UNIVERSIDADE D
COIMBRA

Hugo Duarte Oliveira

Evaluation and Prediction of Multiple Sclerosis Disease Progression

Dissertation presented to the University of Coimbra in order to complete the necessary requirements to obtain the Master's degree in Biomedical Engineering, supervised by Professor Cesar Alexandre Domingues Teixeira and Professor Mauro Filipe da Silva Pinto

Outubro de 2020

1 2 9 0

FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE D COIMBRA

Hugo Duarte Oliveira

Evaluation and Prediction of Multiple Sclerosis Disease Progression

Dissertion presented to the University of Coimbra in order to complete the necessary requirements to obtain the Master's degree in Biomedical Engineering.

Supervisors:
César Alexandre Domingues Teixeira (CISUC)
Mauro Filipe da Silva Pinto (CISUC)

Coimbra, 2020

Agradecimentos

Em primeiro lugar, gostaria de agradecer aos meus orientadores Professor Doutor César Teixeira e Mauro Pinto por toda a disponibilidade e orientação que me deram desde o início deste projeto. Agradeço igualmente todo o apoio, motivação e aprendizagens que me proporcionaram ao longo deste ano letivo.

Quero agradecer também à Doutora Sónia Batista e aos Neurorradiologistas Luís Rito Cruz e Mafalda Mendes Pinto pelo trabalho e tempo dispendido a recolher todos os dados clínicos e ressonâncias, sem os quais este trabalho não seria possível.

A todos os meus amigos, colegas de curso e colegas de laboratório, obrigado não só por todo o apoio e motivação na realização deste projeto, como também por todos os momentos vividos e aprendizagens que me proporcionaram ao longo deste ano letivo.

Por último, deixo um especial agradecimento aos meus pais e à minha irmã por toda a motivação e apoio incondicional que me deram ao longo destes anos. Obrigado por todo o esforço que fizeram para este momento fosse possivel.

Resumo

Avaliar e prever a progressão de doentes com Esclerose Múltipla pode ser uma tarefa árdua devido à heterogeneidade de sintomas associada à doença. No presente trabalho foi usada uma base de dados do Centro Hospitalar e Universitário de Coimbra, contendo dados clínicos dos doentes, com o intuito de prever a progressão da doença. Embora a base de dados contivesse uma grande variedade de dados clínicos, foram usados apenas dados relativos à identificação, visitas e ataques sofridos pelos doentes, já que os demais continham bastante informação em falta.

Os primeiros anos de informação da base de dados foram usados para prever o subtipo e a severidade da doença, avaliada através de uma escala de quantificação da condição neurológica (EDSS). Para tal, houve necessidade de proceder a uma seleção dos doentes para cada uma das previsões efetuadas, de forma a incluir apenas doentes com informação suficiente e um subtipo de Esclerose Múltipla de interesse.

Para cada uma das base de dados efectou-se um processo de engenharia de features, no qual foram extraídas diversas outras características através dos campos da base de dados. Para tal, procedeu-se à aplicação de uma segmentação anual que permitiu aumentar significativamente a quantidade de informação disponível.

A cada conjunto de preditores constituintes de cada base de dados, foram aplicados algoritmos de machine learning, incluindo um balanceamento dos dados, a imputação dos valores em falta e a seleção dos preditores mais importantes, bem como uma avaliação do desempenho obtido com cada classificador.

Os resultados obtidos foram avaliados e comparados com resultados de estudos existentes, usando como métricas a área debaixo da curva ROC, a sensibilidade, a especificidade, a G-mean e o F1-Score. Para além disto foram ainda construídos perfis de distribuição das features, de forma a identificar quais as features responsáveis por uma severidade mais acentuada da doença e por cada um dos subtipos de Esclerose Múltipla. Desta forma foi possível comparar os desempenhos usando várias definições de severidade da Esclerose Múltipla, bem como criar um modelo capaz

de prever tanto casos mais severos como casos identificativos de um subtipo de Esclerose Múltipla progressivo. Em termos das features de um subtipo de Esclerose Múltipla foram identificados a *via piramidal*, o valor atribuído ao *sistema funcional piramidal*, o *impacto nas atividades do dia a dia*, a *via do intestino e bexiga* e o valor EDSS como era expectável, embora tenham sido identificadas outras features não esperadas tais como as vias sensoriais, cerebral, cerebelar e tronco cerebral. No que concerne à previsão da severidade, foram identificadas várias features esperadas, de entre as quais o valor de EDSS e vários sistemas funcionais.

Palavras-chave: Esclerose Múltipla; Machine Learning; Previsão da progressão da doença;

Abstract

Assessing and predicting the progression of patients with Multiple Sclerosis can be a difficult task due to the heterogeneity of symptoms associated with the disease. In this work, a database from the Centro Hospitalar e Universitário de Coimbra, containing clinical data of the MS patients, was used to predict the progression of the patients. Although the database contained a wide variety of clinical data, only data related to the identification, visits and relapses were used since the remaining contained a lot of missing data.

The first years of information from the database were used to predict the patients' Multiple Sclerosis subtype and the disease severity, assessed using a neurological condition quantification scale (EDSS). For this, it was necessary to select patients for each of the predictions made, to include only patients with sufficient information and a subtype of Multiple Sclerosis of interest.

For each of the databases, a feature engineering process was carried out, in which several other characteristics were extracted through the fields of the database. To this end, an annual segmentation was applied, which significantly increased the amount of information available.

Machine learning algorithms were applied to each set of features of each database, including several steps such as balance of data, the imputation of missing values and the selection of the most important features, as well as an evaluation of the performance obtained with each classifier.

The results obtained were evaluated and compared with results of existing studies, using as metrics the area under the ROC curve, sensitivity, specificity, G-mean and F1-Score. Besides, features distribution profiles were also constructed to identify those features responsible for a more severe case and each of the Multiple Sclerosis subtypes.

In this way, it was possible to compare the performances using several definitions of

Multiple Sclerosis severity, as well as to create a model capable of predicting more severe cases and identifying cases of a progressive Multiple Sclerosis subtype. In terms of features of a Multiple Sclerosis subtype, the *pyramidal pathway*, the *Score pyramidal*, the *impact in activities of daily living*, the *bowel and bladder pathway* and the EDSS value were identified as expected, although other unexpected features such as *sensory*, *cerebral*, *cerebellar* and *brainstem related features* have been identified. Regarding the severity prediction, several expected features were identified, among which the EDSS value and several functional systems.

Keywords: Multiple Sclerosis; Machine Learning; Disease progression prediction;

Contents

List of Figures

List of Tables

1

Introduction

This chapter, divided into 5 sections, presents the motivation for the development of this master thesis, as well as its context and goals. In the first section, the motivations for this work supported by some statistics concerning Multiple Sclerosis (MS) are presented. The context of this condition, including the prediction benefits, methodologies and difficulties of MS, is described in section 1.2. In section 1.3 the main goals of this master thesis are enumerated and in section 1.4 its contribution is described. Finally, the structure of this document is found in the last section of this chapter.

1.1 Motivation

MS is the most common neurological disorder affecting young adults. This disease affects more than 2 million people worldwide, having an increase higher than 10% in the prevalence over the past 20 years [1]. It is also one of the most frequent causes of paralysis and wheelchair use, negatively affecting the quality of life of the families and patients with this disease in health, social and economic aspects [2]. This disease has still no cure, and the symptoms vary widely for each person, which constitutes a challenge to evaluate and predict the course of a condition with such heterogeneity of manifestations and characteristics.

In addition to that, some patients face a severe condition during the disease course, that leads to different disease characteristics and manifestations. The reasons that cause the transition to a more severe stage remains unknown, and an effective prediction of patients that are susceptible for such modification remains a challenge. It is essential to understand and predict such disease alteration, once the medication prescribed is dependent on disease development, and this uncertainty limits the efficacy of the disease-modifying agents administered.

From these facts, one may conclude that prediction models of MS development using the initial years of follow-up could contribute to an accurate identification of

patients who may benefit of more aggressive disease-modifying agents and closer monitoring, providing in this way valuable assistance to physicians.

1.2 Context

MS is a chronic immune disease, affecting the Central Nervous System (CNS), leading to neurological disabilities. It is characterized by demyelination and axonal loss, which leads to long term functional impairment and disability [3]. Myelin is a layered tissue that surrounds the axons, whose function is to provide faster communication between the neurons.

The course of this disease is quite different and unpredictable for each patient, presenting symptoms that depend on the site and type of lesions. Those symptoms include optic and motor deficits, tremors, difficulty walking, balance disorders, speech difficulties, memory and concentration problems, fatigue, or even paralysis and complete loss of vision in more severe cases [4].

The majority of the patients begins with a relapse remitting (RR) course, characterized by attacks and relapses with full or partial recovery and no disease progression. The disease phenotype tends to change over time, and several patients face an increase of the neurologic deterioration, with or without relapses, which defines the secondary progressive (SP) course. The conversion of RR to SP course remains a challenge due to the nonexistence of clear clinical and imaging criteria defining the beginning of the progressive phase. Furthermore, the uncertainty recognizing this transition also limits the efficacy of the medications administered that are dependent on the disease phenotype [5].

Another characteristic of MS is the different outcomes in different patients, that creates a wide spectrum of the disease, ranging from benign to malignant cases. The benign MS identifies the cases that present little or no disease progression and minimal disability decades after the first manifestations, while malignant scenarios refer to the most severe cases of MS [6]. Even though those definitions are not consensual to every author, and not considered a course of the disease, they are commonly used to identify the clinical characteristics that lead to less severe cases. Recognizing the benign cases also provides valuable information in terms of medication and follow up, once the patients that fit in this group do not need such a closer monitoring and severe treatment when compared to the malignant cases.

Prediction benefits

The main goal of MS prediction is to explore the information of the initial years of the disease to predict the progression in the following years. Such data can assist the physicians to identify the patients who are more likely to gain an increased disability and who would benefit from more rigorous monitoring. With the appropriate algorithm will be possible to accurately predict the disease course and worsening scenarios, identifying those who would take advantage of an early aggressive medication.

Prediction methodologies

Over the past years, the majority of research in MS involves inferential statistics to predict or to test a hypothesis about biological processes involving MS courses, and retrieve clinical interpretation of the results. Statistical inference is quite limited when applied to a wide amount of data, as opposed to Machine Learning (ML) models that are essentially developed for prediction, and are more effective and present a higher performance and accuracy with a large amount of data. Furthermore, ML prediction's allows to identify the most appropriate medical decisions, supporting the choice of the disease-modifying therapies administered, which constitutes a huge improvement in this area [7].

Considering those reasons, prediction models using ML approaches started recently to be applied to MS to obtain a more quick and reliable classification of MS courses, in a standardized way. Those computer-based technologies have emerged in the medical fields, proven to have a good performance in several areas and demonstrating their usefulness when applied to complex datasets, such as those of MS.

The majority of the authors using ML, use the clinical data from the initial years to predict the outcome after several years, in terms of disease progression outcome, measured by the Expanded disability status scale (EDSS), even though the used criteria vary from author to author due to the lack of consensus on a definition of which constitutes a benign/malignant case. There are also some others evaluating if the initial years could be a good predictor of SP development.

Nowadays, prediction models started to use clinical data other than MRI and lumbar punction. Although lumbar punction provides valuable information for diagnosis and that could be used to predict the disease course, this procedure is very invasive and painful. On the other hand, MRI is not painful although the costs of those tests are quite high. For both these reasons, a non-invasive procedure such

as an algorithm capable of predict MS course in an efficient way is the goal of the recent models that started to be used in MS.

Lastly, it is crucial to test the models developed in a real-world scenario. A simulation, in practical conditions, of the algorithms developed, is essential to prove their efficacy and evaluate the different conditions where they can be used and where the medical knowledge extrapolated from the model is maximized.

Prediction difficulties

The nonexistence of a significant number of studies using clinical data constitutes one of the biggest challenges in the approach considered. The majority of the investigation in this area use MRI data or integration of clinical data with MRI, while the number of studies using ML techniques applied to clinical data is quite limited. Moreover, the lack of coherence between studies, which concerns the identification of significant predictors, constitute a major difficulty in the validation of the results.

The definitions used by different authors are not consensual, leading to distinct benign MS definitions used, which causes diversity and inconsistency in the results obtained.

Furthermore, some failures to MS investigation could be related to the quality of the data sets used. The heterogeneity of characteristics considered in the different data sets, the different number of patients and regions where they were selected and even the imbalanced data, due to the predominance of RR type over progressive types, could lead to inconsistencies in the existing studies. Besides, it is essential to mention that the creation of a database of MS is a complex process once information from different sources must be registered, and the follow-up time necessary is long.

The disease itself, characterized by a different course and multiple symptoms depending on the patient, lead to an arduous development of a model adapted to every case and a complex interpretation of the results.

1.3 Main goals

The generic goal of this thesis is to evaluate and predict disease progression, that can be subdivided into several specific goals:

1. Exploration and curation of an MS database from Coimbra Hospital and University Centre (CHUC);

2. Development of an algorithm capable of predict whether a RR patient will progress to SP or not;
3. Development of a model capable of predicting whether a patient will have a more reserved prognosis or not, and explore all the possibilities of severe cases identified in the literature to retrieve coherent conclusions among the studies;
4. Evaluate the real-world applicability of the algorithms created by simulating the performance of the model in a real clinical scenario;
5. Retrieving clinical interpretation from the results/provide new medical insights.

1.4 Contribution

Prior research, regarding the Multiple Sclerosis prediction, has demonstrated the features and model performance of both predictions of conversion to secondary progressive course and disease severity, concerning the 6th and 10th years of progression. This research advances theory on the identification of more reliable features predictive of a severe condition promoted by the consideration of more disease severity classification problems retrieved from literature and the comparison performed between those. The gap identified in the literature on a consensus definition of disease severity, reflects the important insights provided by this study on the effects of different definitions on the performance and features obtained. Finally, to contributing to the goal of increase model performance of both conversion to secondary progressive course and disease severity, it was extracted and included more information regarding patients clinical data on the model. The following thesis contributed partially to the recently accepted paper *Prediction of Disease Progression and Outcomes in Multiple Sclerosis with Machine Learning* accepted for publication in Scientific Reports Journal.

1.5 Structure

This dissertation contains 8 chapters.

Chapter 2 presents background information related to the disease that will be mentioned throughout the document;

Chapter 3 refers to the state of the art of machine learning approaches in MS;

Chapter 4 presents a detailed description of the database used;

Chapter 5 describes the methods used to obtain the results of this master thesis;

Chapter 6 exhibits the results obtained;

Chapter 7 supplies a detailed discussion of the dataset, experimental procedure used and results obtained;

Chapter 8 presents a conclusion of the work.

2

Background Concepts

In this chapter, the main concepts to understand this thesis are introduced. In section 2.1 the risk factors of developing MS are described, including epidemiologic, lifestyle and environmental factors. An overview of the actual diagnosis tests and expanded disability status scale is presented in section 2.2 and 2.3, respectively. The different courses of MS are explained in section 2.4. In section 2.5 the most common therapies used are described while a ML description is provided in the last section.

MS is a disease of the central nervous system affecting more than 2.3 million people worldwide. In multiple sclerosis the body's autoimmune system destroys the myelin sheath, that is an electrical insulation layer that surrounds the nerves. This layer is essential in the CNS, once it is responsible for increase the velocity at which neurons communicate [8]. The demyelination resultant from MS decreases the efficiency and velocity at which the nerves conduct electrical impulses to and from the brain, which causes physical disability and cognitive impairment leading to a substantial reduction in the quality of life of those patients. This condition, that affects young adults in the prime of life, is disabling and irreversible, even though it's progress can potentially be diminished if an early and appropriate treatment is administered [2]. Similarly to other autoimmune diseases, MS is more common in females.

2.1 Risk factors

The cause of MS remains unknown even though the combination of several factors are admitted to trigger MS.

There is evidence that genetic factors are associated with a higher risk of MS. The prevalence is higher in monozygotic twins than in dizygotic twins and persons with both parents having MS presents a higher risk than those with only one parent with MS. The vulnerability is also greater when the maternal side is the one having MS, which is according to the ratio of prevalence of MS dependent on the gender

[9].

Genetic factors only explain a part of the risk growth, whereby lifestyle and environmental factors are believed to trigger MS in combination with a genetic predisposition. The limited sun exposure is commonly associated with an increased risk of occurrence of MS. Ultraviolet radiation (UVR) is used to convert vitamin D to an active metabolite, and several studies address that a lower level of vitamin D is related with an increased risk of incidence of MS [10]. A relation between several infectious agents and MS risk was also suggested in which the epstein-barr virus was identified as a potential risk factor, especially in adolescence [11]. Moreover, unhealthy habits such as smoking habits and adolescent obesity are associated with MS. There has been suggested that there is a clear cumulative dose of smoking and a high body mass index value that presents a strong association with an increase in MS risk [11].

This disease is also highly affected by geographic factors, depending on the latitude considered. The regions farthest from the equator are identified as high-risk regions. North America and some Northern European countries stand with the higher risk, as opposed to North Africa, the Middle East, Latin America, Asia, Oceania, the Caribbean, and sub-Saharan Africa that exhibits the lower risk. Some countries in Europe and Australia present a moderate risk. The lower prevalence in low-risk countries can be a result of the absence of data on those places due to inferior access to medical facilities and less life expectancy time [12].

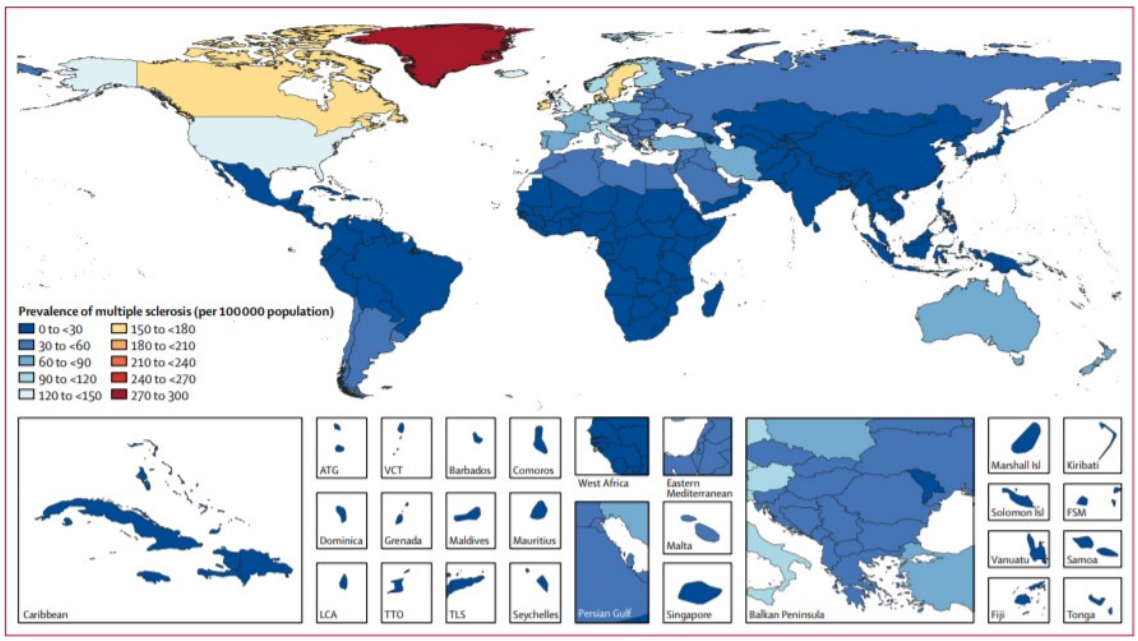

Figure 2.1: Worldwide prevalence of MS [1].

The global occurrence of this condition is substantially different depending on sex, with a clear predominance in females. The prevalence starts to diverge during the adolescence, reaching a ratio of 2:1 in favour of woman around the sixth decade of life. The age is also a crucial factor, once patients with ages between 35 and 64 are more affected, even though this condition affects patients from a wide range of ages [13].

2.2 Diagnosis

The heterogeneity of symptoms and imaging manifestations between MS patients, and the similarities between clinical features of MS and other diseases lead to difficulties in the diagnostic of MS. There isn't still a single diagnostic test, even though an early diagnosis of MS is essential to figure out the medications that should be applied and that can act to prevent relapses and future disabilities. The diagnosis of MS relies on the inclusion of history and physical examinations, imaging and laboratory findings [14].

The gold standard for the diagnosis of MS is the McDonald Criteria 2017 [14] whose objective is to identify the first clinical symptom suggestive of MS and to provide an early categorization of patients as fulfilling the criteria or not. This criteria is based on the number of attacks/relapses, the examination of the Cerebrospinal Fluid (CSF) and identification of dissemination on time and space [15]. To correctly apply the criteria, correct identification of the symptoms is essential. *An attack, relapse or exacerbation is defined as a focal or multifocal inflammatory demyelinating event, occurring in the CNS, that lasts for at least 24h, with or without recovery, and in the absence of fever or infection* [14]. The dissemination in time is the appearance of new CNS lesions over time while the dissemination in space is the occurrence of a multifocal CNS process, characterized by the formation of the lesions in distinct anatomical locations within the CNS. Moreover, the CSF examinations are used to identify CSF-specific oligoclonal bands that could be an indicator of possible MS, although they are not specific for MS [14].

The full McDonald Criteria 2017 is present in Table 2.1.

It is important to note that three different cases could arise from the application of the 2017 McDonald Criteria:

1. Confirmed Multiple Sclerosis - if the criteria are fulfilled and there isn't evidence of a better clarification for the clinical evidence;
2. Possible Multiple Sclerosis - if the 2017 McDonald Criteria are not totally fulfilled, even though Clinically Isolated Syndrome (CIS) lead to the suspicious

Table 2.1: The 2017 McDonald Criteria for Diagnosis of MS [14].

2017 Macdonald Criteria		
Number of attacks	**Number of lesions with objective clinical evidence**	**Additional data needed for a diagnosis of multiple sclerosis**
≥2	≥2	No further tests are required to demonstrate dissemination in space and time
≥2	1 (as well as clear-cut historical evidence of a previous attack involving a lesion in a distinct anatomical location)	No further tests are required to demonstrate dissemination in space and time
≥2	1	Dissemination in space demonstrated by an additional clinical attack implicating a different CNS site or by MRI
1	≥2	Dissemination in time demonstrated by an additional clinical attack or by MRI OR demonstration of CSF-specific oligoclonal bands
1	1	Dissemination in space demonstrated by an additional clinical attack implicating a different CNS site or by MRI AND Dissemination in time demonstrated by an additional clinical attack or by MRI OR demonstration of CSF-specific oligoclonal bands

of MS;

3. Not Multiple Sclerosis - if there is evidence of a better diagnosis for the clinical evidences.

2.3 Expanded disability status scale (EDSS)

The EDSS scale, proposed by Kurtzke [16], is the most frequently used scale in MS, allowing to measure the progression of the disability of the MS patients, varying from 0 (no symptoms) to 10 (death by MS) by half steps. This scale is composed of grades attributed to the 8 different functional systems, in which the impairment of each one is evaluated and used to calculate the final score of disability. The functional systems are networks of neurons with distinct physiological functions and symptoms. The group of systems considered in the EDSS scale and the corresponding functions are described below:

1. **Pyramidal** - Involved in muscle weakness and voluntary movements;
2. **Cerebellar** - Related with coordination of movements and balance. It's also related to symptoms such as ataxia and tremor;
3. **Brain Stem** - Responsible for problems with speech and swallowing, normally associated with the influence of cranial nerves;
4. **Sensory** - Reflects problems associated with loss of sensations below the head;
5. **Bowel and Bladder** - Responsible for incontinence and retention;
6. **Visual** - Impairment of the visual acuity;
7. **Cerebral** - Associated to problems with thinking and memory. Also reflects mood perturbations and concentration problems;

8. **Other or Miscellaneous** - Related to other symptoms that do not fit in any of the other functional systems such as fatigue.

Those functional groups are graded from 0 (normal) to 5/6 (maximal impairment) except for other or Miscellaneous system that is dichotomous, meaning that higher grades correspond to a higher disability [16].

