UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

# Predicting multiple sclerosis conversion in CIS patients: A pilot study combining MRI-derived measures and clinical data through a machine learning approach

**A Degree Thesis**
**Submitted to the Faculty of the**
**Escola Tècnica d'Enginyeria de Telecomunicació de Barcelona**
**Universitat Politècnica de Catalunya**
**by**
**Pau Vidal Bordoy**

**In partial fulfilment**
**of the requirements for the degree in**
**Telematics Engineering**

**Advisors: Verónica Vilaplana Besler and**
**Deborah Pareto Onghena**

**Barcelona, May 2018**

## Abstract

The objective of this project is to investigate the feasibility of machine learning algorithms in predicting the conversion to multiple sclerosis (MS) of patients with a clinically isolated syndrome (CIS), using MRI-derived features. Many CIS patients only present a mild degree of affectation, making the radiological diagnostic not easy. MRI-derived features included in this project are related with brain atrophy and with the inflammatory component through brain lesion volume and topography (density of lesions in each brain lobe). These features have been used together with clinico-demographic data. Several experiments have been performed using three different definitions of conversion to MS. The results vary according to the definition used, with balanced accuracies of 0.68, 0.61 and 0.73 (recall = [0.51, 0.60, 0.68], specificity = [0.85, 0.62, 0.78]). The proposed approach does not overcome current radiological diagnosis, but further investigation including more MRI-derived features is needed.

## Resum

L'objectiu d'aquest projecte és investigar la factibilitat d'utilitzar algoritmes d'aprenentatge automàtic per predir la conversió a Esclerosi Múltiple (EM) de pacients amb síndrome clínic aïllat (CIS), utilitzant característiques derivades d'imatges de ressonància magnètica (RM). Molts pacients amb CIS presenten un nul o lleu grau d'afectació, fent que el diagnòstic radiològic no sigui fàcil. Les característiques extretes de les RM estan relacionades amb l'atròfia cerebral i amb la component inflamatòria a través del volum de lesió cerebral i de la densitat de lesions a cada lòbul (topografia). Aquestes característiques es combinen amb dades demogràfiques i clíniques dels pacients. S'han realitzat experiments utilitzant tres definicions diferents de conversió a EM. Els resultats varien per cada definició, amb precisions equilibrades de 0.68, 0.61 i 0.73 (sensibilitat=[0.51, 0.60, 0.68], especificitat=[0.85, 0.62, 0.78]). La solució proposada no supera la precisió del mètode actual de diagnòstic de EM, però es requereix més investigació.

## Resumen

El objetivo de este proyecto es investigar la factibilidad de utilizar algoritmos de aprendizaje automático para predecir la conversión a Esclerosis Múltiple (EM) de pacientes con síndrome clínico aislado (CIS), utilizando características derivadas de imágenes de resonancia magnética (RM). Muchos pacientes con CIS presentan un nulo o leve grado de afectación, haciendo que el diagnóstico radiológico no sea fácil. Las características extraídas de las RMs están relacionadas con la atrofia cerebral y con la componente inflamatoria a través del volumen de lesión cerebral i de la densidad de lesiones en cada lóbulo cerebral (topografía). Estas características se combinan con datos clínicos y demográficos de los pacientes. Se han realizado experimentos utilizando tres definiciones diferentes de conversión a EM. Los resultados varían para cada definición, con precisiones balanceadas de 0.68, 0.61 y 0.73 (sensibilidad=[0.51,0.60, 0.68], especificidad =[0.85, 0.62, 0.78]). La solución propuesta no supera la precisión del método actual de diagnóstico de EM, pero se requiere más investigación.

## Acknowledgements

First of all, I want to thank Verónica Vilaplana for giving me the opportunity to work on this project. I would like to show my gratitude to her and Deborah Pareto for the advice and support they have given me throughout the project. Their deep knowledge in their respective fields has been very helpful and has allowed me to learn a lot.

I also want to thank the MRI Unit of Vall d'Hebron University Hospital for the help provided for the creation of the database.

I am very grateful to have been given the opportunity of working in a real world problem, where I have been able to apply the skills I have acquired during my studies.

Finally, I want to thank my family for the support they have given me during these years.

# Revision history and approval record

| Revision | Date | Purpose |
|---|---|---|
| 0 | 10/04/2018 | Document creation |
| 1 | 23/04/2018 | Document revision |
| 2 | 03/05/2018 | Document revision |
| 3 | 07/05/2018 | Document revision |
| 4 | 09/05/2018 | Document revision |
| 5 | 11/05/2018 | Document revision |

DOCUMENT DISTRIBUTION LIST

| Name | e-mail |
|---|---|
| Pau Vidal Bordoy | pvidalbordoy@gmail.com |
| Verónica Vilaplana Besler | veronica.vilaplan@upc.edu |
| Deborah Pareto Onghena | deborah.pareto@idi.gencat.cat |

| Written by: | | Reviewed and approved by: | |
|---|---|---|---|
| Date | 11/05/2018 | Date | 11/05/2018 |
| Name | Pau Vidal Bordoy | Name | Verónica Vilaplana Besler<br>Deborah Pareto Onghena |
| Position | Project Author | Position | Project Supervisors |

# Table of contents

# List of Figures

## List of Tables:

# 1. Introduction

This project was proposed and has been supervised by Deborah Pareto from the MRI Unit of Vall d'Hebron University Hospital. It has been co-supervised by Veronica Vilaplana from the Image and Video Processing Group (GPI) from the Signal Theory and Communications Department (TSC) of the Technical University of Catalonia (UPC).

## 1.1. Multiple sclerosis

Multiple sclerosis is a chronic, autoimmune, demyelinating and inflammatory neurodegenerative disease. It is one of the most frequent causes of non-traumatic neurological disability, affecting almost 2.5 million people worldwide. The disease affects more women than men, with a ratio of 3 to 1, affecting one in every thousand women and one in every three thousand men.

There are treatments being used to reduce the sequels after an attack and to prevent the progression of the disease, but there is no known cure for it.

### 1.1.1. Causes

The cause of the multiple sclerosis (MS) is unknown, although evidence shows that it has some relation with genetic and environmental factors.

MS is not considered a hereditary disease but the genetic factors have an important role. The probability of having MS for the general population is approximately 0.2%, the probability for first-degree relatives of a MS patient is 3%, the probability for a non-identical twin is 5% and 30% for an identical twin [1].

MS is less common in people living near the equator. A lack of vitamin D due to getting less sun exposure has been studied as a cause of the disease, and even though the studies support a protective effect of vitamin D, it is not proved how does vitamin D help and how genetic variations modify the effect [2].

The influence of stress, toxins, solvents, vaccines or diet has also been studied, but no direct relationship has been found. One factor that has been shown to be harmful is the consumption of tobacco, with smokers being 1.5 times more likely to suffer from the disease. In addition, tobacco consumption also influences the progression of the disease. According to a study in which 179 patients were analyzed, the risk of progressing to a more advanced stage of the disease was 3.6 times higher in smokers [4].

### 1.1.2. Disease progression

The disease begins in 85% of cases as a Clinically Isolated Syndrome (CIS). 45% of CIS patients have a first attack in which they suffer motor or sensory problems, 20% have optic neuritis, 10% have symptoms related to brainstem dysfunction, and the remaining 25% have two or three of the above symptoms [5].

Of the patients with CIS, between 30% and 70% will suffer a second attack, which means that they will convert to Relapsing-Remitting MS (RRMS) [6]. Patients with RRMS have periods without symptoms of the disease, interrupted by new attacks from which they may recover or not. The remaining CIS patients will not develop the disease. Identifying those who are going to develop MS and those who are not is extremely relevant for patient management and treatment decision.

In the remaining 15% of cases, the disease begins with Primary Progressive MS (PPMS). Patients with PPMS do not recover after the first attack, and have a gradual progression of the disease.

### 1.1.3. Diagnosis

The diagnosis of MS is not easy. The different neurological symptoms can be the manifestation of different diseases that must be ruled out before reaching a definitive diagnosis.

The biomedical technique that has more relevance to diagnose MS is magnetic resonance imaging (MRI). Based on the information gathered from the MRI images, the McDonald criteria have been developed and are widely used [7].

According to the McDonald criteria, after the first attack, a patient is considered to have MS if one or more lesions are observed in the MR image in at least two of the following regions of the central nervous system: periventricular, juxtacortical, infratentorial and spinal cord.

McDonald criteria are the most commonly used method for diagnosis of MS. In a study presented in [8], the criteria showed a sensitivity of 74%, specificity of 86%, and accuracy of 80%, but results may vary depending on the cohort.

The disease, in addition to causing white matter lesions (WML), also causes gray matter (GM) atrophy [9]. Unlike WML, changes in the GM volume are difficult to observe with the naked eye, but it is possible to calculate it using image processing techniques. Therefore, a possible way of overtaking the accuracy of the McDonald criteria is to use both WML and GM atrophy to make the diagnosis.

## 1.2. Objectives

The main objectives of the project are:

- To elaborate a classification method to predict the conversion to MS of patients with CIS using machine learning techniques, determining the most informative feature combination.

- Evaluate the quality of the classifier with metrics such as accuracy, sensitivity and specificity, and compare the results obtained with other studies.

## 1.3. Requirements and specifications

This project studies the possibilities that machine learning offers to solve this problem. There are no requirements on the accuracy that should be achieved. At the end of the project, the metrics will be evaluated to assess the viability of the proposed solution.

Project requirements:

- Allow to future users, the possibility of training the model with their database.

Project specifications:

- Work with the database provided by Vall d'Hebron University Hospital, which consists in 141 images of CIS patients acquired following standardized MRI protocols in a 3.0T system (Trio, Siemens).

## 1.4. Methods and procedures

The MRI Unit of Vall d'Hebron University Hospital and The Image and Video Processing Group from the UPC have had an important weight in the project, but no algorithms or software previously developed by them has been used.

The MR images were scanned by the MRI Unit of hospital, and almost all the images had already been processed by them using the following software:

- Lesion Segmentation Toolbox (LST).
- Statistical Parametric Mapping (SPM).
- Freesurfer.

As part of this project, only a few images have had to be processed, using the SPM software package and FreeSurfer.

The machine learning scripts have been written in Python, using the machine learning library called Scikit-learn. Other libraries like NumPy, Pandas or matplotlib, among others, have also been used.

## 1.5. Work plan, deviations and incidences

The work plan with tasks, milestones and the Gantt diagram can be found in Appendix A.

Preparing the database including the clinical data required more time than expected and this caused a delay in the development of the project.

In addition, in the first experiments the classifier performed poorly. To improve the performance additional features were extracted from the MR images related to lesion topography. This took an additional time that was not planned in the initial work plan.

For these two reasons it was decided to extend the project and postpone the deadline from February to May.

## 2. State of the art of the technology used or applied in this thesis

In recent decades there have been many advances in magnetic resonance imaging (MRI). MRI is used to obtain information about the structure and composition of the body. This information is processed by computers and transformed into images.

These advances have allowed its use in medicine, to observe alterations in tissues and detect cancer and other pathologies. In the case of multiple sclerosis, the MRI technology, together with the McDonald criteria, allow the diagnosis of patients who have typical symptoms of multiple sclerosis.

Another technology that has advanced a lot in the last decades is machine learning (ML). ML is a branch of artificial intelligence whose objective is to develop techniques that allow computers to learn. Specifically, it is about creating programs capable of generalizing behaviors from given information.

The branch of ML that is most commonly used in medicine is supervised ML. Supervised ML algorithms try to find patterns in labeled samples, to be able to predict the label of new unlabeled samples. In medicine, it has many applications such as predicting the diagnosis or predicting the disease stage of patients.

In several studies, MRI has been used to observe and analyze how multiple sclerosis affects the human brain. Classically, the approach used has been voxel-based morphometry [10]. It has been used to investigate the differences in volumetry or tissue integrity caused by multiple sclerosis [11]-[12]. However, one limitation of this type of approach is that it highlights common features among the groups compared, and multiple sclerosis is highly heterogeneous.

