Novel Morphological and Appearance Features for Predicting Physical Disability from MR Images in Multiple Sclerosis Patients

Jeremy Kawahara, Chris McIntosh, Roger Tam and Ghassan Hamarneh

Abstract Physical disability in patients with multiple sclerosis is determined by functional ability and quantified with numerical scores. In vivo studies using magnetic resonance imaging (MRI) have found that these scores correlate with spinal cord atrophy (loss of tissue), where atrophy is commonly measured by spinal cord volume or cross-sectional area. However, this correlation is generally weak to moderate, and improved measures would strengthen the utility of imaging biomarkers. We propose novel spinal cord morphological and MRI-based appearance features. Select features are used to train regression models to predict patients' physical disability scores. We validate our models using 30 MRI scans of different patients with varying levels of disability. Our results suggest that regression models trained with multiple spinal cord features predict clinical disability better than a model based on the volume of the spinal cord alone.

1 Introduction

Multiple sclerosis (MS) studies have found that a patient's physical disability correlates with spinal cord atrophy [1, 7, 8, 12, 16]. Measuring spinal cord atrophy is potentially useful for monitoring the progression of diseases or the effectiveness of therapies [12]. Spinal cord atrophy is defined as a loss of tissue and commonly measured by cross-sectional area (CSA) or spinal cord volume [7, 8, 12]. To quantify the CSA, user-guided computer software is often used to assist in delineating the spinal cord from a 3D MRI (e.g. using one of several recently developed approaches

Jeremy Kawahara, Ghassan Hamarneh, e-mail: {jkawahar, hamarneh}@sfu.ca
Medical Image Analysis Lab., Simon Fraser University, Burnaby, Canada

Chris McIntosh, e-mail: cmcintos@sfu.ca
Princess Margaret Cancer Centre, University Health Network, Toronto, Canada

Roger Tam, e-mail: roger.tam@ubc.ca
MS/MRI Research Group, University of British Columbia, Vancouver, Canada

[4, 5, 10, 11, 15]). The segmented cord's volume or averaged CSA is computed and correlated with the patient's clinical disability score.

To quantify the clinical disability of a patient with MS, clinicians commonly rely on the Expanded Disability Status Scale (EDSS) [6] which assigns the patient a number between zero (a normal neurological exam) and ten (death from MS). Although commonly used, the EDSS score suffers from reproducibility issues, focuses largely on a patient's ambulatory impairment, and is restricted to an ordinal scale. This motivated the development of the Multiple Sclerosis Functional Composite (MSFC) score [3], which we discuss in section 2.5.

While the CSA of the spinal cord has been shown to correlate with clinical score, this correlation is generally moderate with some studies failing to show the expected reduction in CSA [9]. This may be because a reduction in cord size is only one global aspect of atrophy, and few other features that capture more subtle aspects have been explored. Schnabel et al. [13] explored local and global shape measurements across scales and concluded that the spinal cord shape should be measured across a range of scales. In conventional and diffusion tensor (DT) MRIs, Benedetti et al. [1] identified the brain T2 lesion volume, CSA and the mean fractional anisotropy of the cervical cord as features that independently influenced the EDSS score using a multivariate regression model. Composite scores, obtained by combining these three features, improved the correlation with clinical scores when compared to the correlations of a single feature. However, DT-MRI is much less commonly acquired than structural MRI. Valsasina et al. [16] explored the regional atrophy of the cervical cord by applying voxel-wise statistics on registered spinal cord segmentations. They used the determined regional atrophy in a multiple regression model, adjusted for age, sex, and cord volume, and showed correlations with clinical scores and patterns of atrophy.

Although a number of composite MRI biomarkers for MS have been proposed, computing morphological features to capture atrophy and combining these features in *linear* and *non-linear regression models* has not been well studied. As well, few works have testing whether combining *multiple spinal cord features* into a single model will provide a better indicator of disability than just using a single feature. Introducing new atrophic features and methods to combine them may assist clinicians in diagnosis, provide insights into disease progression, and serve as a useful composite biomarker.