The first steps (0 to 3.5) in the scale are calculated according to the grade of the functional systems (FS) while the following steps take also in consideration the impairment of mobility. The different steps of the EDSS scale are described below and represented in Figure 2.2 [16]:

- **EDSS 0** - Normal neurological exam, in which all FS graded 0 (except for Cerebral where grade 1 is accepted);
- **EDSS 1** - No disability and minimal signs in one of the FS (grade 1), excluding Cerebral grade 1;
- **EDSS 1.5** - No disability and minimal signs are presented in more than one functional system (grade 1), excluding Cerebral grade 1;
- **EDSS 2** - Minimal disability in one FS, in which one of the FS is graded with 2 and the others with 0 or 1;
- **EDSS 2.5** - Minimal disability in two FS, in which two of the FS are graded with 2 and the others with 0 or 1;
- **EDSS 3** - Moderate disability in one FS (one FS grade 3 while the others 0 or l), or mild disability in three or four FS (three or four FS grade 2 while the others 0 or 1) though fully ambulatory;
- **EDSS 3.5** - Fully ambulatory but presenting moderate disability in one of the FS (one grade 3) and one or two FS grade 2; or two FS grade 3; or five FS grade 2 (others 0 or 1);
- **EDSS 4** - Fully ambulatory without aid, self-sufficient. One of the FS graded with 4 while the others with 0 or 1 or combinations of lesser grades that exceed the limits of the previous steps. Able to walk without aid or rest some 500 meters;
- **EDSS 4.5** - Fully ambulatory without aid, up and about much of the day, able to work a full day, may otherwise have some limitation of full activity or require minimal assistance; one of the FS graded with 4 while the others with 0 or 1 or combinations of lesser grades that exceed the limits of the previous steps. Able to walk without aid or rest for some 300 meters;
- **EDSS 5** - Ambulatory without aid or rest for about 200 meters; disability severe enough to impair full daily activities; one of the FS graded with 5 while

the others with 0 or 1 or combinations of lesser grades that exceed the limits of step 4;

- **EDSS 5.5** - Ambulatory without aid or rest for about 100 meters; disability severe enough to preclude full daily activities. One of the FS graded with 5 while the others with 0 or 1 or combinations of lesser grades that exceed the limits of step 4;
- **EDSS 6** - Intermittent or unilateral constant assistance required to walk about 100 meters with or without resting. More than two FS are graded with more than 3;
- **EDSS 6.5** - Constant bilateral assistance required to walk about 20 meters without resting. More than two FS are graded with more than 3;
- **EDSS 7** - Unable to walk beyond about 5 meters even with aid, essentially restricted to wheelchair; wheels self in standard wheelchair and transfers alone; up and about in wheelchair some 12 hours a day. More than one FS is graded with more than 4; Pyramidal grade 5 alone, even though this case is rarer
- **EDSS 7.5** - Unable to take more than a few steps; restricted to wheelchair; may need aid in transfer; wheels self but cannot carry on in standard wheelchair a full day; may require a motorized wheelchair. More than one FS is graded with more than 4;
- **EDSS 8** - Essentially restricted to bed or chair or perambulated in wheelchair, but may be out of bed itself much of the day; retains many self-care functions; generally have effective use of arms. Several FS are graded with more than 4;
- **EDSS 8.5** - Essentially restricted to bed during the majority of the day; has some effective use of arm(s) and retains some self-care functions. Usually, several FS are graded with more than 4;
- **EDSS 9** - Helpless bed patient that is able not only to communicate effectively but also to neither to eat and swallow. Almost all FS are graded with more than 4;
- **EDSS 9.5** - Totally helpless bed patient that is not able to communicate effectively, neither to eat nor swallow. Almost all FS are graded with more than 4;
- **EDSS 10** - Death due to MS.

2.3.1 Limitations of EDSS

Even though EDSS is specifically defined for MS and is the most used scale in assessing the disability of patients with MS, this scale has some limitations. The

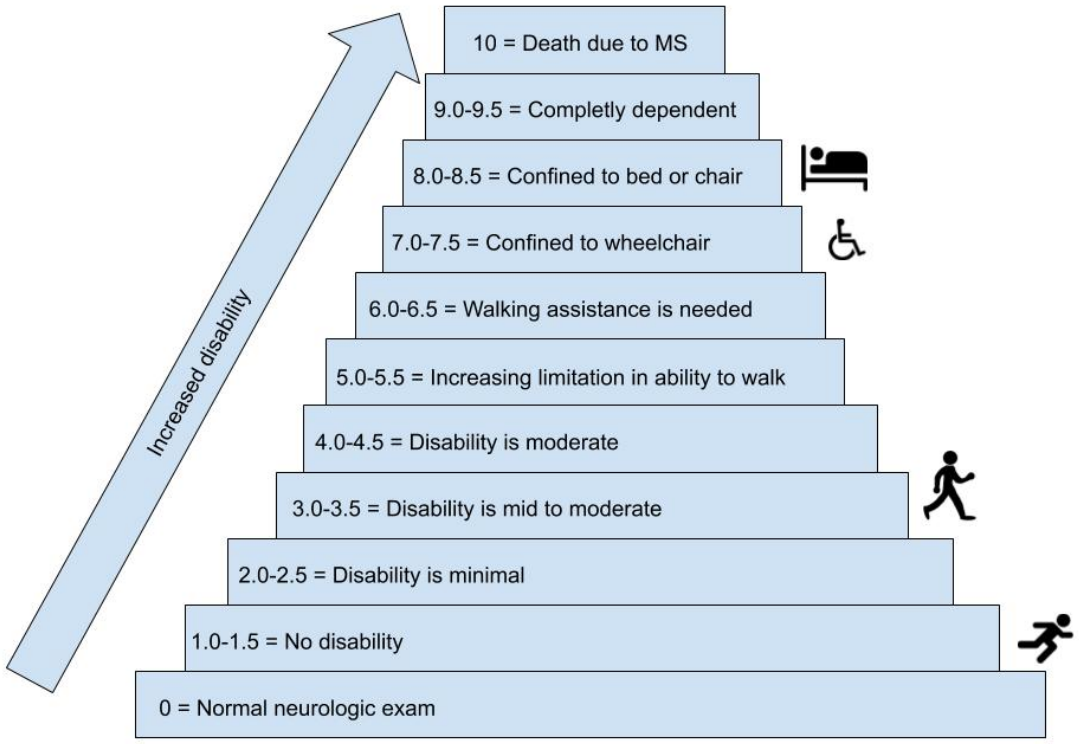

Figure 2.2: Representation of the EDSS scale [16].

scale is not homogeneous, because the steps are not representative of the same impairment, and is not linear in which concerns the time that each patient stays in each of the steps. Moreover, the scale is not fully objective once there is a variation in the scores attributed by the same doctor or different doctors to each situation. Also, the scale is biased for motor functions, since those functions are more considered in the scale than cognitive tasks [17].

2.4 Courses

Due to the heterogeneous manifestations of MS, a distinction of the clinical courses of MS patients was defined to allow easier communication between physicians and more accurate classification of the patients into different courses. In 1996, the neurologists grouped the clinical courses of MS into four groups :

1. **Relapse-Remitting (RR)**: The RR type is the most common form, affecting around 85 % of the patients with MS. The course is characterized by several relapses, that consist of attacks or development of new symptoms, followed by periods of complete or partial recovery, called remission periods [18].

2. **Secondary Progressive(SP)**: SP patients present an initial relapsing-remitting course with several relapses and remissions followed by a progressive state characterized by possible occurrence of occasional relapses, minor remissions and plateaus. [19].
3. **Primary Progressive (PP)**: This type of MS affects around 10-15 % of the patients. It is characterized by a gradual increase in the disease progression from the onset, with possible plateaus and minor improvements, but without any relapses or remissions. In these cases, the disease tends to occur later in life, around 10 years afterwards, and the predominance in the female sex is not verified. It is important to emphasize that in these cases it is harder to distinguish this condition from other neurological diseases. In which concerns to the treatment of MS, the patients with PP type tend to have resistance to the treatments, and doesn't exist a treatment capable of slow the progression of such condition [20] [21].
4. **Progressive Relapsing (PR)**: The PR is the least common manifestation of MS, affecting less than 5 % of the patients with MS. It's defined by an initial progression with periods of relapses, although there are no remissions on this type [22].

Over the past years, there was a rise in the knowledge of MS pathology and improvements in imaging and biological marker research that lead to the necessity of reviewing the existing phenotypes of MS. In 2013, a revision [23] of the disease courses was proposed, in which MRI findings, biomarkers and clinical factors were used to achieve a more accurate and homogeneous classification. Although the main MS phenotype remained, there was an introduction of new characterizations of the MS clinical courses, described in Table 2.2 [24].

Table 2.2: New definitions introduced in 2013 revision criteria.

New definitions introduced in 2013 revision criteria		
Clinical Isolated Syndrome (CIS)		First clinical presentation of the disease that demonstrates signs of inflammatory demyelination
Disease activity	Active	Evidence of relapses or episodes of new or increasing neurological dysfunction or the presence of new T2 or gadolinium-enhancing lesions over a specified period of time
	Not active	No evidence of disease activity
Disease progression	Progressive	Evidence of disease worsening during a defined amount of time (at least one year) in which an increase in neurological dysfunction/disability occurs without clear recovery, even though phases of stability may occur
	Not progressive	No evidence of disease worsening during a specified period of time (at least one year)

One of the modifications was the introduction of the CIS, that could be considered as MS, although for that it is necessary to further satisfy the criteria of

dissemination in time and space. Although a CIS may not lead to a diagnosis of MS, usually those who are diagnosed with MS start with a CIS.

Furthermore, the concepts of disease activity and disease progression were introduced. The disease activity was established in all MS courses, resulting in a division of both CIS, RR, SP and PP into active or non-active. Differently, disease progression is a concept that only describes progressive phases of MS, segregating SP and PP into progressive or not progressive. [25][24][23].

Analysing the Figure 2.3 it is possible to observe the representation of the increased disability in function of time, to each one of the courses after the 2013 revision, which allows to clearly distinguish the different MS courses. It is important to note that the previously considered PR course type of MS has been eliminated because those patients are now classified as PP with disease activity [24][23].

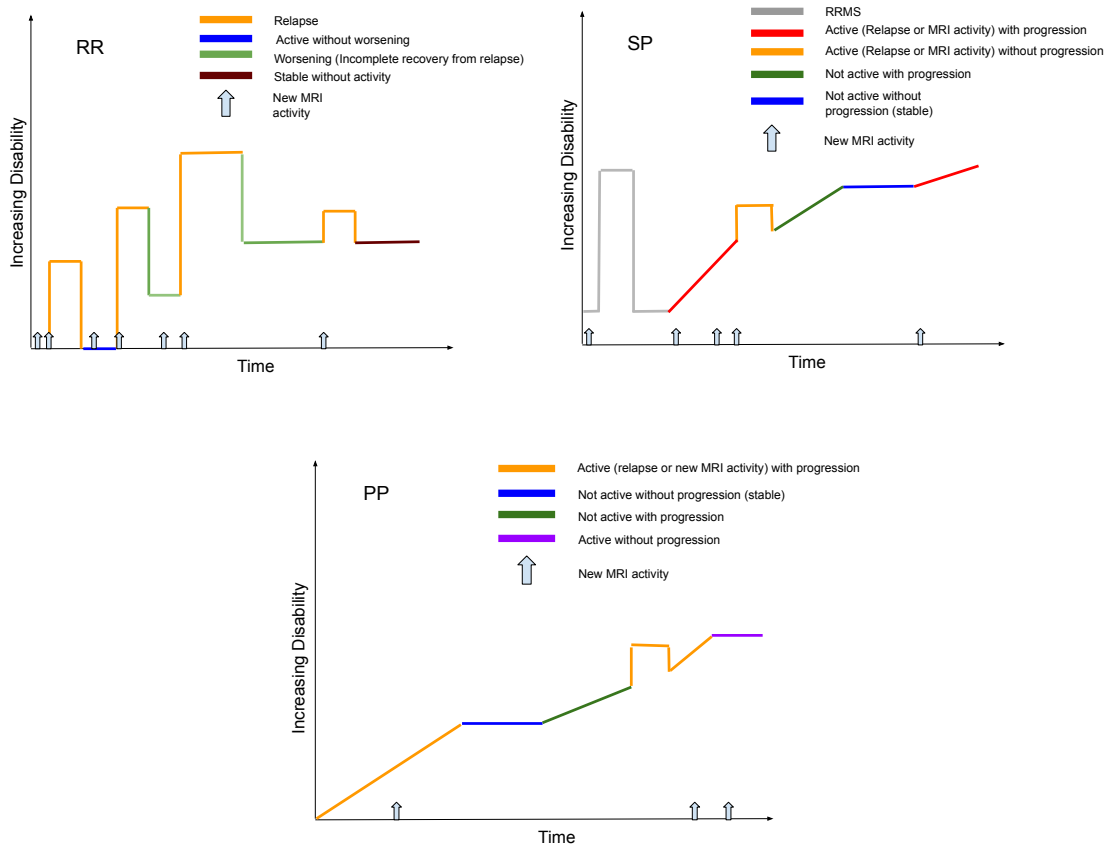

Figure 2.3: MS courses after 2013 revision [25].

The CIS, that is part of the relapse-remitting spectrum, can be either active or not active. The not active CIS will be considered with such designation until the next change in the MRI or episode occurrence, while the active CIS can be considered RR if the Macdonald criteria are verified. If a patient is diagnosed with

RR, it can also be either considered an active or not active type, depending on the existence of clinical relapse rate and image findings during a defined period. Moreover, if a progressive state of the disease is identified, the patients can be diagnosed with secondary progressive (having an initial relapse phase followed by a progressive state) or primary progressive (having a progressive state from the onset) [24]. In these cases, four different situations can occur throughout the disease duration as observed in Figure 2.3:

- **Active and with progression** - occurs when a gradually worsening is verified and a new attack happens;
- **Active but without progression** - occurs when a patient has a new attack but is condition is not worsening;
- **Not active but with progression** - occurs when the patient is gradually worsening even though a new attack is not verified;
- **Not active and without progression** - refers to the stable form of MS.

Moreover, in this master thesis only patients with RR, SP and PP courses are admitted in order to simplify and consider courses with a reasonable amount of patients included.

2.4.1 Benign/malignant MS

The mild course of MS defined as benign MS is characterized by a little or even absence of progression and symptoms after several years of progression [26]. This form of MS is different from both progressive and non-progressive categories, and there is still no consensus on a global definition for benign MS. The terminology benign/malignant is advised to be used with caution, once the terms are not consensual and even after several years of apparent benign MS a worsen can be verified.

Furthermore, the terms worsening and progressing were distinguished to clarify the meaning of each one. In the case of the patients whose disease is advancing due to the frequent occurrence of relapses and incomplete recovery of those, the term disease worsening should be used. On the opposite, the term disease progressing should be used for the cases of progressive disease where the patient's condition is worsening gradually over time [24].

2.5 Therapy

The therapy of MS is based on disease-modifying agents whose objective is to diminish the duration and frequency of relapses, provide symptomatic relief and prevent the advance of the disease to a progressive state, even though there isn't still any drug capable of curing MS [18].

The medication can be divided into acute relapses management and disease-modifying treatments. The acute relapses are commonly treated with intravenous (IV) methylprednisolone or dexamethasone, that are corticosteroids with a rapid onset of action and few adverse effects. These corticosteroids are administered from 3 to 5 days and provide not only treatment for acute exacerbations, but also for reducing the corresponding relapse duration. Such drugs are used to treat relapses of both RR and SP patients [18].

In terms of disease-modifying treatments, there are several treatment options commonly used. Eight Food and Drug Administration (FDA) approved medications are presented in Table 2.3.

Table 2.3: FDA approved drugs [18].

Name	Administration	Dosing Frequency	Disease type	Action
Avonex (Interferon beta-1a)	Intramuscular	Once weekly	RR	Diminish the incidence of relapses
Rebif (Interferon beta-1a)	Subcutaneous	Three times weekly	RR	Diminish the incidence of relapses
Betaseron (Interferon beta-1b)	Subcutaneous	Every other day	RR	Diminish the incidence of relapses
Extavia (Interferon beta-1b)	Subcutaneous	Every other day	RR	Diminish the incidence of relapses
Glatiramer acetate	Subcutaneous	Once daily	RR	Reduce the rate of relapses
Mitoxantrone	Intravenous	Short infusion (about 5 to 15 minutes) every 3 months	RR and SP	Prevents worsening course
Natalizumab	Intravenous	1-hour infusion every 4 weeks	RR	Reduces the rate of relapses
Fingolimod	Oral	Once daily	RR	Reduce the rate of relapses and delay the progression

From the Table 2.3 it is possible to note that the medications administered are dependent on the clinical situation, varying between medications used to diminish the relapse rate or to prevent disease progression. In fact, for patients diagnosed with RR course, the goal of the treatment is to reduce the rate and severity of relapses, while for SP patients the medication is used to prevent progressive worsening of the disease. It should be noted that a patient with SP could also benefit of medications to diminish the incidence of relapses [27].

Although those drugs take several benefits to patients with MS, some of them demonstrate several side effects. For instance, any of the medications containing

interferon beta agents can lead to liver function abnormalities, glatiramer acetate causes several injection-site reactions and chest pain among other side effects. It is important to highlight the fact that several other options of medications are commonly used by physicians, even though those medications have not been approved yet by FDA [18].

The drugs administered for each patient are not usually included in the models for prediction of the type of MS because the medication reflects a personal opinion of the physician on the course of the disease. Besides, if the medication prescribed was considered, it was not possible to know if a given recovery was the result of the drug administration. Even though the non-admission of medication leads to the impossibility of know whether the outcome would have been the same with or without medication and considering that an investigation of the cumulative effect of the medication on the outcome of each patient would also be an interesting approach, in this case, the medication was not included by medical advice. Moreover, considering medication could also lead to a bias in the results. Furthermore, the importance of predicting the disease course consists of the possibility to provide adequate medication to the patients according to the predicted type of MS and ideally, before any medication is given to the patient.

2.6 Machine learning

ML is a class of methods based on computer algorithms that learn from data to make predictions, identify patterns and make decisions. Over the past years, this subset of artificial intelligence started to be applied to several areas.

Supervised ML algorithms are used to learn a target function that best fits input variables to an output variable, to predict new outcomes with the introduction of new inputs. The main goal of the algorithms is to optimize the estimation of the target function, which allows more accurate predictions [28].

This technology usually involves a systematic process to obtain accurate predictions in developed algorithms, consisting of data preparation, classification and performance evaluation.

As can be seen in Figure 2.4 the first step consists of importing the raw data. Such data is then pre-processed, i.e., transformed to a reliable format that may be further used to extract relevant features. Those processes are important to have proper data, in a standardized format, without missing values and outliers. The selected features are further used to classify each sample into one of the labels. Lastly, the classification performance is then evaluated using appropriate metrics

[29].

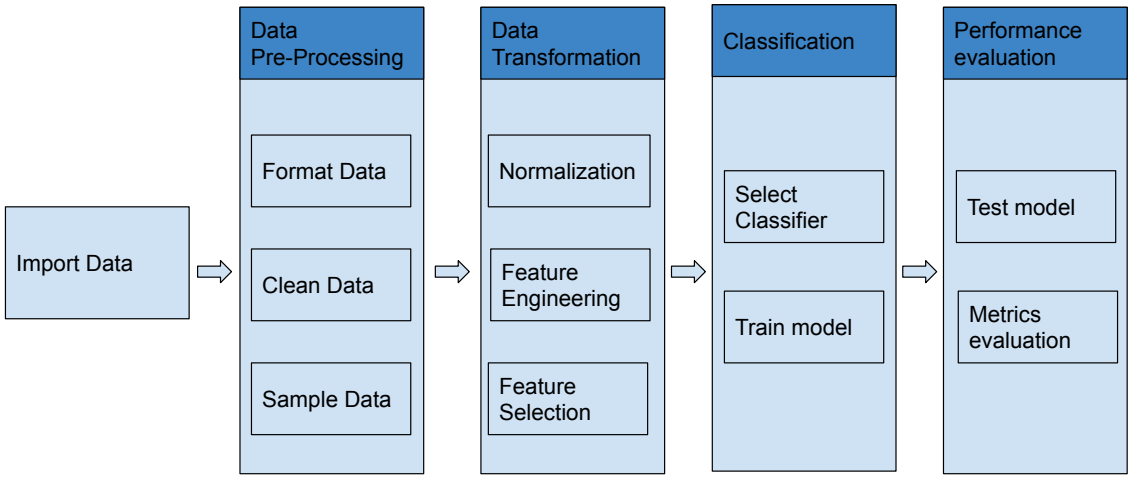

Figure 2.4: Machine learning pipeline.

2.6.1 Data preparation

ML models make predictions based on characteristics learnt from data, where its appropriate use is essential to obtain good performance. After having a defined problem that we pretend to solve using ML predictions, is crucial to identify and select the data that is important for such problem [29]. The ML algorithms can be distinguished in terms of the way they model a problem and learn from the input data, consisting of supervised learning that uses labelled data and unsupervised learning the uses unlabeled data. In this master thesis, there will be used only supervised learning, where an appropriate labeling must be chosen, regarding the definition of the problem. Moreover, it is also important to analyze and select the raw data, defined as data in its original form, that can be used for prediction. The supervised learning algorithms use data to identify patterns and builds logic through the training process, making some predictions and correcting themselves when the predictions are wrong [30].

Raw data is usually presented in unwanted formats, containing missing values and with an extended amount of information, whereby the data need to be pre-processed, to be formatted, cleaned and sampled. Furthermore, outlier detection and removal must be performed. The pre-processed data is usually further transformed using feature engineering [31].

2.6.1.1 Feature engineering

A feature is an individual independent characteristic of the data, that describes an attribute and is used as an input in the model [28]. The features are associated to a process of feature engineering that is responsible to improve model performance by converting raw data into features that better represent the underlying problem [32]. When the data contains attributes with different scales and quantities, a process of normalization is essential to create features with uniform scales. The data must also be analyzed to segregate features representing complex concepts into more features or aggregate several features into one more representative. Such processes allow creating more appropriate features from the data, leading to more flexibility and less complexity on the model [33]. Feature engineering is an important process once better features mean higher accuracy and an improvement in model performance.

The feature selection is part of features engineering and is a process to decide the features that are meaningful for the considered problem. In fact, from all considered attributes, there are some redundant and irrelevant for the context while some others are more important for the model. Feature selection algorithms are methods that solve this problem by automatically elect a group of features with more importance due to a higher predictive power [32].

Furthermore, if all features are used, an overfitting situation may occur, in which the model learns the noise of training data, corresponding to particular details that do not provide generalization ability, and leading to a negative impact in terms of model performance. To overcome this question, feature selection methods should be applied to the data, to reduce the complexity of the model by obtaining a reduced amount of features capable of generalizing the problem, providing outputs to sets of inputs that the model has never seen before [34]. In this way, the model is also easier to understand.

Selection methods can be categorized into 3 distinct groups: filter methods, wrapper methods and embedded methods. Those classes reduce the number of features, even though they use different algorithms to achieve that goal.

Filter methods select features from a dataset based on the characteristics of the variables, once the features are filtered before the learning begins. Those methods usually evaluate the features individually and do not take into account the relationship between them. The filter methods rely on a score, calculated using measures such as correlation, distance metrics and consistency metrics, that is attributed to each feature, indicating whether the feature should be selected (if contains a high score) or removed [35]. Those methods are computationally inexpensive. Examples of filter methods include chi-square test and correlation coefficient scores [36].

As opposite, wrapper methods evaluate a subset of features by analyzing all possible feature subsets based on the performance of a given ML algorithm. For each of the possible subsets, a new model needs to be trained to select the best performing model according to a performance metric [37]. This fact leads wrapper methods, that detect the interaction between variables, to high performances even though they are computationally expensive. Examples of wrapper methods are BFS (Breadth First Search), SFS (Sequential Forward Selection), SBS (Sequential Backward Selection) and SVM RFE (Support Vector Machine Recursive Feature Elimination) [36].

Finally, the embedded methods perform feature selection automatically during the model training. Embedded methods take into consideration the interaction of features like wrapper methods do, are less computationally expensive than wrapper methods and more accurate than filter methods [34]. Those methods are subdivided into regularization (lasso regression and ridge regression) and tree-based methods (decision trees). Regularization methods add a penalty to the different parameters of a model to reduce its freedom, once the penalty is applied depending on the number of features. In this way, regularization methods make the model more robust to noise and increase its generalization because it shrinks many features to almost zero and find the most coherent group of features [36].

2.6.1.2 Missing values imputation

In datasets, it is common to encounter several fields without any information. This fact is denominated as missing data, and it is important to impute the missing values using the existent information, in order to be possible to use the samples having missing data.

The first aspect that should be considered is the reason why the data is missing. In fact, data can be missing completely at random (MCAR) if there isn't any association between a missing value and any other value of the dataset, and is considered as Missing at random(MAR) if the data is missing due to other variables of the dataset [38]. It can also be admitted that the values are missing not at random (MNAR) if the values are missing due to a relationship between the reason because the value is missing and it's values [39].

There are several methods for the imputation of missing data. Some approaches involve an imputation using the mean or the median of the existing values of that variable / feature. It is also common to use the most common value among the variable values to replace the missing data.