Some studies have tried to predict the clinical outcome of clinically isolated syndrome patients, focusing on the number of WML and their location [13]-[14]. A recent study [15] proposed a machine learning approach to predict outcome in CIS, based on WML size and location and clinical features of patients. In this study, information about the GM volume was not used, and therefore the results may have room for improvement.

In some studies [16]-[17], machine learning has been used to classify between patients at a more advanced stage of the disease and healthy patients. The accuracies obtained in these studies were very high, but this is because the brains of patients with more advanced disease are badly damaged, and the difference with healthy brains is visible, even to the naked eye.

# 3. Materials and preprocessing

## 3.1. Subjects

The database of this project consists of 141 CIS patients followed at the Cemcat (Centre d'Esclerosi Múltiple de Catalunya), with at least three years of clinical follow-up. All patients were scanned after the onset of a CIS and one year later. In all patients, clinical and demographic information was recorded after the first attack: sex, age, type of CIS presentation, and Expanded Disability Status Scale (EDSS) score.

| | |
|---|---|
| Sex (F/M) | 84 / 57 |
| Mean age (range) | 33.6 (18-49) |
| Type of CIS presentation | brainstem/cerebellum: 32<br>spinal cord: 32<br>optic neuritis: 63<br>other: 14 |
| Mean EDSS score (range) | 1.7 (0-4.5) |

Table 1: Demographic and clinical characteristics of patients with CIS.

### 3.1.1. Definition of conversion to multiple sclerosis

The purpose of this project is to predict between the patients who have a first attack (CIS), those who will convert to multiple sclerosis. But first, it is important to define the meaning of conversion to multiple sclerosis.

After the first attack, patients are diagnosed by the radiologist using the McDonald criteria. Patients that do not meet the criteria at baseline, will be diagnosed with MS if any of the following two conditions are fulfilled:

- They suffer a second attack: when a patient has a second attack, it is considered to have Clinically Definite Multiple Sclerosis (CDMS).
- After twelve months they meet the McDonald criteria: one year after the first attack, an MRI is performed again, and if there are symptoms of disease progress that meet the McDonald criteria, patients are diagnosed with MS.

In this project, three definitions of conversion have been used:

First def.: Patients who meet McDonald criteria in the first MRI are considered to have MS, this means that they are not considered CIS patients and therefore are discarded from the classification. Using this definition, the objective is to predict, among patients that after a first attack are not considered to have MS, which ones convert to MS. In other words, the goal is to find the false negatives of the baseline diagnosis.

Second def.: For this definition, the McDonald criteria are not taken into account. All patients that have a first attack are considered CIS patients and those who suffer a second attack are considered converters.

Third def.: All patients who have a first attack are considered CIS patients. Those that meet at least one of the following conditions are considered converters:

- Fulfill McDonald criteria after the first attack.
- Fulfill McDonald criteria one year after the first attack.
- Have a second attack in the following three years.

Those who do not meet any of these three conditions are considered non-converters.

The following table shows the number of converters and non-converters according to each definition of conversion.

| | **Converters** | **Non-converters** | **Total** |
|---|---|---|---|
| **First definition** | 21 | 76 | **97** |
| **Second definition** | 32 | 109 | **141** |
| **Third definition** | 65 | 76 | **141** |

Table 2: Number of available subjects depending on the definition of conversion used.

It is difficult to decide which one is the best definition and, in the end, it will depend on the purpose of each project. And probably, some of the patients that have been considered as non-converters in the three definitions may have had a second attack after the observation time of 3 years.

### 3.2. MRI acquisition and pre-processing

The MR images were captured by the MRI Unit of Vall d'Hebron University Hospital, using a 3.0-T MRI magnet with a 12-channel phased-array head coil (Trio Tim, Siemens, Germany).

The following pulse sequences were obtained:

- Transverse proton density (PD). PD-weighted images provide good contrast between GM and WM. Pathological processes, such as demyelination or inflammation, often increase water content in tissues, which increases the intensity of the tissues affected because water appears whiter.
- T2-weighted fast spin-echo. T2-weighted images provide good contrast between cerebrospinal fluid (CSF) and brain tissue. Tissues with demyelination or inflammation often appear whiter.

- Sagittal 3D T1 magnetization prepared rapid gradient-echo (MPRAGE). T1-weighed images provide good contrast between GM and WM tissues, and CSF appears black. Tissues with demyelination or inflammation often appear darker.
- Transverse fast T2-FLAIR. In T2-FLAIR images only WML appear bright.

All the images were normalized to the Montreal Neurological Institute standard space using the SPM tool.

# 4. Methods

## 4.1. Feature extraction

Information about WML and GM atrophy was extracted from the MR images.

To extract the GM features, the FreeSurfer suite has been used. Two types of features have been extracted:

- Cortical thickness (CT). CT features, as the name suggests, measure the thickness of different cortical regions. Each hemisphere is divided into 34 regions, for a total of 68 regions. The list with all the regions can be found in Appendix B.

- Subcortical deep gray matter (SDGM). The SDGM features measure the volume of seven GM regions: thalamus, caudate, putamen, pallidum, hypocampus, amygdala and accumbens.

According to a study [18], increasing head size is associated with larger volumes but not CT. For this reason, SDGM features were normalized by the total intracranial volume of each patient, but CT features were not normalized.

To obtain features related to WML, the LST software was used. The LST estimates a lesion mask based on 3D T2-FLAIR images. These masks were used to calculate the total WML volume, counting the number of lesion voxels.

In the second part of the project, to obtain more information about WML, a WML lesion topography was performed. To do so, the lesion masks were used together with masks of different regions of interest (ROIs).

Each hemisphere was divided in eight ROIs: frontal lobe, parietal lobe, occipital lobe, temporal lobe, limbic lobe, insular cortex, cingulate cortex, cerebellum. The lesion masks of each patient were multiplied by each ROI mask, obtaining 16 images for each patient. Finally, the lesion volume of each region was calculated, counting the number of lesion voxels in each image.

On the other hand, to perform voxel-based experiments, GM and WM segmentations of each brain have been used. The images corresponding to the segmentations were provided by the Hospital MRI Unit and, therefore, this feature extraction is not considered a part of this project.

## 4.2. Data analysis

Before starting to perform machine learning experiments, the features were analyzed using histograms and boxplots.

The left histogram in figure 1 shows the average CT. Converters are shown in orange and non-converters in blue. It is clearly observed that the classes are heavily overlapped, with the exception of one outlier that has very low average CT. The same happens with the total SCGM volume, the classes are overlapped too, and in this case there are two outliers.

The right histogram in figure 1 shows the total WML volume. Although the two classes are also quite overlapping, the distribution of the two classes is different. While converters have sparse values, most non-converters have low lesion volumes. A few patients have very high values of WML volume, in order to show them in the histogram, they were added in the last bin.

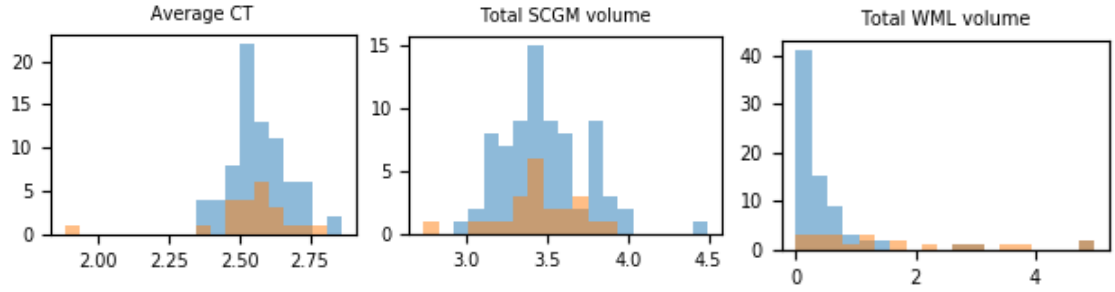


Figure 1: Histograms of average CT (left), total SCGM volume (center) and total WML volume (right), using the first definition of conversion.

In figure 2 below, there are nine histograms corresponding to the CT of nine different regions. As can be seen, the two classes are overlapped in all the regions.

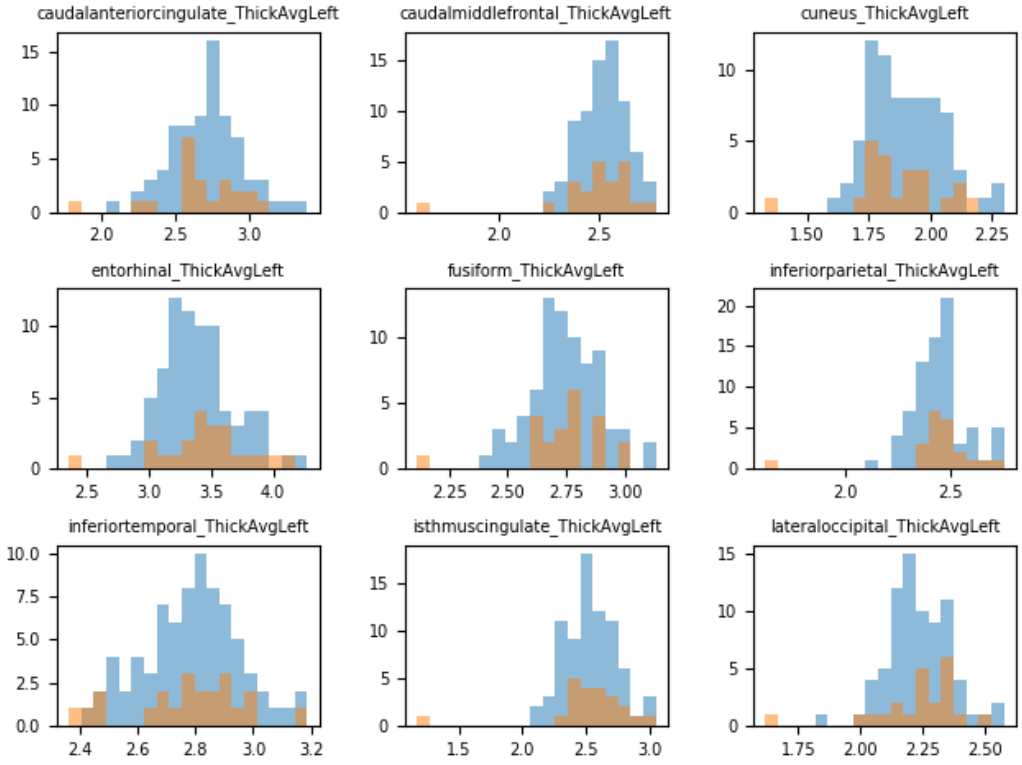


Figure 2: Histograms of the CT of nine different regions, using the first definition of conversion. Converters are represented in orange and non-converters in blue.

In figure 3, there are nine histograms corresponding to the WML volume of nine different regions. Almost all of them are similar to the total WML volume histogram in figure 2. Comparing these histograms with GM histograms, it seems that these features will be more relevant for the classification with machine learning.