We propose novel features extracted from MRI and the corresponding spinal cord segmentation that are potentially more specific to the clinical status than pure area or volume. Using these extracted features, we employ different regression models ranging in complexity and intuitiveness, starting with simple linear regression models, then multiple linear regression models and finally, non-linear non-parametric regression forests. To determine which of our proposed candidate features are useful biomarkers, we explore our data for features that are consistently associated with clinical state. Our results suggest that our proposed features and the more complex regression models are capable of outperforming the predictive abilities of a linear regression model using only spinal cord volume as the explanatory variable.

2 Methods

In this section we describe our data and the regression problem, examine the new candidate spinal cord features, outline the different types of regression models used, describe our cross-validation set-up, and finally discuss how the clinical scores are computed.

2.1 The Data and the Problem

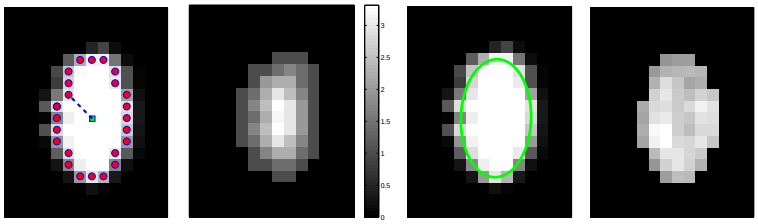
We are given a set of n MRI scans $I = \{I_1, \ldots, I_n\}$ where each 3D MRI scan I_i has a corresponding real number clinical score $y_i \in Y$, and a corresponding spinal cord segmentation $S_i \in S$. The dimensions of I_i and S_i are the same. Each voxel in S_i has a value between 0 and 1, where 0 represents the background and 1 represents the spinal cord. Voxels in S_i that are on the boundary of the spinal cord are assigned a fuzzy value between 0 and 1 that represents an estimated percentage of the voxel that contains spinal cord (i.e., partial volume) [15].

Our objective is to create a model M, using the images I and segmentations S, capable of predicting the patients' clinical scores Y from novel MR images. We extract a set of features X from I and S that are transformed by model M into values $\hat{Y}$, such that these predicted values $\hat{Y} = M(X)$ estimate the corresponding clinical scores Y.

One approach is to set M as a simple linear regression model with the spinal cord volume as the single explanatory variable X. This is similar to the existing literature where a Pearson's correlation coefficient is computed to measure the linear dependency between the spinal cord volume and clinical score. However, as mentioned in the introduction, this linear dependency using spinal cord volume does not always reveal a strong clinical relationship. We improve on this by deriving new morphological and MRI-based appearance features X and examining ways to combine them in more descriptive models M.

2.2 Candidate Features

We describe simple candidate morphological and appearance features X that are potentially sensitive to spinal cord changes. This is not meant to be a comprehensive set of features, but is sufficient to explore the potential of going beyond measuring cord size to predict disability. We first define the commonly used spinal cord volume, which is computed by summing all voxels, including the partial volumes $S_i(j) \in [0,1]$, in the segmentation, $vol = \sum_{j=1}^{J} S_i(j)$, where J is the total number of voxels in S_i. While spinal cord volume captures a global measure of spinal cord atrophy, we are also interested in features that vary at least partly independently from area or volume, and that are sensitive to spinal cord changes at a local scale.



(a) (b) (c) (d)

Fig. 1: Illustrations of the proposed features. (a) The distances (*dashed line*) from the center-of-mass (*center box*) to the boundary voxels (*circles*) make up per_k. (b) The distances to the nearest boundary point from the voxels inside the cord give $dist_k$ (brighter implies farther). (c) An ellipse is fit to the cord. (d) The normalized intensities of the cord are considered in int_k.

Our first proposed feature is designed to be more sensitive to local changes in the spinal cord's boundary. On each 2D axial slice of the segmentation S_i, we find voxels on the boundary between the spinal cord and background by considering voxels in S_i with a partial volume greater than 0.5 to be spinal cord. This results in a 2D binary image that we use to extract the cord's boundary voxels. For the k^{th} 2D axial slice of the spinal cord, we take the Euclidean distance between the center-of-mass c_k of the cord's k^{th} cross section, and the spinal cord boundary/perimeter voxels b computed as, $per_k = (d(c_k, b_k^1), \ldots, d(c_k, b_k^{m(k)}))$, where b_k^i represents the i^{th} boundary voxel on the k^{th} slice, and $d(c,b)$ computes the Euclidean distance between the two coordinates (Fig. 1a). The number of boundary voxels $m(k)$ can change for each 2D slice. We find the *minimum* distance from the center-of-mass to the boundary voxels in each 2D slice averaged over K 2D slices,

$$per_{\text{min}} = \frac{1}{K} \sum_{k=1}^{K} \min(per_k). \tag{1}$$

In a similar way, to compute additional features we replace the "min" function from (1) with the mean (per_{mean}), standard deviation (per_{std}), and the max (per_{max}) functions.