Although there are several other imputation methods it is worth to mention

the imputation using KNN, that is a simple approach that looks for the k closest neighbours and uses the mean of those points to impute the missing values.

2.6.2 Classification

After the data pre-processing and transformation, further steps are required to accomplish the objective of developing an accurate classification model.

An important stage is the data partitioning, that in its standard form consists of a subdivision of the original dataset into a training, validation and testing subset. In this step, a percentage of the data is used to learn the data and fit the parameters (train the model), in order to apply such model to new data afterwards (test the model) and obtain the results of its performance. Furthermore, there are usually cases of imbalanced datasets, consisting of a big difference in the number of samples for each class. In order to overcome such difference, methods as oversampling and undersampling are usually applied.

The final step consists of selecting an appropriate classifier to our problem.

2.6.2.1 Data preparation

In some classification problems, data is imbalanced, i.e, an unequal distribution of the examples exists between classes. Such a difference can influence the model to learn the patterns of the majority class while ignoring the minority class, which constitutes a problem because the minority class is usually the class whose predictions provide more interest [40].

Although it could be thought that the imbalanced dataset result from a poor data acquisition, it commonly reflects the real scenario.

To balance the classes, it is usual to create a new version of the training dataset, in which the distribution of the classes is modified to obtain a more balanced distribution. Such approach is named resampling, and it is subdivided into oversampling and undersampling. The oversampling is a method in which the examples of the minority class are duplicated to achieve more equality between classes. An example of such method is random oversampling that involves a random selection of the minority class that will be duplicated and added to the training dataset. On the opposite, undersampling methods work by deleting samples of the majority class, in which the samples are randomly selected for random undersampling [41].

A different approach consists of weight balancing where a higher weight is attributed to the less represented classes. This type of balancing can be used if it is desirable to attribute a higher weight to the minority class due to their higher

importance.

It is important to partition the dataset into training, validation and test sets. Data partitioning consists in choosing the percentage of data that is used in each one of those sets. A certain amount of data should be used to identify relationships between the target function and the outcome (training set), but there is also the need to verify if the relationships are accurate and if the model fit new data. For such reasons, a percentage of data is used to provide an unbiased evaluation of a model fit on the training dataset while tuning model hyperparameters (validation set), and the remaining data is used to a final estimation of the performance after model structure and parameters are completely fixed (testing set) [42].

The partition of the data is also essential to evaluate the fit of the model and avoid overfitting. In fact, Cross-Validation (CV) methods are commonly used to partition the data. K-fold CV [43] is an example of CV, mostly used in small datasets, that divides a dataset into k parts, using k-1 parts to train the model and the remaining part to validate the model. The algorithm performs k iterations in order to guarantee that all k parts will be used in for test in one iteration. Such algorithm ensures that the data is completely explored once all the samples are used for training and after CV the best models, if sufficient data is available, are applied in a testing set.

2.6.2.2 Classifiers

Several classification algorithms can be used and the appropriate choice depends on the existing problem and available data. Some classifiers are listed below [44] [45]:

1. **Decision tree** - This classifier is based on a construction of a flowchart-like tree structure in which the internal nodes represent features, the branches represent a decision rule and the leaf nodes represent the classification into one of the classes. This algorithm classifies each point by moving from the top of the tree (root) until the bottom of the tree, choosing in each case a side of the branch [46].
2. **Logistic regression** -This type of regression analysis, that is usually applied when the dependent variable is binary, fit the data to a logistic function and use such function to predict the occurrence of an event. This method allows to obtain the probability $p(\mathbf{x})$ of a sample x_i belonging to a certain class. The function used to obtain the probability is indicated in the equation (2.1),

where w_i is the regression coefficient value concerning the feature x_i [47].

$$p(\mathbf{x}) = \frac{1}{1 + e^{-(w_0 + \sum_{i=1}^{n} w_i x_i)}}. \tag{2.1}$$

3. **Fisher Linear Discriminant Analysis (LDA)**- This method is a classification and dimensionality reduction technique that transform the features from higher-dimensional space into a lower-dimensional space. The maximum number of dimensions that is possible to project data is C-1, where C is the number of classes. Dimensionality reduction is achieved by projecting samples in a lower-dimensional space that is chosen by considering the projection that maximizes classes separability and minimizes the variance within the class, known as Fisher criterion [48]. Classification is then possible by defining a linear classifiers over the projected data.
4. **K-Nearest Neighbors (KNN)** - This algorithm stores the instances corresponding to training data points, and use feature similarity to predict the values of new data points based on how closely it matches the points stored. The label is determined by the most prevalent class among the corresponding k nearest neighbours [49].
5. **Support Vector Machine (SVM)** - This algorithm is commonly used for binary classification and prediction, and it works by assuming that each instance corresponds to a point in space to find a hyperplane that maximizes the separation margin between samples of the classes. Even though there are several possible hyperplanes, the algorithm selects the one that provided the maximum separation margin between data points of both classes, to obtain a more confident classification for future data points [50]. The new instances are classified based on their position, in which they are classified according to the side of the hyperplane they are placed. The classification model can be expressed as follows [51]:

$$f(z) = sgn(w^T z + b).$$

Where w is the normal vector to the decision hyperplane, z the new sample to be classified, and b is a bias term. w and b are obtained by minimizing:

$$\Psi = \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{N} \xi_i, \text{ subjected to}$$

$$yi \left(w^T x + b\right) \geq 1 - \xi_i; i = 1, ..., M.$$

Where $\{x_i, y_i\}$ (with i = 1; $\cdots$, M) is the training data, being x_i input feature vectors and $y_i \in \{-1, +1\}$ the class labels. ξ is a quantification of the degree of misclassification, and C defines the influence of ξ on the minimization criterion Ψ.

2.6.3 Performance evaluation

The classification models developed need to be further evaluated and analyzed. In order to evaluate how well the model fit data, several metrics are used to evaluate prediction [52]:

1. **Confusion matrix** - The confusion matrix allows to evaluate the model performance, where four measures are obtained by assuming that the problem is binary and there exists a class named as positive class, and there exists other class named as negative.
   - **True Positives (TP)** : number of samples correctly classified as the positive;
   - **True Negatives (TN)** : number of samples correctly classified as negative;
   - **False Positives (FP)** : number of samples classified as positive that are negative;
   - **False Negatives (FN)** : number of samples classified as negative that are positive.
2. **AUC** - The area under the ROC curve is also used for binary classification problems. As can be seen in Figure 2.5, the ROC curve is a graphic plot of the TPR (true positive ratio or sensitivity) in the function of the FPR (False Positive Ratio or 1-specificity), used to classify each event into one class. In order to classify based on a given classifier decision function, a threshold needs to be defined, in which if the value is above the threshold it belongs to one class and if it is below the threshold it belongs to the other class [53].

   Several values of thresholds can be used to develop a ROC curve, and the AUC is the area under the curve. Basically, if the value of AUC obtained by the equation 2.2 is close to 1, the model classify correctly each class, if the value is close to 0, it classify the classes in an opposite way, and if the value is close to 0.5 means that the model has no class separation.

$$AUC = \int_{x=0}^{1} ROC(x)dx. \tag{2.2}$$

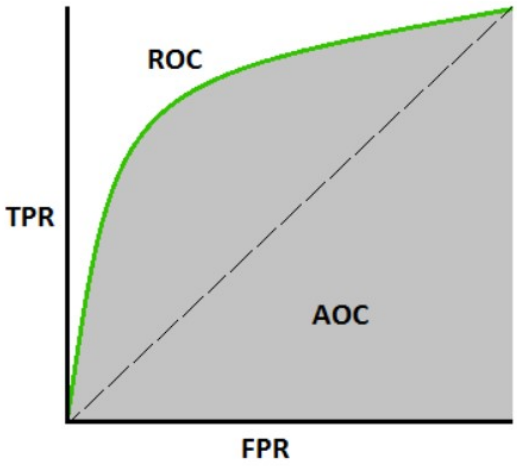

Figure 2.5: ROC curve [53].

3. **Accuracy** - Accuracy represents the number of predictions correctly made out of all predictions made. For this reason, the accuracy of 100% corresponds to the highest accuracy possible, and the number of events incorrectly classified make this number decrease.

$$Accuracy = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{2.3}$$

   Even though this metric is commonly used, it is not appropriate in imbalanced classification once if the model attributes all examples to one class, the accuracy is high, but the model is not effective [54].

4. **Sensitivity** - This metric corresponds to the true positive rate, identifying the proportion of positives that are correctly identified as such

$$Sensitivity = \frac{TP}{TP + FN} \tag{2.4}$$

5. **Specificity** - This metric corresponds to the true negative rate, identifying the actual negatives that are correctly identified as such

$$Specificity = \frac{TN}{TN + FP} \tag{2.5}$$

6. **F1-score** - This metric is the harmonic mean of precision and recall (sensitivity) as can be seen in the equation 2.7. The precision corresponds to the ratio of correctly predicted positive observations to the total predicted positive observations as can be seen in equation 2.6. This metric takes both false positives and false negatives into account and is more realistic than accuracy

in imbalanced class distributions, once accuracy usually benefits the classifiers with poor performance for the minority class [54].

$$Precision = \frac{TP}{TP + FP} \tag{2.6}$$

$$F_1\text{-}Score = \frac{2 * Precision * Recall}{Precision + Recall}. \tag{2.7}$$

7. **G-mean** - This metric is the geometric mean of the sensitivity and specificity, as can be seen in the equation 2.8

$$G\text{-}Mean = \sqrt{\text{sensitivity } * \text{ specificity}} \tag{2.8}$$

Similarly to the F1-Score, this metric is more appropriate to imbalanced classification than accuracy [54].

It is important to evaluate the results obtained, testing it in real-life situations, with data unseen by the model in both pre-processing, feature selection and classification. In fact, there are models that learn the training data in detail and perform too well in the training data, but very poorly in new data. The performance evaluation allows to identify such problems and validate the model.

3

State of the art

This chapter provides a brief explanation of the current state of the art on the prediction of MS progression. Firstly, there are presented different definitions used for benign MS. Afterwards, an overview of the different features identified as predictive of a benign course are listed with the correspondent study where they were identified. Studies identifying demographic features and the existing ML approaches to predict worsening cases are also described and the results achieved are enumerated. Moreover, studies focusing on the Secondary Progressive development prediction are presented, along with a description of the identified features and best results.

3.1 Definitions of benign MS

The definition of benign MS had varied throughout the years. As reviewed by Ramsaransing GSM et al. [6] it started to be defined as "It is not rare to encounter complete remission which is hoped to be definitive" in 1872. Although the definitions of benign MS are different from author to author, the introduction of the EDSS scale led to a higher consensus among authors once they started to use such scale to define benign MS. In literature, definitions such as variations of the disability score from 0 to 2 [55], [56], from 0 to 4 [57] and with a variation in the length of disease duration from 5 years to 20 years are encountered. Even though the benign course in MS was defined in a consensus meeting in 1996 as a disease in which the functional systems remain fully functional after 15 years of disease onset, the most common definition used is an EDSS score ≤ 3.0, after a disease duration of at least 10 years [58], [59], [26], [60], [61],[62]. A progression index defined by the division of EDSS score by disease duration is also addressed by some authors [63] as a meaning for benign course, in which a progression index <0,2 corresponds to a benign course

$$\text{progression index} = \frac{\text{EDSS score}}{\text{disease duration}}. \tag{3.1}$$

In the Table 3.1 the different definitions of benign MS used by each author are defined, as the number of patients used in each case and the frequency of patients identified as benign cases in each study. It's possible to identify that the benign frequency varied from 6 to 71 %, which emphasize the fact that different definitions change considerably the interpretation of the problem, and consequently the distribution of the patients in each group.

Table 3.1: Definitions of benign multiple sclerosis proposed by several authors organised by definition and chronological order, respectively.

Author; year	Definition of Benign used	Benign Frequency(%)	Nº Patients
McAlpine ; 1961 [64]	Without restriction of activity for normal	32	241
Leibowitz and Alter; 1970 [65]	employment and domestic purposes	20.7	266
Sheperd ; 1979 [66]	but not necessarily symptom-free after a	7.2	557
Amato et al. ; 1999 [67]	follow-up period of more than 10 years	29	224
Kurtzke et al. ; 1977 [55]	Kurtzke DSS between 0 and 2 after more than 10 years of disease duration	20.1	527
Lauer and Firnhaber ; 1987 [56]		19	363
Poser et al. ; 1982 [63]	Deterioration of not more than one grade of disability during a 5-year period (progression index $\leq$0.2)	21.9	2056
Hutchinson ; 1986 [58]	EDSS $\leq$ 3 after 10 years of illness	54	60
Thompson et al. ; 1986 [59]		42	400
Hawkins et al. ; 1999 [26]		19.9	181
Cabre et al. ; 2001 [60]		19.4	62
Ramsaransing et al ; 2001 [61]		26,7	2204
Mandriol et al. ; 2008 [62]		59,4	64
Lublin et al. ; 1996 [68]	Disease in which the patient remains fully functional in all neurologic systems 15 years after disease onset		
Krisztina et al. ; 2001 [69]	EDSS between 0 and 3 points	15	248
Perini et al. ; 2001 [70]	EDSS score 3 with normal neurophysiological examination in a period of 15 or more years after clinical onset of the disease	6	500
Benedikz et al. ; 2002 [57]	EDSS <4 after 15 years	64	372
Kalanie et al. ; 2003 [71]	EDSS score $\leq$ 3.0 after 10 years and score $\leq$ 2 after 5 years	14	265
Portaccio et al. ; 2009 [72]	EDSS $\leq$ 3 and a disease duration of 15 years	71	63
Rovaris et al. ; 2011 [73]		49.3	369
Hviid et al. ; 2011 [74]		14,6	1265
Calabrese et al. ; 2013 [75]		32,1	140
Bueno et al. ; 2017 [76]	EDSS $\leq$ 3 and a disease duration of 20 years	49	61
Yijun Zhao et al. ; 2017 [77]	Increased EDSS 1.5 at up to five years after baseline visit		1693

From those studies, different methods were used and several predictive features were identified.

Hawkins et al. [26] developed a study to evaluate both the characteristics and prognostic factors capable of identifying the patients that follow a benign course of MS. Clinical and demographic variables were considered and t-test and χ^2 test were used to compare the different patient's groups. There weren't clear conclusions about the symptoms related with benign MS course, even though optic neuritis and sensory disturbance were identified as the most common features characterizing that

course.

Poser et al. [63] used statistical tests to identify factors predictive of a benign MS, defined by a progression index ≤ 0.2 (given by the quotient between the present disability and the disease duration). Patients with optic neuritis as the initial symptom, and with an age ≤ 39 years were identified as patients with more probability to develop a benign course. The RR course was also identified as a feature predictive of a benign course.

G. S. M. Ramsaransing et al. [6] developed a review about the existing studies addressing the benign course of MS. From the studies analysed, several of them found clinical factors associated with a favourable MS course namely low age at onset, relapse-remitting course, short-duration onset symptoms, complete first remission, and long time between onset and first relapse while a high number of functional systems at onset and high EDSS at onset were identified as possible indicators of not benign MS.

Tatjana Reynders et al. [78] also performed a literature review in which the features predictive of a benign course were identified. Features such as early relapse rate, disease phenotype at onset and low EDSS values at 5-10 years from the onset in RR course are considered as factors with strong evidence of being features predictive of a benign course, while there was no evidence of age and gender predict benign MS independently. It is suggested that both age and gender are interacting with the disease phenotype. Moreover, it is mentioned that mental and cognitive features are not usually identified in studies, which may be caused due to the biased EDSS scale to motor functions.

It was also studied cases of worsening condition, in which worsening was defined as an EDSS increase ≥ 1.5 at up to five years after the baseline visit. Yijun Zhao et al. [77] used ML approaches to classify the patients into worsening or non-worsening groups. SVM and logistic regression were selected as classifiers. It was used demographic, clinical and MRI data of the first two years to predict the EDSS increase up to five years. The best result was achieved for 1 year of follow up with MRI data having a sensitivity of 0.71, a specificity of 0.68 and an accuracy of 0.69. Some features such as EDSS score, disease activity score, sensory, cerebellar, visual, mental, bowel/bladder and brainstem FS scores were identified as features with predictive power to progressive MS while race, ethnicity and family history of MS are determined as features predictive of non-progressive cases.

In a different study Yijun Zhao et al. [79] evaluated different models using SVM, logistic regression, random forest, XGBoost, LightGBM, Meta-L and compared the results obtained with the different classifiers. It was used the first two years to pre-

dict up to five years and classify the patients into worsening/ non-worsening group. Using LightGBM a sensitivity of 0.78, a specificity of 0.68 and overall accuracy of 0.70 was achieved. Features such as pyramidal functions, variations in the EDSS, disease category and some others were also identified as having predictive power.

In the Table 3.2 an overview of the different features identified in literature with the predictive power of benign or malignant MS (used to describe MS patients who reach significant level of disability in a short period of time) is present.

Table 3.2: Predictive features of a benign/malignant course identified in literature.

Feature	Benign MS Authors	Non Benign MS Authors
Age at onset:		
High:	----------	----------
Low:	Poser et al. [63] , Thompson et al. [59], Hawkins et al. [26], Ramsaransing et al. [61]	----------
Gender:		
Male:	----------	Portaccio et al. [72]
Female:	Hawkins et al. [26]	----------
Symptoms at onset:		
Visual:	McAlpine [64], S A Hawkins et al. [26] , Poser et al. [63] , Yijun Zhao et al. [77], Ramsaransing et al. [61]	Amato et al. [67],
Pyramidal:	Yijun Zhao et al.[79]	Amato et al. [67], Kurtzke et al. [55], Ramsaransing et al. [61], Mandrioli et al. [62]
Cerebellar:	Leibowitz and Alter [65], Yijun Zhao et al. [77]	Amato et al. [67], Kurtzke et al. [55], Lauer and Firnhaber[56]
Sensory:	S A Hawkins et al [26]., Sheperd [66], Amato et al. [67], Hawkins et al. [26], Mandrioli et al. [62], Yijun Zhao et al. [77]	----------
Bowel & Bladder:	Yijun Zhao et al. [79]	Amato et al.[67]
Brainstem:	McAlpine [64], Yijun Zhao et al.[77]	----------
Motor:	----------	----------
Spinal:	----------	----------
Mental:	Yijun Zhao et al. [77]	----------
Onset characteristics:		
Low number of functional systems involved:	Kurtzke et al. [55]	----------
High number of functional systems involved:	----------	Amato et al. [67], Kurtzke et al. [55], Lauer and Firnhaber [56]
High EDSS:	Tatjana Reynders et al. [78]	Amato et al. [67], Kurtzke et al. [55]
RR course	Sheperd [66], Poser et al. [63], Tatjana Reynders et al. [78]	----------
Progressive course	----------	Amato et al. [67], Lauer and Firnhaber [56]
Relapse:		
Long time between onset symptoms and next relapse:	Kurtzke et al. [55], Thompson et al. [59]	----------
Low relapse rate:	McAlpine et al [64], Amato et al. [67], Ramsaransing et al. [61]	----------

3.2 Early prediction of SP cases

Due to the initial similarities between RR and SP patients, some authors evaluated the clinical factors capable of distinguishing the patients who will evolve to SP course and those who will remain with the RR course.

Adrian Ion-Margineanu et al. [80] developed a study to classify MS patients into one of the existing MS types. A ML algorithm was applied to clinical data combined with lesion loads and magnetic resonance metabolic features. Using an LDA, it was achieved an F1-Score of 87% for the distinction of RR vs SP and an F1-Score of 85 % for the distinction of RR vs PP. It was possible to verify that the lesion loads were better when used to differentiate between RR and SP forms.

Boiko et al. [81] established a correlation between the course of MS and the number of relapses during the first year of the disease. Moreover, this author identified that for subjects with later first relapse, the transition to SP course is longer.

Bergamaschi et al. [82] developed a bayesian model to estimate the risk of evolving to SP course, in which polisyntomatic onset and late age at onset were identified as unfavourable factors and female gender was associated with lower risk. Furthermore, factors such as type of onset, motor and sphincter relapses and an early increase in disability were considered as key features for assessing the risk of advance to SP course.

3.3 Overview of MS research

Analyzing several studies, it is possible to verify that the research approaches and models developed have changed over time. Older studies focus their efforts to infer about associations between biological processes and disease manifestations, or relations between disease courses and symptoms. In these studies, inferential statistics are used to infer about such relations and try to identify possible mechanisms or symptoms of interest to understand how the disease manifestations occur and how benign cases and SP development can be identified in an early stage.

In more recent research, it can be noted that the inferential statistics have been mostly replaced by ML approaches that find generalizable predictive patterns capable of predicting the disease course [77][79] [80]. Models of pattern recognition, neuronal networks and fuzzy logic, for example, have been used for prognosis of disease course, focusing on forecasting future outcomes [83]. Accompanied by this change in research objectives, several new features with predictive power for benign cases and disease courses classification have arisen. ML also allowed to work with

larger quantities of clinical data in a standardized way.

Considering the actual knowledge regarding the MS pathology, a consensual and complete definition for benign/malignant MS should be created. Actually, there is no possibility to compare some results from different studies predicting benign cases, due to the use of diverse definitions. It is relevant to note that some models can present better results because the definition used is less rigorous. Such criteria heterogeneity causes ambiguity in results obtained. In fact, using the EDSS scale, the new knowledge about MS pathology, the new improvements in clinical factors and the most recent imaging techniques, a more concrete definition of benign MS should be proposed to standardize future MS investigation.

4

Dataset Description

A database, from the Neurology Department of Centro Hospitalar e Universitário de Coimbra, whose use for research proposal has been approved by the Ethical Committee of the Faculdade de Medicina da Universidade de Coimbra was used in this master thesis, containing information from 1134 patients. The data, for each patient, includes identification, demographic information, relevant events in MS course (relapses, the onset of the progressive course, symptoms identified in each visit) and treatments administered. Among the total data is possible to distinguish between static and dynamic data.

The patients are periodically monitored, with regular visits to the clinic every 3 months or every 6 months, depending on the severity of the cases. Besides, if a patient feels bad or forecast that he will be unwell, he or she can go to the clinic in an unplanned way, which corresponds to an irregular visit. The patients still go to the clinic when they have relapses, where they can be hospitalized or not. Although it is administered continuous treatments on the patients, in those cases where relapse occurs, is also administered corticosteroids.

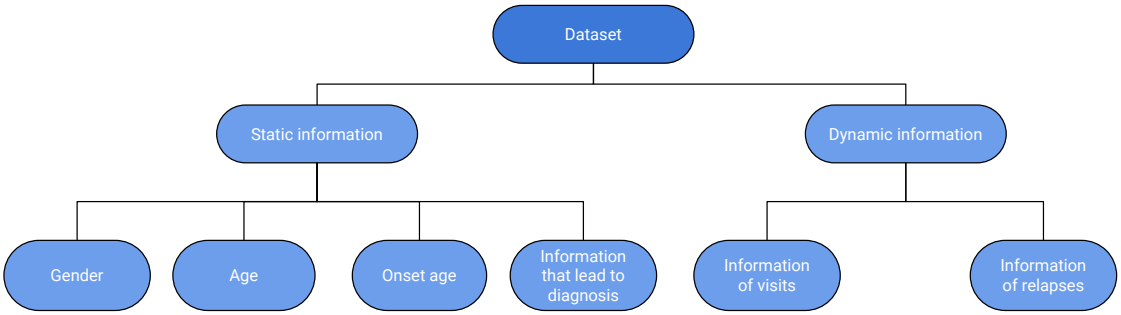

Figure 4.1: Type of information contained in the dataset.

As observed in Figure 4.1 the static information corresponds to the data that do not change with time, including demographic data such as gender, age and age at onset and clinical information that lead to the diagnosis, including CSF examinations, the measure of evoked potentials and MRI exams. As opposite, the dynamic

information is the data that varies with time, and variations are observed in different periods. This data contains information about visits and relapses that can be different from visit to visit and from relapse to relapse.

The entire information provided in the database, was organized in 12 groups, as described in Table 4.1

Table 4.1: Description of the original database xls sheets.

