NeuroImage: Clinical 24 (2019) 102011

Contents lists available at ScienceDirect

NeuroImage: Clinical

journal homepage: www.elsevier.com/locate/ynicl

SVM recursive feature elimination analyses of structural brain MRI predicts near-term relapses in patients with clinically isolated syndromes suggestive of multiple sclerosis

Viktor Wottschel[a,b,*], Declan T. Chard[b,c], Christian Enzinger[d], Massimo Filippi[e], Jette L. Frederiksen[f], Claudio Gasperini[g], Antonio Giorgio[h], Maria A. Rocca[e], Alex Rovira[i], Nicola De Stefano[h], Mar Tintoré[i], Daniel C. Alexander[j], Frederik Barkhof[a,b,c,k], Olga Ciccarelli[b,c], for the MAGNIMS study group and the EuroPOND consortium

[a] Department of Radiology and Nuclear Medicine, Amsterdam University Medical Centers, Location VUmc, Amsterdam, The Netherlands
[b] Queen Square MS Centre, University College London, London, United Kingdom
[c] National Institute of Health Research (NIHR), University College London Hospitals, Biomedical Research Centre, London, United Kingdom
[d] Research Unit for Neuronal Repair and Plasticity, Department of Neurology, Medical University of Graz, Graz, Austria
[e] Neuroimaging Research Unit, Institute of Experimental Neurology, Division of Neuroscience, San Raffaele Scientific Institute, Vita-Salute San Raffaele University, Milan, Italy
[f] Rigshospitalet-Glostrup and University of Copenhagen, Copenhagen, Denmark
[g] San Camillo-Forlanini Hospital, Rome, Italy
[h] University of Siena, Siena, Italy
[i] Hospital Vall d'Hebron, Barcelona, Spain
[j] Centre for Medical Image Computing, Department of Computer Science, University College London, London, United Kingdom
[k] Institute of Neurology and Healthcare Engineering, University College London, London, United Kingdom

ARTICLE INFO

Keywords:
Multiple sclerosis
Machine learning classification
Feature selection

ABSTRACT

Machine learning classification is an attractive approach to automatically differentiate patients from healthy subjects, and to predict future disease outcomes. A clinically isolated syndrome (CIS) is often the first presentation of multiple sclerosis (MS), but it is difficult at onset to predict who will have a second relapse and hence convert to clinically definite MS. In this study, we thus aimed to distinguish CIS converters from non-converters at onset of a CIS, using recursive feature elimination and weight averaging with support vector machines. We also sought to assess the influence of cohort size and cross-validation methods on the accuracy estimate of the classification.

We retrospectively collected 400 patients with CIS from six European MAGNIMS MS centres. Patients underwent brain MRI at onset of a CIS according to local standard-of-care protocols. The diagnosis of clinically definite MS at one-year follow-up was the standard against which the accuracy of the model was tested. For each patient, we derived MRI-based features, such as grey matter probability, white matter lesion load, cortical thickness, and volume of specific cortical and white matter regions. Features with little contribution to the classification model were removed iteratively through an interleaved sample bootstrapping and feature averaging approach. Classification of CIS outcome at one-year follow-up was performed with 2-fold, 5-fold, 10-fold and leave-one-out cross-validation for each centre cohort independently and in all patients together.

The estimated classification accuracy across centres ranged from 64.9% to 88.1% using 2-fold cross-validation and from 73% to 92.9% using leave-one-out cross-validation. The classification accuracy estimate was higher in single-centre, smaller data sets than in combinations of data sets, being the lowest when all patients were merged together.

Regional MRI features such as WM lesions, grey matter probability in the thalamus and the precuneus or cortical thickness in the cuneus and inferior temporal gyrus predicted the occurrence of a second relapse in patients at onset of a CIS using support vector machines. The increased accuracy estimate of the classification achieved with smaller and single-centre samples may indicate a model bias (overfitting) when data points were

* **Corresponding author at:** Department of Radiology and Nuclear Medicine, Amsterdam University Medical Centers, Location VUmc, Postbus 7057, 1007 MB Amsterdam, the Netherlands.
E-mail address: v.wottschel@amsterdamumc.nl (V. Wottschel).

https://doi.org/10.1016/j.nicl.2019.102011
Received 15 April 2019; Received in revised form 6 September 2019; Accepted 17 September 2019
Available online 22 October 2019

limited, but also more homogeneous. We provide an overview of classifier performance from a range of cross-validation schemes to give insight into the variability across schemes. The proposed recursive feature elimination approach with weight averaging can be used both in single- and multi-centre data sets in order to bridge the gap between group-level comparisons and making predictions for individual patients.

1. Introduction

Multiple sclerosis (MS) is a disease of the central nervous system that is characterised by neuroinflammation, demyelination and neurodegeneration. The first clinical episode of MS is referred to as a clinically isolated syndrome (CIS). A majority of CIS patients (>80%) will eventually develop a second episode over a course of 20 years (Miller et al., 2012), which then defines clinically definite MS (CDMS). A shorter time to conversion from CIS to CDMS is associated with a faster disease progression and higher disability subsequently (Miller et al., 2012). The number of lesions on the MRI scan at onset of CIS is a clinically highly relevant prognostic factor for the development of CDMS and disability (Tintore et al., 2015).

Machine learning offers tools for learning how to distinguish two or more groups based on their features and subsequently assign new, previously unseen, cases to one of the groups. The idea of supervised learning is to identify common characteristics in the individual groups (i.e., patients with a known diagnosis or clinical outcome) that can be generalised to a larger population. This supervised classification has become increasingly popular in neuroimaging over the last decade with a few applications also in MS (Weygandt et al., 2011; Bendfeldt et al., 2012; Wottschel et al., 2015). However, only few studies have been performed on the prediction of conversion to CDMS in CIS patients (Wottschel et al., 2015; Muthuraman et al., 2016; Bendfeldt et al., 2018), and these have often been limited to one centre (Wottschel et al., 2015; Muthuraman et al., 2016).