We define a related measure that focuses on local changes in 3D by calculating a 3D distance transform from the surface of the segmented spinal cord masked by (or restricted to) the interior region of the cord. To compute the distance transform, we calculate the Euclidean distance between voxels inside the spinal cord and the nearest boundary voxel in 3D. To further differentiate this feature from the *per* features, we consider voxels that contain any partial volume to be spinal cord, which changes the boundary voxels. The distance transform for slice k with q voxels inside the cord is represented as $dist_k = (t_k^1, \ldots, t_k^{q(k)})$ where t_k^i is the distance from the i^{th} voxel inside the cord on the k^{th} slice to the nearest 3D boundary coordinate

(Fig. 1b). The number of voxels inside the cord, $q(k)$, can change for each 2D slice. In a similar fashion to (1), we replace per_k with $dist_k$ and the "min" function with the mean ($dist_{\text{mean}}$), max ($dist_{\text{max}}$), standard deviation ($dist_{\text{std}}$) and the max divided by the mean distance ($dist_{\text{mean}}^{\text{max}}$) function averaged over the K 2D slices. For clarity we formally define,

$$dist_{\text{mean}}^{\text{max}} = \frac{1}{K}\sum_{k=1}^{K}\frac{\max(dist_k)}{\text{mean}(dist_k)}, \tag{2}$$

which averages the ratio of the furthest boundary distance by the mean distance.

To compute features that are more robust to local noise, such as small segmentation errors, we fit an ellipse (Fig. 1c) to each 2D cross-sectional slice of the segmented spinal cord and compute the eccentricity (ecc), minor axis (ax_{min}), and major axis (ax_{maj}), averaged over the length the cord.

All the features proposed so far are dependent on the geometrical characteristics of the cord, but we also include features based on the intensities found within the MRI. As the intensity values can vary widely in different MRI scans, we normalize a scan's intensities by its overall 3D scan intensities to produce z-scores. We extract the z-scores of those voxels that are labelled as spinal cord (partial volume > 0.5) and take the mean (int_{mean}) and standard deviation (int_{std}) of the spinal cord intensity values averaged over the K 2D slices (Fig. 1d).

2.3 Regression Models

Linear regression employs a linear function to model the relationship between the explanatory variable (e.g. spinal cord volume) and a response variable (clinical score). The parameters of this model are the coefficients β of the explanatory variables and the error term ε. These coefficients can be estimated from the data by applying a *least-squares* fitting that minimizes the differences between the response variable and the fitted explanatory variable. A model with only a single explanatory variable x_1, is known as *simple linear regression*, and is one of the simplest models to analyze. Given a dataset with n observations, this produces a straight line, $y_i = \beta_1 x_{i1} + \varepsilon_i, i = 1,\ldots,n$. *Multiple linear regression* builds on this by adding r explanatory variables to the model, $y_i = \beta_1 x_{i1} + \ldots + \beta_r x_{ir} + \varepsilon_i$.

While these models assume a linearity of the underlying relations, we also explore a more flexible, non-linear, non-parametric model, known as a *regression forest*. A regression forest significantly differs from the previously described models as it is completely learned from the data and makes no assumptions about the underlying distributions [2].

2.4 Training and Testing the Models

The models in section 2.3, are described in order of increasing complexity. With this added complexity, we increase the potential to accurately model the underlying function, but also increase the difficulty in intuitively understanding the model and increase the likelihood of over-fitting the model to the training data. To reduce the possibility of over-fitting, we divide our data into a training and testing set. Given the relatively small size of our dataset, we use leave-one-out cross-validation. This is repeated for all samples to give us an indication of the robustness and generalizability of our regression model and chosen features.