Information groups	Description	Used / Not used - cause
Identification	Information of patient such as demographic information	Used
Concomitant Diseases	Information of other diseases of the patients	Not used - Available for few patients
Family History	Information of family history diseases	Not used - Available for few patients
Visits	Information regarding visits at hospital	Used
MRI	Information of a reduced number of patients	Not used - Available for few patients
CSF	Includes important information regarding the patient condition retrieved from CSF examination	Not used - Plenty of missing data
Evoked Potential	Includes important information regarding the patient condition retrieved from the evoked potential examination	Not used - Plenty of missing data
Laboratory Test	Includes important information regarding the patient condition retrieved from laboratory tests	Not used - Plenty of missing data
Relapses	Information about relapses	Used
Adverse Event	Information of adverse events suffered by the patients	Not used - Available for few patients
Pregnancy	Information of the pregnancy such as complications occurred	Not used - Available for few patients
Treatments	Information of the treatments administered to the patients	Not used - Could lead to bias

Table 4.1 demonstrates that several groups were excluded due to a significant amount of missing data verified. Some others were not used since the information available concerned only a small number of patients. Among used groups, it was possible to identify several cases of patients containing missing data in core fields, such as onset and diagnosis date and symptoms from visits and relapses. Furthermore, there were also verified cases of patients not diagnosed with MS.

To provide more details about the fields used to model development, a description of each feature contained in the curated database is described in the following sections, divided by group.

4.1 Identification

In this section, the fields of the Identification group are presented and explained:

1. **Birth Date**: Birthdate of patients used with diagnosis date, onset date and secondary progressive diagnosis date to calculate the patient's age at these events;
2. **Gender**: Patient's gender;
3. **Age at onset**: Calculated using birth date and onset date to obtain the mean age at onset of the patients considered;
4. **Diagnosis Date**: Corresponds to the moment when MS was for the first time diagnosed. It is used along with the birth date and secondary progressive diagnosis date;
5. **Secondary progressive diagnosis date**: Date at which patient having RR was diagnosed with SP course;
6. **Supratentorial**: Boolean field, indicating the existence of MS manifestations related to the supratentorial region at the time of diagnosis;
7. **Optic pathways**: Boolean field, indicating the existence of MS manifestations related to the optic pathways at the time of diagnosis;
8. **Brainstem-Cerebellum**: Boolean field, indicating the existence of MS manifestations related to the brainstem and/or cerebellum at the time of diagnosis;
9. **Spinal cord**: Boolean field, indicating the existence of MS manifestations related to the spinal cord at the time of diagnosis;
10. **Clinical findings**: Boolean field, indicating the existence of clinical evidence in the initial MS manifestations at the time of diagnosis;
11. **MRI**: Boolean field, indicating the existence of visible MS manifestations in the MRI exams at the time of diagnosis;
12. **Evoked potentials**: Boolean field, indicating the existence of MS manifestations in evoked potentials test at the time of diagnosis;
13. **CSF**: Boolean field, indicating the existence of MS manifestations in the lumbar puncture exam at the time of diagnosis;
14. **MS course**: Boolean field, indicating whether the actual MS course of the patient is RR, SP or PP.

4.2 Visits

In this section, the fields of the Visits group are presented and explained:

1. **Visit date**: Corresponds to the date in which the visit to the hospital occurred;
2. **Routine**: Boolean field, indicating whether it was a routine visit or an emer-

gency one;

3. **Suspected relapse**: Boolean field, indicating if a relapse is expected or not;
4. **Score pyramidal**: Field with values varying from 0-6, corresponding to the momentaneous score for the Pyramidal FS;
5. **Score cerebellar**: Field with values varying from 0-5, corresponding to the momentaneous score for the Cerebellar FS;
6. **Cerebellar weakness**: Boolean field, indicating if there were MS manifestations of cerebellar weakness;
7. **Score brainstem**: Field with values varying from 0-5, corresponding to the momentaneous score for the Brainstem FS;
8. **Score sensory**: Field with values varying from 0-5, corresponding to the momentaneous score for the Sensory FS;
9. **Score bowel**: Field with values varying from 0-5, corresponding to the momentaneous score for the Bowel and Bladder FS;
10. **Score visual**: Field with values varying from 0-5, corresponding to the momentaneous score for the Visual FS;
11. **Visual symptom**: Boolean field, indicating whether a patient had visual symptoms or not;
12. **Score mental**: Field with values varying from 0-5, corresponding to the momentaneous score for the Mental FS;
13. **Score ambulation**: Field with values varying from 0-11, corresponding to the ambulatory capacity of the patient;
14. **Gdataxia**: Boolean field, indicating if gait disturbances of ataxia (lack of voluntary coordination of muscle movements) occurred. It is related to the dysfunctions in the Cerebellum;
15. **Dysaesthesiae**: Boolean field, indicating if a symptom of dysaesthesia (Abnormal and disagreeable sensations experienced in the absence of stimulation) occurred;
16. **Ataxia lower extremity**: Boolean field, indicating if a symptom of ataxia in the lower extremities occurred. It is related to dysfunctions in the Cerebellar functional system;
17. **Paresthesiae**: Boolean field, indicating if a symptom of paresthesiae (burning or prickling sensation) occurred;
18. **CognitionPB**: Boolean field, indicating if cognition perturbation occurred due to MS. It is related to dysfunctions in the Cerebral functional system;
19. **GdParesis**: Boolean field, indicating if gait disturbances of paresis (partial

loss of voluntary movement) occurred. It is related to dysfunctions in the Pyramidal functional system;

20. **GdSpasticity**: Boolean field, indicating if gait disturbances of spasticity occurred;
21. **MwUpperExtremity**: Boolean field, indicating if a symptom of muscular weakness in the upper extremities occurred. It is related to dysfunctions in the Pyramidal functional system;
22. **MicturitionPb**: Boolean field, indicating if perturbations in the micturition (involuntary release of urine) were verified. It is related to dysfunctions in the Bowel and Bladder functional system;
23. **Fatigue**: Boolean field, indicating if patients felt fatigue;
24. **mwLowerExtrimity**: Boolean field, indicating if a symptom of muscular weakness in the lower extremities occurred. It is related to dysfunctions in the Pyramidal functional system;
25. **MoodPb**: Boolean field, indicating if perturbations in the mood were verified;
26. **EDSS**: Value of EDSS varying from 0 to 10 attributed by the physician at each visit.

4.3 Relapses

In this section, the fields of the Relapses group are presented and explained:

1. **Relapse date**: Corresponds to the date in which the relapse occurred;
2. **Impact ADL Functions**: Indicates the impact of five different activities of daily living. Those ADL include bathing (personal hygiene and grooming), dressing (dressing and undressing), transferring (movement and mobility), toileting (continence-related tasks including control and hygiene), eating (preparing food and feeding);
3. **Recovery**: Corresponds to the amount of recovery after the relapse, admitting values of complete recovery, partial recovery and none recovery;
4. **Severity**: Indicates the intensity of the relapse, admitting values of mild, moderate and severe;
5. **CNS Pyramidal Tract**: Boolean field, indicating if MS manifestations related to Pyramidal tract occurred;
6. **CNS Brainstem**: Boolean field, indicating if MS manifestations related to Brainstem occurred;
7. **CNS Bowel Bladder**: Boolean field, indicating if MS manifestations related

to Bowel and Bladder tract occurred;

8. **CNS Neuropsycho Functions**: Boolean field, indicating if neuropsychologic MS manifestations occurred;
9. **CNS Cerebellum**: Boolean field, indicating if MS manifestations related to Cerebellum occurred;
10. **CNS Visual Functions**: Boolean field, indicating if MS manifestations related to visual functions occurred;
11. **CNS sensory functions**: Boolean field, indicating if MS manifestations related to sensory functions occurred.

4.4 Patient selection

Considering the objective of predicting benign/malignant cases and RR to SP transition, and taking into account the limitations of the database used, several steps were performed to obtain an optimal selection of patients to each prediction scenario.

As can be seen in 4.2 a), the original database contained information of 1134 patients. Among those patients, 258 were not diagnosed with MS, which cause their automatic exclusion, once they were not useful to fulfil the objective of this study. Moreover, the fields with complete missing data were also excluded.

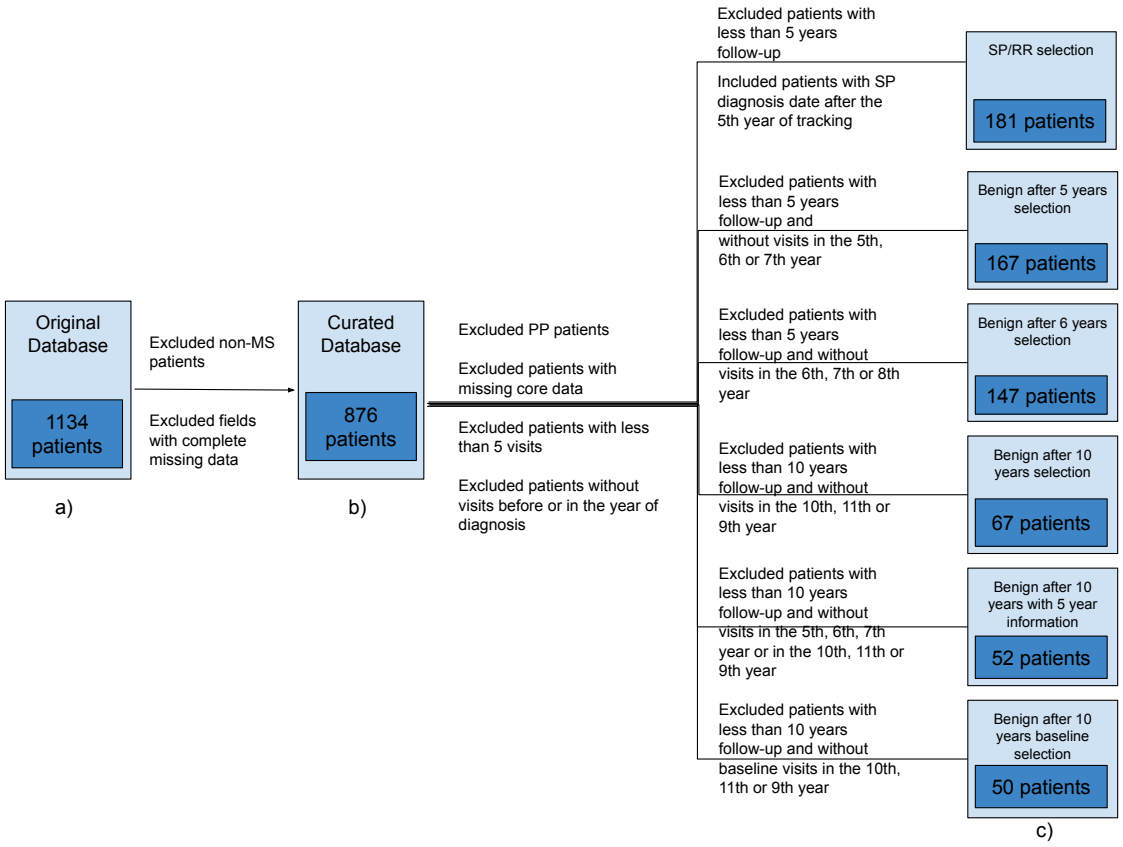

Figure 4.2: Iterative steps of the selection criteria of the patients.

The exclusion of the patients due to the aforementioned reason lead to the creation of a curated database identified in Figure 4.2 b). In total, the curated database contained clinical data from 876 patients, in which 752 were diagnosed with RR sub-type, 85 with SP sub-type and 39 with PP sub-type. The PP patients were not included in further steps once they represent a minority of the cases and are characterized by clinical manifestations different from other courses in an initial phase. The patients with missing data in both onset date, diagnosis date and SP diagnosis date (only considered for RR/SP selection) were also not included due to the impossibility to retrieve critical information to predict the outcomes.

It was verified that the number of visits and relapses was quite variable from patient to patient, existing a significant difference in the amount of data between some patients. Moreover, the majority of the patients had no information of visits previous to or in the year of the diagnosis date, which makes it impossible to retrieve the information from the initial years. Lastly, the years of follow-up are also not the same for every patient. Since different labels will be compared it is necessary to select an appropriate subset of patients for each case, that contains enough data and a required number of years of follow-up.

To integrate information from different sources and uniform the amount of data considered to each patient, several inclusion criteria were defined to guarantee that sufficient information is used in each case. The inclusion criteria common to every label is enumerated below:

1. Diagnosed with MS;
2. Diagnosed with RR or SP sub-types;
3. Information of at least 5 visits;
4. Records of patient follow-up at least since the diagnosis date.

Such criteria select the patients most likely to have information from the onset of the disease, by selecting those with visits before or in the year of the diagnosis date. In the patients where the diagnosis date was missing, the onset date was used to perform such selection. If both dates were missing, the patients were not included due to the impossibility to retrieve critical information to predict the outcomes. For the RR/SP selection, there was introduced an exception to the previous statement, once the patients containing the first annotated visit only after RR diagnosis, but the SP diagnosis date after the fifth year of tracking were also included. In this way it is guaranteed that the information used still belonged to an RR subtype. Moreover, only patients with at least 5 visits were considered to standardize the amount of data between patients and guarantee that each patient has a minimum of information to be included.

Some additional requirements were added to ensure that for each case the patients had sufficient information to predict the outcomes. Once the predictions varied in terms of years, in some cases the patients were excluded for not having at least 5 years of follow-up, while in others the patients were excluded for not having at least 10 years of follow-up. Furthermore, some predictions use a value of EDSS in a specific year, and for such reason, the existence of at least one visit in that year or consecutive years were defined as inclusion criteria.

Analysing the Figure 4.2 c) it is possible to identify the 6 different datasets obtained by the application of common inclusion criteria and the additional requirements for each case, which lead to a different selection of patients. The differences in the selection arise from the follow-up years, the inclusion of non-baseline information, and the existence of at least one visit in specific years. Although some datasets in the Figure 4.2 c) are quite similar, they had to be considered as distinct once in the present study it was used several definitions of benign MS from the literature and it was desirable to recreate as closely as possible such definitions and evaluate the differences in the results obtained by those tiny differences. In this way, 6 datasets were created containing the information from the selected patients in each

case.

For every case, the demographic characteristics and the number of visits and years of follow up of the patients that fulfil the inclusion criteria are presented in the Table 4.2. It can be seen, in all selected groups, that the patients included are predominantly women, which is according to the fact that the disease affects more females. In terms of the mean age at onset, the value is around 31 years which is also expected once the disease start usually at young adult period. Furthermore, it can be seen that the predictions with a follow-up of 10 years present the highest number of visits, with exception to prediction after 10 years using only baseline information (information of EDSS values retrieved from the routine visits) that presents a lower number due to the exclusion of non-routine visits.

Table 4.2: Characteristics of the patients that fulfil the inclusion criteria for the different sets.

	SP/RR	**Benign after 5 years selection**	**Benign after 6 years selection**	**Benign after 10 years selection**	**Benign after 10 years with 5 year information**	**Benign after 10 years baseline selection**
Number of patients:	181	167	145	67	52	50
Female:	136 (75.14%)	118 (70.66%)	101 (69.66 %)	52 (77.61 %)	40 (76.92 %)	39 (78%)
Mean age at onset:	31.10 ± 10.54	30.87 ± 10.46	30.28 ± 10.11	32.30 ± 11.84	32.63 ± 11.89	31.52 ± 11.36
Follow-up years:	10.01 ± 8.18	8.86 ± 7.66	9.67 ± 7.98	14.52 ± 10.35	13.87 ± 11.36	11.76 ± 2.30
Number of visits:	13.22 ± 4.87	13.2 ± 4.73	13.82 ± 4.76	15.90 ± 4.83	16.02 ± 4.42	8.6 ± 3.34

5

Experimental Procedure

The experimental procedure involved a ML approach in order to predict the MS disease course. Two different aspects regarding the MS disease course were predicted, specifically the SP development/not development and the disease severity, evaluated by the identification of benign/malignant cases.

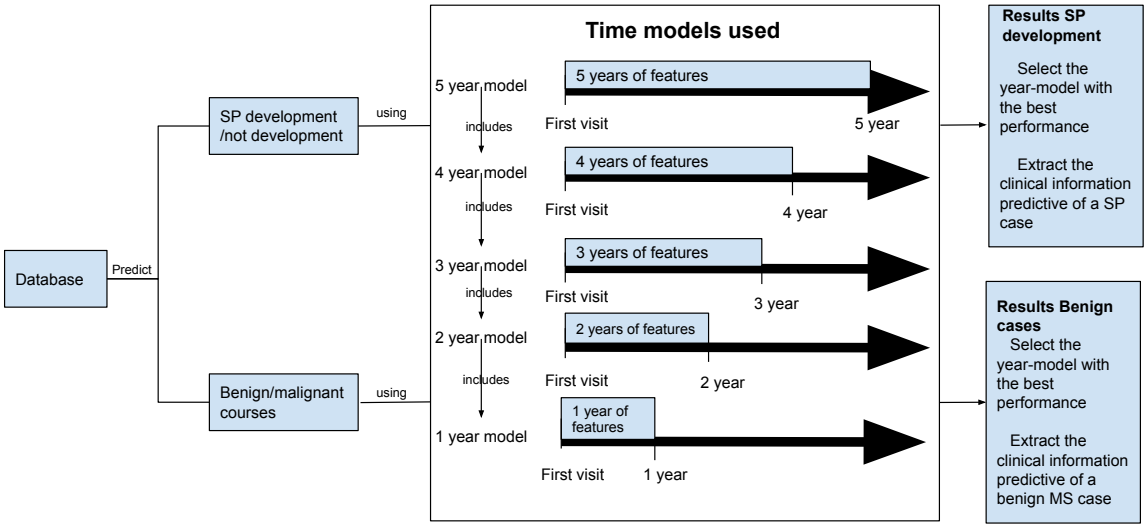

Figure 5.1: Methodology used.

As can be seen in the Figure 5.1 the same database was used to predict SP development/not development and benign/malignant cases.

In both problems were applied 5 time models, each constituted by a different amount of years of follow-up and features considered. The time models represent part of the database used for each prediction, that correspond to the years of data used as information for the prediction. One of the objectives is to identify the number of years of follow-up that are needed to obtain a good model performance, and for such reason the model was tested using a different amount of years of follow-up. For example, when 1 year of follow-up was chosen, only the features from the first year were considered for the prediction while when 2 years were picked the features from both first and second years were chosen. The same logic was applied

for 3, 4 and 5 years.

Moreover, the application of such time models also allow to extract the most important predictors, and the corresponding years where they were important, for both SP development and benign/malignant predictions. Although the methodology is similar for all cases, the differences between them are caused by the different subsets of patients selected, and the different labels admitted as mentioned in the next section.

5.1 Labels

Since two different classification problems are considered, namely the classification in RR/SP and the classification in benign/malignant, the labels for each problem are logically distinct once the outputs predicted in each model are different.

Regarding the RR/SP model, the label considered was binary, indicating if the patient was diagnosed with RR (0) or with SP (1). This label was created using the identification group that contained the information of the disease course for each one of the individuals considered.

In the case of benign/malignant model, a more complex situation was defined, once several labels were created and tested. From the literature, it was possible to verify that the nonexistence of a consensus definition of benign MS, utilized by all authors, is leading to an ambiguity in the results and an arduous comparison of the results achieved in the different papers. To overcome such problem, several labels were retrieved from literature and some others were introduced in order to evaluate how the performance is affected by the selected label and whether different labels lead to different predictive features.

In this way, 9 distinct binary labels indicating if the patient had a benign case (0) or malignant case (1) were tested. Each label is described in the Table 5.1.

Table 5.1: Description of the labels used.

Label	Description	From state of art	Dataset used	Number of patients	Benign percentage
Label SP	Development of SP course	Yes	RR/SP selection	181	-
Label1	EDSS 0–2 in the 10 years using baseline visits [55] [56]	Yes	Benign after 10 years baseline selection	50	42
Label2	EDSS <=3 in the 10th year using baseline visits [58] [59] [26] [60] [61] [62]	Yes	Benign after 10 years baseline selection	50	58
Label3	EDSS <=3 in the 10 year	No	Benign after 10 years selection	67	55.22
Label4	EDSS <=3 during the initial 10th years	No	Benign after 10 years selection	67	28.53
Label5 *	EDSS <=3 after 10 years and <=2 after 5 years [71]	Yes	Benign after 10 years with 5 year information	52	52.17
Label6 **	EDSS <=4 in the 10th year using baseline visits [57]	Yes	Benign after 10years baseline selection	50	76
Label7	EDSS <=3 in the 6th year [84]	Yes	Benign after 6 years selection	145	73.80
Label8 *, ***	Increase EDSS <1.5 after 5 years [77]	Yes	Benign after 5 years selection	167	86.82
Label9 *, ***	Progression index <0,2 after a duration of 5 years [63]	Yes	Benign after 5 years selection	167	77.24

* In literature only the EDSS baseline was considered

** In literature were considered 15 years

*** In literature were considered that the limit cases were benign (<= instead of <)

Analysing the Table 5.1 it is possible to recognize the meaning of each label and verify those that were retrieved from the bibliography. Firstly, there were labels admitting only baseline information (EDSS retrieved from the routine visits), while some others included baseline and non-baseline information (EDSS retrieved from both routine and non-routine visits, including the EDSS from relapses). Regarding the label 1 and 2, both labels use the same selection of patients and consider only the baseline information, leading to labels equal to literature. Once the database had several limitations, slight changes had to be performed to the other labels retrieved from literature. In terms of label 5, it should be used only baseline information, but it lead to an even reduced number of patients and for such reason it was decided to use both baseline and non-baseline information. Regarding label 6, in literature it is used an EDSS ≤ 4 in the 15th year, but it were used only 10 years in this case once the number of patients with 15 years of follow-up was quite reduced. In terms of label 8 and label 9, both labels presented an number of malignant cases quite reduced. To overcome this problem and obtain an higher number of patients of the minority class, it were admitted that in the limit cases the patient was malignant (it was used $<$ instead of $\leq$) and it were also considered baseline and non-baseline information, although in literature only baseline information was used.

Furthermore, it can be noted in Table 5.1 that the number of patients used varied depending on the label considered, because different labels required different amount of years of follow-up and visits occurring in different specific years. To each

case, all patients with a sufficient follow-up period and with visits in the year that the EDSS is retrieved were considered. The percentages of benign in each case were also different from label to label, which is according to literature and demonstrate that different definitions provide a enormous variance in the amount of patients identified as benign.

It is important to note that there were identified in the state-of-the-art other definitions that were not used because a longer period of follow-up (15 and 20 years) was admitted and in the present database the number of patients with that amount of years of follow-up is quite reduced.

5.2 Feature engineering

A process of feature engineering was used in the available data, in order to increase the number of features used by ML algorithms. Such technique included the creation of features by applying statistic operators to previous segmented data in time windows, to extract more information from raw data.

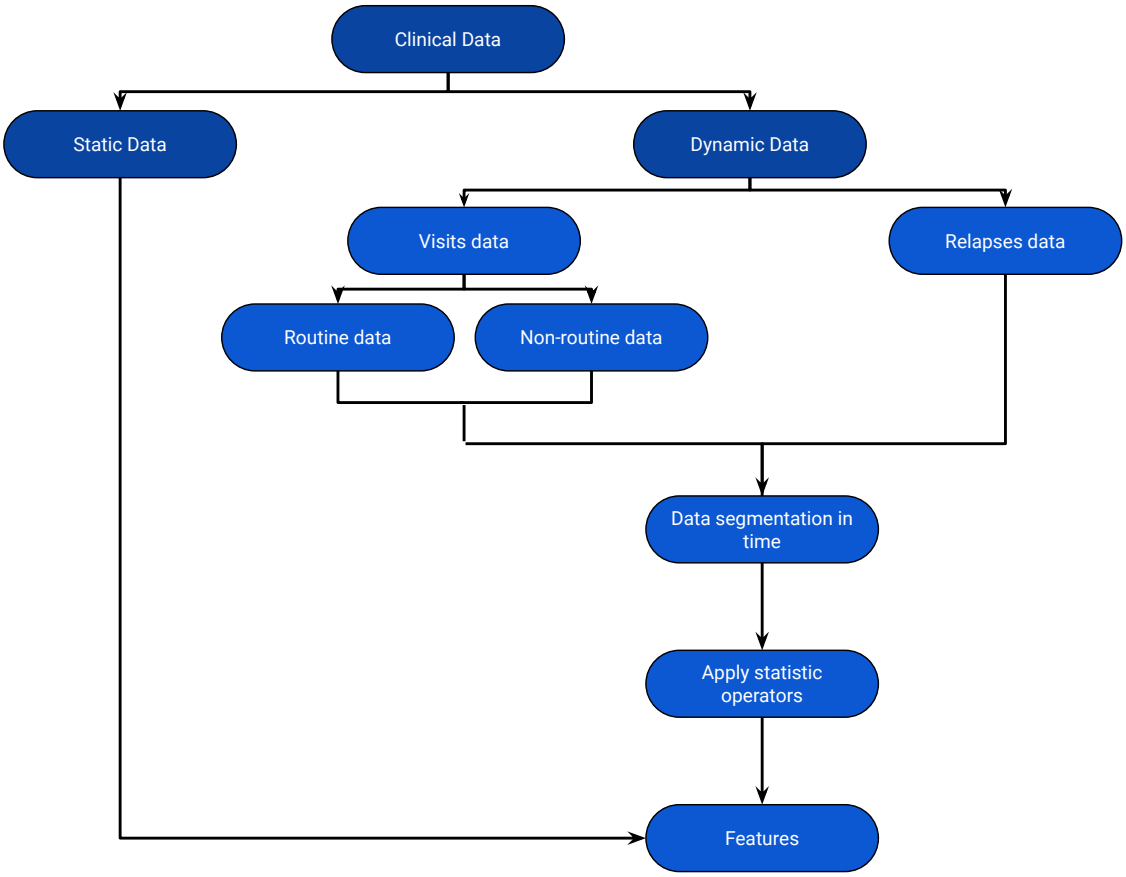

Figure 5.2: Procedure used to create features.