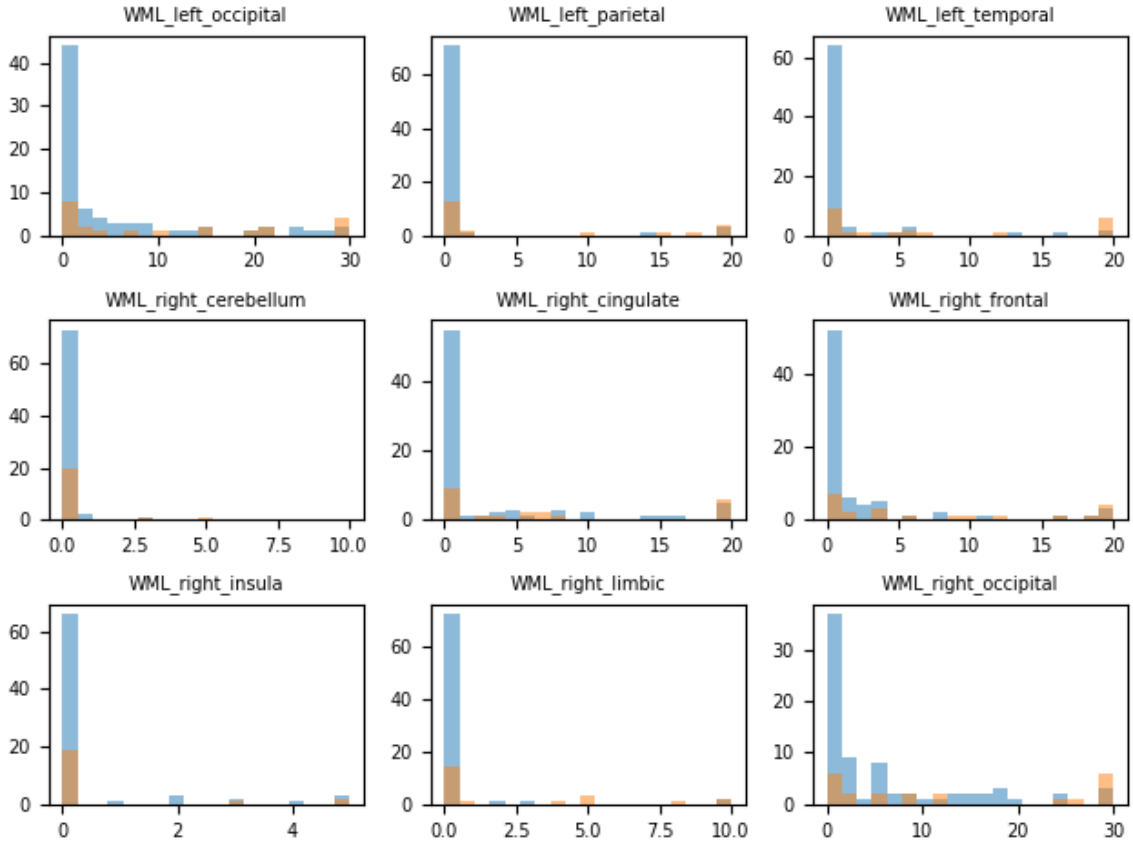


Figure 3: Histograms of WML of nine different regions. Using the first definition of conversion. Converters are represented in orange and non-converters in blue.

To summarize the degree of spread and skewness in the data, boxplots have been used. In the boxplots below, the rectangular box contains 50% of the values, the orange line represents the median of the dataset, and the individual data points represent the outliers.

The previous histograms are all generated using the first definition of conversion. As can be seen in the boxplots in figures 4, 5 and 6, total CT distributions of converters and non-converters are very similar with the three definitions of conversion, total SCGM volume distributions are slightly different only using the third definition of conversion, and total WML volume distributions are clearly different with any of the three definitions, especially with the second and third definitions (note that the scale of the vertical axis changes).

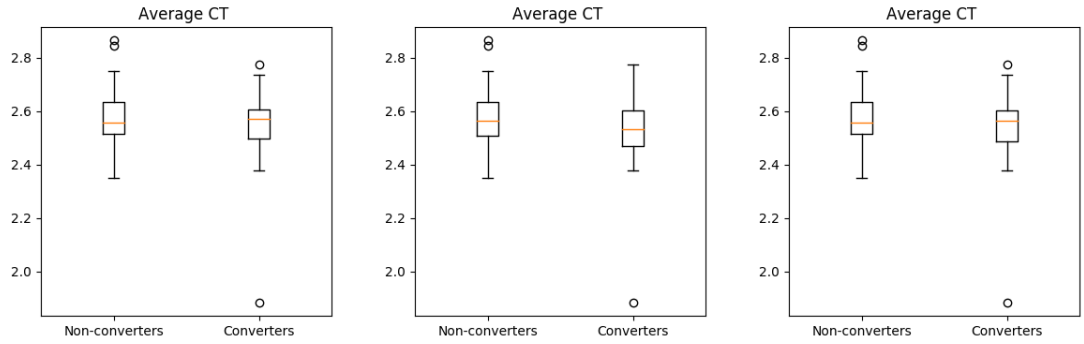


Figure 4: Boxplots of average CT using the first (left), second (center) and third (right) definitions.

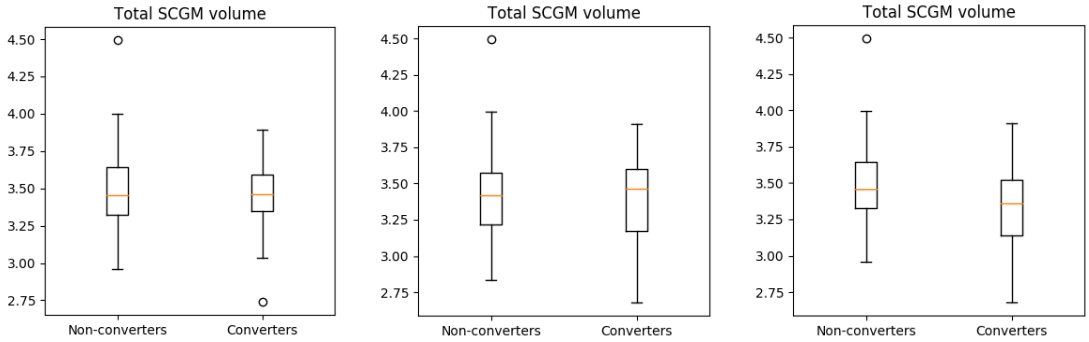


Figure 5: Boxplots of total SCGM volume using the first (left), second (center) and third (right) definitions.

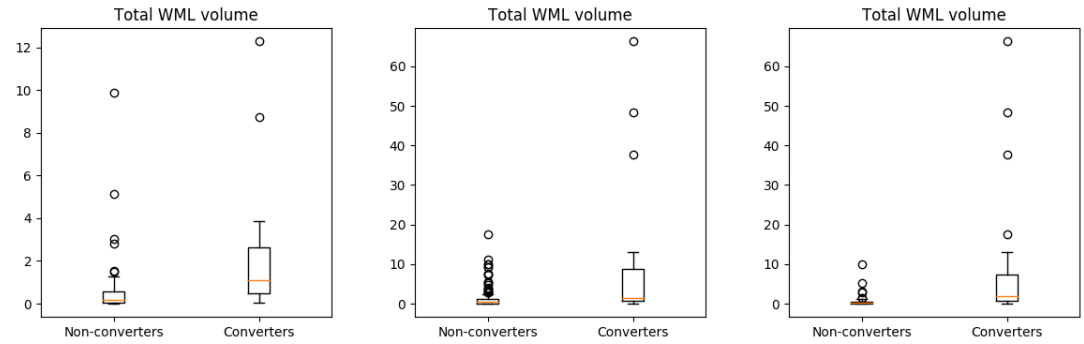


Figure 6: Boxplots of total WML volume, using the first (left), second (center) and third (right) definitions.

## 4.3. Machine learning

In this project supervised machine learning algorithms have been used to classify between the two classes (converters and non-converters).

This classification problem is challenging due to several reasons:

- Small dataset. Especially with the first definition of conversion, the minority class contains only 21 patients. For this reason, it is important to use a simple model to avoid overfitting.

- Class overlap. As observed in histograms and boxplots from individual features in the previous section, classes are heavily overlapped. Machine learning algorithms combine information from different features and hence in some cases it is possible to classify between overlapping classes, but it is not always possible.

- High intraclass variance. As observed in the histograms and boxplots in the previous section, intraclass variance is high, especially in converters. Together with a small dataset, this usually causes that the results of the classifier vary a lot depending on the splits (train / test) of the dataset. To solve this problem, a method called repeated holdout cross-validation has been used.

- Large number of features. The number of features available in this project is high compared to the size of the dataset. As a rule of thumb, it is recommended to have at least 5 training samples for each feature [19], and in this project there is approximately one training sample for each feature. To solve this problem, two different feature selection methods have been implemented.

### 4.3.1. Performance metrics

To evaluate the quality of the classifiers, the following metrics have been used:

- Accuracy. It is the ratio of correctly predicted samples to all tested samples.
- Recall / sensitivity. It measures the probability that a positive sample is correctly classified as positive.
- Specificity. It measures the probability that a negative sample is correctly classified as negative.
- Balanced Accuracy (BA). It is calculated as the average of the recall and the specificity. When the classes are balanced, the BA equals the accuracy.
- Precision. It is the probability that a sample classified as positive has been correctly classified.
- $F_1$-Score. It is the harmonic average of the precision and recall.

The formulas to calculate these metrics are the following:

$$Acccuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$Recall\ /\ Sensitivity = \frac{TP}{TP + FN} \quad (2)$$

$$Specificity = \frac{TN}{TN + FP} \quad (3)$$

$$Balanced\ accuracy = \frac{Recall + Specificity}{2} \quad (4)$$

$$Precision = \frac{TP}{TP + FP} \quad (5)$$

$$F_1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \quad (6)$$

where
True positives ($TP$): positive samples (converters) correctly classified.
False negatives ($FN$): negative samples (non-converters) classified incorrectly as converters.
False positives ($FP$): positive samples (converters) classified incorrectly as non-converters.
True negatives ($TN$): negative samples (non-converters) correctly classified.

### 4.3.2. Feature selection

In machine learning experiments, contrary to what may seem intuitive, increasing the number of features does not imply improving the predictive capacity of the classifier. When the number of features is very large in proportion to the number of samples, the so-called "curse of dimensionality" occurs.

Typically, as the number of features increases, the predictive power increases to a certain point but then begins to decrease [20]. This is known as Hughes phenomenon or peaking phenomena. This typical behavior can be seen in figure 7.

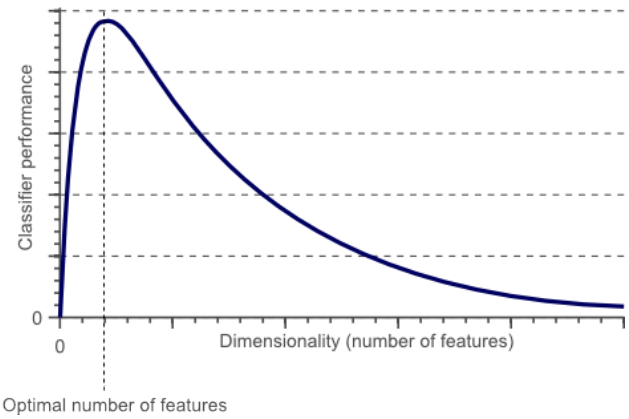


Figure 7: Typical classifier performance when the number of features increases

A database with M samples and N features, can be represented as M points in an N-dimensional space. Given M points, if the number of features increases, the data points become sparse, and this sparsity sometimes reduces the performance of the machine learning algorithms.

Models created with a high number of features tend to overfit, that is, they generalize worse and therefore have poor performance on unseen data.

Reducing the number features has some benefits:

- Redundant features and features that do not provide information are removed.
- With fewer features, the training time of the classifier is significantly reduced.
- The classifier model is easier to interpret.

This last point is especially important since it is interesting to be able to understand which are the most relevant features for classification, to see if they are consistent with medical findings.

There are three different methods to perform feature selection: filters, wrappers and embedded methods [21]. Wrapper methods have not been used in this project because they require a higher number of samples. Filters and embedded methods are explained in the following sections.

#### 4.3.1.1 Filter methods

In filter methods, the features are selected based on the score obtained in a scoring function. Once the scores are calculated, features are ranked according to the score obtained, and the best ones in the ranking are selected. The irrelevant features are filtered before the classification stage.

The procedure is as follows:

1. Features are evaluated using a scoring function.
2. Features are sorted according to the value obtained.
3. The k best features are selected.

Four different score functions have been used in this project:

- Welch's t-test. This statistical test is used to test if two populations have the same mean. Unlike the typically used Student's t-test, Welch's t-test does not assume that the two populations have equal variances, and it is more reliable when the two populations are unbalanced. For each feature, this test has been used to calculate the difference between the mean value of the two classes. It is assumed that features with more different means are more relevant and therefore, they are selected.