2.5 Clinical Scores

As discussed in the introduction, the EDSS and the MSFC scores, which we aim to predict from X, are commonly used to quantify clinical disability. We choose to focus on the MSFC score rather than the EDSS score because the MSFC captures disability to which the EDSS score is relatively insensitive, such as arm/hand function. In addition, the EDSS scores tend to exhibit a poor distribution due to the non-linearity of the scale, with many patients clustered between 4.5 and 6.5 (Fig. 2a).

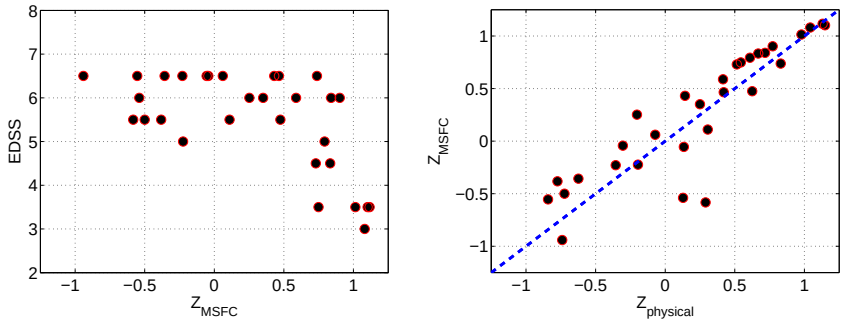
The MSFC score tests for: upper extremity function, determined by a 9-hole peg test (9-HPT); walking speed, measured by a timed 25-foot walk (T25W); and cognitive function, evaluated by a paced auditory serial addition test (PASAT). These three tests are shown to vary relatively independently, be sensitive to changes over time, and capture aspects of MS that are not captured in the EDSS score [3]. These components averaged together compose the MSFC score,

$$Z_{\text{MSFC}} = (Z_{\text{9-HPT}} - Z_{\text{T25W}} + Z_{\text{PASAT}})/3 \tag{3}$$

where the scores are normalized to produce z-scores using a reference population that includes healthy controls [3].

While this composite score is used to give an overall indication of the progression of multiple sclerosis, we do not expect the cognitive component, Z_{PASAT}, to have a strong causal relation with spinal cord atrophy as the spinal cord is not directly related to cognitive function. We test this by computing the Pearson's correlation coefficient with the cognitive test Z_{PASAT} and spinal cord volume vol, and do not find a significant correlation (r = -0.016, p-value = 0.93). For this reason, we remove Z_{PASAT} and only include the physical disability tests to define a new clinical measurement of physical disability,

$$Z_{\text{physical}} = (Z_{\text{9-HPT}} - Z_{\text{T25W}})/2. \tag{4}$$



(a) EDSS vs. Z_{MSFC} (b) Z_{MSFC} vs. Z_{physical}

Fig. 2: The distributions of scores are shown. (a) The Z_{MSFC} scores have a wider distributions than EDSS scores. As expected, as EDSS decreases, there is a trend for Z_{MSFC} to increase. (b) We remove the cognitive component from Z_{MSFC} to form Z_{physical}, slightly changing the distributions (*deviations from dashed line*).

This combined physical score, Z_{physical}, is the clinical score we use as the response variable for this work. The distribution of values and the changes in correlation between Z_{MSFC} and Z_{physical} are shown in Fig. 2b.

3 Results

We validate our proposed features and models using 30 3D T1-weighted MRIs acquired with a spoiled gradient echo sequence and an MR field strength of either 1.5 tesla or 3.0 tesla. These scans were gathered from multiple centers and parameters varied by site. Each scan is from a different patient (age ranged from 34 to 64) with secondary progressive MS. For each 3D MRI, we have its corresponding clinical score as described in section 2.5 and a segmentation of the spinal cord. To ensure reasonably accurate segmentations, we use a seeded semi-automatic method similar to Tench et al. [15] where a user-guided region growing algorithm marks the spinal cord voxels with a 1 and the background voxels with a 0. Due to the limited resolution of the MRIs and the small size of the cord, voxels on the boundary of the spinal cord, composed both of spinal cord and background, make up approximately 25% of the total voxels in the cross-sectional area [15]. To give an estimate of the spinal cord area contribution these boundary voxels make, the boundary voxels are assigned a fuzzy value between 0 and 1, computed as a function of the cord, boundary and cerebrospinal fluid intensities, based on equation (2) in [15]. The original MRI voxel resolutions were either $0.976 \times 0.976 \times 1$ mm or $0.976 \times 0.976 \times 1.3$ mm, but are normalized via trilinear interpolation to $1 \times 1 \times 1$ mm. When computing

our features X, we only consider the first 20 2D slices starting from and including the C3 region and moving inferior, i.e. K=20 in (1) and (2).