The process used to create features is presented in Figure 5.2. As previously mentioned, the data can be divided in static data, not dependent on time and dynamic information that changes with time. For this reason, a different process was used to create features from static data and dynamic data.

The static information, comprising age, age at onset, gender, clinical information that lead to diagnosis, and even diagnosis date was automatically used to obtain the features, once the information is invariable with time. Furthermore, some additional static features retrieved from data were included, concretely the EDSS at onset, the number of functional systems involved at onset and the amount of years from diagnosis/onset/birth to diagnosis of progressive course. The features created are presented in the Table 5.2

Table 5.2: Features created using static information.

Static Data	All time models
	Used
CSF	✓
Evoked Potentials	✓
Contains CSF and Evoked Potentials	✓
RR/SP/PP	✓
Age	✓
Gender	✓
Deceased	✓
Age of Onset	✓
Supratentorial	✓
Optic Pathways	✓
Brainstem-Cerebellum	✓
Spinal Cord	✓
Progression from onset	✓
From onset to diagnosis	✓
Clinical Findings	✓
MRI	✓
From diagnosis to diagnosis progressive	✓
From onset to diagnosis progressive	✓
From birth to diagnosis progressive	✓
Family History	✓
Number of functional systems involved at onset	✓
EDSS at onset	✓

As opposite, the features created using dynamic data involved a more complex process. The dynamic data was initially divided into visits data and relapses data. It is important to note that the visits data were further segregated into routine

data, corresponding to a routine appointments of a 3-6 months frequency and non-routine data corresponding to non-scheduled visits. To each one of those, it were applied time models that allowed to retrieve all the clinical information from one point in time into another point in time. The data was segmented annually, and was treated as time-series. To the data it were applied statistical operators such as mean, median, mode, standard deviation and ratios. This methodology lead to the creation of the features.

The time segmentation performed is presented in Figure 5.3

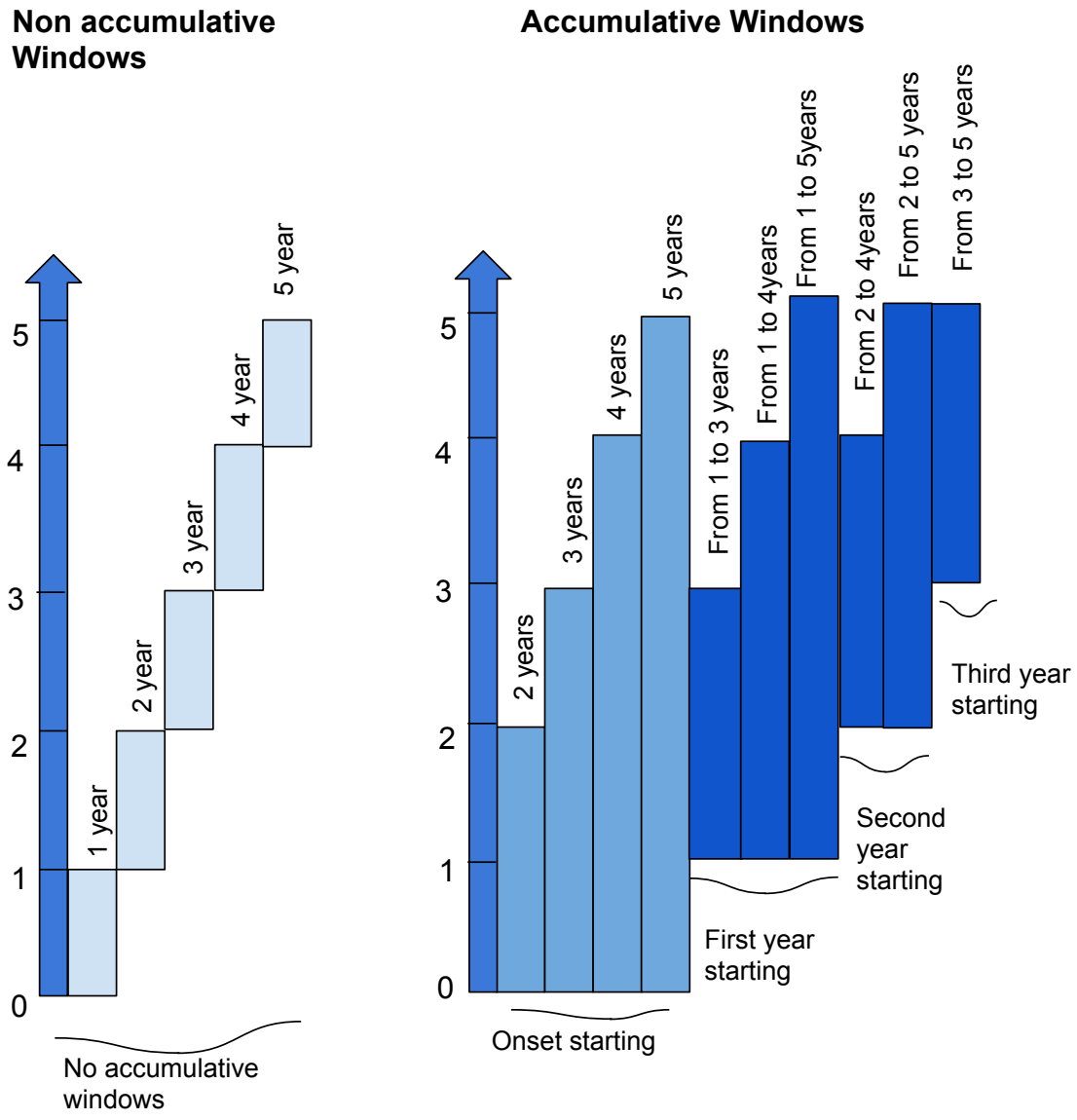

Figure 5.3: Accumulative and non accumulative windows used for feature creation.

The temporal segmentation involved accumulative and non-accumulative windows. First of all, it is important to note that the year of onset was admitted to

correspond to the year of the first visit registered on both accumulative and non-accumulative windows, once it is expected that the patient was already with MS when he first went to the hospital. By admitting this fact, the first year corresponds to the first year of symptoms and not to the first year after the disease was diagnosed, which lead to a more accurate identification of the years in which the symptoms occurred.

The non-accumulative windows represent the first five years, in which each one of the years was considered independently of the others. By applying such segmentation, it is possible to retrieve all dynamic features of a single year.

As opposite, accumulative windows constitute an aggregation of information from several years, by considering a succession of at least 2 years. The start and duration of accumulative windows was variable. It were considered onset starting windows, that included data from onset until 2, 3, 4, 5 and windows starting in the second, third and fourth year ranging from those years until 3, 4, 5 years, depending on the case.

The importance of the application of such time segmentation, relies in the fact that the information of routine visits of the 1st year after onset can be different from the information of routine visits 2 years after onset if more visits had occurred in the second year. With this rational it is possible to create several features using raw data, and identify the amount of years that provide optimal predictions. Furthermore, such segmentation allows to reduce significantly the amount of missing data and it is also important to evaluate if the features with more predictive power vary between different windows considered.

All features created can be identified using Tables 5.3 and 5.4. The first column corresponds to the raw features, the following columns corresponds to the statistical operators used, and the final columns (3 last columns for visits group and last column for relapses), indicate the cases in which the features were used. It is important to note that the same features were created for all time windows. For instance, for the Score Pyramidal, considering three years of features, the application of the statistical operator mean allow to retrieve the mean of all registered values of that feature until the third year. The remaining statistical operators follow the same rationale. Furthermore, the ratio events is defined as the proportion of occurrences of a feature, which means that if the feature fatigue is considered, the ratio events represent the number of times that fatigue occurred in a visit among all visits considered. The ratio-non events are calculated by 1-ratio events.

The assignment of the name to each feature involved a methodological process: <statistical operator><feature name><time model><Information group>. Ap-

plying this rationale, an example of a feature created is *mean_ gdataxia_ from2to5years_ visits_routine*.

Table 5.3: Features created from the visits information group.

Visits Data	All time models									
	mean	median	mode	std	maximum	ratio events	ratio non-events	Used in total data	Used in routine data	Used in non-routine data
Suspected Relapse	✓	✓	✓	✓		✓	✓	✓	✓	✓
Score Pyramidal	✓	✓	✓	✓				✓	✓	✓
Score Cerebellar	✓	✓	✓	✓				✓	✓	✓
Cerebellar Weakness	✓	✓	✓	✓		✓	✓	✓	✓	✓
Score BrainStem	✓	✓	✓	✓				✓	✓	✓
Score Sensory	✓	✓	✓	✓				✓	✓	✓
Score Bowel	✓	✓	✓	✓				✓	✓	✓
Score Visual	✓	✓	✓	✓				✓	✓	✓
Visual Symptom	✓	✓	✓	✓		✓	✓	✓	✓	✓
Score Mental	✓	✓	✓	✓				✓	✓	✓
Score Ambulation	✓	✓	✓	✓				✓	✓	✓
gdAtaxia	✓	✓	✓	✓		✓	✓	✓	✓	✓
dysaesthesiae	✓	✓	✓	✓		✓	✓	✓	✓	✓
ataxiaLowerExtrem	✓	✓	✓	✓		✓	✓	✓	✓	✓
paresthesiae	✓	✓	✓	✓		✓	✓	✓	✓	✓
cognitionPb	✓	✓	✓	✓		✓	✓	✓	✓	✓
gdParesis	✓	✓	✓	✓		✓	✓	✓	✓	✓
gdSpasticity	✓	✓	✓	✓		✓	✓	✓	✓	✓
mwUpperExtrem	✓	✓	✓	✓		✓	✓	✓	✓	✓
micturitionPb	✓	✓	✓	✓		✓	✓	✓	✓	✓
fatigue	✓	✓	✓	✓		✓	✓	✓	✓	✓
mwLowerExtrem	✓	✓	✓	✓		✓	✓	✓	✓	✓
moodPb	✓	✓	✓	✓		✓	✓	✓	✓	✓
EDSS	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Nº visits in the X year								✓	✓	✓

X= 1, 2, 3, 4, 5 year-model

Table 5.4: Features created from relapses information group.

Relapses Data	All time models							
	mean	median	mode	std	maximum	ratio events	ratio non-events	Used
Impact ADL Functions	✓	✓	✓	✓		✓	✓	✓
Recovery	✓	✓	✓	✓				✓
Severity	✓	✓	✓	✓				✓
CNS Pyramidal Tract	✓	✓	✓	✓		✓	✓	✓
CNS Brainstem	✓	✓	✓	✓		✓	✓	✓
CNS Bowel Bladder	✓	✓	✓	✓		✓	✓	✓
CNS Neuropsycho Functions	✓	✓	✓	✓		✓	✓	✓
CNS Cerebellum	✓	✓	✓	✓		✓	✓	✓
CNS Visual Functions	✓	✓	✓	✓		✓	✓	✓
CNS Sensory Functions	✓	✓	✓	✓		✓	✓	✓
EDSS	✓	✓	✓	✓	✓	✓	✓	✓
Nº relapses in the X year								✓

X= 1, 2, 3, 4, 5 year-model

5.3 Machine learning pipeline

The ML pipeline used, was the same for both SP development and disease severity predictions, consisting on an initial process of feature extraction, using

static and dynamic information, and a choice of the amount of years of features to integrate the dataset used.

As demonstrated in Figure 5.4 the obtained dataset was uniformly shuffled and then partitioned into training and testing set using a k-Fold Cross-Validation, where 70% of the data was selected for training and 30% for testing. Once for the majority of the labels considered, the dataset was highly imbalanced, 70 % of the samples of the minority class were selected and an equal number of samples of the majority class were extracted from the entire shuffled set, to constitute the training set, while the remaining samples compose the test set. The partition of 70/30 was selected once it constitutes a classic partition. Furthermore, several fields of both training and testing set were missing, whereby an imputation of the missing values using the mean of the values of the training set was performed and that mean was then used to imputate the missing values of the test set.

The missing values imputation was proceeded by an standardization using a z-score. The feature selection was the next step, where two filter methods (correlation and AUC) were applied followed by a Least Absolute Shrinkage and Selection Operator (LASSO) Regression. For both standardization and feature selection steps it can be noted that the result of the training parameters were used to perform the same operations in the testing set. This means that, for example, the mean and standard deviation from the training data are used to standardize the testing data. Several different classifiers (KNN-3, LDA, Linear Regression and SVM) were trained and the obtained models were tested in unseen data (testing set)

Finally the performance of the model was evaluated using several metrics. It is important to note that the entire process was repeated 100 times (10 different k-fold CV, each one with k=10) which allowed to explore the entire data and guarantee that the results did not change significantly once the dataset was significantly explored. In this way, the mean of the results of all iterations that constitute the final results do not changed significantly from execution to execution which lead to a good approach to the associated stochasticity.

Regarding the implementation, was utilized MATLAB for all phases of this master thesis. All the functions used for this work are built-in MATLAB, except for the standardization [85] and missing data imputation [86]. Each step of the pipeline is explained in the next subsections.

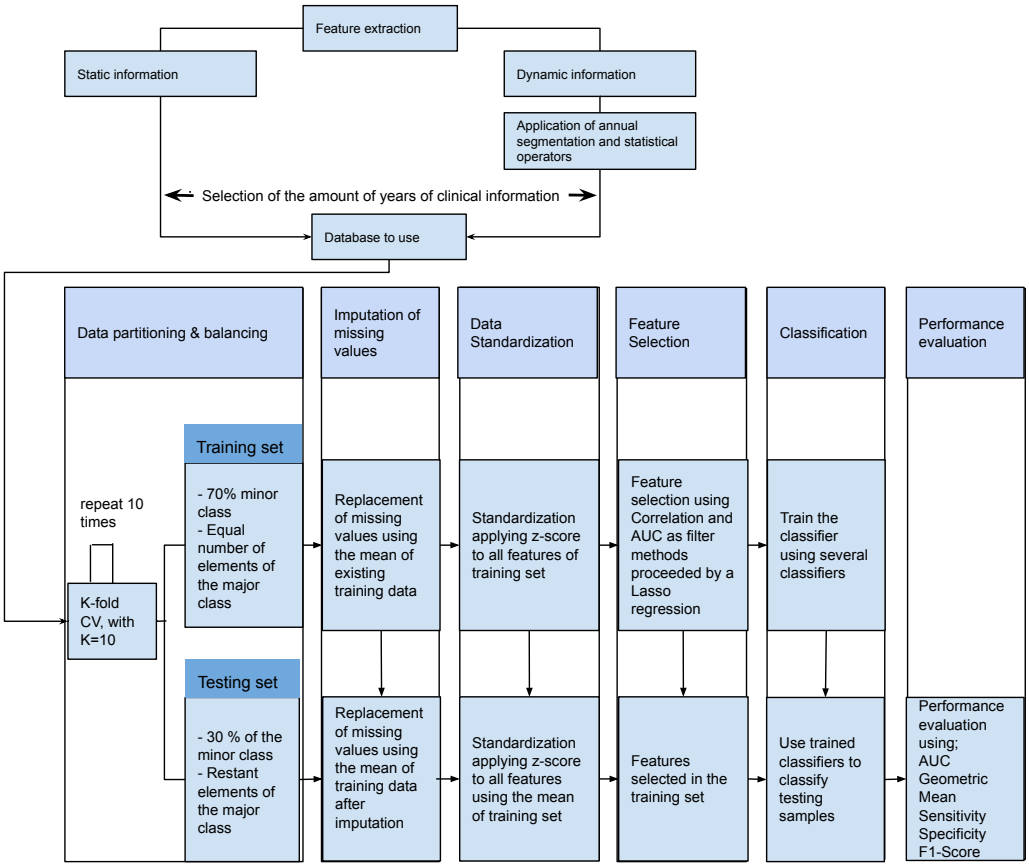

Figure 5.4: Pipeline used for both RR/SP and benign/malignant predictions.

5.3.1 Partitioning and balancing methods

Considering the dataset composed by the features previously created respecting the number of years of selected, it is necessary to partition the data. The partition involves segregation of an amount of data to estimate the parameters for the ML model (training data) and the remaining to evaluate how well the ML models work (testing data).

A k-fold CV with k = 10 was used to shuffle randomly the patients and separate them into k uniform groups, wherein each group contained the real proportion of patients from both classes (RR/SP and benign/malignant). Once the dataset used is imbalanced for both SP development and disease severity predictions, an undersampling approach was performed in the uniformly shuffled data, in which 70 % of the patients of the minority class were selected for the training group together with an equal number of samples of the majority class. The remaining samples, including the 30% samples of the minority class and the remaining samples of the majority class not used for the training set, were selected for the testing set. With

this approach is guaranteed that an acceptable number of samples of the minority class is used for training and that an equal number of samples of both classes is used to train the model which solve the imbalanced situation.

Furthermore, once a k-fold CV was used to guide the split into training and testing sets, it is ensured that in each iteration performed, the group of patients allocated to training and testing is different. All the process was repeated 10 times, resulting in 100 iterations (10 different k-fold CV, each one with k=10), which allow to better explore all data and increase the confidence in the evaluation results. Lastly, the performance was obtained by the average of the results from all the 100 iterations. It is important to mention that the *cvpartition* function from MATLAB was used to perform the k-fold CV.

5.3.2 Imputation of missing values

In the dataset considered there were missing data in several fields, that were imputated using the existing data. The process by which the data is imputated can vary in the assumption made, diverging between assuming the mean of the existing values, finding similar points, use the most frequent value, among others.

An aspect that is important to be mentioned is that the missing values are assumed to be missing completely at random, because it seems to not exist any relation between the missing values and the outcome or the values of the other features [87]. When analyzing the fields that contain plenty of missing data, it appears not to exist any relation between the missing data and the benign, malignant, RR or SP samples, once the missing data occurs similarly in all those outcomes. In addition, the visits to the hospital are usually scheduled from 3 to 3 or 6 to 6 months, and in this database there are missing several of this visits in all patients. Consequently, the gaps regarding the patient clinical absence are assumed to be caused by the non-annotation of visits. Furthermore, it also appears to not exist any case of a feature that is missing due to the information retrieved from the remaining features.

In the current methodology, the missing values were imputed using the mean of the existing values. This is a simple approach, that is computationally fast. It was performed an imputation in the training set, using exclusively the values of such set. The values of the mean for each field obtained with the imputation on the training set were then used to perform the imputation on the testing set.

5.3.3 Feature selection

The dataset used contained an huge amount of features when compared to the number of samples used. It is known that if all features were used for prediction, the model could learn the noise of the data, using irrelevant and redundant data that only contributed to an overfitting of the model to the training data. The feature selection methods allow to select a subset of features that likely will lead to an increase in the capability of generalization. The goal of selecting features is to keep those that maximize performance, and removing those that are irrelevant and lead to a decrease in the model generalization capacity [88].

The approach used consist in the application of two filter methods followed by an embedded one. Initially the features were filtered based on a Pearson's linear correlation coefficient between each feature and the output. The correlation for all features was retrieved and the 100 features containing the higher correlation were selected. A correlation filter was admitted due to the fact that a higher correlation tend to demonstrate a relationship between the attributes considered.

Furthermore, a second filter evaluating which of the 100 features perform better in terms of ROC analysis was applied. The AUC was calculated for the 100 features, and the 50 features with the higher value were selected. It was used *corr* and *perfcurve* functions for calculating correlation coefficient and AUC, respectively. Those 50 features resulting from AUC filter were then standardized using a z-score to convert such data into a common range.

The two initial filters were applied in order to reduced substantially the number of features that will be used in the embedded method. The embedded method selected is the LASSO that is a computationally expensive method with a long period to converge, and by decreasing the amount of features the time to compute was also reduced.

The LASSO is a feature selection method that was used to select an optimal set of features. This method performs both feature selection and regularization, which allows to enhance both the prediction performance and the interpretability of the model, while the redundancy is removed. The LASSO forces the sum of the absolute values of the model parameters to be inferior to a fixed value, which is achieved using a L1 regularization process, where some of the coefficients of the regression variables (features) are penalized and shrinked to zero. The features whose coefficients are non-zero after the shrinking process are then selected to be part of the model, while the remaining features are eliminated [89]. This method does not select the best features individually, but rather the best group of features. The cost function of

this regression, for M samples and P features, is described in Eq. (5.1), where λ controls the L1 penalty influence, $\mathbf{w}$ is the regression coefficients vector, $\mathbf{y}$ the data to be fitted, $\hat{\mathbf{y}}$ the fitted regression values according to features $\mathbf{x}$:

$$\sum_{i=1}^{M}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{M}(y_i - \sum_{j=0}^{P} w_j \times x_{ij})^2 + \lambda \sum_{j=0}^{P} |w_j| \,. \tag{5.1}$$

The higher the value of λ, the greater the regularization action, which means that more features are shrinked to zero and eliminated.

In this case, was calculated a LASSO regression for a geometric sequence of 100 values, in which only the highest one can lead to a model with 0 features. The value of λ was selected by a classical rule of thumb [90], guaranteeing that there should be at least 5 training samples for each feature. For example, if for a given training dataset there exists approximately 30 samples (2*0.70*21 samples minority class), a maximum number of 5 allowed features is obtained. This logic was applied to all considered labels. It was decided to use the rule of thumb instead of a CV procedure (that is a more standard approach) to get the value of lambda, once the CV procedure leads to the selection of the highest value of lambda that creates a model with 0 features. For the implementation of LASSO was used the lasso function of MATLAB.

5.3.4 Classifiers

For the predictive model, several classifiers were tested to evaluate the results using different alternatives and realize if the performance of the model is relatively good regardless of the chosen classifier. Moreover, only simple classifiers were selected once it was used a reduced number of samples, and more complex algorithms such as neuronal networks could lead to overfitting.

The first classifier used was the KNN. It was used a value of k equal to 1,3 and 5. Only odd numbers were selected for k value once the algorithm predicts the label based on the most common k nearest points, and with even numbers, there is a possibility to exist equally common labels. Moreover, low values of k were selected once the dataset is small and the low k values lead to increase sensitivity to the points localizations. The best results were obtained for a k=3, although they were not substantially different from the results obtained with the other values of k. For training the classifier was used *fitcknn* function from MATLAB.

The LDA is a dimensionality reduction technique that is used for supervised classification that project data to a maximum of C-1 dimensions, where C is the

number of classes. This means that for binary problems, the features are projected into 1 dimension (1 axis). Having as base the projected one dimensional data, a linear classifier is defined, originating a Fisher LDA classifier. It was used *fitcdiscr* function from MATLAB for training the classifier with pseudolinear selected as discriminant type.

The linear regression is a classifier that establish a linear relation between the input variables and the output values, in order to obtain a line that best fits the points. In this case, if the predicted response values of the linear regression model obtain a value higher than 0.5, the value is admitted to belong to the class 1 while if the value is lower or equal to 0.5 the sample is classified as 0. For training the classifier was used *fitlm* function from MATLAB.

The SVM was also used as classifier, in which linear function was used as kernel function. It was used *fitcsvm* function from MATLAB.

5.3.5 Evaluation metrics

To evaluate the results obtained with the several classifiers considered, several evaluation metrics were considered. Some of the used metrics are described below:

1. Sensitivity - This metric measures the true positive rate, which stands as the proportion of the actual positives that are correctly identified as such. In this specific case, the sensitivity reflects the proportion of the SP cases correctly identified for the RR/SP label and malignant cases correctly identified for the benign/malignant labels;
2. Specificity - This metric measures the true negative rate, which stands as the proportion of the actual negatives that are correctly identified as such. In this specific case, the specificity reflects the proportion of the RR cases correctly identified for the RR/SP label and benign cases correctly identified for the benign/malignant labels;
3. F1-Score - This metric presents the number of correctly identified SP cases, among the total of cases classified as SP, for the RR/SP classification problem. Regarding the severity classification problem, this metric presents the number of correctly classified malignant cases, among the total of cases classified as malignant.