- ANOVA F-test. This statistical test is similar to the t-test, but instead of comparing the means of two populations, the variances are compared. In this case, the features with more different variances between converters and non-converters are selected.

- Pearson correlation coefficient. It measures the linear correlation between two variables. When used in feature selection, the correlation between each feature and the labels is measured. It provides a value between -1 (negative linear correlation) and 1 (positive linear correlation), and a coefficient of 0 means no correlation between the two variables. The features with higher absolute values are selected.

- Spearman's rank correlation coefficient. It is similar to the Pearson correlation coefficient, but instead of measuring a linear correlation, it assesses monotonic relationships. For this reason, it is more recommended when the relation of the two variables is not linear. The features with higher absolute values are selected.

#### 4.3.1.2. Embedded methods

In embedded methods, the feature selection is performed as part of the training of the model. In this project, two different methods have been used.

The Least Absolute Shrinkage and Selection Operator (LASSO) method has been used as a constraint to construct the linear model. The LASSO penalizes the coefficients with the L1 penalty, causing many of the coefficients to be 0. The features with coefficients different from 0 are the selected features.

The dataset used in this project contains highly correlated features. For this reason, the regularization technique called Elastic Net has also been used, since it is recommended for datasets that contain highly correlated variables [22]. The Elastic Net is a regularized

regression method that combines the L1 and L2 penalties of the LASSO and ridge methods. It produces a sparse model with good prediction accuracy.

### 4.3.3. Classifiers

In this project two classifiers have been used: Support Vector Machines (SVMs) and Logistic Regression (LR).

#### 4.3.3.1. Support Vector Machines

SVMs are supervised learning algorithms that can be used for classification or regression. When an SVM is trained for classification, the algorithm constructs a hyperplane that tries to separate the samples of the two classes, maximizing the distance to the nearest data point. Figure 8 shows a simple example of a hyperplane that separates the samples of two classes. The highlighted data points, called support vectors, define the boundary between the classes.

This example is very simple since it has only two features and the classes are linearly separable, but SVMs can also perform non-linear classification, implicitly mapping their inputs into high-dimensional feature spaces. However, with small datasets non-linear models tend to overfit. For this reason, a linear SVM has been used in this project.

A standard SVM tries to build a hyperplane that separates all the samples of the two classes. However, this can lead to poorly fit models if classes are non-separable. To solve this, soft-margin SVMs allow some samples to be placed in the wrong side of the hyperplane [23]. The hyperparameter C is the parameter of the cost function of the soft-margin, and it controls the influence of each individual support vector.

For large values of C, the algorithm constructs a hyperplane with a smaller margin, trying to classify all the training samples correctly. On the other hand, for small values of C, the algorithm constructs a hyperplane with larger margins, even if that hyperplane misclassifies more training samples.

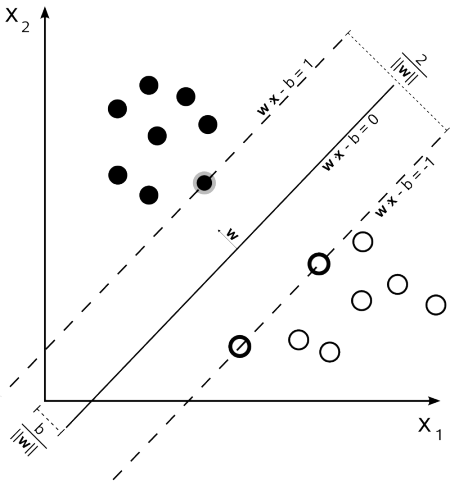


Figure 8: SVM hyperplane representation

#### 4.3.3.2. Logistic regression

Logistic regression (LR), despite its name, is a generalized linear model that can be used for binary classification. Generalized linear models expect the predicted value to be a linear combination of a set of independent variables, in other words, the outcome always depends on the sum of the inputs and parameters.

Like SVMs, LR also has a regularization parameter. In case of LR, the cost function is regularized by adding a penalty, with $\lambda$ as the regularization parameter. The parameter C of SVMs, has a similar effect of inversed $\lambda$, $1/\lambda$. It is commonly said that LR has a hyperparameter C, with $C = 1/\lambda$. In the following section, this definition is going to be used.

Like any other classification algorithm, it has to be trained to learn to classify new unlabeled data. When a new sample is classified, the algorithm calculates the probability that the sample belongs to the positive class. If this probability is greater than 0.5, the sample will be classified as positive and if it is lower, as negative.

This can be useful for medical decisions, where mistakes can be costly. For example, a threshold of 90% could be applied, and in patients with lower probability, the medical test could be repeated or a different test could be performed.

#### 4.3.3.3. Hyperparameter tuning

The optimal C value varies according to each dataset; actually, it also varies for each different division of the same dataset into train and test subsets. To increase the performance of the classifier, this hyperparameter has to be optimized.

In order to avoid choosing an optimal C based on only one division of the dataset, a stratified cross-validation has been implemented. The dataset is split into K subsets using a stratified K-fold, seeking to ensure that each subset is representative of all the data, by having the same proportion of converters and non-converters in each subset. Then, the classifier is trained and evaluated K times, using a different test subset every time and the remaining K-1 subsets to train.

To optimize the regularization parameter, the scheme in figure 9 has been implemented. The procedure is as follows:

1. The training set is divided into K subsets using a stratified K-fold.

2. K-1 subsets are used to train the classifier and the other one to test. This process is repeated using different values of C, saving the results of the classifier for each value.

3. Step 2 is repeated K more times, changing the training and test subsets until all the subsets have been used to test.

4. The average results for every value of C are calculated, and the value of C that achieves the highest $F_1$-Score is selected.

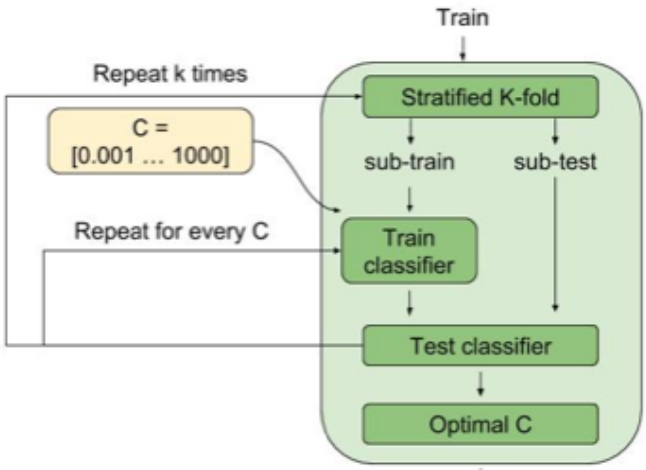


Figure 9: Parameter optimization scheme.

### 4.3.4. Repeated hold-out cross-validation

The dataset of this project is small and there is high intraclass variance. For this reason, the classifier performance varies greatly depending on the division that is made of the database.

To solve this problem, a method called repeated hold-out has been used in all the experiments. It consists in repeating each experiment N times, with different train and test partitions, and averaging the results.

The graph in figure 10 illustrates how the performance of the classifier is stabilized as more repetitions are averaged. The number of repetitions is shown on the horizontal axis. The first data point shows the metrics obtained in the first repetition. The second data point shows the average of the first two repetitions, and so on until the last point that shows the average metrics of 1000 repetitions.

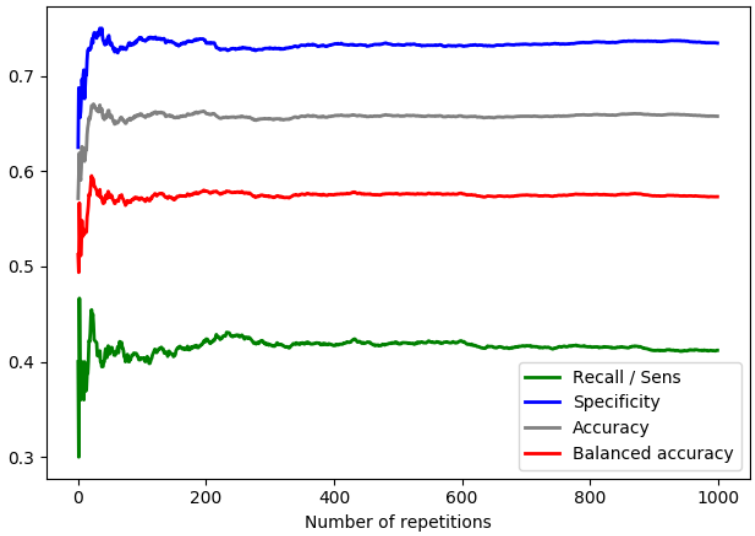


Figure 10: Average results using repeated holdout cross-validation.

### 4.3.5. Experiments

In this project many experiments have been carried out, which can be grouped into two types of experiments. The two types of experiments differ in the method used for the feature selection. They are explained in the following sections.

#### 4.3.4.1. Experiments using filter-based feature selection

The first set of experiments use filter-based methods to perform the feature selection. The scheme of these experiments can be seen in figure 11. Different configurations, changing the feature selection scoring function, the classifier and the definition of conversion have been carried out.

The procedure is as follows:

1. The dataset is split in two subsets, using p% for training and the remaining (1-p)% for test. The dataset is shuffled randomly in order to have different training and test sets in each repetition.

2. The m best features are selected using a filter method.

3. The regularization parameter C is optimized using the optimization scheme explained in the previous section.

4. Using the optimal C, the classifier is trained using the whole training subset and evaluated on the test subset, and the results are saved.

5. Steps 2, 3 and 4 are repeated for every value of m.

6. Steps 1 through 5 are repeated N times.

In the experiments that use filter-based feature selection methods, instead of optimizing the number of features, the classification process is repeated with all the possible number of features, starting with one until all the features are used. The reason is that the objective of this project is not to create production-ready solution, but to analyze and understand how machine learning can solve this classification problem. In these experiments, the results are shown in graphs. The number of features is shown on the horizontal axis and the metrics on the vertical axis.

This scheme has also been used to perform the voxel-based experiments. The only difference is that instead of selecting the m best features, the m best voxels are selected.

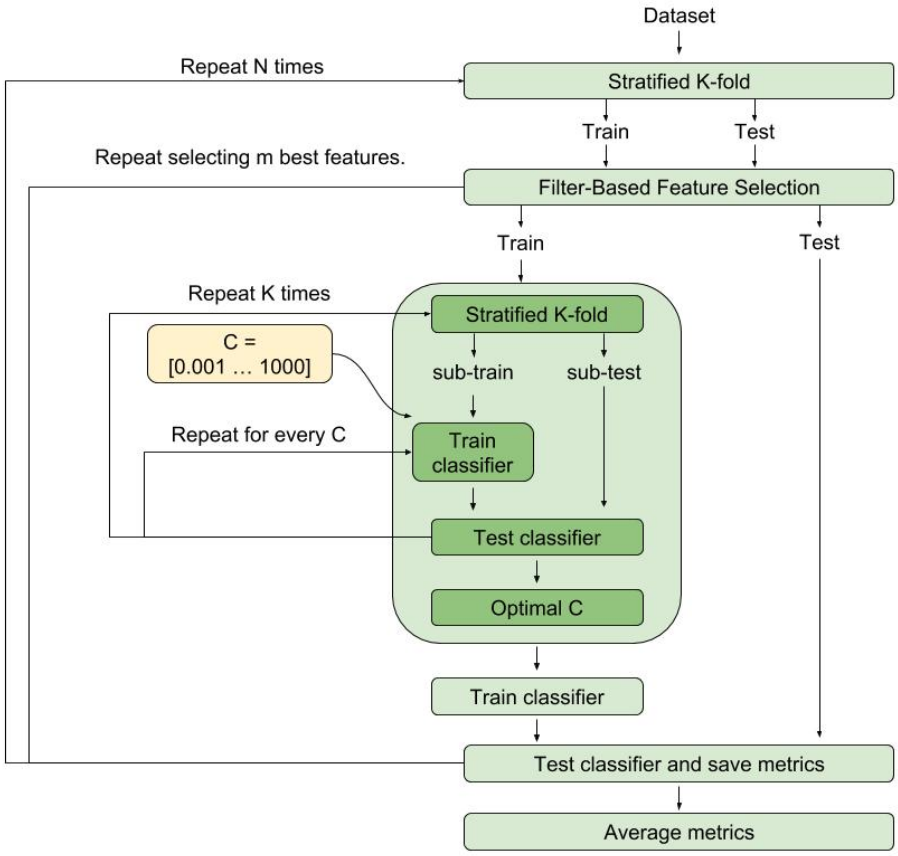


Figure 11: Scheme of the experiments using filter-based feature selection.