A common issue is the selection of relevant features to perform a classification. Some studies in MS and Alzheimer's disease have used voxelwise grey-matter (GM) probability (Bendfeldt et al., 2012; Klöppel et al., 2008), which works well when patient groups can be distinguished based on their extent of (regional) brain atrophy. Other studies used hand-picked features that potentially provide predictive information (Wottschel et al., 2015; Bendfeldt et al., 2018). In a previous single-centre study (Wottschel et al., 2015), we showed that support vector machine-based classification predicted clinical outcome in CIS patients with an accuracy score of 71.4% using leave-one-out cross validation. We found that a specific subset of features, mostly related to MS lesions, performed better than individual or all available features. However, as we note in (Wottschel et al., 2015) leave-one-out cross-validation may overestimate classification performance on unseen test data.

Here, we aimed to identify CIS patients developing CDMS within the first year of their symptoms, using data collected in six European centres. We introduce a recursive feature elimination scheme, based on weight averaging with support vector machines, in a large set of imaging measures, including GM probability, cortical thickness, T2 white matter lesion load, and volume of specific GM and white matter (WM) regions. These features can be easily and robustly extracted from MRI scans, and we investigated whether our model automatically identified informative features with respect to the classification task. We examined the influence of the cross-validation partitioning on the estimated classification accuracy by using 2-fold, 5-fold, 10-fold and leave-one-out cross-validation on all data sets to provide an overview of the bias introduced by the different schemes. The model was run in each centre's cohort independently and then in combinations of data sets, including all patient data together, in order compare different levels of heterogeneity in the data.

2. Methods

2.1. Data

This is a retrospective study performed on data obtained by six European centres, which are members of the MAGNIMS (Magnetic Resonance Imaging in Multiple Sclerosis, www.magnims.eu) network (Barcelona/Spain (B), Copenhagen/Denmark (C), Graz/Austria (G), London/UK (L), Milan/Italy (M) and Siena/Italy (S)). The total number of CIS patients included was 400, and 91 (22.8%) of them converted from CIS to CDMS within one year. All baseline scans were performed within 14 weeks (SD 7 weeks) of CIS onset. We do not have information on treatment in this retrospective cohort. A more detailed overview of patient characteristics is given in Table 1.

This project was approved locally by the ethics committees and patient consent was obtained prior to data collection.

The inclusion criteria were as follows: (1) Patients with a CIS were examined within three months from symptoms onset; (2) T1-weighted MRI sequences of the brain were obtained at onset of a CIS, using standard-of-care local protocols; (3) Demographic (age, sex) and clinical information (e.g. type of CIS) at baseline and the presence/absence of a second relapse at one year follow-up was available; (4) presence of T2-hyperintense WM brain lesions as outlined in each centre on PD/T2-weighted or FLAIR MRI by experienced researchers, resulting in binary lesion masks.

2.2. Image processing

Due to the heterogeneity of the MRI data, we used derived measures such as GM probability or cortical thickness (CT) which we believe to be more robust to inter-centre variation compared to direct intensity information. To calculate the features used in the classification

Table 1
Patient demographic and clinical characteristics per each MAGNIMS centre.

	Barcelona	Copenhagen	Graz	London	Milan	Siena	All
No. of patients	175	24	47	72	35	47	400
Age [y]	31.9(16–50)	36.6(24–54)	33.8(21–50)	34.2(19–49)	29.5(20–43)	32.3(20–54)	32.7
Sex	124F/51M	14F/10M	34F/13M	44F/28M	24F/11M	25F/22M	265F/135M
Median EDSS (range)	2 (0–6)	3.75 (1–5)	1 (0–3.5)	1 (0–8)	2 (0–6)	1.5 (0–2)	2 (0-8)
Median global lesion load (range) [mm^3]	1299 (14–25220)	461 (23–3270)	82875 (6630–779726)	849 (29–25581)	1511 (38–19383)	1868 (77–22796)	1597 (14-779726)
Converters to CDMS at 1y follow-up	19.4%	25%	23.4%	30.6%	22.9%	21.3%	22.8%
Type of CIS onset (brainstem / optic nerve / spinal cord / other)	52/45/48/30	0/24/0/0	10/15/10/12	6/62/4/0	10/10/3/12	10/8/17/12	88/164/82/66

experiments, a comprehensive image processing pipeline was created as follows.

1 Bias field correction: all MRI scans were initially corrected for bias field inhomogeneities using the N4 algorithm (Tustison et al., 2010).
2 Lesion filling: WM lesions can have intensities similar to GM on T1-weighted MRI, which can cause problems in registration and segmentation. To reduce this effect, we used a patch-based approach (Prados et al., 2016) to fill the lesion voxels with intensities similar to their neighbourhood.
3 Registration: lesion masks were created from PD/T2- or FLAIR-weighted images whereas most other image processing is performed in T1 space. Therefore, the PD/T2 or FLAIR MRI scans were affinely registered to T1 space using reg_aladin from the NiftyReg toolbox (Modat et al., 2010). Lesion masks were subsequently resampled using the obtained transformation parameters.
4 Brain parcellation: we performed a fine-grained brain parcellation of all T1 scans using the GIF (geodesic information flows) algorithm (Cardoso et al., 2012). This tool segments the brain into 143 ROIs based on the Neuromorphometrics atlas (Klein and Tourville, 2012), of which most are cortical areas as shown in Fig. 1.
5 Merging hemispheres: Measurements from the left and right hemisphere are highly correlated, which is undesirable for machine learning analyses (Bolón-Canedo et al., 2014). Therefore, corresponding contralateral ROI values were averaged in order to reduce the noise in the data and reduce collinearity of features. (Please note that we show results with unmerged contralateral features in the supplementary material section 'Unmerged Hemispheres'.)
6 Grouping: ROIs were merged into nine larger areas according to their anatomical location. Most of these areas correspond to the anatomical brain lobes, and, therefore we refer to all of them as 'lobes' in the context of this study. These 'lobes' were limbic, insular, frontal, parietal, temporal, occipital, cerebellum, GM and WM. Deep grey matter is defined as thalamus, hippocampus, nucleus accumbens, amygdala, caudate nucleus, pallidum, putamen and basal ganglia.
7 Segmentation: In addition to the 143 ROIs, the GIF algorithm also provides a probabilistic segmentation of GM and WM, as well as binary masks of brain tissue and intracranial volume.
8 Cortical thickness: this was calculated using DiReCT, a registration-based algorithm (Das et al., 2009). It has been shown to have the same degree of reproducibility as the more commonly used Freesurfer method (Tustison et al., 2014) but is faster if WM and GM probability maps are already available.
9 ROI masking. We used the ROIs from steps 4 and 6 to calculate local information from GM probability maps, CT maps and lesion masks.