Those metrics, along with AUC and G-mean, were selected once they reflect more realistically the results obtained. Other common evaluation metrics such as accuracy were not included since the dataset is highly imbalanced and a high value of accuracy could represent a bad classification if the model is classifying all samples

as the majority class.

6

Results

This chapter is divided into sections representing each classification problem (labeling) defined and a final section where all results of benign/malignant prediction are compared. For each classification problem, the performance of the model including the results obtained with four classifiers, and the features with most predictive power in each year are presented.

In terms of performance of the model, in each case, the best results achieved for each year are in bold.

Regarding the identification of predictive features, several figures containing the predictive power of the features in each year model are presented. The predictive power corresponds to the number of times that a given feature was identified among the 100 iterations. Only the features that were present in at least 10 iterations were exhibited in the figures. It is important to note that only the name is presented, and it does not contain the statistical operator used once it was verified that the features were identified using several statistical operators in which there wasn't a clear predominance of one of them.

Moreover, the features were represented in graphs divided by time models, in which each colour represent an accumulation of years. The features retrieved from non-accumulative windows were counted as onset starting accumulative windows once the percentage of features from non-accumulative windows was quite reduced (for example, only 9 % of the features were from non-accumulative windows for RR/SP classification problem and only 6 % were from non-accumulative windows for classification problem 1). With this approach, the visualization of the graphs obtained is simpler, once the year 3, for example, represent all features from non-accumulative windows and accumulative windows starting on onset until that year. It was also used coding procedure to every graph containing the predictive features to facilitate the interpretation of the graphs

In the Figures 6.1 until 6.11 it is also represented in each feature a signal (+) indicative of benign or RR case or (-) that indicates a malignant or SP case. Such

information was retrieved by the correlation that each feature had with the outcome.

Lastly, it is important to note that for the case of severity classification problems, the definitions of benign considered were all using a value of EDSS scale as a threshold, which enables a comparison between them.

6.1 RR / SP classification problem

Table 6.1: Results of the performance obtained for the classification problem RR/SP.

Best results		KNN 3	LDA	SVM Linear	Linear Regression
1 year model	**AUC:**	0.70 ± 0.08	0.70 ± 0.08	**0.69 ± 0.08**	0.58 ± 0.10
	G-Mean:	0.63 ± 0.06	0.65 ± 0.08	**0.63 ± 0.09**	0.64 ± 0.09
	Specificity:	0.51 ± 0.09	0.65 ± 0.09	**0.61 ± 0.13**	0.65 ± 0.09
	Sensitivity:	0.79 ± 0.13	0.66 ± 0.15	**0.68 ± 0.18**	0.64 ± 0.16
	F1-Score:	0.11 ± 0.02	0.13 ± 0.04	**0.12 ± 0.04**	0.13 ± 0.04
2 year model	**AUC:**	0.74 ± 0.08	0.77 ± 0.08	**0.76 ± 0.09**	0.62 ± 0.14
	G-Mean:	0.68 ± 0.07	0.67 ± 0.14	**0.67 ± 0.14**	0.66 ± 0.14
	Specificity:	0.62 ± 0.10	0.75 ± 0.07	**0.71 ± 0.10**	0.75 ± 0.07
	Sensitivity:	0.76 ± 0.14	0.63 ± 0.21	**0.67 ± 0.21**	0.62 ± 0.21
	F1-Score:	0.14 ± 0.03	0.16 ± 0.04	**0.16 ± 0.04**	0.16 ± 0.05
3 year model	**AUC:**	**0.74 ± 0.08**	0.76 ± 0.07	0.76 ± 0.08	0.61 ± 0.12
	G-Mean:	**0.70 ± 0.06**	0.66 ± 0.12	0.69 ± 0.08	0.66 ± 0.12
	Specificity:	**0.67 ± 0.07**	0.75 ± 0.06	0.70 ± 0.08	0.75 ± 0.06
	Sensitivity:	**0.75 ± 0.13**	0.61 ± 0.18	0.71 ± 0.17	0.60 ± 0.18
	F1-Score:	**0.15 ± 0.03**	0.16 ± 0.04	0.16 ± 0.04	0.15 ± 0.04
4 year model	**AUC:**	**0.73 ± 0.08**	0.78 ± 0.06	0.76 ± 0.07	0.70 ± 0.08
	G-Mean:	**0.70 ± 0.10**	0.66 ± 0.11	0.64 ± 0.11	0.66 ± 0.11
	Specificity:	**0.78 ± 0.07**	0.81 ± 0.06	0.79 ± 0.07	0.81 ± 0.06
	Sensitivity:	**0.65 ± 0.16**	0.56 ± 0.18	0.55 ± 0.18	0.56 ± 0.18
	F1-Score:	**0.19 ± 0.05**	0.18 ± 0.06	0.16 ± 0.06	0.18 ± 0.06
5 year model	**AUC:**	**0.84 ± 0.08**	0.82 ± 0.10	0.81 ± 0.11	0.77 ± 0.13
	G-Mean:	**0.78 ± 0.11**	0.72 ± 0.11	0.73 ± 0.12	0.72 ± 0.11
	Specificity:	**0.78 ± 0.06**	0.80 ± 0.09	0.77 ± 0.11	0.80 ± 0.09
	Sensitivity:	**0.78 ± 0.18**	0.65 ± 0.16	0.71 ± 0.18	0.65 ± 0.16
	F1-Score:	**0.23 ± 0.07**	0.22 ± 0.08	0.21 ± 0.09	0.22 ± 0.09

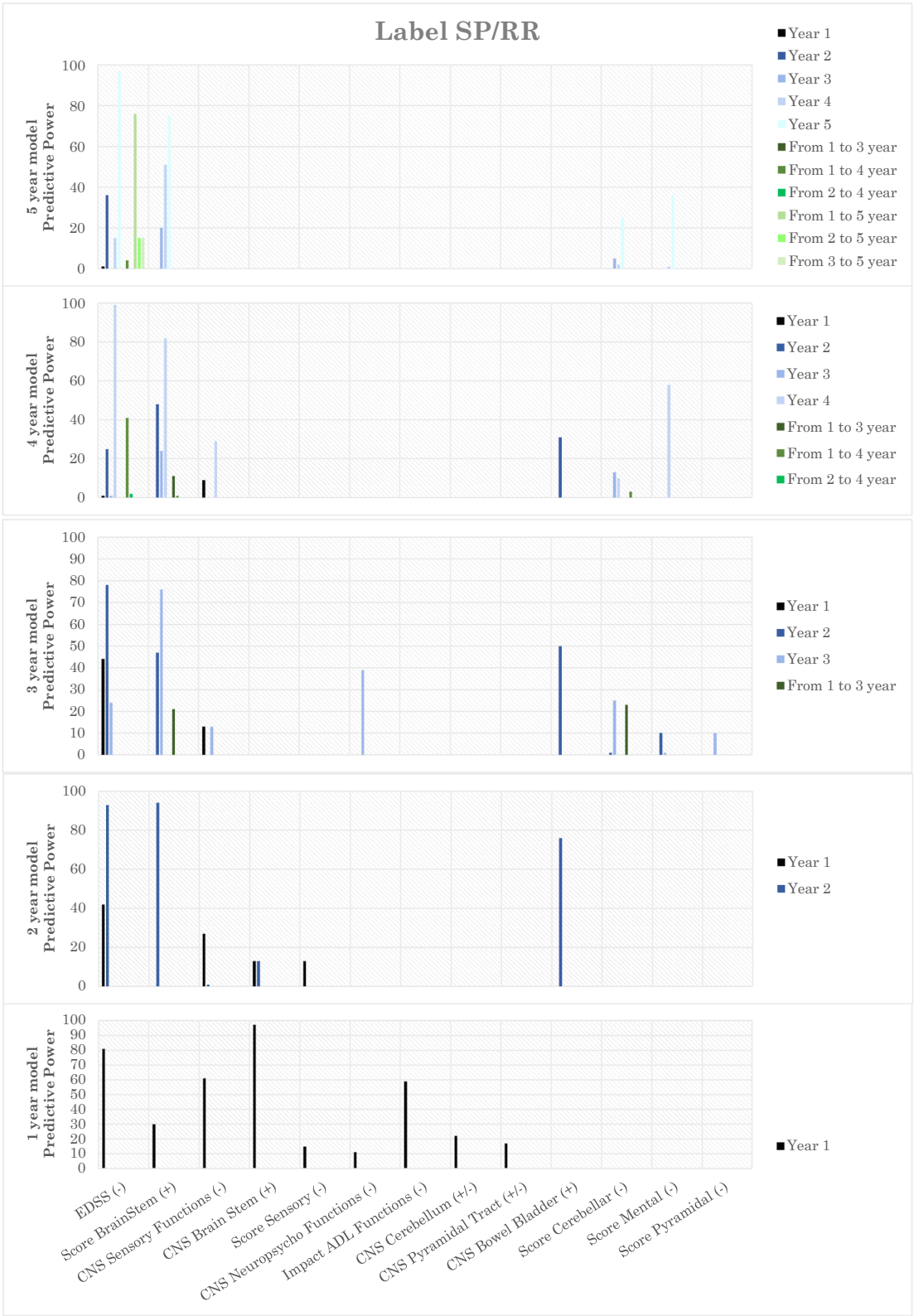

Figure 6.1: Features with highest predictive power identified in classification problem RR/SP. The darker the shade of blue, the closer to the diagnosis is the year when the feature was retrieved and the darker the shade of green the closer to the diagnosis is the accumulation of the years (non-starting on onset) where the feature was identified.

As observed in Table 6.1 the best results to predict SP development were obtained using an SVM linear as a classifier in the two initial years and using a KNN-3 in the remaining ones.

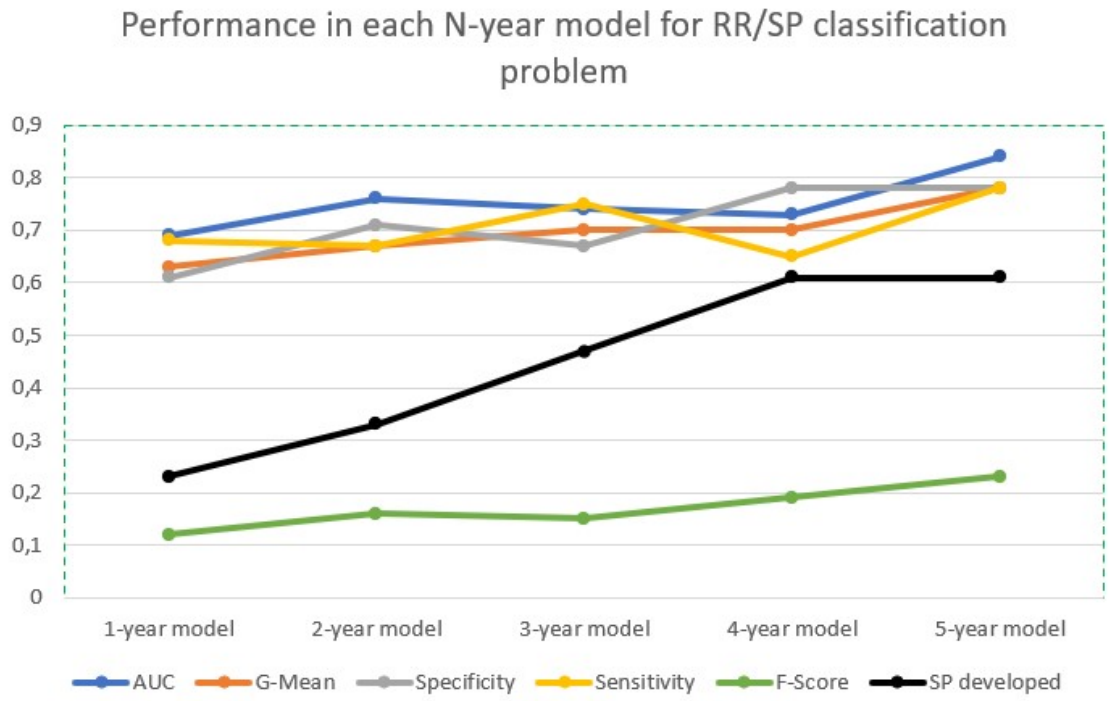

Figure 6.2: Performance in each N-year model for RR/SP classification problem.

As can be seen in the Figure 6.2, a clear evolution in the performance of the model can be observed over time. It can also be noted that sensitivity outperformed significantly the control performance (SP developed line), which indicates that in each year model the model developed correctly identified more SP patients when compared to the accumulative ratio of patients that developed SP in that year. The 5 year model, using a KNN-3, was selected as the best model, with a satisfactory performance obtained: an AUC of 0.84 $\pm$ 0.08, a geometric mean of 0.78 $\pm$ 0.11, a specificity of 0.78 $\pm$ 0.06, a sensitivity of 0.78 $\pm$ 0.18 and an F1-Score of 0.23 $\pm$ 0.07. Overall, the results obtained for F1-Score were low for every year considered with every classifier.

The evolution of the features identified as predictive can be observed in Figure 6.1. It can be noted that the features identified are different in distinct years. In the first-year model, the EDSS, Impact ADL Functions, CNS Brainstem, CNS Sensory Functions, CNS Cerebellum, Score Sensory, Score Brainstem, CNS Neuropsycho Functions and CNS Pyramidal tract were identified as predictive. Among those features, only the EDSS and Score Brainstem remain predictive in all other years.

It can be observed that the Score Cerebellar and Score Mental were not predictive in the two initial years, but started to be predictive in the third and remain as predictive features until the 5-year model. This may be due to the fact that they are functional systems that take longer to show clinically evident symptoms.

6.2 Disease severity classification problem

The disease severity was predicted using several definitions of benign MS. Each definition of benign was used as a different label, which constituted an different classification problem. The results obtained, including the performance of the model and the most predictive features, for each of the classification problems considered are presented in the sections from 6.2.1 until 6.2.9. In the section 6.2.10 is presented a comparison of the results obtained, in each year-model, for all disease severity classification problems.

6.2.1 Classification problem 1 - EDSS 0-2 in the 10th year using baseline visits

Table 6.2: Results of the performance obtained for the classification problem 1.

	Classifier	KNN-3	LDA	SVM Linear	Linear Regression
1 year model	AUC	0.57 ± 0.10	**0.62 ± 0.08**	0.60 ± 0.09	0.58 ± 0.09
	G-Mean	0.50 ± 0.09	**0.56 ± 0.08**	0.54 ± 0.10	0.56 ± 0.08
	Specificity	0.51 ± 0.17	**0.61 ± 0.14**	0.60 ± 0.19	0.61 ± 0.14
	Sensitivity	0.54 ± 0.14	**0.52 ± 0.11**	0.51 ± 0.12	0.52 ± 0.11
	F1-Score	0.60 ± 0.10	**0.61 ± 0.09**	0.60 ± 0.10	0.61 ± 0.09
2 year model	AUC	**0.62 ± 0.09**	0.64 ± 0.08	0.61 ± 0.08	0.60 ± 0.09
	G-Mean	**0.59 ± 0.09**	0.60 ± 0.07	0.57 ± 0.08	0.60 ± 0.07
	Specificity	**0.54 ± 0.15**	0.68 ± 0.15	0.57 ± 0.16	0.68 ± 0.15
	Sensitivity	**0.66 ± 0.12**	0.55 ± 0.12	0.59 ± 0.13	0.55 ± 0.12
	F1-Score	**0.71 ± 0.08**	0.64 ± 0.08	0.66 ± 0.09	0.64 ± 0.08
3 year model	AUC	**0.68 ± 0.06**	0.71 ± 0.06	0.67 ± 0.06	0.63 ± 0.07
	G-Mean	**0.63 ± 0.06**	0.64 ± 0.07	0.58 ± 0.08	0.64 ± 0.07
	Specificity	**0.66 ± 0.14**	0.69 ± 0.14	0.55 ± 0.14	0.69 ± 0.14
	Sensitivity	**0.61 ± 0.10**	0.60 ± 0.09	0.63 ± 0.09	0.60 ± 0.09
	F1-Score	**0.69 ± 0.07**	0.69 ± 0.07	0.69 ± 0.06	0.69 ± 0.07
4 year model	AUC	**0.76 ± 0.08**	0.76 ± 0.08	0.74 ± 0.10	0.72 ± 0.10
	G-Mean	**0.70 ± 0.08**	0.71 ± 0.08	0.66 ± 0.10	0.71 ± 0.08
	Specificity	**0.76 ± 0.14**	0.79 ± 0.10	0.65 ± 0.14	0.79 ± 0.10
	Sensitivity	**0.66 ± 0.12**	0.65 ± 0.15	0.68 ± 0.10	0.65 ± 0.15
	F1-Score	**0.74 ± 0.08**	0.74 ± 0.11	0.74 ± 0.08	0.74 ± 0.11
5 year model	AUC	0.75 ± 0.09	0.77 ± 0.09	**0.76 ± 0.08**	0.67 ± 0.11
	G-Mean	0.69 ± 0.08	0.62 ± 0.11	**0.71 ± 0.09**	0.62 ± 0.11
	Specificity	0.88 ± 0.12	0.93 ± 0.10	**0.86 ± 0.12**	0.93 ± 0.10
	Sensitivity	0.54 ± 0.09	0.42 ± 0.16	**0.58 ± 0.09**	0.42 ± 0.16
	F1-Score	0.68 ± 0.08	0.57 ± 0.15	**0.71 ± 0.08**	0.57 ± 0.15

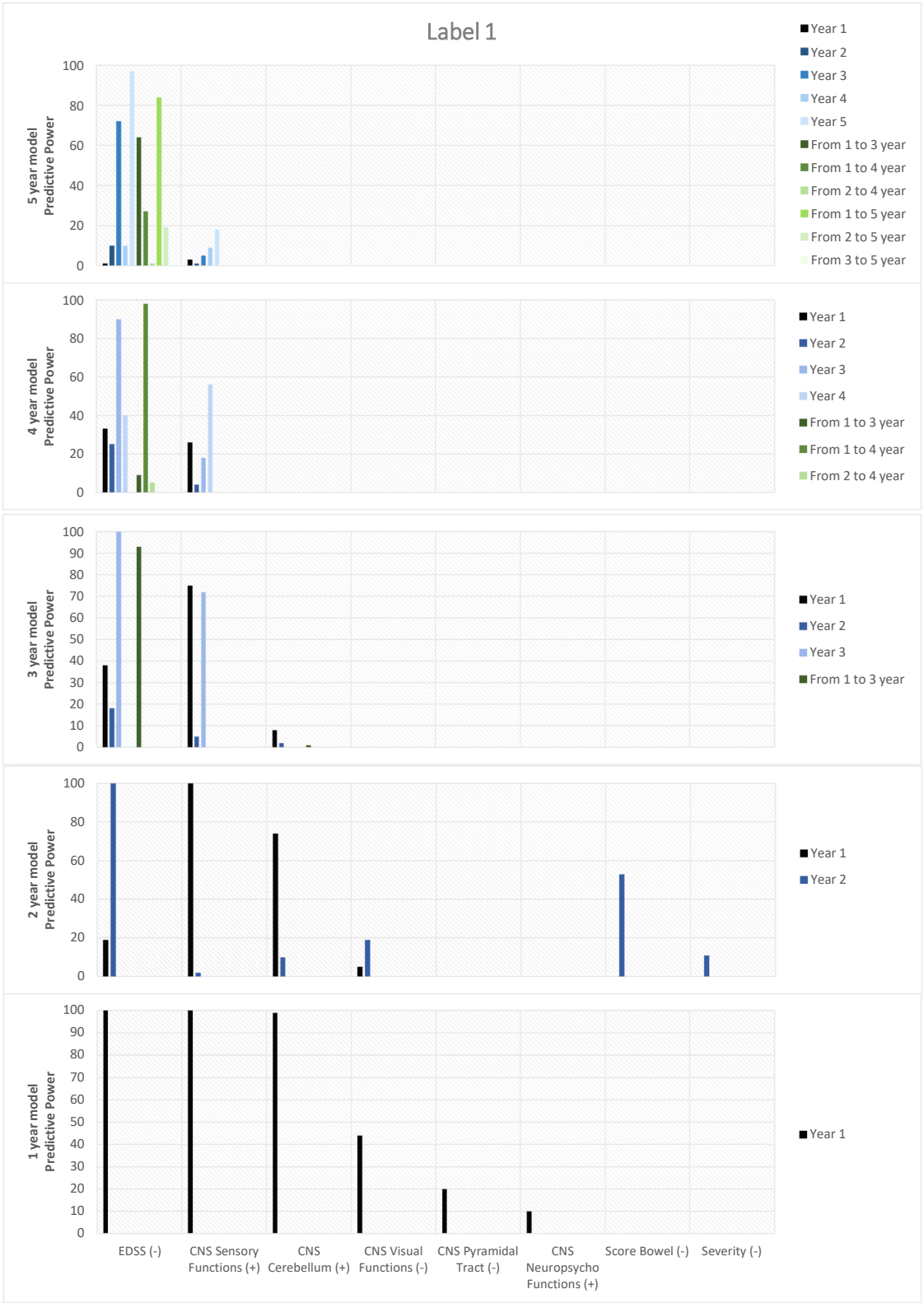

Figure 6.3: Features with highest predictive power of a benign/malignant course identified in classification problem 1. The darker the shade of blue, the closer to the diagnosis is the year when the feature was retrieved and the darker the shade of green the closer to the diagnosis is the accumulation of the years (non-starting on onset) where the feature was identified.

It can be observed in the Table 6.2 that the best result for the prediction of benign/malignant cases using the classification problem 1 was achieved using a KNN-3 as a classifier and 4 years of features obtaining an AUC of 0.76 ± 0.08, an geometric mean of 0.70 ± 0.08, a specificity of 0.76 ± 0.14, an sensitivity of 0.66 ± 0.12 and an F1-Score of 0.74 ± 0.08. Overall, it was observed an increase in the performance of the model from the first to the second year and from the third to fourth year. The opposite occurs from second to the third year and from fourth to the fifth year in which a decrease in the value of sensitivity was noted. It is worth mentioning that the best performing classifiers varied depending on the year considered, although the performances of the classifiers were quite similar in the majority of the cases. The KNN-3 was the best classifier for the second, third and fourth year, the LDA was selected for the first year and the SVM linear was identified as the best performing in the 5 year model.

Regarding the features identified as predictive, it can be noted in the Figure 6.3 that every feature was identified in the first year with exception to Score Bowel and Severity that only appear in the second year. Furthermore, the evolution of the features over the years is clear, being noted that from all features identified in the first year only the EDSS and CNS Sensory Functions remain predictive in all year models. The CNS cerebellum was identified in the three first models, the CNS visual functions in the two first models and the CNS pyramidal Tract and CNS Neuropsycho Functions only in the first year model.

In terms of predictive power, the EDSS, CNS Sensory Functions and CNS Cerebellum (in the two first-year models) represent the cases with a higher value of predictive power. It is also important to note that the features from the year of the time model are the most predictive in such year model for the majority of the cases, although the features from the initial year remain important until later years.

6.2.2 Classification problem 2 - EDSS $\leq$ 3 in the 10th year using baseline visits

Table 6.3: Results of the performance obtained for the classification problem 2.