#### 4.3.5.1. Experiments using embedded feature selection

The second set of experiments use embedded methods to perform the feature selection. In figure 12 the scheme of this experiments is shown. This scheme is simpler than the scheme of the previous section because the feature selection is performed internally in the classifier.

The procedure is as follows:

1. The dataset is split in two subsets, using p% for training and the remaining (1-p)% for test. The dataset is shuffled randomly in order to have different training and test sets in each repetition.

1. The regularization parameter C is optimized.

2. Using the optimal C, the classifier is trained and evaluated, and the results are saved.

3. Steps 1 through 3 are repeated N times.

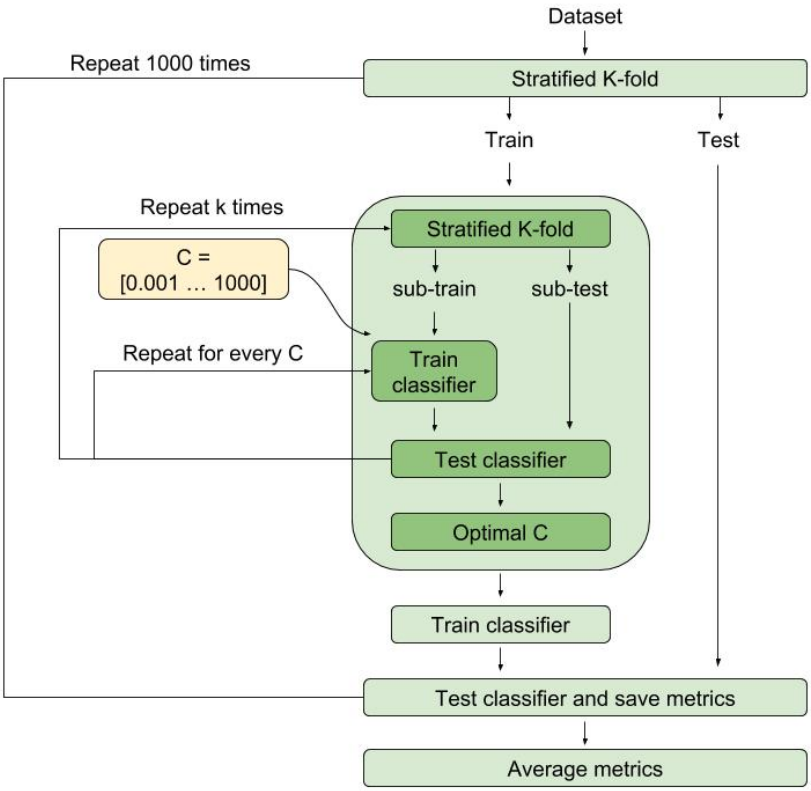


Figure 12: Scheme of the experiments that use embedded feature selection.

# 5. Results

## 5.1. Results using only WML features

Experiment configuration:

- Filter-based feature selection using the Spearman correlation.
- Using an SVM with parameter optimization.
- Experiment repeated 1000 times.
- Using only WML features: 16 WML volumes of each brain lobe.

Figure 13 shows the evolution of the performance of the classifier as a function of the number of features included, using the first definition of conversion. The best result is obtained when all the 16 WML features are used, the BA is 0.68 and the F1-Score is 0.52. Using the second and the third definitions of conversion, the best result is also obtained when all the 16 WML features are used. With the second definition of conversion, the results are worse (BA = 0.58, F1-Score = 0.38). And with the third definition the results are better (BA = 0.74, F1-Score = 0.70).

Patients that were diagnosed with MS at baseline MRI, have high WML values. For this reason, the results are better with the third definition of conversion.

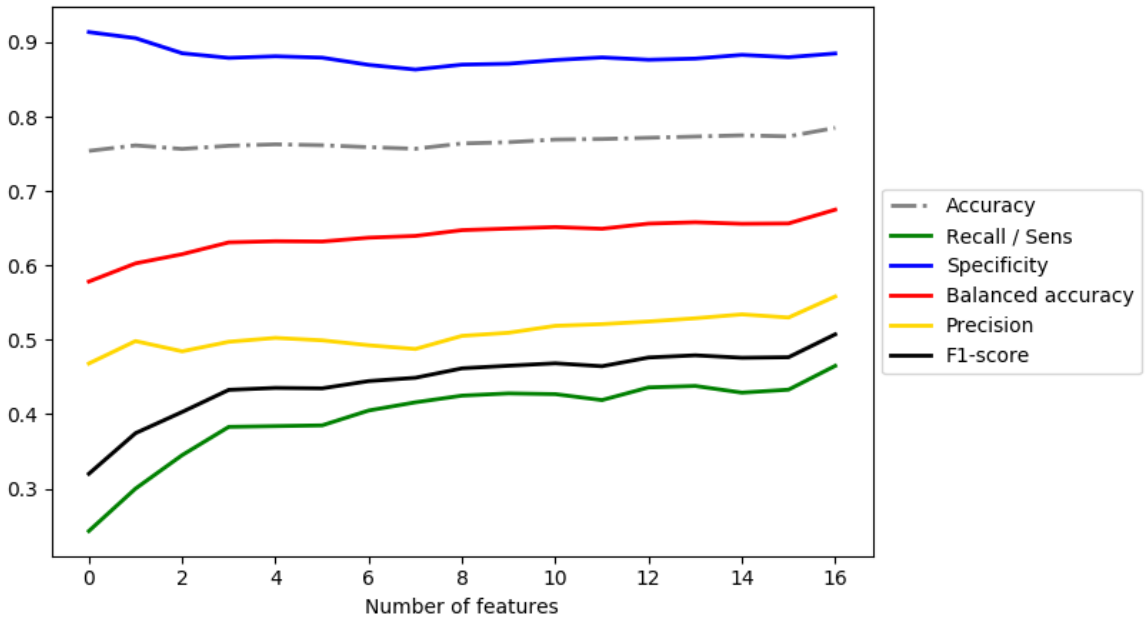


Figure 13: Classifier performance using WML features

## 5.2. Results using only GM features

Experiment configuration:

- Filter-based feature selection using the Spearman correlation.
- Using an SVM with parameter optimization.
- Experiment repeated 1000 times.
- Using only GM features: CT of 68 regions and 7 SCGM volumes.

As can be seen in figure 14, using the first definition of conversion, the balanced accuracy is greater than 0.5. Even though the metrics are very low, these results show that GM features contain information useful for classification.

Using the first definition of conversion, patients who meet the McDonald criteria at baseline MRI are discarded, because they are diagnosed with MS, and are not considered CIS patients anymore. These patients are not used for evaluation, but they can be added to the training set to train the classifier.

The same experiment was repeated, adding the MS patients to the training set and the results improved remarkably, especially when all the features were used (BA = 0.65, F1-Score = 0.48). For the second definition of conversion, the results decrease considerably (BA = 0.55, F1-Score = 0.37). But with the third definition, the results improved notably, reaching a balanced accuracy of 0.68 and an F1-Score of 0.65.

This makes sense because, as seen in the data analysis, SCGM features differ more between converters and non-converters using the third definition of conversion.

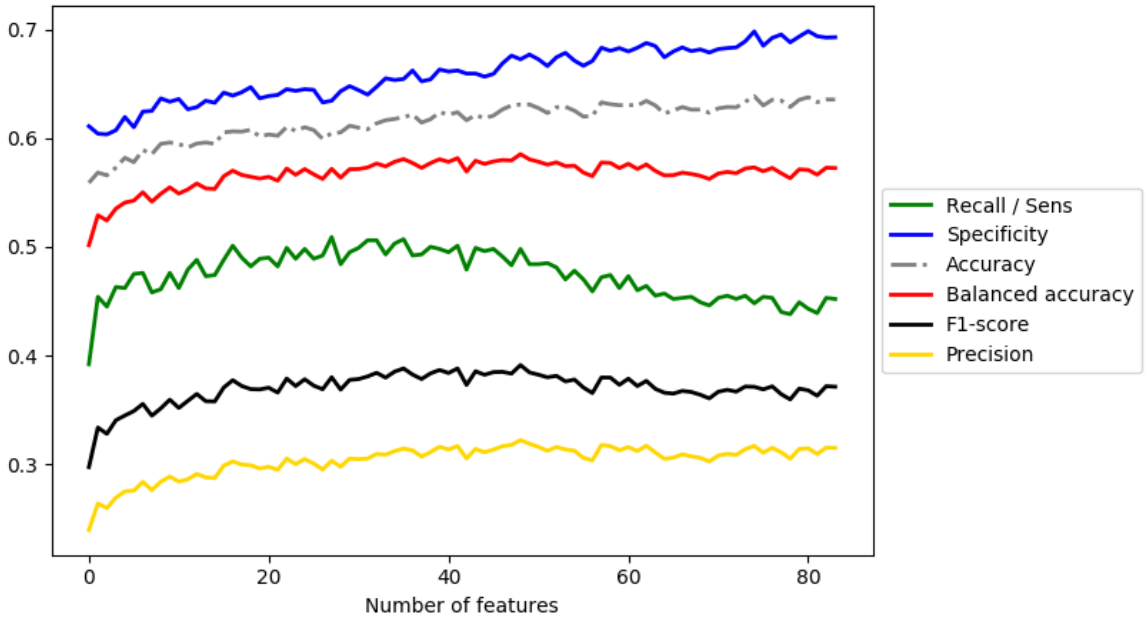


Figure 14: Classifier performance using GM features.

## 5.3. Experiments using all the features

In the experiments explained in this section, all the features have been used: GM and WM features, together with clinico-demographic features.

As can be seen in figure 15, the results are very similar to the results obtained using only WML features. The results only improve with the second definition of conversion (max. BA = 0.61, max. $F_1$-Score = 0.42).

This experiment was repeated using the four different scoring functions in the filter-base feature selection. The best performance was achieved using Spearman correlation. The results can be found in Appendix C.

This experiment was repeated with logistic regression instead of an SVM and the results are almost identical (max. BA=0.61, max. $F_1$-Score=0.42).

Using embedded feature selection methods, the classifier performance decreased. Using LASSO, the results were slightly worse (BA=0.60, $F_1$-Score=0.40), and using elastic net the performance decreased even more (BA=0.58, $F_1$-Score=0.39).

The experiments on the previous section have shown that GM features have information, but they are almost redundant when combined with WML. This seems to indicate that the inflammatory component of the disease, reflected through the WML features, plays a major role as compared to atrophy (GM features) in the conversion of these patients.