2.3. Feature definitions

Following the image processing, an extensive list of features has been defined on different ROI scales as follows.

1 Global features: these features describe whole-brain measures such as overall GM volume, WM volume and brain volume as a percentage of the intracranial volume. In addition, we added demographic and clinical measures such as age, sex, CIS type and EDSS.
2 ROI features: these features refer to the brain parcellations obtained from GIF (see Section 2.2, point 4). Each ROI from the brain parcellation was used to mask each patient's GM probability map, CT map, lesion segmentation and T1 scan (to estimate the volume). We excluded ROIs describing ventricles, skull and background because they are not expected to be discriminative.
3 Lobe features: we merged ROIs based on their anatomical location into larger coherent regions, which mostly correspond to brain lobes as described above.

Eventually, we concatenated the global features, the ROI features for GM probability, CT, and volume, as well as the lobe features for GM probability, CT, volume and lesion load. ROI lesion load was not used because we only included WM lesions, which are found mostly in only two very large ROIs (WM in left and right hemisphere). Due to misregistration of subjects, features such as 'WM lesion load - dGM' can occur and should be interpreted as WM lesion load on the border of deep GM structures.

This concatenation of all features led to a vector with 213 or 214 entries for each of the 400 subjects depending on the centre. All features were included in the initial models and were subject to the recursive feature elimination approach.

Due to differences in scanning protocols and MRI resolutions, there were centre-specific offsets for some features. Therefore, the feature matrix for each centre was feature-wise transformed to z-scores in order to improve comparability and SVM performance (Juszczak et al., 2005). The transformation centres the data to zero mean with unit variance following $x'=(x-x?\ ))/\sigma_x x' = (x-\bar{x}))/\sigma_x$ where x' is the normalised vector, $\bar{x}$ the mean value of feature vector x, and σ_x the feature's standard deviation.

2.4. Classification model

One aim of the classification was to identify CIS patients who will convert to CDMS based on the previously described features, which were derived from baseline data. The classifier used for this study was a linear SVM, with which we employ a novel iterative feature selection process.

The SVM algorithm assigns a weight to each feature and this weight vector defines the hyperplane (i.e. the multi-dimensional extension of lines and planes) separating the two classes. One interpretation of these weights is as measures of feature strength for informing group membership (Bendfeldt et al., 2012; Klöppel et al., 2008). A common problem, however, is instability of this weight vector across different samples, even from the same data set. While the weights of some features remain relatively similar, others can vary substantially, even alternating between positive and negative signs (i.e. pointing to different classes for the same problem). Such behaviour indicates overfitting to

Fig. 1. Illustration of the Neuromorphometrics atlas used for brain parcellation in this study.

Table 2
Results for single centres using 2-fold cross-validation.

	Accuracy (95% CI) [%]	Sensitivity [%]	Specificity [%]
Individual data set			
Barcelona (B)	64.9 (64.5–65.3)	63.7	66.1
Copenhagen (C)	88.1 (87.4–88.8)	79.4	96.8
Graz (G)	74.3 (73.7–75.0)	63.8	84.9
London (L)	75.8 (75.3–76.3)	74.3	77.3
Milan (M)	88.1 (87.5–88.7)	79.4	96.8
Siena (S)	82.9 (82.3–83.4)	68.9	96.8
Combinations of data sets (first letter of sites)			
BCGLMS	64.8 (64.6–65.1)	64.1	65.6
BLMS	65.2 (64.9–65.4)	64.0	66.4
BLM	66.9 (66.6–67.2)	66.4	67.4

features that offer little or no problem-specific information.

Here, we propose an algorithm to select only informative features and avoid such overfitting. The algorithm runs a SVM on 1000 bootstrap samples of patients and averages the resulting weight vectors to define a mean weight vector descriptive of the whole cohort. By doing this, the weights with alternating signs average to values close to zero, while stable features maintain higher absolute values. The 20% of all included features with average weights closest to zero are subsequently removed and the process is repeated iteratively until the estimated classification accuracy (mean across bootstraps) does not improve further. The choice of 20% maintains accuracy while minimising computation time: smaller percentages increase computation time for the same result due to smaller step sizes, while larger percentages may remove relevant features in early iterations due to the larger step size. Additional example results for percentages of 15% and 25% can be found in the Supplementary Material section 'Variation of feature removal parameters'.

2.5. Class imbalance and patient sampling

Imbalanced class sizes tend to bias the SVM classifier performance towards the majority class. To avoid this, we used down-sampling (also known as undersampling), which is a common approach to avoid class imbalance (Anand et al., 2010). An equal number of subjects to the size of the minority class was selected at random from the majority class. In our study, the minority group was represented by the converters, and the majority class by the non-converters. This approach can potentially introduce a sampling bias, meaning that the random sample is not representative of the whole class. We mitigate this problem by repeating the process 1000 times with different majority class samples so the whole cohort will be represented overall.

The main measure of classifier performance in this study was accuracy, which is the proportion of correctly classified cases (i.e., converters and non-converters) relative to the total cohort size. The 95% confidence interval with respect to the 1000 repetitions was reported. Additionally, the mean sensitivity and specificity of the classifier (where converters are defined as positive samples and non-converters as negative) were also reported.

2.6. Cross-validation

Cross-validation is an important tool in machine learning for testing generalisability of a classifier. In *k*-fold CV, the data is split into *k* parts so that *k*-1 parts are used for training and one part for testing. A separate classifier is trained on each of the *k* training sets and evaluated on the corresponding test set. Typically average performance metrics over all *k* folds are reported together with range of variation (Geisser, 1993). 10-fold CV is sometimes suggested as a compromise between bias and training sample size (Kohavi, 1995).

It is important to note, however, that the accuracy estimates arising from the different CV approaches are only indications of classifier performance with different levels of bias from training set size and classifier correlations (Arlot and Celisse, 2010). The real accuracy can only be estimated with two sufficiently large independent data sets for training and testing. We refer to the cross-validation results as *accuracy estimates* throughout this manuscript for this reason.

