in the CLIMB study at Brigham and Women's Hospital. We employed three baseline machine learning models and three ensemble learners in our study. We further addressed the data imbalance issue by increasing the weight for the minority class. Our experimental results demonstrate that XGBoost and LightGBM offer comparable predictive power for our task, and their performances are superior to the other four models in our study.

In addition, we examined the top risk factors identified by our linear and tree-based models. We conclude that a patient's change in EDSS scores (Δ*EDSS*), pyramidal function measure (*PYRAMIDAL_FUNCTION*), MS disease category (*DISEASE_CATEGORY*), disease activity (*DISEASE_ACTIVITY*) and ambulatory index (*AMBULATORY_INDEX*) are the top predictive indicators to forecast a patient's disability level in five years.

For future work, we plan to explore time series models such as recurrent neural networks (RNN) to better model the temporal aspect of the data. We also plan to incorporate genetic information and additional biomarkers from patients' medical records.

Acknowledgment

The CLIMB study is supported by the Ann Romney Center for Neurological Diseases, Serono and the National MS Society.

References

[1] J. Kurtzke, "Rating neurologic impairment in multiple sclerosis: An expanded disability status scale (edss)," *Neurology 33 (11)*, pp. 1444–1452, 1983.

[2] S. Gauthier, B. Glanz, M. Mandel, and H. W. HL, "A model for the comprehensive investigation of a chronic autoimmune disease: The multiple sclerosis climb study," *Autoimmun Rev.*, vol. 5(8), pp. 532–536, 2006.

[3] C. Cortes and V. Vapnik, "Support vector machine," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[4] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[5] S. W. Menard, *Applied logistic regression analysis*, 1995, no. 04; e-book.

[6] T. G. Dietterich, "Ensemble methods in machine learning," in *International workshop on multiple classifier systems*. Springer, 2000, pp. 1–15.

[7] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. ACM, 2016, pp. 785–794.

[8] D. H. Wolpert, "Stacked generalization," *Neural networks*, vol. 5, no. 2, pp. 241–259, 1992.

[9] "Xgboost - ml winning solutions (incomplete list)," *https://github.com/dmlc/xgboost/tree/master/demo#machine-learning-challenge-winning-solutions*, 2016.

[10] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems*, 2017, pp. 3146–3154.