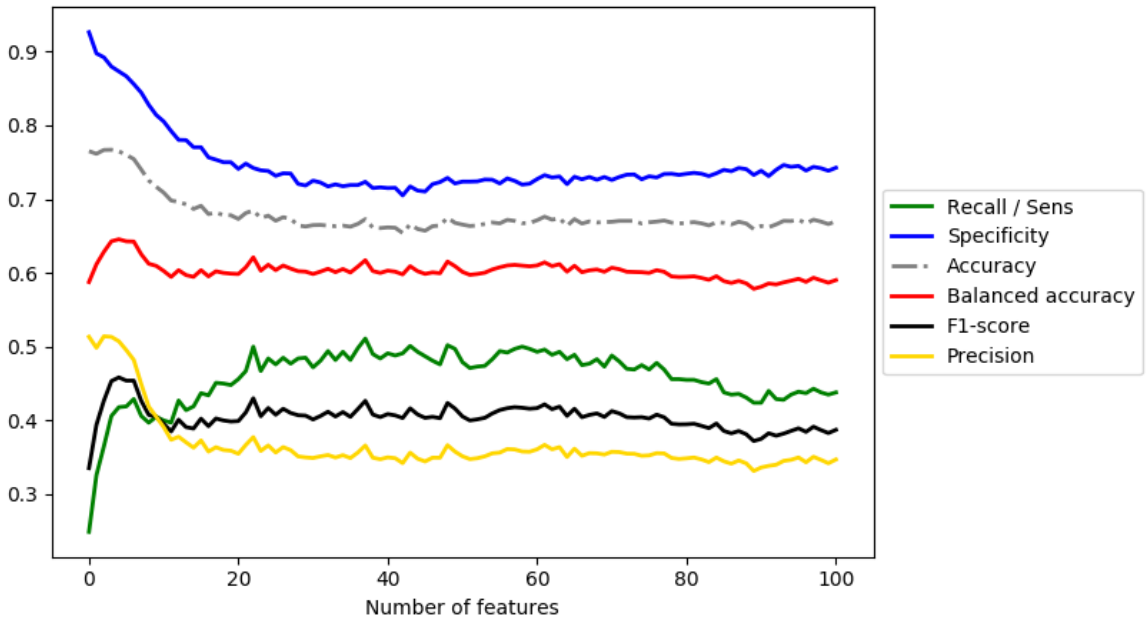


Figure 15: Classifier performance using all the features. Using the first definition of conversion.

## 5.4. Relevant features

In order to translate the results obtained in this project to a clinical setting, it is important to identify which features contribute more.

When a linear SVM is trained, the algorithm computes a coefficient for each feature. Features with more weight in the model have higher absolute coefficients. Therefore, features can be sorted by its coefficient to create a relevance ranking.

The ranking in Table 2 contains the twenty features with higher coefficients using the first definition of conversion (similar rankings are obtained with the other definitions). WML topography is the major contributor, followed by cortical GM features. It is interesting that both left and right paracentral appear in high positions in the raking (10 and 17). It should be noted that in position 12, there is the EDSS score at baseline (EDSS0), a clinical value that reflects the disability stage of the patient (the higher the value, the higher the disability degree). In our study, a higher EDSS score at baseline increases the probability of converting to MS.

| Position | Feature name | Coefficient |
|---|---|---|
| 1 | WML_left_cingulate | 45.99 |
| 2 | WML_left_temporal | 36.69 |
| 3 | WML_right_limbic | 31.34 |
| 4 | WML_left_parietal | -26.19 |
| 5 | WML_right_temporal | -24.01 |
| 6 | WML_left_frontal | -20.92 |
| 7 | WML_right_occipital | 20.26 |
| 8 | WML_left_insula | 9.38 |
| 9 | WML_right_parietal | 8.97 |
| 10 | paracentral_ThickAvgRight | -8.25 |
| 11 | WML_right_cingulate | -7.43 |
| 12 | edss0 | 6.74 |
| 13 | WML_left_occipital | 6.72 |
| 14 | lateraloccipital_ThickAvgRight | 6.36 |
| 15 | caudalanteriorcingulate_ThickAvgRight | -5.17 |
| 16 | WML_right_frontal | 4.76 |
| 17 | paracentral_ThickAvgLeft | -4.07 |
| 18 | precuneus_ThickAvgRight | -3.00 |
| 19 | pericalcarine_ThickAvgRight | -1.91 |
| 20 | caudate | -1.09 |

Table 3: Top 20 features ranked according to its coefficient in the classifier

## 5.5. Voxel-based classification

Experiment configuration:

- Filter-based feature selection using the F-test.
- Using an SVM with parameter optimization.
- Experiment repeated 1000 times.
- Using the voxels as features.

In this context, the feature is the intensity value of the voxel. Masks are included to evaluate only selected portions of voxels. In this case, the GM and WM segmented masks were used to assess the two components independently. The GM component mostly reflects brain atrophy, while WM reflect mostly inflammatory activity.

Using the GM segmentation of the brain, the BA was approximately 0.50 for the three definitions of conversion, which means that the classification was done randomly.

Using the WM segmentation, with the first and the second definitions of conversion, the balanced accuracy is also approximately 0.50. But with the third definition the results improve, reaching a balanced accuracy of 0.59 and an F1-Score of 0.54, when the 100 best voxels are selected.

The 100 best voxels selected by the feature selection method can be seen in figure 16. These voxels are located in the superior corona radiata. This finding is consistent with a study [24] in which it was concluded that most typical WML locations are the superior and posterior regions of the corona radiata.

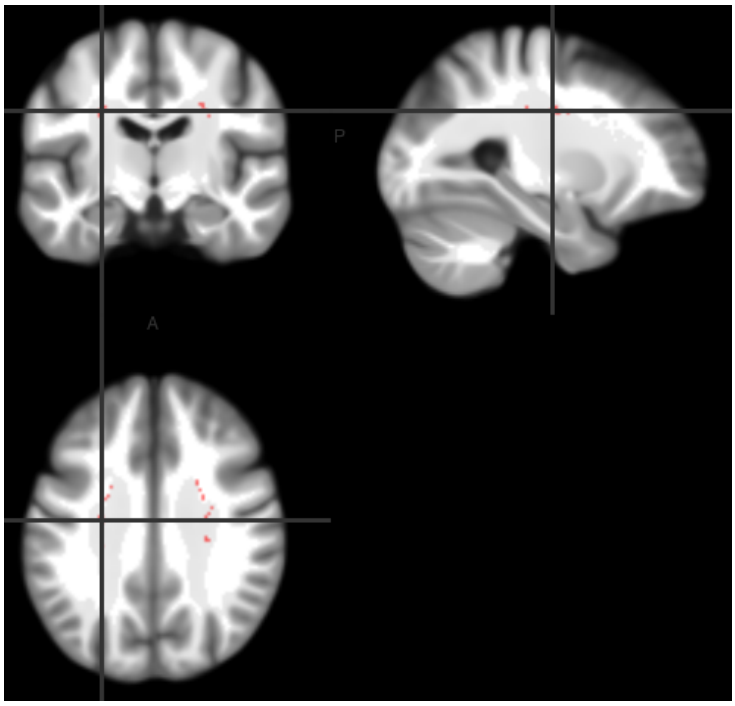


Figure 16: In red, the voxels in the image that contribute to the group classification overlaid over the mean MPRAGE MRI of the whole cohort (skull-stripped).

# 6. Budget

For the realization of this project, no hardware has been required apart from the computer. The computer that has been used is a Macbook Pro. It has an approximate cost of 1200 € and lasts at least three years. The proportional part of the cost during the time that has been used is 300€.

The only applications used that require a license are MATLAB and Microsoft Office. The cost of a Microsoft Office license is 7 euros per month, so this equates to a cost of 63€. The price of a standard MATLAB annual license is 800€. This license has been provided by the UPC but it will be taken into account in the budget.

Regarding labor, taking into account that this is a 24 ECTS project, it corresponds to 600 hours of work. Calculating approximately that the cost per hour of a junior engineer is 12 €/h, the total cost is:

$$600 \text{ h} \cdot 12 \text{ €/h} = 7200 \text{ €}$$

Additionally, the work of the advisors must be taken into account. This project has been directed by two co-directors with an average dedication of 1.5 hours per week during 36 weeks. Considering a cost per hour of 40 €/h, the total cost is:

$$2 \text{ co-directors} \cdot 36 \text{ weeks} \cdot 1.5 \text{ h/week} \cdot 40 \text{ €/h} = 4320 \text{ €}$$

The following table summarizes the costs of the project.

| Item | Cost |
|---|---|
| MATLAB | 800 € |
| Microsoft Office | 63 € |
| MacBook Pro | 300 € |
| Work hours (junior engineer) | 7200 € |
| Work hours (advisors) | 4320 € |
| **TOTAL** | **12683 €** |

Table 4: Budget

# 7. Conclusions and future development:

The objective of this project was to explore the possibilities of machine learning in predicting the conversion to MS of patients with a CIS, using MRI-derived features. Three different definitions of conversion to MS have been used, repeating the experiments for each definition.

Different methods of feature selection have been implemented and two classifiers have been used (SVM, LR). The best performance has been achieved with a filter-based feature selection, using the Spearman correlation, and the results obtained with SVMs and LR are very similar.

The results of the experiments are within the expected range, although the performance varies depending on the definition of MS conversion used (BA = [0.68, 0.61, 0.73], recall = [0.51, 0.60, 0.68], specificity = [0.85, 0.62, 0.78]).

The results show that GM atrophy starts in an early stage of MS, although in most patients it is imperceptible. GM features are informative, but when combined with WML features, they become less relevant to predict MS conversion.

Once the experiments using MRI-derived features were finished, we performed experiments using the voxels as features, but the performance is very poor, specially using the first and the second definitions (BA = [0.51, 0.49, 0.59], recall = [0.07, 0.13, 0.52], specificity = [0.95, 0.85, 0.66]).

When the voxels are used as features, the 3D matrix that represents the image is flattened into a 1D array and hence the spatial information is lost. Moreover, the MR images used are made up of approximately two million voxels and therefore, the number of features is very high for the number of samples. For these two reasons, the results of the voxel-based experiments were already expected to be poor.

In conclusion, the results obtained are not good enough to confirm the feasibility of developing a ML solution into routine clinical practice, in order to predict conversion to MS in CIS patients.

Since MS is a heterogeneous disease, it is difficult to find patterns among patients, specially with a small number of subjects. Therefore, it would be interesting to use the proposed approach in this project with a larger dataset. Also, with a large database it would be suitable to use deep learning.

In ML, when a new sample is predicted, the classifier returns the most probable class, and some ML classifiers can also provide the probability of belonging to each class. In our problem, this means that we can obtain the probability that a patient will convert or not.

There is the possibility, to only take into account the prediction when the probability exceeds a certain threshold, and use the current diagnosis of MS when the probability is below the threshold. It would be interesting to study the viability of this approach.

# Bibliography:

[1] Willer, C. J., Dyment, D. A., Risch, N. J., Sadovnick, A. D., Ebers, G. C., & Canadian Collaborative Study Group. (2003). Twin concordance and sibling recurrence rates in multiple sclerosis. Proceedings of the National Academy of Sciences, 100(22), 12877-12882.

[2] Ascherio, A., Munger, K. L., & Simon, K. C. (2010). Vitamin D and multiple sclerosis. The Lancet Neurology, 9(6), 599-612.

[3] Ascherio A, Munger KL (June 2007). "Environmental risk factors for multiple sclerosis. Part II: Noninfectious factors". Annals of Neurology. 61 (6): 504–13.

[4] Hernán, M. A., Jick, S. S., Logroscino, G., Olek, M. J., Ascherio, A., & Jick, H. (2005). Cigarette smoking and the progression of multiple sclerosis. *Brain*, *128*(6), 1461-1465.

[5] Tsang, Benjamin KT, and Richard Macdonell. "Multiple sclerosis: diagnosis, management and prognosis." Australian family physician 40.12 (2011): 948.

[6] 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. *The Lancet Neurology*, *4*(5), 281-288.

[7] Polman, C. H., Reingold, S. C., Banwell, B., Clanet, M., Cohen, J. A., Filippi, M., ... & Lublin, F. D. (2011). Diagnostic criteria for multiple sclerosis: 2010 revisions to the McDonald criteria. Annals of neurology, 69(2), 292-302.

[8] Tintore, M., Rovira, A., Rio, J., Nos, C., Grive, E., Sastre-Garriga, J., ... & Montalban, X. (2003). New diagnostic criteria for multiple sclerosis application in first demyelinating episode. *Neurology*, *60*(1), 27-30.

[9] Fisher, E., Lee, J. C., Nakamura, K., & Rudick, R. A. (2008). Gray matter atrophy in multiple sclerosis: a longitudinal study. Annals of neurology, 64(3), 255-265.