In this study, we performed a variety of experiments to show the effect of sample size and cross-validation partitioning in the proposed classification pipeline using multi-centre data. Our goal was to show a) that the classifier is able to identify relevant features to differentiate the two groups, b) the effect of data set size and composition, and c) variability in accuracy estimates arising from the choice of cross-validation scheme. To do this, we used data from six individual centres with varying number of patients (see Table 1) as well as multi-centre combinations of these six centres, including a combination of all patients together. In order to explore how the classifier performance changes with varying cross-validation schemes, we ran 2-fold, 5-fold, 10-fold and LOO cross-validation on all data sets where it was possible (i.e. centres Copenhagen and Milan had less than 10 converters so that a stratified 10-fold CV was not feasible). Similarly, we ensured centre-stratification for the multi-centre experiments so that the settings involving centre C or M could not use 10-fold CV.

Our multi-centre experiments focus on the combinations with more heterogeneous imaging protocols (B, L, M and S) but we also explore the combination of all centres together. An overview of all experiments is given in Supplemental Table 1.

3. Results

We performed SVM classifications to predict the occurrence of a second clinical episode in patients with a CIS suggestive of MS using a large cohort of 400 patients studied in a multi-centre setting and used an iterative RFE feature selection approach that removed the least-contributing 20% of features at each iteration. There were differences in estimated classification accuracy between the individual, centre-specific data sets as well as between the different cross-validation schemes (Tables 2–5). As expected (Kohavi, 1995), the classification accuracy score was higher when using higher fold cross-validations than when using lower fold methods, and was the highest with the leave-one-out method. In particular, the mean accuracy estimates ranged from 64.9% to 88.1% across individual centres when using a 2-fold cross-validation, and from 73% to 92.9% when using a leave-one-out scheme. The classification accuracy estimate was higher with smaller data sets than with larger data sets, which might indicate overfitting or a spurious selection bias. Multi-centre data sets lead to the lowest accuracy estimates, which is likely due to the heterogeneity in the data. When combining all centres' data to only one multi-centre data set, we obtained accuracy estimates between 64.8% for 2-fold cross-validation and 70.8% for the leave-one-out method.

3.1. Recursive feature elimination

The proposed recursive feature elimination approach led to an initial increase in accuracy score from early iterations of the procedure, when features were removed that are not relevant to the classification task. However, the predictive power of the model reduced in later iterations, once relevant features were eliminated and only a low number of features were left in the model (Fig. 2). The trajectory of accuracy estimates was similar across all cross-validation schemes and data sets. This behaviour is illustrated in the Fig. 2 for the whole multi-centre data set.

3.2. Cross-validation

The accuracy estimates were very similar for all cross-validation schemes when using all features (Fig. 2). With a reduction of features, however, the difference in accuracy estimates among the cross-

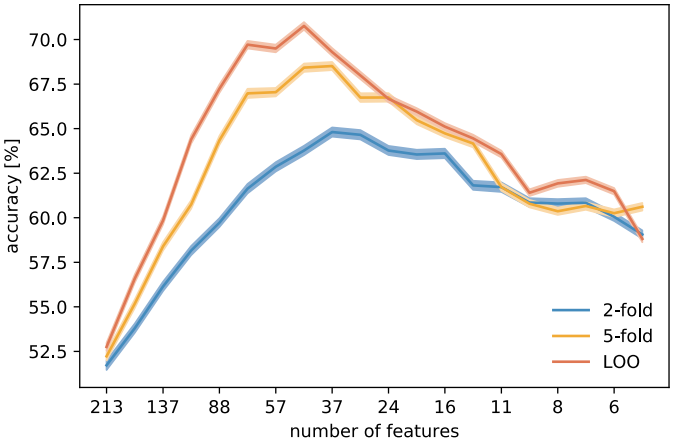

Fig. 2. Accuracy estimates achieved at different iterations of the recursive feature selection when using all centres' data sets combined together (BCGLMS)). The accuracy estimates increase with the first steps of the RFE, and the accuracy estimates generally increase with the number of folds. The shaded areas indicate 95% confidence intervals over 1000 bootstraps.

validation approaches increased (Fig. 2). While the difference between 2-fold and leave-one-out method is only 2% when using all features, it rose to 9.2% when using the selected feature sets, that maximised the accuracy estimates. An overview of the performance estimates for all data sets and all cross-validation schemes is given in Fig. 3 and Tables 2–5.

3.3. Class size

The data sets from the individual centres differed in size and ranged from 24 CIS patients with 6 converters in the Copenhagen data set, to 175 patients with 34 converters in the Barcelona cohort (Table 1). In total, there were 400 patients of which 91 converted to CDMS within one year of follow-up (see top of Fig. 4). Because SVMs are susceptible to class imbalance and tend to introduce a bias towards the majority class, we down-sampled the majority class (i.e. the non-converters) to match the size of the minority class.

The accuracy estimates increased with decreasing class size (see bottom of Fig. 4). In single-centre data sets, a size of 34 patients per group leads to an accuracy estimate of 64.9% using 2-fold cross-validation and 73.0% using leave-one-out scheme, while the smallest class size of 6 led to an accuracy estimate of 88.1% and 91.9% for 2-fold and leave-one-out validation, respectively (Fig. 4). In multi-centre data sets, there was a small increase from 64.8% with 91 patients per class to 66.9% with 64 patients when using 2-fold cross-validation. Similarly, the accuracy estimate was 70.8% with 91 patients and 73.3% with 64 patients when using the leave-one-out method (Fig. 4).

3.4. Most relevant features

The recursive feature elimination algorithm selected features from all domains, but the exact composition of feature sets at peak accuracy score differed slightly between the data sets. When using all data with a conservative 2-fold cross-validation the following features had the highest absolute weights at peak accuracy: (i) White matter lesion load in the whole brain, WM, deep GM, and the frontal, temporal and limbic lobes, (ii) GM probability features in the cerebellum, deep GM regions (such as the thalamus), and across the cortex, especially in the occipital and temporal lobes; (iii) CT of the occipital, frontal and temporal lobes; (iv) Whole brain volume and volumes of the limbic lobe, middle temporal gyrus and supramarginal gyrus. The type of CIS was selected as the only non-imaging feature relevant to the classification.