3.1 Error Metrics

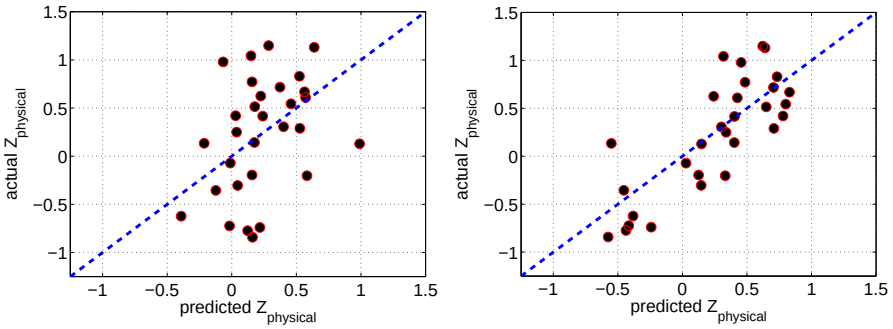
To quantify how closely the predictions $\hat{Y}$ produced by our model are to the true clinical scores Y, we use the following metrics. We compute the *mean absolute error* (MAE) by taking the mean of the absolute difference between the predicted score and the true clinical score, $\text{MAE} = \frac{1}{n}\sum_i^n |\hat{y}_i - y_i|$, giving equal weight to all errors. To get an indication of the variability in the error, we compute the *standard deviation of absolute error* as, $\text{SAE} = \text{std}(|\hat{Y} - Y|)$. To give a higher weight to larger errors, we report the *root mean square error*, $\text{RMSE} = \sqrt{\frac{1}{n}\sum_i^n (\hat{y}_i - y_i)^2}$. MAE, SAE, and RMSE values closer to zero indicate a better model. To indicate the consistency of our predictions, we also compute the *Pearson's correlation coefficient* and its corresponding p-value between the predicted clinical scores $\hat{Y}$ and the true clinical scores Y.

3.2 Simple Linear Regression with Spinal Cord Volume

To establish a baseline test on which we aim to improve, we use a simple linear regression model with spinal cord volume as the explanatory variable similar to what is done by Losseff et al. [8]. We compute the volume of the segmented cord (*vol*) and use leave-out-one cross-validation to train our model and test on the omitted volume. As expected from the existing literature [1, 7, 8, 12, 16], we detect a moderate yet statistically significant correlation between volume and clinical score (*vol*: r=0.473, p=0.00824). The predictive ability for a linear regression model using volume as the explanatory variable is reported in Table 1 (row 1) and shown in Fig. 3a.

3.3 Simple Linear Regression with Proposed Features

In our second test, we examine each proposed feature's ability to act as the explanatory variable in a linear model. For each proposed feature in section 2.2, we compute the Pearson's correlation coefficient between the proposed feature and the clinical scores. We find that $ax_{\text{min}}, per_{\text{mean}}, per_{\text{min}}, dist_{\text{max}}^{\text{mean}}$ all provide a slight increase in correlation when compared to *vol*. Of these features, per_{min} shows the strongest improvement in Pearson's correlation (per_{min}: r=0.565, p=0.00115; vs. volume *vol*: r=0.473, p=0.00824) and the p-value of per_{min} survives the Bonferroni correction for multiple testing ($0.00115 < \frac{0.05}{13}$).



(a) Linear regression using cord volume

(b) Regression forest using two features

Fig. 3: Actual vs. predicted clinical scores are shown. (a) Spinal cord volume *vol* is used as the explanatory variable in a simple linear regression model. (b) A regression forest trained on two selected features, ax_{maj} and per_{min}, demonstrates an improved correlation. Deviations from the *dashed line* are errors.