[10] Ashburner, J., & Friston, K. J. (2000). Voxel-based morphometry—the methods. *Neuroimage*, *11*(6), 805-821.

[11] Audoin, B., Zaaraoui, W., Reuter, F., Rico, A., Malikova, I., Confort-Gouny, S., ... & Ranjeva, J. P. (2010). Atrophy mainly affects the limbic system and the deep grey matter at the first stage of multiple sclerosis. *Journal of Neurology, Neurosurgery & Psychiatry*, *81*(6), 690-695.

[12] Raz, E., Cercignani, M., Sbardella, E., Totaro, P., Pozzilli, C., Bozzali, M., & Pantano, P. (2010). Gray-and white-matter changes 1 year after first clinical episode of multiple sclerosis: MR imaging. *Radiology*, *257*(2), 448-454.

[13] Giorgio, A., Battaglini, M., Rocca, M. A., De Leucio, A., Absinta, M., Van Schijndel, R., ... & Enzinger, C. (2013). Location of brain lesions predicts conversion of clinically isolated syndromes to multiple sclerosis. *Neurology*, *80*(3), 234-241.

[14] Tintore, M., Otero-Romero, S., Río, J., Arrambide, G., Pujal, B., Tur, C., ... & Vidal-Jordana, A. (2016). Contribution of the symptomatic lesion in establishing MS diagnosis and prognosis. *Neurology*, *87*(13), 1368-1374.

[15] Wottschel, V., Alexander, D. C., Kwok, P. P., Chard, D. T., Stromillo, M. L., De Stefano, N., ... & Ciccarelli, O. (2015). Predicting outcome in clinically isolated syndrome using machine learning. *NeuroImage: Clinical*, *7*, 281-287.

[16] Muthuraman, M., Fleischer, V., Kolber, P., Luessi, F., Zipp, F., & Groppa, S. (2016). Structural brain network characteristics can differentiate CIS from early RRMS. *Frontiers in neuroscience*, *10*, 14.

[17] Kocevar, G., Stamile, C., Hannoun, S., Cotton, F., Vukusic, S., Durand-Dubief, F., & Sappey-Marinier, D. (2016). Graph theory-based brain connectivity for automatic classification of multiple sclerosis clinical courses. *Frontiers in neuroscience*, *10*, 478.

[18] Barnes, J., Ridgway, G. R., Bartlett, J., Henley, S. M., Lehmann, M., Hobbs, N., ... & Fox, N. C. (2010). Head size, age and gender adjustment in MRI studies: a necessary nuisance?. *Neuroimage*, *53*(4), 1244-1255.

[19] Johnstone, I. M., & Titterington, D. M. (2009). Statistical challenges of high-dimensional data.

[20] Hughes, G.F. (January 1968). "On the mean accuracy of statistical pattern recognizers". IEEE Transactions on Information Theory. 14 (1): 55–63. doi:10.1109/TIT.1968.1054102

[21] Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. Journal of machine learning research, 3(Mar), 1157-1182.

[22] Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(2), 301-320.

[23] Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine learning, 20(3), 273-297.

[24] Bodini, B., Battaglini, M., De Stefano, N., Khaleeli, Z., Barkhof, F., Chard, D., ... & Rovira, A. (2010). T2 lesion location really matters: a 10 year follow-up study in primary progressive multiple sclerosis. Journal of Neurology, Neurosurgery & Psychiatry, jnnp-2009.

# Appendix A

## Work packages

<table>
<tr><td>Project: Project proposal and work plan</td><td colspan="2">WP ref: 1</td></tr>
<tr><td>Major constituent: Documentation</td><td colspan="2">Sheet 1 of 7</td></tr>
<tr><td rowspan="2">Short description:<br>Documentation on project description and project work plan.</td><td colspan="2">Planned start date: 25/09/2017<br>Planned end date: 05/10/2017</td></tr>
<tr><td colspan="2">Start event: T1<br>End event: T3</td></tr>
<tr><td>Internal task T1: Project description.<br>Internal task T2: Work Plan.<br>Internal task T3: Document review and approval.</td><td>Deliverables:<br>TFG Proposal and Work Plan</td><td>Dates:<br>05/10/2017</td></tr>
</table>

Table 5: Work package 1

<table>
<tr><td>Project: Information research</td><td colspan="2">WP ref: 2</td></tr>
<tr><td>Major constituent: Documentation and learning.</td><td colspan="2">Sheet 2 of 7</td></tr>
<tr><td rowspan="2">Short description:<br>State-of-the-art analysis.</td><td colspan="2">Planned start date: 01/09/2017<br>Planned end date: 10/10/2017</td></tr>
<tr><td colspan="2">Start event: T1<br>End event: T7</td></tr>
<tr><td>Internal task T1: Study how MRI images are acquired.<br>Internal task T2: Study general traits of MS.<br>Internal task T3: Study how MS diagnostic is done currently.<br>Internal task T4: Study papers in which machine learning is used to predict MS conversion.<br>Internal task T5: Familiarization with Pattern Recognition for Neuroimaging Toolbox (PRoNTo).<br>Internal task T6: Familiarization with Python.</td><td>Deliverables:</td><td>Dates:</td></tr>
</table>

Table 6: Work package 2

| Project: Software development | WP ref: 3 | |
|---|---|---|
| Major constituent: Software | Sheet 3 of 7 | |
| Short description:<br>Implement the algorithm able to predict MS conversion. | Planned start date: 10/10/2017<br>Planned end date: 25/12/2017 | |
| | Start event: T1<br>End event: T3 | |
| Internal task T1: Create a database with the features of every patient.<br>Internal task T2: Study and implementation of different machine learning approaches.<br>Internal task T3: Testing the different approaches. | Deliverables: | Dates: |

Table 7: Work package 3

| Project: Critical Review | WP ref: 4 | |
|---|---|---|
| Major constituent: Documentation | Sheet 4 of 7 | |
| Short description:<br>Document the development of the project and review the initial work plan. | Planned start date: 27/11/2017<br>Planned end date: 01/12/2017 | |
| | Start event: T1<br>End event: T3 | |
| Internal task T1: Writing about progress to date.<br>Internal task T2: Review of the work plan.<br>Internal task T3: Document review and approval. | Deliverables:<br>Critical review | Dates:<br>01/12/2017 |

Table 8: Work package 4

| Project: Test and results assessment | WP ref: 5 | |
|---|---|---|
| Major constituent: Documentation | Sheet 5 of 7 | |
| Short description:<br>Compare results to state-of-the-art techniques. | Planned start date: 10/12/2017<br>Planned end date: 31/12/2017 | |
| | Start event: T1<br>End event: T2 | |
| Internal task T1: Test the algorithm implemented with the testing database.<br>Internal task T2: Compare the results obtained with state-of-the-art techniques. | Deliverables: | Dates: |

Table 9: Work package 5

| Project: Final report | WP ref: 6 | |
|---|---|---|
| Major constituent: Documentation | Sheet 6 of 7 | |
| Short description:<br>This report will describe the entire project, from the theoretical background to the results obtained | Planned start date: 01/01/2018<br>Planned end date: 25/01/2018 | |
| | Start event: T1<br>End event: T2 | |
| Internal task T1: Write the document<br>Internal task T2: Document review and approval. | Deliverables: | Dates: |

Table 10: Work package 6

| Project: Project presentation | WP ref: 7 | |
|---|---|---|
| Major constituent: Documentation | Sheet 7 of 7 | |
| Short description:<br>Prepare an oral presentation to explain the project. | Planned start date: 22/01/2018<br>Planned end date: 01/02/2018 | |
| | Start event: T1<br>End event: T4 | |
| Internal task T1: Prepare the document and the speech.<br>Internal task T2: Review.<br>Internal task T3: Practice the presentation.<br>Internal task T4: Project presentation. | Deliverables: | Dates: |

Table 11: Work package 7

## Milestones

| WP# | Task# | Short title | Milestone / deliverable | Date (week) |
|---|---|---|---|---|
| 1 | 3 | Project Proposal and Work Plan | Project Proposal and Work Plan | 4 |
| 3 | 1 | Create database with the features selected | Database | 9 |
| 3 | 2 | Machine learning classifier | Machine learning classifier | 12 |
| 4 | 3 | Critical Review | Documentation | 13 |
| 5 | 1 | Test of classifier results | Documentation | 16 |
| 6 | 2 | Final report | Documentation | 19 |
| 7 | 4 | Project presentation | Documentation | 20 |

Table 12: Milestones

# Gantt diagram

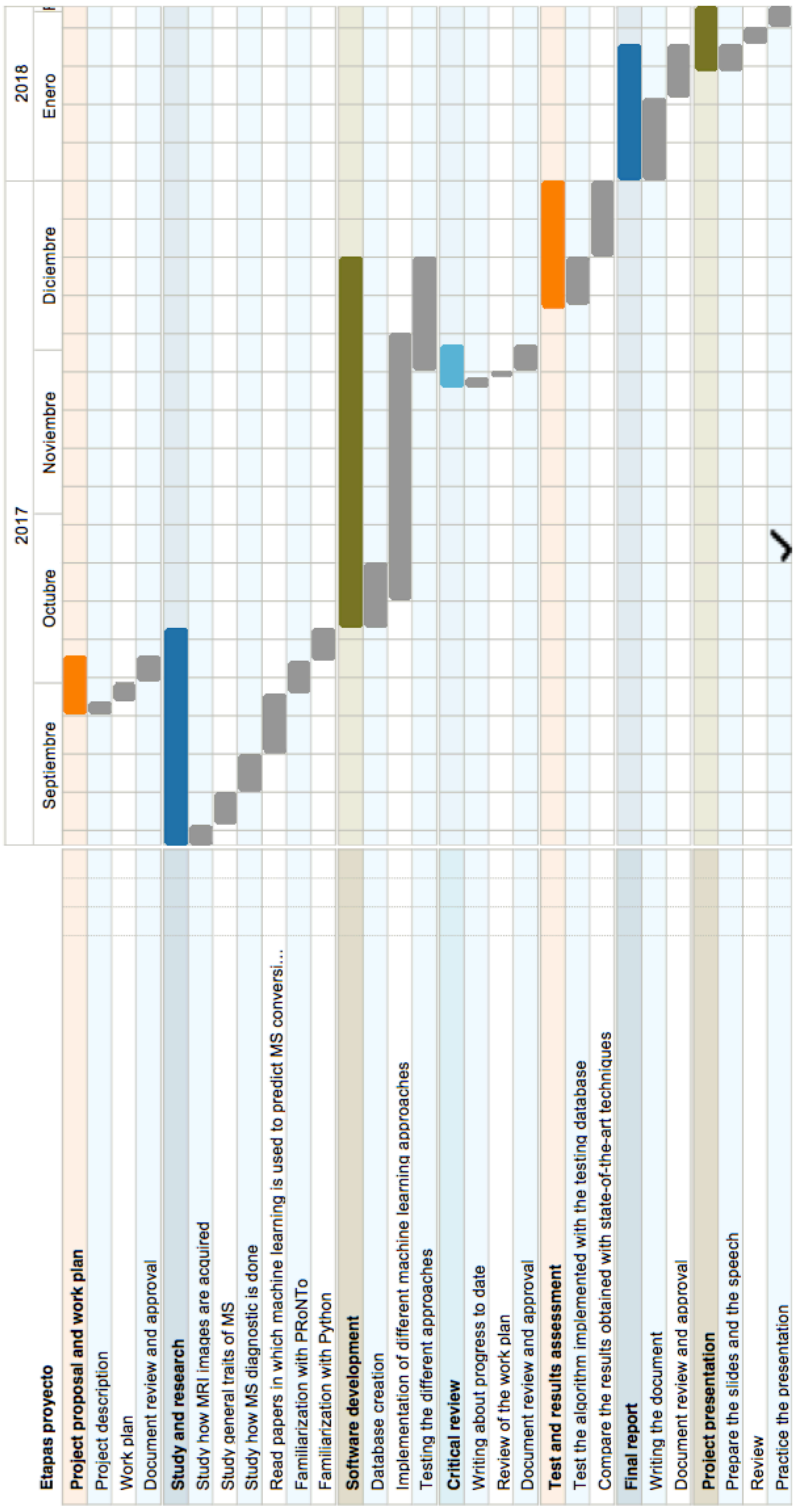


Figure 17: Gantt diagram

# Appendix B

## Frontal

- Superior Frontal.
- Rostral and Caudal Middle Frontal.
- Pars Opercularis, Pars Triangularis, and Pars Orbitalis.
- Lateral and Medial Orbitofrontal.
- Precentral.
- Paracentral.
- Frontal Pole.