An illustration of the non-lesional imaging features is given in Fig. 5. A complete list of selected features, as well as all candidate features, for this experiment is given in the Supplementary Material. The final feature sets were not identical between experiments but we observed large overlap suggesting consistency and inherently meaningful feature selection.

4. Discussion

Our proposed recursive feature elimination approach and weight averaging, with support vector machines, classifies the progression from CIS to CDMS. The estimated accuracy for this task ranged between 64.8% and 70.8% over the cross-validation schemes in a multi-centre setting including all patient data, and between 64.9% and 92.9% in single-centre data sets. However, there were large differences between the individual centres, and between the applied cross-validation schemes. In a previous study we used a small set of 12 'hand-picked' features associated with MS progression in order to predict the conversion from CIS to CDMS with an accuracy estimate of 71.4% when using support vector machines and LOO-CV (Wottschel et al., 2015). However, it remained unclear if the initially selected features and cross-validation setting were optimal, and the experiments were performed on a single-centre data set. Here, we have extended this approach to show that a) there were differences between centres and data set sizes, b) features can be selected in a more automated way and c) that the cross-validation scheme had a strong influence on the outcome.

4.1. Recursive feature elimination

The classifier did not perform well when all 214 features were used for classification, but performance improved subsequently with each iteration until a locally *optimal* number of features was reached. Once the classifier started removing features crucial to the classification, the accuracy score dropped again. This is in line with previous studies where a certain subset of features performed better than single features or all features together (Wottschel et al., 2015). The interleaved weight-averaging across bootstraps to select redundant features for removal is a novel and viable option to identify relevant markers of disease progression.

4.2. Class size

The accuracy estimate was generally higher in data sets with fewer samples, so that the highest accuracy score was achieved in the smallest centre's cohort and the lowest accuracy score in the largest cohort. This could be explained by the fact that small samples represent less disease variability and are therefore easier for a classifier to learn but also by

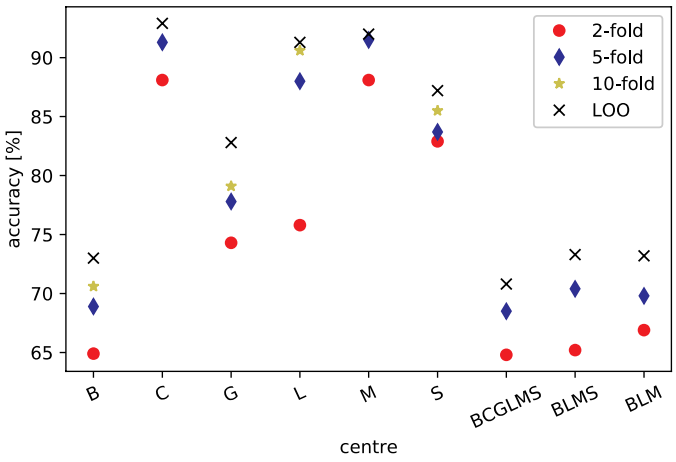

Fig. 3. Accuracy estimates per centre or combination of centres for each cross-validation method. Corresponding values for confidence intervals, sensitivity and specificity can be found in Tables 2–5.

Table 3
Results for single centres using 5-fold cross-validation.

	Accuracy (95% CI) [%]	Sensitivity [%]	Specificity [%]
Individual data set			
Barcelona (B)	68.9 (68.5–69.3)	66.9	70.9
Copenhagen (C)	91.3 (90.8–91.8)	83.3	99.4
Graz (G)	77.8 (77.3–78.4)	63.1	92.6
London (L)	88.0 (87.7–88.3)	87.4	88.6
Milan (M)	91.5 (91.1–91.9)	84.0	99.1
Siena (S)	83.7 (83.2–84.2)	70.2	97.2
Combinations of data sets (first letter of sites)			
BCGLMS	68.5 (68.3–68.7)	67.8	69.3
BLMS	70.4 (70.1–70.6)	69.9	70.85
BLM	69.8 (69.5–70.1)	69.7	69.9

Table 4
Results for single centres using 10-fold cross-validation (note that centres C and M had less than 10 converters and thus could not be used for 10-fold CV).

	Accuracy (95% CI) [%]	Sensitivity [%]	Specificity [%]
Individual data set			
Barcelona (B)	70.6 (70.2–71.0)	70.0	71.2
Copenhagen (C)	NA	NA	NA
Graz (G)	79.1 (78.5–79.7)	67.9	90.3
London (L)	90.6 (90.3–90.9)	88.9	92.2
Milan (M)	NA	NA	NA
Siena (S)	85.5 (85.0–86.0)	73.0	98.1
Combinations of data sets (first letter of sites)			
BCGLMS	NA	NA	NA
BLMS	NA	NA	NA
BLM	NA	NA	NA

Table 5
Results for single centres using leave-one-out cross-validation.

	Accuracy (95% CI) [%]	Sensitivity [%]	Specificity [%]
Individual data set			
Barcelona (B)	73.0 (72.6–73.3)	72.6	73.3
Copenhagen (C)	92.9 (92.4–93.3)	86.0	99.7
Graz (G)	82.8 (82.2–83.3)	71.2	94.2
London (L)	91.3 (91.0–91.6)	91.6	91.0
Milan (M)	92.0 (91.6–92.4)	84.4	99.6
Siena (S)	87.2 (86.7–87.7)	75.8	98.6
Combinations of data sets (first letter of sites)			
BCGLMS	70.8 (70.6–71.0)	70.3	71.3
BLMS	73.3 (73.0–73.5)	73.2	73.3
BLM	73.2 (72.9–73.4)	73.5	72.8

cross-validation bias exacerbated by the small sample size. In addition to this, it is more likely to observe spurious correlations between small data sets and large features sets. It must be noted, however, that this lack of variability led to an overfitted model that works well for the data set in question, but cannot be generalised to a larger population. The overfitting found in the smaller single-centre data sets was not observed in the multi-centre setting due to an increase in sample size, and therefore also variability.