We test if per_{min} is a stronger explanatory variable than volume by performing the same cross-validation procedure. We report our results in Table 1 (row 2), which demonstrates that not only does per_{min} correlate better than volume, but it gives a more consistent score and is less susceptible to outliers. This is shown by the lower MAE, SAE, and RMSE scores, and higher Pearson's correlation when compared to a model using volume. This suggests that per_{min} may be a better indicator of physical disability than spinal cord volume.

3.4 Multiple Linear Regression with Proposed Features

To explore the use of multiple explanatory variables in a linear regression model, using the 13 candidate features described in section 2.2, we form separate models where each feature can either be included or excluded from the model, for a total of $2^{13} = 8192$ possible combinations. To get a sense of which variables generalize well, we test each model using leave-one-out cross-validation. We correct for multiple testing by applying the positive False Discovery Rate (pFDR) [14] to reduce the likelihood that a positive result is a Type I error. As our goal is to determine if a multiple linear regression model can provide improvements over simple linear regression, we compute how many models result in a RMSE that are less than the RMSE reported using the linear model with the explanatory variable per_{min} (i.e. RMSE < 0.527). There are 292 such models and from this subset of models, we find the maximum p-value to be 0.00684 with a corresponding q-value of 0.000172. Out of all our tests, there are 749 tests with a p-value less than 0.00684, indicating a low number ($749 \times 0.000172 < 1$) of improved models that are potentially false

positives. The features selected from the model with the lowest RMSE are: $best_{7\text{MR}} = \{int_{\text{mn}}, ax_{\text{min}}, per_{\text{mean}}, per_{\text{max}}, per_{\text{min}}, dist_{\text{max}}, dist_{\text{max}}^{\text{mean}}\}$, and the prediction results are reported in Table 1 (row 3). We note that this model with multiple features shows a significant reduction in prediction error when compared to the models using a single explanatory variable.

However, as the issue of how best to correct for multiple testing is still an open one, we further examine our models for a more conservative selection of features. We examine what features were consistently selected in the top 25 models. As can been seen in Fig. 4, the same five features are selected in nearly every model suggesting these features jointly are useful. Based on this trend, we form a linear regression model using only the consistently selected features, $sel_{5\text{MR}} = \{int_{\text{mean}}, per_{\text{mean}}, per_{\text{max}}, dist_{\text{max}}, dist_{\text{max}}^{\text{mean}}\}$, and report the cross-validated results in Table 1 (row 4). While the predictive ability of this model is less than the $best_{7\text{MR}}$ predicting model, this model has two less explanatory variables than the $best_{7\text{MR}}$ model, which may be more generalizable in a novel dataset (even though we cross-validated our dataset). These improvements over the models with a single explanatory variable, suggests that it is useful to combine multiple spinal cord features within a single model.

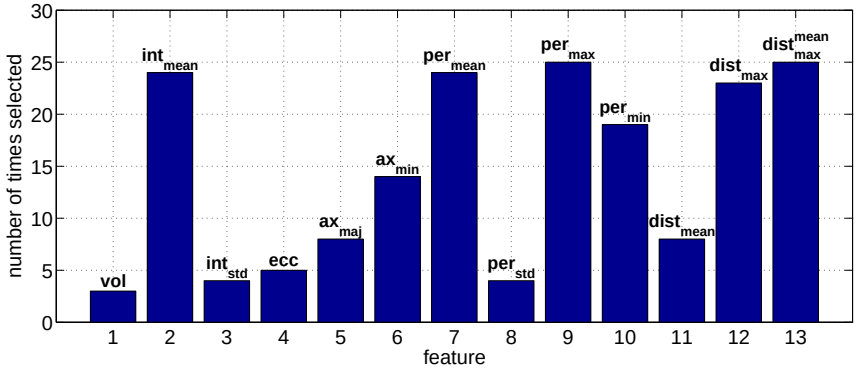


Fig. 4: The number of times a features was selected in the top (lowest RMSE) 25 multiple linear regression models is graphed. The y-axis shows the number of times the feature was selected and the x-axis is the feature selected. We can see that two features were selected in all the top 25 models, $per_{\text{max}}, dist_{\text{max}}^{\text{mean}}$, two were selected in 24 models, $int_{\text{mean}}, per_{\text{mean}}$, and one was selected in 23 models, $dist_{\text{max}}$. These five features are consistently selected which suggests their general importance in forming the model.