## Parietal

- Superior Parietal.
- Inferior Parietal.
- Supramarginal.
- Postcentral.
- Precuneus.

## Temporal

- Superior, Middle, and Inferior Temporal.
- Banks of the Superior Temporal Sulcus.
- Fusiform.
- Transverse Temporal.
- Entorhinal.
- Temporal Pole.
- Parahippocampal.

## Occipital

- Lateral Occipital.
- Lingual.
- Cuneus.
- Pericalcarine.

## Cingulate

- Rostral Anterior (Frontal).
- Caudal Anterior (Frontal).
- Posterior (Parietal).
- Isthmus (Parietal).

# Appendix C

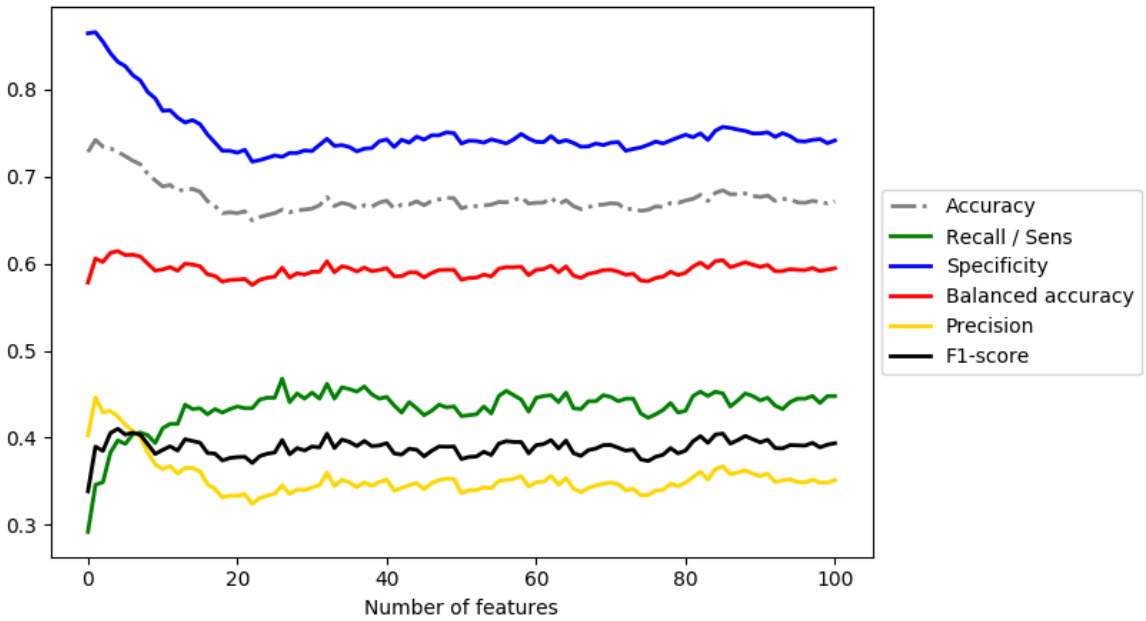


Figure 18: Performance of the classifier using a T-test to rank the features, with the first definition of conversion.

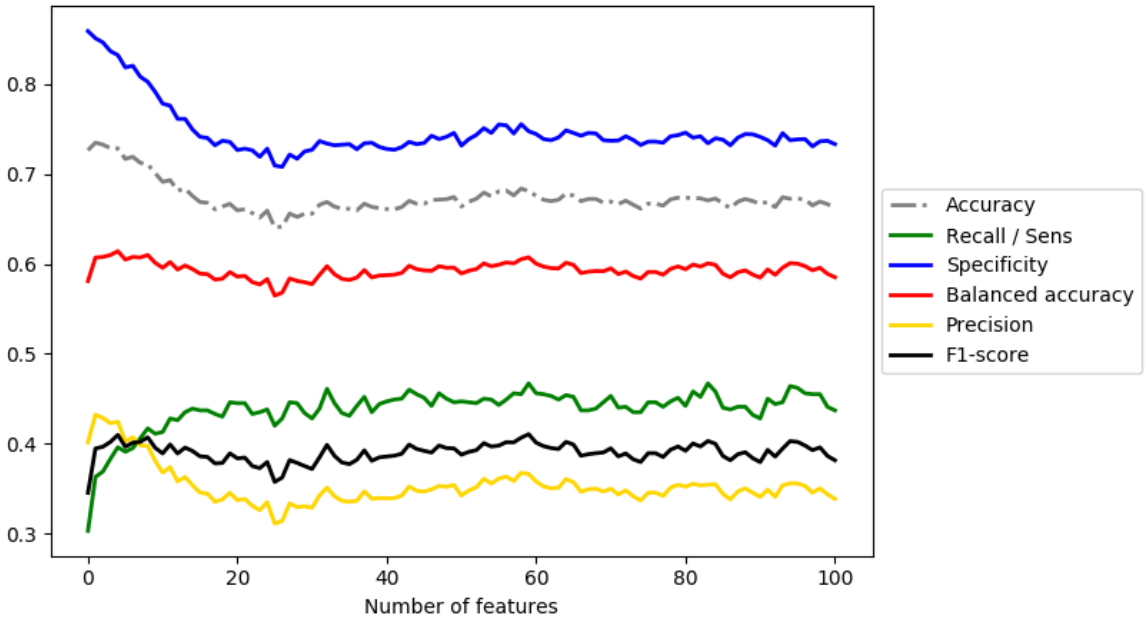


Figure 19: Performance of the classifier using an F-test to rank the features, with the first definition of conversion.

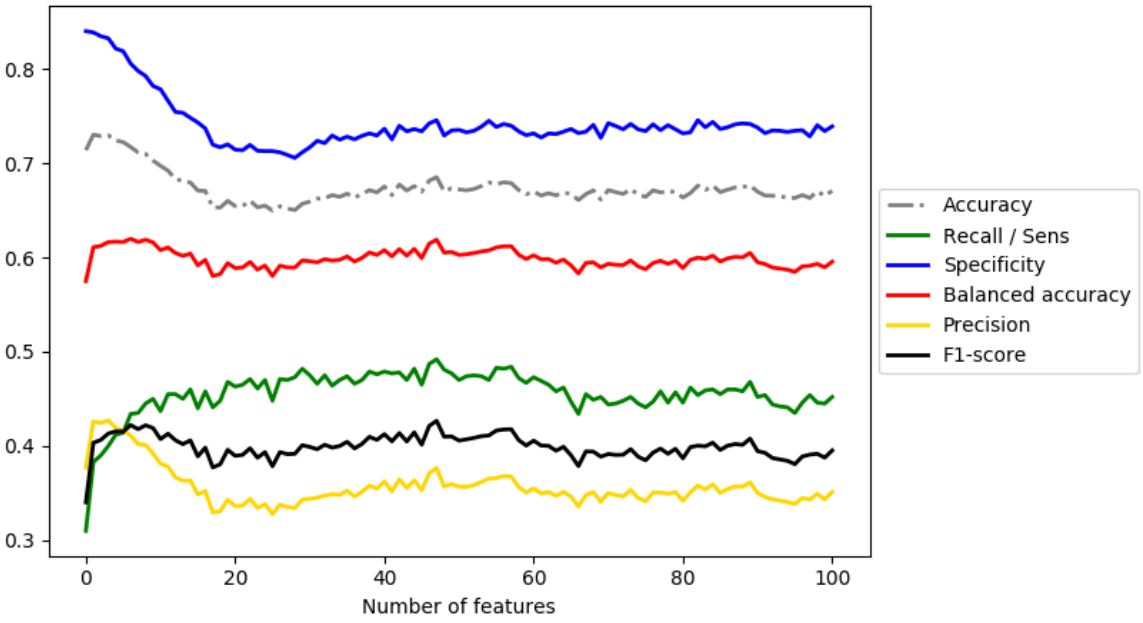


Figure 20: Performance of the classifier using the Pearson correlation coefficient to rank the features, with the first definition of conversion.

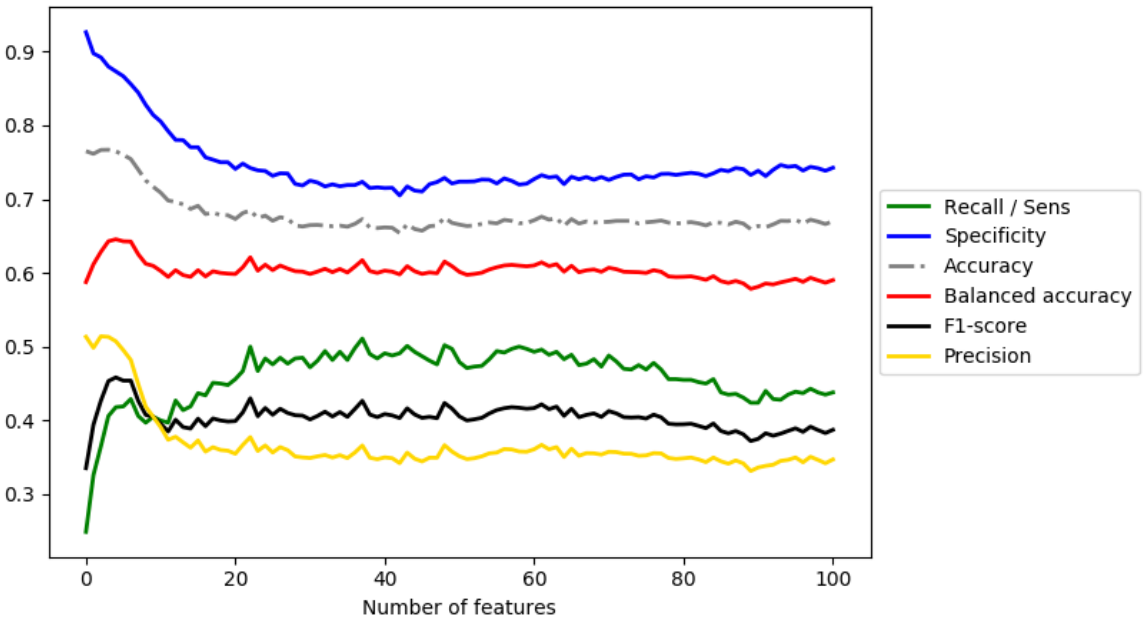


Figure 22: Performance of the classifier using the Spearman correlation to rank the features, with the first definition of conversion.

## Glossary

BA: Balanced Accuracy.
CDMS: Clinically Definite Multiple Sclerosis.
CIS: Clinically Isolated Syndrome.
CSF: Cerebrospinal fluid.
CT: Cortical Thickness.
EDSS: Expanded Disability Status Scale.
GM: Gray Matter.
LASSO: Least Absolute Shrinkage and Selection Operator.
LR: Logistic Regression.
LST: Lesion Segmentation Toolbox.
ML: Machine Learning.
MPRAGE: Magnetization-Prepared Rapid Gradient-Echo.
MRI: Magnetic Resonance Imaging.
PPMS: Primary Progressive Multiple Sclerosis.
ROI: Region of Interest.
RRMS: Relapsing-Remitting Multiple Sclerosis.
SDGM: Dubcortical Deep-Gray Matter.
SPM: Statistical Parametric Mapping.
SVM: Support Vector Machines.
WM: White Matter.
WML: White Matter Lesions.