4.3. Cross-validation scheme

It is well-known that the choice of cross-validation method has an influence on the estimated classification accuracy, so we report statistics from multiple schemes in order to mitigate potential bias arising from correlation between classifiers (Kohavi, 1995). However, many studies use the leave-one-out scheme arguing that it is more suitable for small data sets because more data can be used for training and that it mimics clinical practice where one can learn from large data sets and then apply the findings to new individual cases (Bendfeldt et al., 2012; Wottschel et al., 2015). Here, we performed a direct comparison of

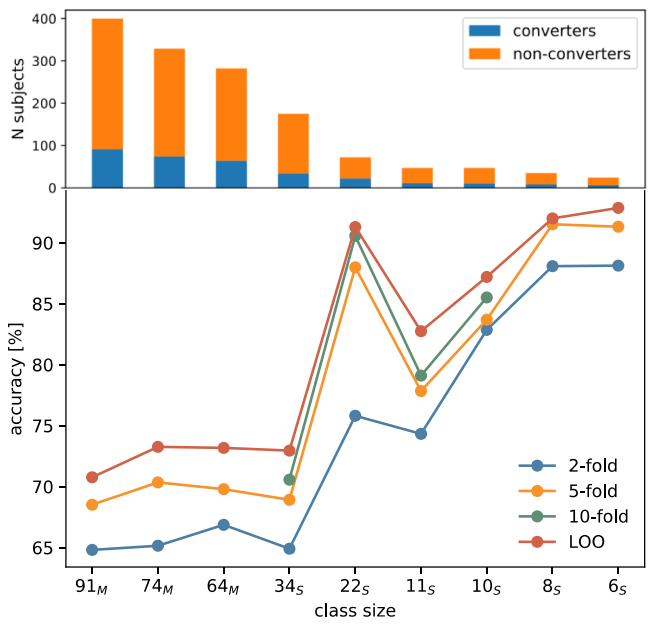

Fig. 4. Top: bar chart of the proportion of converters in the cohort. Bottom: estimated classification accuracy relative to the size of the minority class. There is a general increase of estimated accuracy with a decrease in sample size. The subscript M and S indicate multi-centre and single-centre data sets respectively.

different cross-validation schemes on multiple data sets and showed clearly that there was a difference of up to 20% in estimated classification accuracy between 2-fold and leave-one-out cross-validation. Even though this difference is somewhat artificially high in our experiments, the effect is consistent between data sets and suggests that estimates from experiments with a high number of folds are more inflated than those with a lower number of folds.

The choice of cross-validation partitioning also has a direct effect on the portion of data that is used for training, such that in 10-fold cross-validation 90% of the data is used for training, but only 50% is used in a 2-fold method. A smaller amount of training data leads to a worse and less generalisable model, which is something that may have happened also in our experiments. The pattern of accuracy score change was similar between cross-validation schemes in all data sets independent of size, suggesting that we were observing a fold-size effect rather than a training-size effect – even in data sets with a larger absolute number of subjects per class the differences between the cross-validation schemes are striking. For future studies, we suggest to compare two or more cross-validation schemes to estimate potential biases when it is not possible to use completely distinct data sets for training and testing.

4.4. Most relevant features

The classification in the multicentre setting using all data seems to be strongly driven by the presence of white matter lesions in the whole brain, WM, deep GM, and the frontal, temporal and limbic lobes (see complete list in the Supplementary Materials). Current literature supports these findings as white matter lesion load in different regions is predictive of disease progression in MS (Popescu et al., 2013; Kearney et al., 2015; Filippi, 2001). Additionally, important features were those related to GM probability and cortical measures. These findings extend previous studies, which reported that surrogate measures of atrophy, such as GM probability derived in deep GM regions, like the thalamus, predict cognitive impairment (Batista et al., 2012) and clinical disability in MS (Eshaghi et al., 2018). Similarly, GM probability of the occipital lobe, but also in other parts of the cortex, was associated with the rate of progression to CDMS in CIS (Calabrese et al., 2011). Specific cortical ROIs associated with CT and

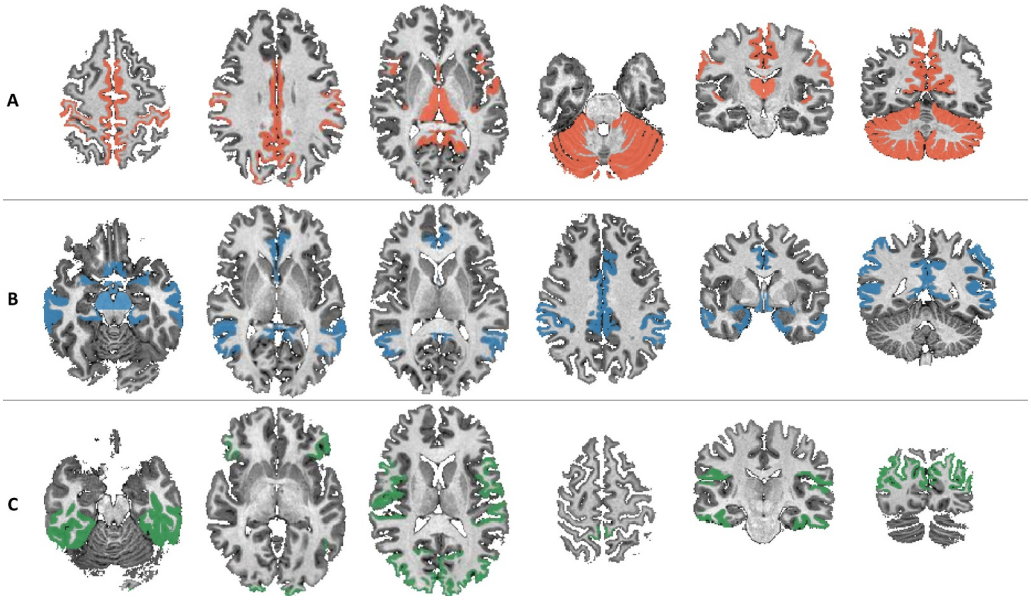

Fig. 5. Location of features relevant to the prediction of CIS conversion at 1-year follow-up. The highlighted areas represent A: GM probability, B: regional volume sizes and C: cortical thickness respectively. Please note that white matter lesion load across the whole brain was also selected but is not shown here for clarity. Type of CIS onset was selected as the only non-imaging feature. A full list of features can be found in the supplementary material.

regional volumes, being additional surrogate measures of atrophy, were also selected (Calabrese et al., 2011). The type of CIS was selected as the only non-imaging feature, which is in line with group-level analyses that showed that a CIS with involvement of the optic nerve (i.e., optic-neuritis) has a better prognosis compared to initial lesions in the spinal cord (Miller et al., 2012, 2005). Overall, the features selected by our proposed approach are well supported by existing literature where the same or similar types of biomarkers have been associated with disease progression in MS. This study, however, allows for combining these features to make predictions of future clinical outcome in individual subjects.