3.5 Non-linear Regression Forest with Proposed Features

In our final tests, we use a non-linear regression forest (RF) implemented with MATLAB's TreeBagger class (R2012a; The MathWorks Inc., Natick, MA). The minimum number of observations per leaf is set to one. All other parameters are left to their default settings except for the number of trees which we describe below. To see if a non-linear model, trained on a single feature can outperform a linear model, we train a RF with 250 trees on each proposed feature from section 2.2. Out of our 13 proposed features, we find that per_{min} on its own returns superior results when compared to the other models that use only a single feature, Table 1 (row 5). To consider multiple features in our RF, as was done in section 3.4, we try all possible combinations of features (2^{13}) in a RF. However, to lower computational cost, we use 80 trees with 6-fold (instead of leave-one-out) cross validation when exploring all the feature combinations. We find those features used in the model that produces the lowest RMSE. Correcting for multiple testing using pFRR (sec. 3.4), returns less than 1 expected number of false positives.

Similar to section 3.4, we also examine a more conservative selection of features by choosing those features that are consistently in the 25 models with lowest RMSE. We find that the features used in the lowest RMSE model and the features consistently chosen in the 25 lowest RMSE models *are the same*. These selected features are the ax_{maj} (chosen in 24 out of 25 models) and the per_{min} (chosen in 25 out of 25 models). We train another RF with 250 trees on $sel_{2\text{RF}} = \{ax_{\text{maj}}, per_{\text{min}}\}$ and show leave-one-out cross-validated results that outperform all our previous regression models, reported in Table 1 (row 6) and shown in Fig. 3b. This demonstrates that select novel morphological features, combined in a non-linear, non-parametric regression model can potentially provide more accurate predictions of MS physical disability than a linear model, and outperforms predictions based on spinal cord volume.

4 Conclusion

We proposed new morphological and appearance features to capture the subtle changes in a patient's spinal cord as it undergoes atrophy due to multiple sclerosis. These proposed features were combined in a regression model and our results indicate that they are potentially useful imaging biomarkers for multiple sclerosis. When only considering any one particular feature, the distance from the cord's center-of-mass to the cord's boundary, per_{min}, provided the strongest results and was an improvement over spinal cord volume at clinical prediction.

Our results also suggest that combining the selected features in a regression model improves the predictive ability over a simple linear regression model using any one of the tested features, including volume, alone. As well, a non-linear regression forest, trained on select morphological features, appears to be a promising approach to improve on the predictive ability of linear models. To ensure generaliz-

Table 1: The *model* column contains the different type of models explored where *linear* represents a linear model, *multiple* represents a multiple linear regression model, and *RF* represents a regression forest model. The *features* column contains the different features the model was trained on, where *vol* represents the volume of the spinal cord, per_{min} represents the minimal distance to the cord's center-of-mass from the cord's boundary, *best* represents the combination of features that gives the lowest RMSE error, and *sel* are the features consistently selected in our top 25 models. The error metrics we report are the *Mean Absolute Error* (MAE), the *Root Mean Squared Error* (RMSE), the *Standard deviation of Absolute Error* (SAE), the *Pearson's* correlation coefficient r and its corresponding p-value before correction for multiple comparisons.

model	features	MAE	SAE	RMSE	r	p-value
linear	*vol*	0.448	0.326	0.551	0.367	0.0460841
linear	per_{min}	0.444	0.290	0.527	0.464	0.0097723
multiple	$best_{7\text{MR}}$	0.379	0.253	0.453	0.667	0.00005645
multiple	$sel_{5\text{MR}}$	0.414	0.233	0.473	0.617	0.00028511
RF	per_{min}	0.381	0.251	0.453	0.682	0.00003277
RF	$sel_{2\text{RF}}$	**0.293**	**0.201**	**0.353**	**0.803**	0.00000009

ability of our results (i.e. that the proposed biomarkers and models are not specific to our data and that our findings are not due to a Type I error), even though our data came from multiple centers, future work must involve larger datasets representing a greater variety in imaging, pathological, and clinical parameters.

Acknowledgements.

JK, RT, and GH were partially supported by NSERC and Biogen Idec Canada. CM was supported by the Canadian Breast Cancer Foundation and the Canadian Cancer Society Research Institute.