	Classifier	KNN-3	LDA	SVM Linear	Linear Regression
1 year model	AUC	0.54 ± 0.13	0.64 ± 0.11	**0.67 ± 0.09**	0.64 ± 0.09
	G-Mean	0.50 ± 0.16	0.58 ± 0.11	**0.60 ± 0.11**	0.58 ± 0.11
	Specificity	0.57 ± 0.17	0.65 ± 0.11	**0.65 ± 0.12**	0.65 ± 0.11
	Sensitivity	0.51 ± 0.23	0.55 ± 0.17	**0.58 ± 0.17**	0.55 ± 0.17
	F1-Score	0.41 ± 0.11	0.46 ± 0.11	**0.48 ± 0.11**	0.46 ± 0.11
2 year model	AUC	0.60 ± 0.12	0.68 ± 0.10	**0.69 ± 0.09**	0.70 ± 0.10
	G-Mean	0.57 ± 0.11	0.59 ± 0.11	**0.62 ± 0.10**	0.59 ± 0.11
	Specificity	0.50 ± 0.13	0.70 ± 0.08	**0.61 ± 0.09**	0.70 ± 0.08
	Sensitivity	0.67 ± 0.20	0.53 ± 0.18	**0.66 ± 0.17**	0.53 ± 0.18
	F1-Score	0.47 ± 0.12	0.46 ± 0.13	**0.51 ± 0.11**	0.46 ± 0.13
3 year model	AUC	0.6 ± 0.14	0.70 ± 0.13	**0.69 ± 0.11**	0.67 ± 0.13
	G-Mean	0.57 ± 0.13	0.61 ± 0.16	**0.62 ± 0.12**	0.61 ± 0.16
	Specificity	0.59 ± 0.16	0.74 ± 0.11	**0.66 ± 0.11**	0.74 ± 0.11
	Sensitivity	0.58 ± 0.18	0.52 ± 0.19	**0.61 ± 0.18**	0.52 ± 0.19
	F1-Score	0.46 ± 0.14	0.50 ± 0.16	**0.51 ± 0.12**	0.50 ± 0.16
4 year model	AUC	0.65 ± 0.11	0.75 ± 0.10	**0.74 ± 0.09**	0.65 ± 0.12
	G-Mean	0.61 ± 0.12	0.62 ± 0.13	**0.66 ± 0.11**	0.62 ± 0.13
	Specificity	0.66 ± 0.12	0.82 ± 0.11	**0.76 ± 0.13**	0.82 ± 0.11
	Sensitivity	0.59 ± 0.18	0.49 ± 0.18	**0.59 ± 0.19**	0.49 ± 0.18
	F1-Score	0.49 ± 0.14	0.49 ± 0.18	**0.55 ± 0.13**	0.51 ± 0.16
5 year model	AUC	**0.81 ± 0.09**	0.86 ± 0.07	0.85 ± 0.08	0.70 ± 0.10
	G-Mean	**0.70 ± 0.11**	0.64 ± 0.12	0.71 ± 0.12	0.64 ± 0.12
	Specificity	**0.84 ± 0.12**	0.91 ± 0.10	0.90 ± 0.09	0.91 ± 0.10
	Sensitivity	**0.61 ± 0.17**	0.46 ± 0.17	0.58 ± 0.18	0.46 ± 0.17
	F1-Score	**0.61 ± 0.13**	0.54 ± 0.16	0.62 ± 0.15	0.54 ± 0.16

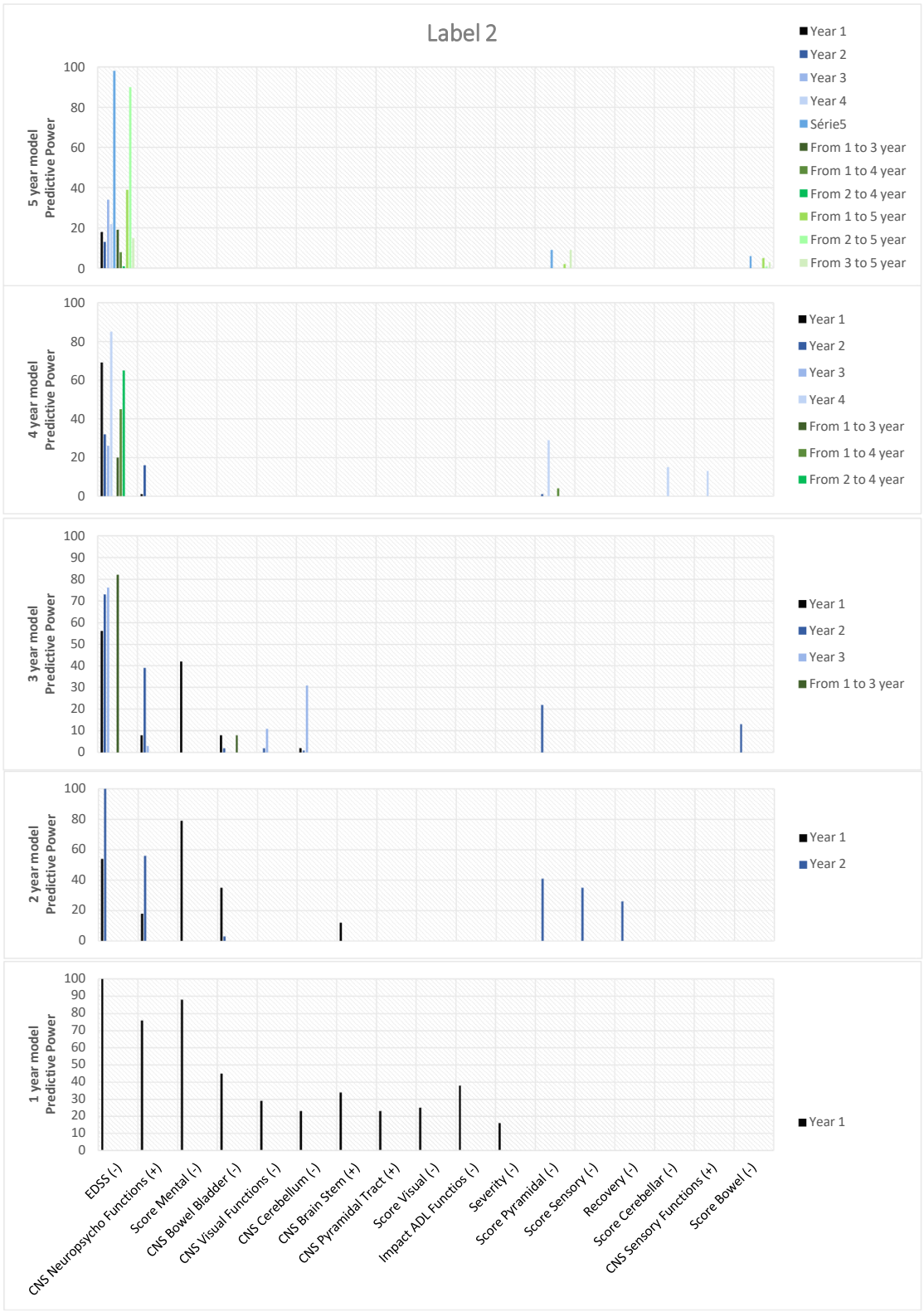

Figure 6.4: Features with highest predictive power of a benign/malignant course identified in classification problem 2. The darker the shade of blue, the closer to the diagnosis is the year when the feature was retrieved and the darker the shade of green the closer to the diagnosis is the accumulation of the years (non-starting on onset) where the feature was identified.

The best classifier to predict the benign cases using the classification problem 2 is the SVM classifier for the four initial year models and the KNN-3 in the 5-year model as observed in Table 6.3. The best results were obtained using 5 years of features with an AUC of 0.81 ± 0.09, an geometric mean of 0.70 ± 0.11, a specificity of 0.84 ± 0.12, a sensitivity of 0.61 ± 0.17 and an F1-Score of 0.61 ± 0.13. Overall, there is an increase in the performance over time, although the rise is not gradual from year to year, once from the first year model to the second year model and from the fourth year model to the fifth year model the growth was more accentuated.

The features identified as predictive using the classification problem 2 are represented in the Figure 6.4. Overall, the number of features identified decreases over time, in which 11 features were identified in the first year model and only 3 were predictive in the 5 year model. Among all features, only the EDSS was predictive in all year models, containing a high predictive power in all cases. The CNS Neuropsycho functions was the unique feature predictive only on the 4 initial year models, while both Score Mental and CNS Bowel Bladder were predictive in the 3 initial year models.

The Score Pyramidal represent an interesting case once it was predictive in all year models with exception to the first-year model. Moreover, the Score Cerebellar, CNS Sensory Functions and Score Bowel appear to be predictive only in later years.

6.2.3 Classification problem 3 - EDSS $\leq$ 3 in the 10th year

Table 6.4: Results of the performance obtained for the classification problem 3.

	Classifier	KNN-3	LDA	SVM Linear	Linear Regression
1 year model	AUC	0.59 ± 0.14	0.80 ± 0.13	**0.81 ± 0.13**	0.75 ± 0.12
	G-Mean	0.53 ± 0.20	0.67 ± 0.19	**0.68 ± 0.17**	0.67 ± 0.19
	Specificity	0.71 ± 0.15	0.85 ± 0.10	**0.81 ± 0.11**	0.85 ± 0.10
	Sensitivity	0.46 ± 0.24	0.58 ± 0.26	**0.62 ± 0.25**	0.58 ± 0.26
	F1-Score	0.43 ± 0.14	0.58 ± 0.18	**0.58 ± 0.18**	0.58 ± 0.18
2 year model	AUC	0.61 ± 0.13	0.83 ± 0.09	**0.78 ± 0.11**	0.73 ± 0.12
	G-Mean	0.57 ± 0.20	0.64 ± 0.23	**0.65 ± 0.19**	0.64 ± 0.23
	Specificity	0.77 ± 0.14	0.85 ± 0.10	**0.80 ± 0.12**	0.85 ± 0.10
	Sensitivity	0.49 ± 0.23	0.55 ± 0.27	**0.59 ± 0.25**	0.55 ± 0.27
	F1-Score	0.48 ± 0.15	0.57 ± 0.17	**0.56 ± 0.15**	0.57 ± 0.17
3 year model	AUC	**0.74 ± 0.11**	0.80 ± 0.11	0.76 ± 0.12	0.72 ± 0.12
	G-Mean	**0.64 ± 0.20**	0.64 ± 0.16	0.64 ± 0.16	0.64 ± 0.16
	Specificity	**0.78 ± 0.12**	0.87 ± 0.10	0.80 ± 0.11	0.87 ± 0.10
	Sensitivity	**0.58 ± 0.27**	0.50 ± 0.20	0.54 ± 0.21	0.50 ± 0.20
	F1-Score	**0.54 ± 0.16**	0.54 ± 0.16	0.52 ± 0.15	0.54 ± 0.15
4 year model	AUC	**0.81 ± 0.12**	0.83 ± 0.15	0.83 ± 0.13	0.79 ± 0.15
	G-Mean	**0.70 ± 0.19**	0.72 ± 0.20	0.70 ± 0.19	0.72 ± 0.20
	Specificity	**0.78 ± 0.11**	0.85 ± 0.10	0.82 ± 0.11	0.85 ± 0.10
	Sensitivity	**0.68 ± 0.27**	0.65 ± 0.26	0.65 ± 0.28	0.65 ± 0.26
	F1-Score	**0.60 ± 0.17**	0.64 ± 0.17	0.61 ± 0.18	0.64 ± 0.17
5 year model	AUC	0.81 ± 0.12	0.88 ± 0.12	**0.89 ± 0.10**	0.84 ± 0.12
	G-Mean	0.72 ± 0.18	0.77 ± 0.18	**0.78 ± 0.18**	0.77 ± 0.18
	Specificity	0.81 ± 0.10	0.90 ± 0.10	**0.88 ± 0.10**	0.90 ± 0.10
	Sensitivity	0.69 ± 0.25	0.70 ± 0.25	**0.73 ± 0.25**	0.70 ± 0.25
	F1-Score	0.63 ± 0.16	0.71 ± 0.17	**0.71 ± 0.17**	0.71 ± 0.17

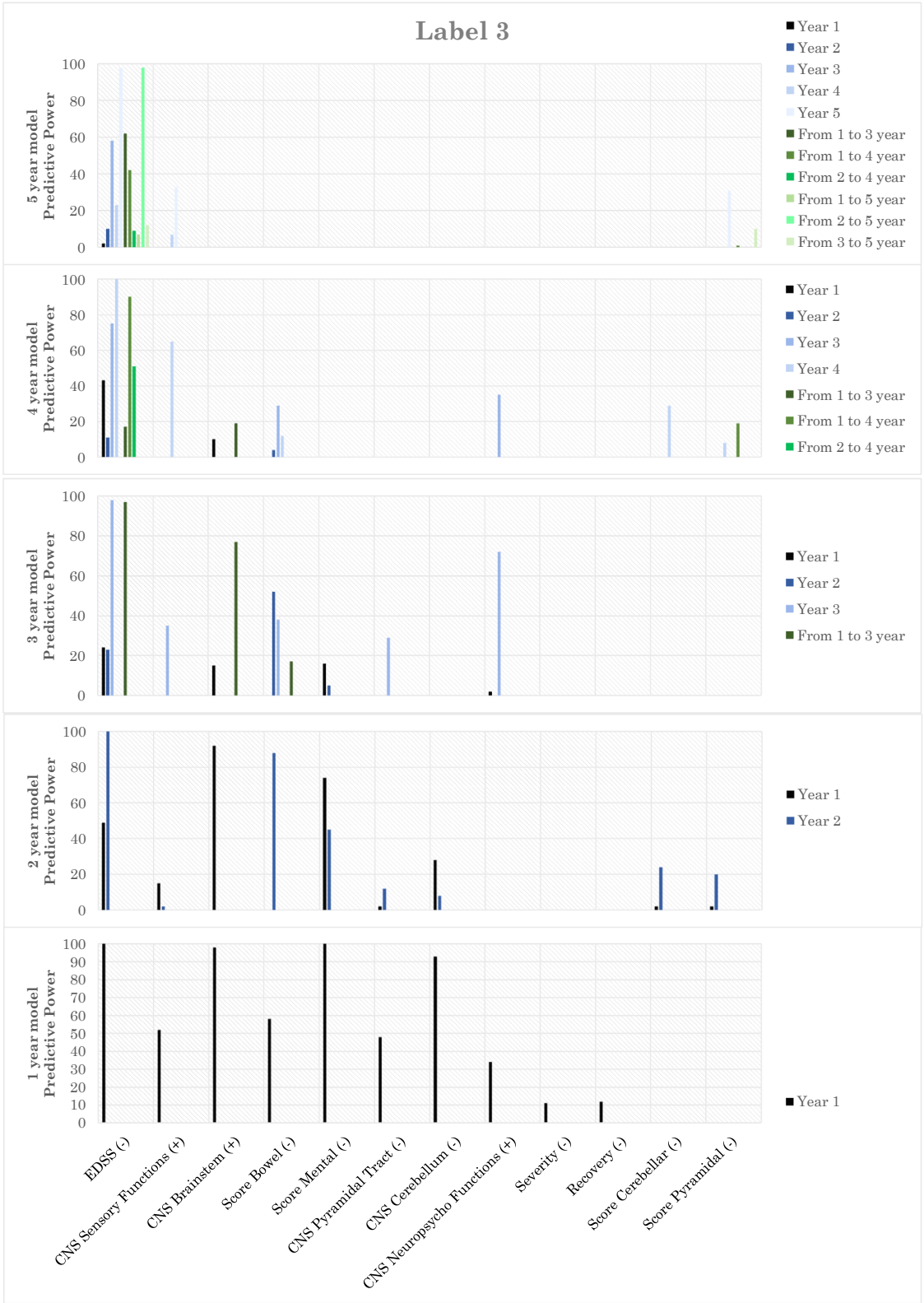

Figure 6.5: Features with highest predictive power of a benign/malignant course identified in classification problem 3. The darker the shade of blue, the closer to the diagnosis is the year when the feature was retrieved and the darker the shade of green the closer to the diagnosis is the accumulation of the years (non-starting on onset) where the feature was identified.

The results obtained using the classification problem 3, presented in the Table 6.4, demonstrate a clear increase in the performance of the model in the two final year models, although the three initial year models present a similar performance. Thus, the best result was obtained using 5 years of features with an SVM linear as the classifier, where was obtained an AUC of 0.89 ± 0.10, a geometric mean of 0.78 ± 0.18, an specificity of 0.88 ± 0.10, an sensitivity of 0.73 ± 0.25 and an F1-Score of 0.71 ± 0.17. It is worth mentioning that the SVM linear was the best performing classifier in the first, second and fifth year models, while the KNN-3 was better for the third and fourth-year models.

In Figure 6.5 it can be noted an evolution of the features over the years. Regarding the features identified in the first year, the Severity and Recovery were the less predictive features once they present the lower predictive power and were the unique features that weren't identified in any other year. Among the remaining features predictive in the first year, the EDSS and CNS Sensory Functions were predictive in all models, the CNS Brainstem and Score Bowel were predictive in the four initial models and the Score Mental and CNS Pyramidal Tract were predictive in the three initial year models. Furthermore, the CNS Cerebellum was predictive in the two initial models and the CNS Neuropsycho Functions was predictive in the first, third and fourth-year model.

Lastly, it is possible to observe that both the Score Cerebellar and the Score Pyramidal appeared for the first time in the second year model and were also predictive in the four year model. The most recent features appear to be the most predictive in this case, which means that in the second year model the features from the second year are the most predictive and in the third year model the features from the third year are the most predictive.

6.2.4 Classification problem 4 - EDSS $\leq$ 3 during the initial 10th years

Table 6.5: Results of the performance obtained for the classification problem 4.

	Classifier	KNN-3	LDA	SVM Linear	Linear Regression
1 year model	**AUC**	0.76 ± 0.08	0.82 ± 0.07	**0.82 ± 0.06**	0.83 ± 0.07
	G-Mean	0.67 ± 0.07	0.70 ± 0.07	**0.70 ± 0.07**	0.69 ± 0.07
	Specificity	0.90 ± 0.14	0.88 ± 0.14	**0.84 ± 0.13**	0.88 ± 0.14
	Sensitivity	0.51 ± 0.09	0.56 ± 0.10	**0.60 ± 0.10**	0.55 ± 0.10
	F1-Score	0.66 ± 0.08	0.70 ± 0.08	**0.73 ± 0.07**	0.70 ± 0.08
2 year model	**AUC**	0.72 ± 0.10	0.81 ± 0.07	**0.80 ± 0.08**	0.81 ± 0.08
	G-Mean	0.64 ± 0.07	0.70 ± 0.07	**0.70 ± 0.08**	0.69 ± 0.07
	Specificity	0.85 ± 0.15	0.87 ± 0.15	**0.83 ± 0.15**	0.87 ± 0.15
	Sensitivity	0.49 ± 0.09	0.58 ± 0.09	**0.61 ± 0.11**	0.57 ± 0.09
	F1-Score	0.64 ± 0.08	0.72 ± 0.07	**0.74 ± 0.09**	0.71 ± 0.07
3 year model	**AUC**	0.73 ± 0.10	0.78 ± 0.10	**0.78 ± 0.09**	0.78 ± 0.09
	G-Mean	0.68 ± 0.10	0.68 ± 0.09	**0.69 ± 0.09**	0.67 ± 0.09
	Specificity	0.87 ± 0.17	0.92 ± 0.13	**0.89 ± 0.15**	0.92 ± 0.13
	Sensitivity	0.55 ± 0.09	0.51 ± 0.10	**0.55 ± 0.10**	0.49 ± 0.10
	F1-Score	0.69 ± 0.08	0.66 ± 0.09	**0.70 ± 0.09**	0.65 ± 0.09
4 year model	**AUC**	0.78 ± 0.08	0.80 ± 0.07	**0.81 ± 0.06**	0.79 ± 0.10
	G-Mean	0.70 ± 0.08	0.68 ± 0.08	**0.72 ± 0.08**	0.67 ± 0.08
	Specificity	0.92 ± 0.10	0.93 ± 0.11	**0.91 ± 0.12**	0.93 ± 0.11
	Sensitivity	0.54 ± 0.09	0.51 ± 0.10	**0.57 ± 0.10**	0.49 ± 0.10
	F1-Score	0.69 ± 0.08	0.66 ± 0.09	**0.72 ± 0.08**	0.65 ± 0.09
5 year model	**AUC**	0.82 ± 0.07	0.83 ± 0.05	**0.84 ± 0.06**	0.82 ± 0.07
	G-Mean	0.72 ± 0.07	0.69 ± 0.05	**0.72 ± 0.07**	0.68 ± 0.05
	Specificity	0.96 ± 0.09	0.98 ± 0.05	**0.95 ± 0.08**	0.98 ± 0.05
	Sensitivity	0.54 ± 0.09	0.49 ± 0.07	**0.55 ± 0.10**	0.47 ± 0.07
	F1-Score	0.70 ± 0.08	0.65 ± 0.06	**0.70 ± 0.09**	0.63 ± 0.06

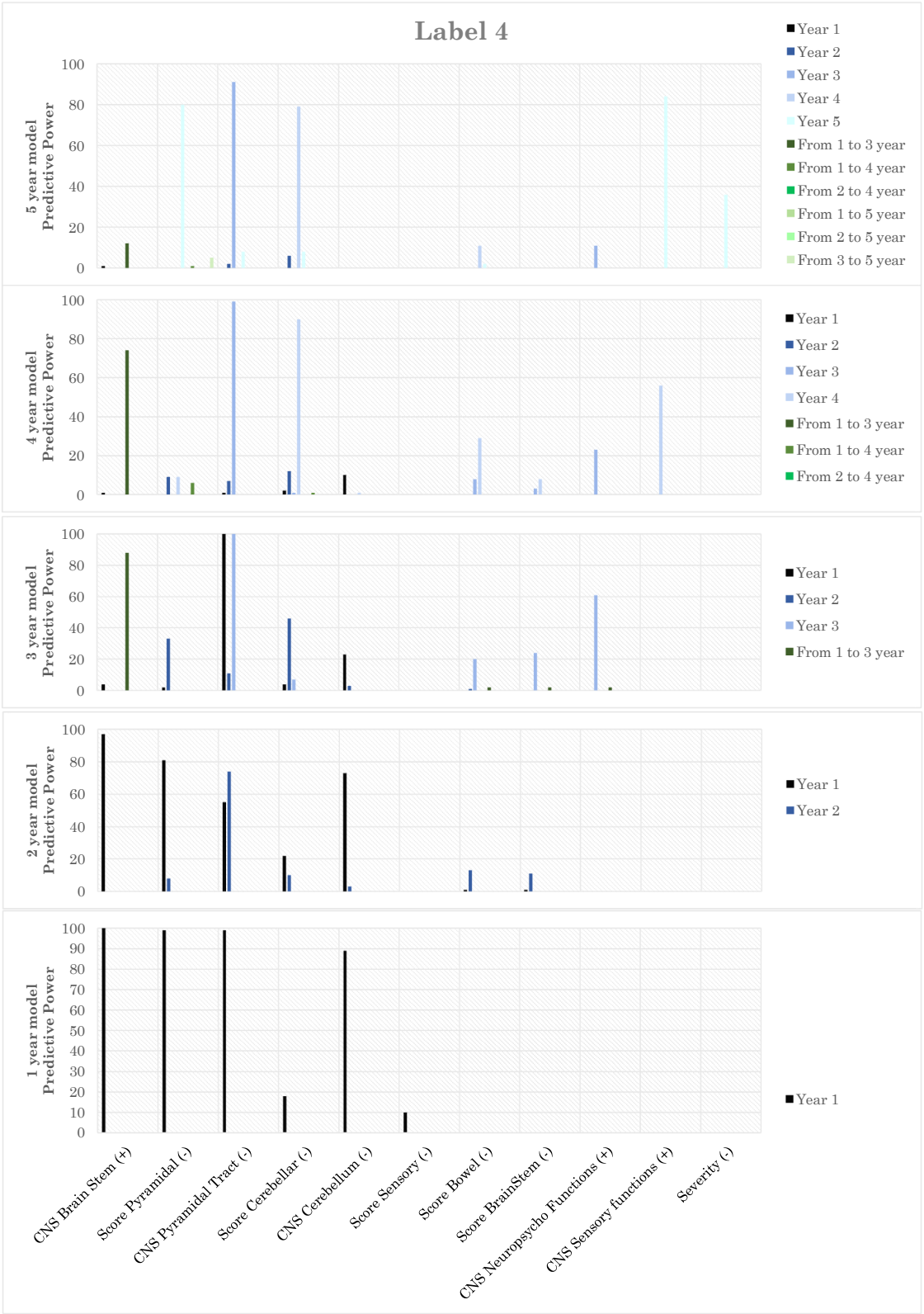

Figure 6.6: Features with highest predictive power of a benign/malignant course identified in classification problem 4. The darker the shade of blue, the closer to the diagnosis is the year when the feature was retrieved and the darker the shade of green the closer to the diagnosis is the accumulation of the years (non-starting on onset) where the feature was identified.

For the classification problem 4, it can be seen in the Table 6.5 that the best classifier is the SVM linear for all year models. In this case, the results were worsening over the years, with the best performances occurring with 1 and 2-year models. Once the two initial models perform similarly, the 1-year model was selected as the most appropriate model, once it is possible to predict the outcome sooner. With features from the first year, it was obtained an AUC of 0.82 $\pm$ 0.06, a geometric mean of 0.70 $\pm$ 0.07, a specificity of 0.84 $\pm$ 0.13, a sensitivity of 0.60 $\pm$ 0.10 and an F1-Score of 0.73 $\pm$ 0.07.

The Figure 6.6 presents the features identified as predictive of a benign/malignant course using classification problem 4. It should be noted that the EDSS was not used as a feature in this classification problem, once the objective is to verify if there was any value of EDSS higher than 3 since the beginning of the disease and for such reason if the EDSS was used, the label was considered as a feature. The CNS Brainstem, the Score Pyramidal, the CNS Pyramidal tract and the Score Cerebellar were predictive in all year models while the CNS Cerebellum were identified in the four initial year models. Among the other features, the Score Bowel were identified as predictive since the second year model, and the Score Brainstem was predictive from second until fourth year model. Moreover, the CNS Neuropsycho Functions was predictive since the third year, the CNS Sensory functions from the fourth year model and the Severity only in the fifth year model.