4.5. Limitations

For this study, we aimed to use a broad range of features that can be derived from structural MRI scans. However, the classification performance could be improved by information from advanced MRI techniques, such as magnetisation transfer imaging (MTR) (Audoin et al., 2010) or double or phase-shifted inversion recovery (DIR/PSIR) (Filippi et al., 2010), which have been shown to express damage outside of WM lesions and GM lesions respectively. Similarly, a large range of non-imaging markers such as genetic (Kelly et al., 1993) or environmental factors (Ebers, 2008) could potentially be very informative in such a study where individual subjects' prognoses are being made. Furthermore, a comparative study using healthy controls and patients with MS with the same features would be desirable. Here, however, we analysed data retrospectively and did not have any of this extra information available. Future work which includes prospective, harmonised imaging protocols, demographic, environmental and genetic factors, and all the other variables that define MS at an individual level, may improve the prediction accuracy of the classifier.

Furthermore, the features included in this retrospective study were by no means a complete set of all possible features that can be derived from MRI scans. Other machine learning studies included information, such as lesion size and shape, for prediction of CIS conversion (Bendfeldt et al., 2018), which was not done here because we limited this study to measures that are more easily obtainable through standard pipelines.

The recursive feature elimination approach is a powerful method to identify relevant features, but it does not guarantee the globally optimal solution, as described in previous studies (Wottschel et al., 2015). This issue is increased here due to the step size of 20% of all available features, which are removed at each iteration. It is possible that a different step size would have led to higher accuracy score values, but a too high percentage would make it more likely to accidentally remove relevant features, whereas a too low percentage would increase computation time and might introduce a significant multiple comparisons problem. There is no strong difference in accuracy estimates when the step size was varied between 15% and 25%, so that 20% was selected as a compromise between computation time and potential loss of valuable features (see also Supplementary Material).

The study used retrospectively selected cross-sectional data that was used to derive regional measures such as regional GM probability, cortical thickness and normalised volume that can be considered surrogate measures for atrophy. Due to the lack of longitudinal MRI follow-up, however, it cannot be confirmed that atrophy is driving the models' predictions because the differences in volume could also be due to normal physiological variability. Future work should investigate this in a large cohort with one or more radiological follow-ups.

5. Conclusion

We have presented a new approach for predicting the near-term conversion from CIS to CDMS within a one-year follow-up. The overview of accuracy estimates from different cross-validation settings revealed a strong influence of the selected scheme and its potential bias on the reported accuracy. Similarly, we showed that small data sets seemed to 'over-perform', which indicates overfitting problems when classifiers did not have a sufficient number of samples to learn and generalise from. Therefore, future neuroimaging studies using machine learning classification need to ensure that data sets are large enough for the classifier to pick up meaningful patterns, and to compare outcomes from multiple cross-validation settings in order to obtain meaningful accuracy estimates.

The proposed recursive feature elimination approach with weight averaging can be used both in single- and multi-centre data sets in order to bridge the gap between group-level comparisons and predicting outcomes for individual patients. It could also be used for automated biomarker selection in various applications as it is not limited to the types of features in this study but could in fact use any sort of information such as genetic or neuropsychological data.

Declaration of Competing Interest

The authors declare no potential conflicts of interest with respect to the research, authorship, or publication of this article.

Acknowledgments

This project received funding from the European Union's Horizon 2020 Research and Innovation Program EuroPOND under grant agreement number 666992, and it was supported by the National Institute for Health Research University College London Hospitals Biomedical Research Centre. We thank all participating partners of the MAGNIMS study group for sharing their data with us.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.nicl.2019.102011.

References

Miller, DH, Chard, DT, Ciccarelli, O, 2012. Clinically isolated syndromes. Lancet Neurol. 11 (2), 157–169. https://doi.org/10.1016/S1474-4422(11)70274-5.

Tintore, M, À, Rovira, Río, J, et al., 2015. Defining high, medium and low impact prognostic factors for developing multiple sclerosis. Brain 138 (7), 1863–1874. https://doi.org/10.1093/brain/awv105.

Weygandt, M, Hackmack, K, Pfüller, C, et al., 2011. MRI pattern recognition in multiple sclerosis normal-appearing brain areas. Kleinschnitz C, ed. PLoS One 6 (6), e21138. https://doi.org/10.1371/journal.pone.0021138.

Bendfeldt, K, Klöppel, S, Nichols, TE, et al., 2012. Multivariate pattern classification of gray matter pathology in multiple sclerosis. Neuroimage 60 (1), 400–408. https://doi.org/10.1016/j.neuroimage.2011.12.070.

Wottschel, V, Alexander, DC, Kwok, PP, et al., 2015. Predicting outcome in clinically isolated syndrome using machine learning. NeuroImage Clin. 7, 281–287. https://doi.org/10.1016/j.nicl.2014.11.021.

Muthuraman, M, Fleischer, V, Kolber, P, Luessi, F, Zipp, F, Groppa, S, 2016. Structural brain network characteristics can differentiate CIS from early RRMS. Front. Neurosci. 10, 14. https://doi.org/10.3389/fnins.2016.00014.

Bendfeldt, K, Taschler, B, Gaetano, L, et al., 2018. MRI-based prediction of conversion from clinically isolated syndrome to clinically definite multiple sclerosis using SVM and lesion geometry. Brain Imaging Behav. 1–14. https://doi.org/10.1007/s11682-018-9942-9.