References

1. Benedetti, B., Rocca, M.A., Rovaris, M., Caputo, D., Zaffaroni, M., Capra, R., Bertolotto, A., Martinelli, V., Comi, G., Filippi, M.: A diffusion tensor MRI study of cervical cord damage in benign and secondary progressive multiple sclerosis patients. Journal of Neurology, Neurosurgery & Psychiatry **81**(1), 26–30 (2010)
2. Criminisi, A., Shotton, J., Konukoglu, E.: Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. Foundations and Trends® in Computer Graphics and Vision **7**(2-3), 81–227 (2011)
3. Fischer, J., Rudick, R., Cutter, G., Reingold, S.: The multiple sclerosis functional composite measure (MSFC): an integrated approach to MS clinical outcome assessment. Multiple Sclerosis **5**(4), 244–250 (1999)

4. Horsfield, M.A., Sala, S., Neema, M., Absinta, M., Bakshi, A., Sormani, M.P., Rocca, M.A., Bakshi, R., Filippi, M.: Rapid semi-automatic segmentation of the spinal cord from magnetic resonance images: Application in multiple sclerosis. Neuroimage **50**(2), 446–455 (2010)
5. Kawahara, J., McIntosh, C., Tam, R., Hamarneh, G.: Globally optimal spinal cord segmentation using a minimal path in high dimensions. In: IEEE ISBI, pp. 836–839 (2013)
6. Kurtzke, J.F.: Rating neurologic impairment in multiple sclerosis an expanded disability status scale (EDSS). Neurology **33**(11), 1444–1452 (1983)
7. Lin, X., Tench, C., Turner, B., Blumhardt, L., Constantinescu, C.: Spinal cord atrophy and disability in multiple sclerosis over four years: application of a reproducible automated technique in monitoring disease progression in a cohort of the interferon β-1a (Rebif) treatment trial. Journal of Neurology, Neurosurgery & Psychiatry **74**(8), 1090–1094 (2003)
8. Losseff, N., Webb, S., O'riordan, J., Page, R., Wang, L., Barker, G., Tofts, P., McDonald, W., Miller, D., Thompson, A.: 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)
9. Mann, R.S., Constantinescu, C.S., Tench, C.R.: Upper cervical spinal cord cross-sectional area in relapsing remitting multiple sclerosis: Application of a new technique for measuring cross-sectional area on magnetic resonance images. J. Magn. Reson. Imaging **26**(1), 61–65 (2007)
10. McIntosh, C., Hamarneh, G.: Spinal crawlers: Deformable organisms for spinal cord segmentation and analysis. In: R. Larsen, M. Nielsen, J. Sporring (eds.) MICCAI 2006, *LNCS*, vol. 4190, pp. 808–815. Springer, Heidelberg (2006)
11. McIntosh, C., Hamarneh, G., Toom, M., Tam, R.: Spinal cord segmentation for volume estimation in healthy and multiple sclerosis subjects using crawlers and minimal paths. In: IEEE HISB, pp. 25–31 (2011)
12. Rocca, M., Horsfield, M., Sala, S., Copetti, M., Valsasina, P., Mesaros, S., Martinelli, V., Caputo, D., Stosic-Opincal, T., Drulovic, J., Comi, G., Filippi, M.: A multicenter assessment of cervical cord atrophy among MS clinical phenotypes. Neurology **76**(24), 2096–2102 (2011)
13. Schnabel, J.A., Wang, L., Arridge, S.R.: Shape description of spinal cord atrophy in patients with MS. Comput Assist Radiol ICS **1124**, 286–291 (1996)
14. Storey, J.D.: A direct approach to false discovery rates. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **64**(3), 479–498 (2002)
15. Tench, C.R., Morgan, P.S., Constantinescu, C.S.: Measurement of cervical spinal cord cross-sectional area by MRI using edge detection and partial volume correction. J. Magn. Reson. Imaging **21**(3), 197–203 (2005)
16. Valsasina, P., Rocca, M.A., Horsfield, M.A., Absinta, M., Messina, R., Caputo, D., Comi, G., Filippi, M.: Regional cervical cord atrophy and disability in multiple sclerosis: A voxel-based analysis. Radiology **266**(3), 853–861 (2013)