Klöppel, S, Stonnington, CM, Chu, C, et al., 2008. Automatic classification of MR scans in Alzheimer's disease. Brain 131 (Pt 3), 681–689. https://doi.org/10.1093/brain/awm319.

Tustison, NJ, Avants, BB, Cook, PA, et al., 2010. N4ITK: improved N3 bias correction. IEEE Trans. Med. Imaging 29 (6), 1310–1320. https://doi.org/10.1109/TMI.2010.2046908.

Prados, F, Cardoso, MJ, Kanber, B, et al., 2016. A multi-time-point modality-agnostic patch-based method for lesion filling in multiple sclerosis. Neuroimage 139, 376–384. https://doi.org/10.1016/j.neuroimage.2016.06.053.

Modat, M, Ridgway, GR, Taylor, ZA, et al., 2010. Fast free-form deformation using graphics processing units. Comput. Methods Programs Biomed. 98 (3), 278–284. http://www.sciencedirect.com/science/article/pii/S0169260709002533 Accessed December 3, 2013.

Cardoso, M, Wolz, R, Modat, M, 2012. Geodesic information flow. Image Comput. http://link.springer.com/chapter/10.1007/978-3-642-33418-4_33 Accessed April 30, 2015.

Klein, A, Tourville, J., 2012. 101 labeled brain images and a consistent human cortical labeling protocol. Front. Neurosci. 6, 171. https://doi.org/10.3389/fnins.2012.0017.

Bolón-Canedo, V, Sánchez-Maroño, N, Alonso-Betanzos, A, Benítez, JM, Herrera, F, 2014. A review of microarray datasets and applied feature selection methods. Inf. Sci. 282, 111–135. https://doi.org/10.1016/J.INS.2014.05.042.

Das, SR, Avants, BB, Grossman, M, Gee, JC, 2009. Registration based cortical thickness measurement. Neuroimage 45 (3), 867–879. https://doi.org/10.1016/j.neuroimage.2008.12.016.

Tustison, NJ, Cook, PA, Klein, A, et al., 2014. Large-scale evaluation of ANTs and FreeSurfer cortical thickness measurements. Neuroimage 99, 166–179. https://doi.org/10.1016/j.neuroimage.2014.05.044.

Geisser, S., 1993. Predictive Inference. Chapman & Hall, New York, pp. 191ff.

Juszczak, P, Tax, DMJ, Duin, RPW, 2005. Feature scaling in support vector data description. In: Proceedings of the Eighth Annual Conference on Advanced School for Computing and Imaging. 95. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.100.2524 Accessed August 22, 2016.

Anand, A, Pugalenthi, G, Fogel, GB, Suganthan, PN, 2010. An approach for classification of highly imbalanced data using weighting and undersampling. Amino Acids 39 (5), 1385–1391. https://doi.org/10.1007/s00726-010-0595-2.

Kohavi, R., 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence. 2. USA: Morgan Kaufmann Publishers Inc., San Francisco, CA, pp. 1137–1143. IJCAI'95. http://dl.acm.org/citation.cfm?id=1643031.1643047.

Arlot, S, Celisse, A., 2010. A survey of cross-validation procedures for model selection. Stat. Surv. 4 (0), 40–79. https://doi.org/10.1214/09-SS054.

Popescu, V, Agosta, F, Hulst, HE, et al., 2013. Brain atrophy and lesion load predict long term disability in multiple sclerosis. J. Neurol. Neurosurg. Psychiatry 84 (10), 1082–1091. https://doi.org/10.1136/jnnp-2012-304094.

Kearney, H, Altmann, DR, Samson, RS, et al., 2015. Cervical cord lesion load is associated with disability independently from atrophy in MS. Neurology 84 (4), 367–373. https://doi.org/10.1212/WNL.0000000000001186.

Filippi, M., 2001. Magnetic resonance imaging findings predicting subsequent disease course in patients at presentation with clinically isolated syndromes suggestive of multiple sclerosis. Neurol. Sci. 22 (8), S49–S51. https://doi.org/10.1007/s100720100033.

Batista, S, Zivadinov, R, Hoogs, M, et al., 2012. Basal ganglia, thalamus and neocortical atrophy predicting slowed cognitive processing in multiple sclerosis. J. Neurol. 259 (1), 139–146. https://doi.org/10.1007/s00415-011-6147-1.

Eshaghi, A, Prados, F, Brownlee, WJ, et al., 2018. Deep gray matter volume loss drives disability worsening in multiple sclerosis. Ann. Neurol. 83 (2), 210–222. https://doi.org/10.1002/ana.25145.

Calabrese, M, Rinaldi, F, Mattisi, I, et al., 2011. The predictive value of gray matter atrophy in clinically isolated syndromes. Neurology 77 (3), 257–263. https://doi.org/10.1212/WNL.0b013e318220abd4.

Miller, D, Barkhof, F, Montalban, X, Thompson, A, Filippi, M, 2005. Clinically isolated syndromes suggestive of multiple sclerosis, part I: natural history, pathogenesis, diagnosis, and prognosis. Lancet Neurol. 4 (5), 281–288. https://doi.org/10.1016/S1474-4422(05)70071-5.

Audoin, B, Zaaraoui, W, Reuter, F, et al., 2010. Atrophy mainly affects the limbic system and the deep grey matter at the first stage of multiple sclerosis. J. Neurol. Neurosurg. Psychiatry 81 (6), 690–695. https://doi.org/10.1136/jnnp.2009.188748.

Filippi, M, Rocca, MA, Calabrese, M, et al., 2010. Intracortical lesions: Relevance for new MRI diagnostic criteria for multiple sclerosis. Neurology 75 (22), 1988–1994. https://doi.org/10.1212/WNL.0b013e3181ff96f6.

Kelly, MA, Cavan, DA, Penny, MA, et al., 1993. The influence of HLA-DR and -DQ alleles on progression to multiple sclerosis following a clinically isolated syndrome. Hum. Immunol. 37 (3), 185–191. https://doi.org/10.1016/0198-8859(93)90184-3.

Ebers, GC, 2008. Environmental factors and multiple sclerosis. Lancet Neurol. 7 (3), 268–277. https://doi.org/10.1016/S1474-4422(08)70042-5.