Overall, the year 3 appear to be important once features from such year are predictive in both third, fourth and fifth-year models.

6.2.5 Classification problem 5 - EDSS $\leq$ 2 after 5 years and $\leq$ 3 after 10 years

Table 6.6: Results of the performance obtained for the classification problem 5.

	Classifier	KNN-3	LDA	SVM Linear	Linear Regression
1 year model	**AUC**	0.69 ± 0.10	**0.74 ± 0.08**	0.73 ± 0.08	0.69 ± 0.11
	G-Mean	0.62 ± 0.12	**0.65 ± 0.09**	0.64 ± 0.10	0.65 ± 0.09
	Specificity	0.78 ± 0.16	**0.78 ± 0.17**	0.78 ± 0.17	0.78 ± 0.17
	Sensitivity	0.52 ± 0.15	**0.56 ± 0.12**	0.56 ± 0.15	0.56 ± 0.12
	F1-Score	0.62 ± 0.12	**0.66 ± 0.10**	0.65 ± 0.11	0.66 ± 0.10
2 year model	**AUC**	0.71 ± 0.07	**0.75 ± 0.05**	0.72 ± 0.07	0.71 ± 0.09
	G-Mean	0.66 ± 0.07	**0.70 ± 0.07**	0.66 ± 0.08	0.70 ± 0.07
	Specificity	0.79 ± 0.14	**0.76 ± 0.13**	0.67 ± 0.15	0.76 ± 0.13
	Sensitivity	0.57 ± 0.11	**0.65 ± 0.11**	0.67 ± 0.10	0.65 ± 0.11
	F1-Score	0.66 ± 0.08	**0.72 ± 0.07**	0.71 ± 0.07	0.72 ± 0.07
3 year model	**AUC**	0.79 ± 0.08	**0.82 ± 0.05**	0.80 ± 0.05	0.76 ± 0.07
	G-Mean	0.71 ± 0.08	**0.75 ± 0.07**	0.71 ± 0.08	0.75 ± 0.07
	Specificity	0.83 ± 0.14	**0.76 ± 0.12**	0.69 ± 0.13	0.76 ± 0.12
	Sensitivity	0.63 ± 0.10	**0.75 ± 0.10**	0.76 ± 0.10	0.75 ± 0.10
	F1-Score	0.72 ± 0.08	**0.79 ± 0.06**	0.77 ± 0.07	0.79 ± 0.06
4 year model	**AUC**	0.82 ± 0.07	0.84 ± 0.06	**0.84 ± 0.05**	0.84 ± 0.07
	G-Mean	0.77 ± 0.08	0.77 ± 0.07	**0.77 ± 0.06**	0.77 ± 0.07
	Specificity	0.89 ± 0.11	0.81 ± 0.08	**0.79 ± 0.09**	0.81 ± 0.08
	Sensitivity	0.67 ± 0.09	0.74 ± 0.11	**0.75 ± 0.09**	0.74 ± 0.11
	F1-Score	0.77 ± 0.08	0.79 ± 0.08	**0.79 ± 0.06**	0.79 ± 0.08

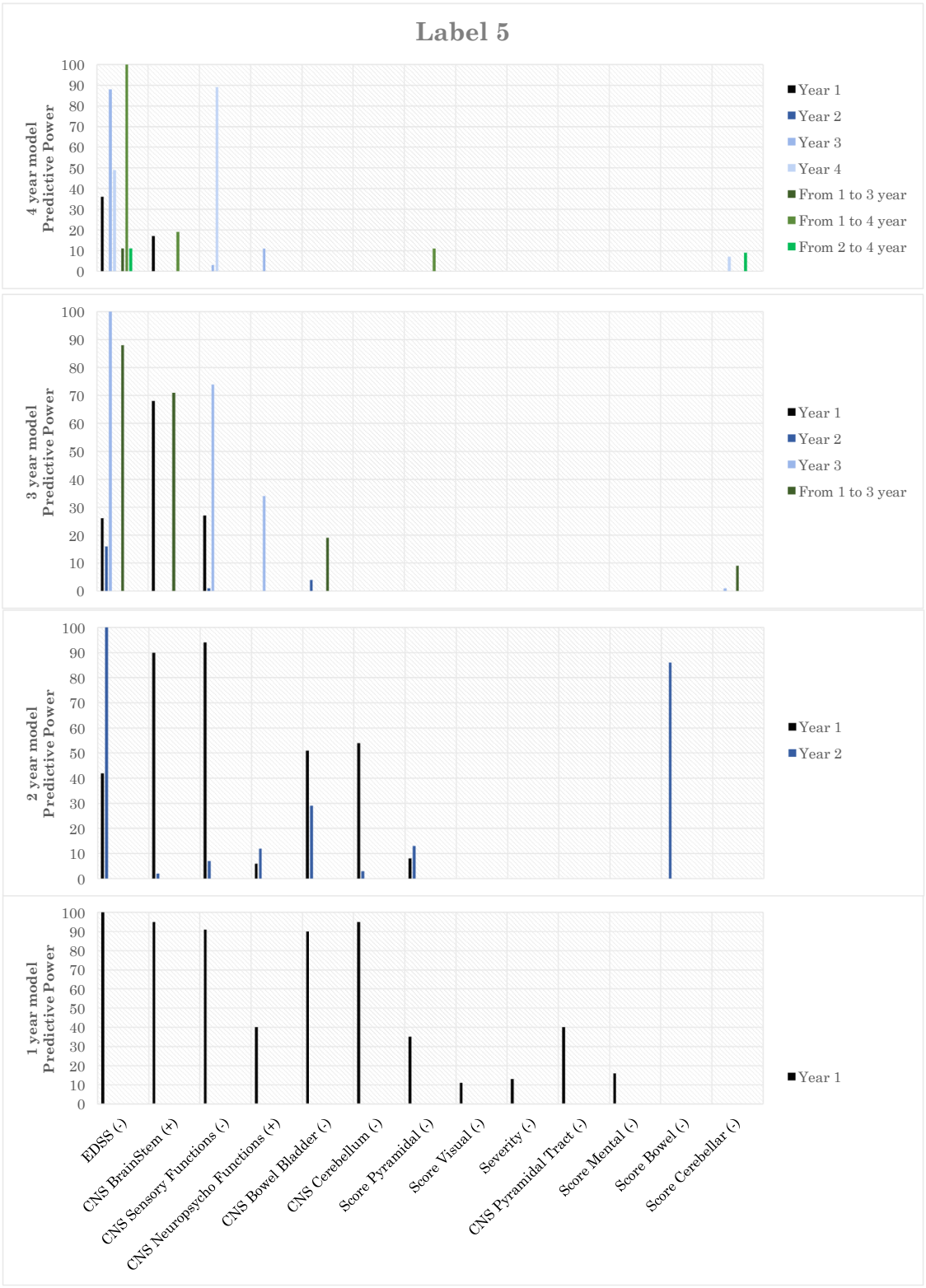

Figure 6.7: Features with highest predictive power of a benign/malignant course identified in classification problem 5. The darker the shade of blue, the closer to the diagnosis is the year when the feature was retrieved and the darker the shade of green the closer to the diagnosis is the accumulation of the years (non-starting on onset) where the feature was identified.

Considering the classification problem 5, the performance of the model increased over time as can be seen in Table 6.6. The LDA was the most appropriate classifier for both first, second and third year models, while the SVM Linear was the best for the fourth year model. The best results were obtained using 4 years of features, where was achieved an AUC of 0.84 $\pm$ 0.05, a geometric mean of 0.77 $\pm$ 0.06, a specificity of 0.79 $\pm$ 0.09, a sensitivity of 0.75 $\pm$ 0.19 and an F1-Score of 0.79 $\pm$ 0.06.

In terms of predictive features, only the Score Bowel and the Score Cerebellar were not identified in the 1-year model. The Score Bowel was an important predictor on the second year model and the Score Cerebellar was identified in the third and fourth year model although it contained a low predictive power in both year models. Among the remaining features, the EDSS, CNS BrainStem, CNS Sensory Functions and CNS Neuropsycho Functions were identified in all year models. From those features, CNS Neuropsycho Functions presented a lower predictive power when compared to the others. Furthermore, the CNS Bowel Bladder was identified in the three initial year models while both CNS Cerebellum and Score Pyramidal were identified only in the two initial year models.

6.2.6 Classification problem 6 - EDSS $\leq$ 4 in the 10th year using baseline visits

Table 6.7: Results of the performance obtained for the classification problem 6.

	Classifier	KNN-3	LDA	SVM Linear	Linear Regression
1 year model	AUC	0.64 ± 0.11	0.69 ± 0.12	**0.68 ± 0.12**	0.67 ± 0.12
	G-Mean	0.38 ± 0.29	0.46 ± 0.26	**0.47 ± 0.24**	0.46 ± 0.26
	Specificity	0.82 ± 0.10	0.84 ± 0.05	**0.85 ± 0.06**	0.84 ± 0.05
	Sensitivity	0.31 ± 0.28	0.34 ± 0.23	**0.34 ± 0.22**	0.34 ± 0.23
	F1-Score	0.31 ± 0.08	0.32 ± 0.10	**0.31 ± 0.10**	0.32 ± 0.10
2 year model	AUC	0.63 ± 0.11	0.68 ± 0.11	**0.68 ± 0.11**	0.66 ± 0.11
	G-Mean	0.40 ± 0.29	0.44 ± 0.29	**0.51 ± 0.20**	0.44 ± 0.28
	Specificity	0.79 ± 0.11	0.82 ± 0.05	**0.84 ± 0.07**	0.82 ± 0.05
	Sensitivity	0.33 ± 0.29	0.34 ± 0.26	**0.37 ± 0.21**	0.35 ± 0.26
	F1-Score	0.30 ± 0.08	0.32 ± 0.09	**0.30 ± 0.10**	0.32 ± 0.09
3 year model	AUC	0.60 ± 0.12	0.65 ± 0.12	**0.64 ± 0.12**	0.65 ± 0.13
	G-Mean	0.40 ± 0.26	0.41 ± 0.28	**0.47 ± 0.23**	0.41 ± 0.28
	Specificity	0.80 ± 0.10	0.82 ± 0.07	**0.81 ± 0.09**	0.82 ± 0.08
	Sensitivity	0.29 ± 0.22	0.29 ± 0.23	**0.33 ± 0.20**	0.30 ± 0.23
	F1-Score	0.29 ± 0.11	0.31 ± 0.12	**0.30 ± 0.14**	0.31 ± 0.12
4 year model	AUC	**0.60 ± 0.12**	0.62 ± 0.10	0.61 ± 0.11	0.61 ± 0.10
	G-Mean	**0.41 ± 0.26**	0.41 ± 0.26	0.41 ± 0.26	0.41 ± 0.27
	Specificity	**0.82 ± 0.10**	0.82 ± 0.07	0.81 ± 0.07	0.81 ± 0.07
	Sensitivity	**0.30 ± 0.23**	0.29 ± 0.22	0.30 ± 0.22	0.30 ± 0.22
	F1-Score	**0.29 ± 0.10**	0.30 ± 0.10	0.29 ± 0.11	0.29 ± 0.10
5 year model	AUC	0.80 ± 0.08	0.82 ± 0.06	0.83 ± 0.05	**0.89 ± 0.07**
	G-Mean	0.79 ± 0.08	0.80 ± 0.08	0.79 ± 0.08	**0.80 ± 0.08**
	Specificity	0.93 ± 0.04	0.93 ± 0.03	0.93 ± 0.04	**0.93 ± 0.03**
	Sensitivity	0.69 ± 0.14	0.69 ± 0.14	0.69 ± 0.14	**0.69 ± 0.14**
	F1-Score	0.61 ± 0.09	0.63 ± 0.09	0.62 ± 0.10	**0.63 ± 0.09**

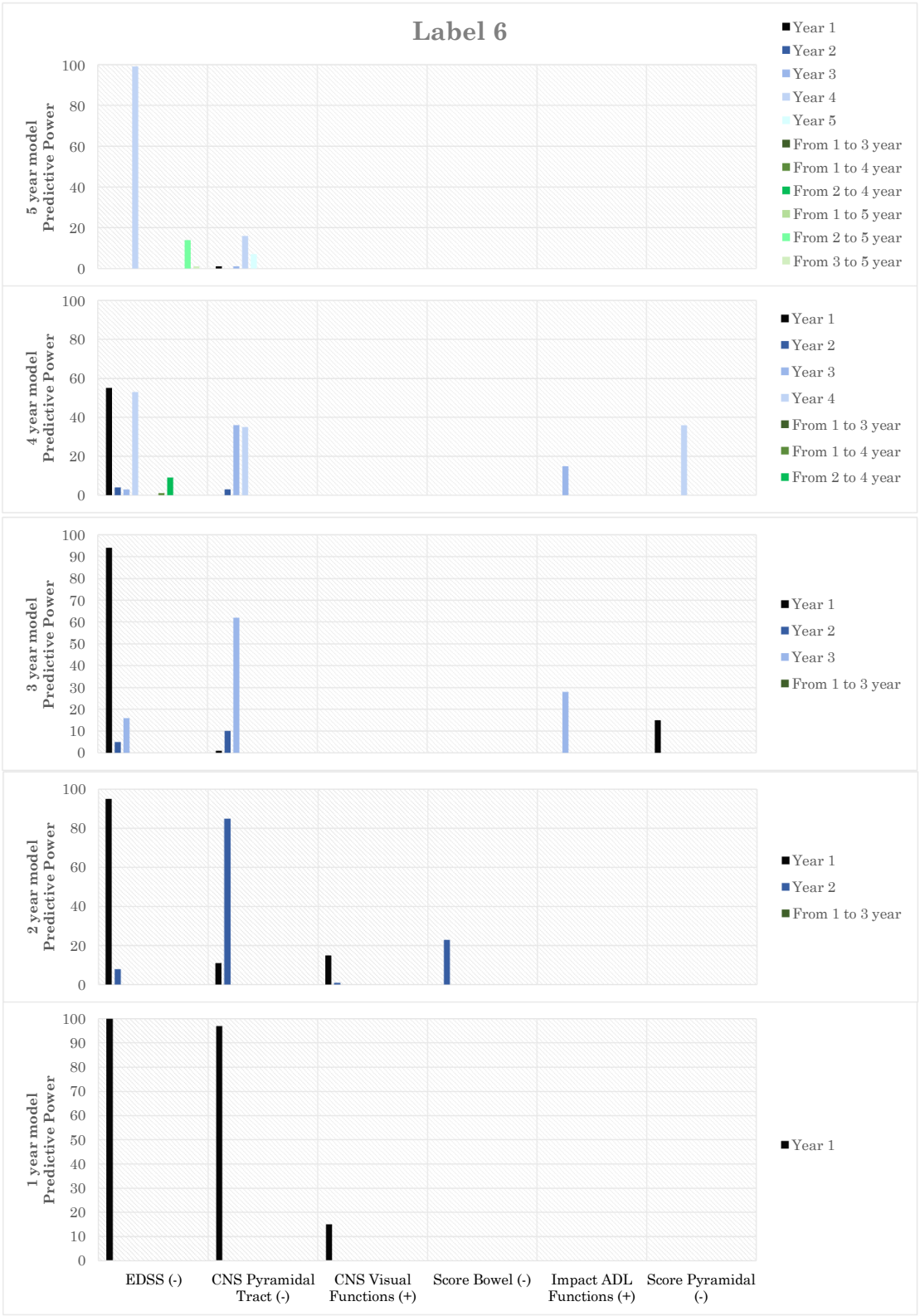

Figure 6.8: Features with highest predictive power of a benign/malignant course identified in classification problem 6. The darker the shade of blue, the closer to the diagnosis is the year when the feature was retrieved and the darker the shade of green the closer to the diagnosis is the accumulation of the years (non-starting on onset) where the feature was identified.

The results of the performance obtained for the classification problem 6, summarized in the Table 6.7, demonstrate that the SVM linear was the best classifier for the three initial year models, the KNN-3 was the appropriate classifier in the fourth year model and the Linear Regression was the best option on the fifth year model. Overall, the performance of the model is poor in the four initial years, presenting low values of sensitivity and F1-Score. In the fifth year model, a different scenario occurs, once the performance is acceptable. The best results were achieved using a Linear Regression on the fifth year with an AUC of 0.89 $\pm$ 0.07, an geometric mean of 0.80 $\pm$ 0.08, a specificity of 0.93 $\pm$ 0.03, a sensitivity of 0.69 $\pm$ 0.14 and an F1-Score of 0.63 $\pm$ 0.09 obtained.

The Figure 6.8 represent the evolution of the predictive features for classification problem 6. Overall, it is possible to note that only 6 features were identified as predictive. The EDSS and CNS Pyramidal Tract were identified in every year model. The CNS Visual Functions was identified in the first two year models and the Score Bowel only in the second year model. Finally, both the Impact ADL Functions and Score Pyramidal were identified in the third and fourth-year models. It is important to note that in terms of predictive power, the EDSS and the CNS Pyramidal Tract appear to be the features with the higher values for the majority of the year models.

Furthermore, the most recent features are also the most predictive in this case.

6.2.7 Classification problem 7 - EDSS $\leq$ 3 in the 6th year

Table 6.8: Results of the performance obtained for the classification problem 7.

Best results	Classifier	KNN-3	LDA	SVM Linear	Linear Regression
1 year model	AUC:	0.73 ± 0.05	**0.78 ± 0.05**	0.79 ± 0.05	0.75 ± 0.05
	G-Mean:	0.66 ± 0.06	**0.73 ± 0.06**	0.72 ± 0.05	0.73 ± 0.06
	Specificity:	0.61 ± 0.06	**0.74 ± 0.05**	0.68 ± 0.08	0.74 ± 0.05
	Sensitivity:	0.73 ± 0.11	**0.73 ± 0.11**	0.76 ± 0.12	0.73 ± 0.11
	F1-Score:	0.32 ± 0.05	**0.41 ± 0.06**	0.38 ± 0.06	0.41 ± 0.06
2 year model	AUC:	0.79 ± 0.05	0.87 ± 0.03	**0.88 ± 0.04**	0.86 ± 0.04
	G-Mean:	0.73 ± 0.05	0.81 ± 0.05	**0.80 ± 0.05**	0.81 ± 0.05
	Specificity:	0.65 ± 0.07	0.85 ± 0.03	**0.82 ± 0.05**	0.85 ± 0.03
	Sensitivity:	0.82 ± 0.09	0.77 ± 0.10	**0.79 ± 0.11**	0.77 ± 0.10
	F1-Score:	0.38 ± 0.05	0.54 ± 0.06	**0.51 ± 0.06**	0.54 ± 0.06
3 year model	AUC:	0.84 ± 0.06	0.90 ± 0.04	**0.91 ± 0.04**	0.89 ± 0.04
	G-Mean:	0.79 ± 0.06	0.83 ± 0.06	**0.83 ± 0.05**	0.83 ± 0.06
	Specificity:	0.75 ± 0.05	0.85 ± 0.03	**0.80 ± 0.04**	0.85 ± 0.03
	Sensitivity:	0.83 ± 0.11	0.81 ± 0.11	**0.86 ± 0.11**	0.81 ± 0.11
	F1-Score:	0.46 ± 0.07	0.56 ± 0.06	**0.52 ± 0.06**	0.56 ± 0.06
4 year model	AUC:	0.87 ± 0.05	0.90 ± 0.05	**0.90 ± 0.04**	0.90 ± 0.04
	G-Mean:	0.82 ± 0.06	0.81 ± 0.07	**0.82 ± 0.05**	0.81 ± 0.07
	Specificity:	0.84 ± 0.04	0.88 ± 0.04	**0.83 ± 0.05**	0.88 ± 0.04
	Sensitivity:	0.80 ± 0.12	0.75 ± 0.12	**0.82 ± 0.10**	0.75 ± 0.12
	F1-Score:	0.54 ± 0.07	0.58 ± 0.09	**0.54 ± 0.06**	0.58 ± 0.09
5 year model	AUC:	**0.91 ± 0.04**	0.87 ± 0.05	0.93 ± 0.02	0.89 ± 0.04
	G-Mean:	**0.82 ± 0.06**	0.76 ± 0.07	0.81 ± 0.06	0.76 ± 0.07
	Specificity:	**0.90 ± 0.04**	0.89 ± 0.03	0.89 ± 0.03	0.89 ± 0.03
	Sensitivity:	**0.76 ± 0.13**	0.66 ± 0.11	0.73 ± 0.11	0.66 ± 0.11
	F1-Score:	**0.61 ± 0.07**	0.54 ± 0.08	0.59 ± 0.07	0.54 ± 0.08

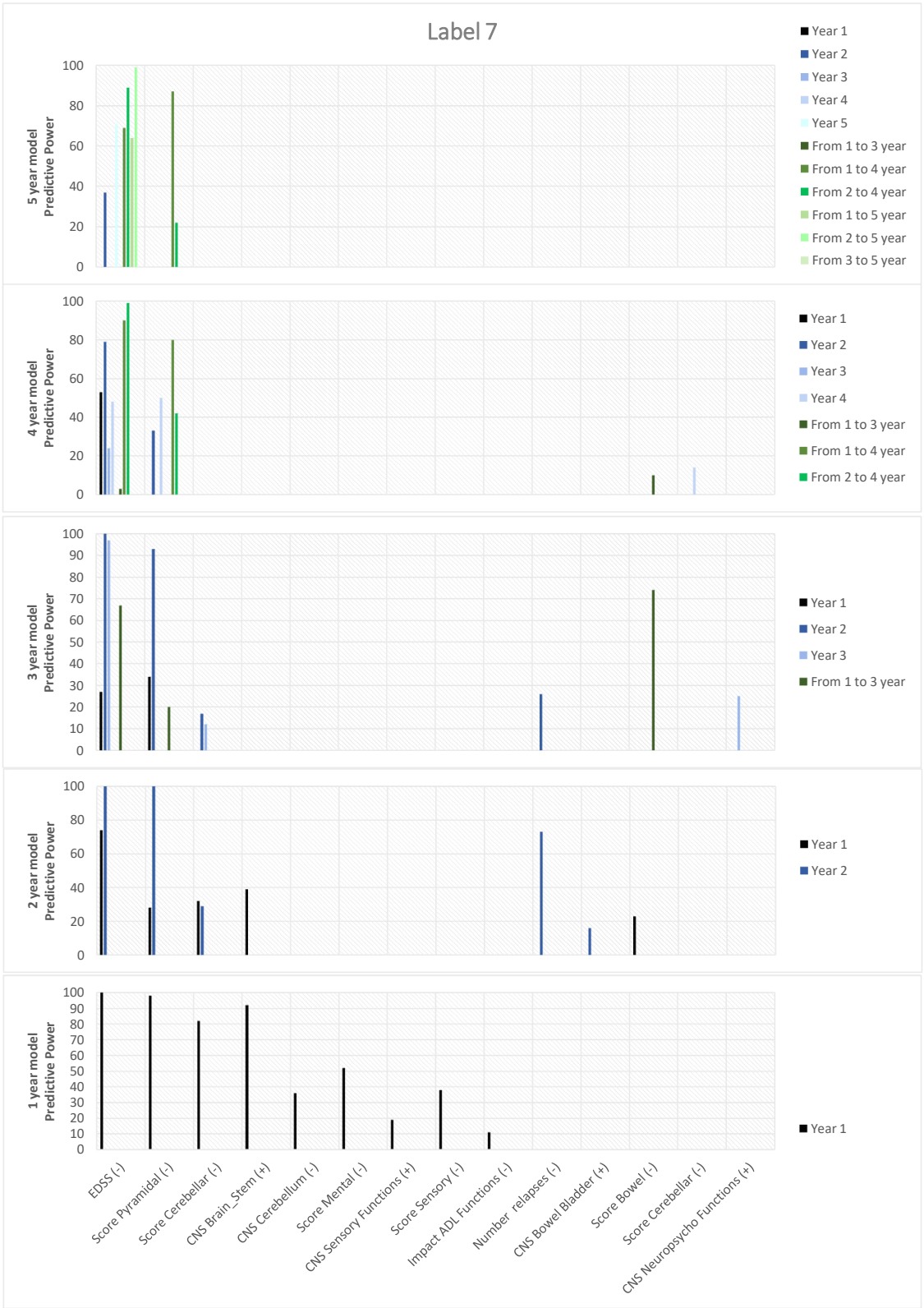

Figure 6.9: Features with highest predictive power of a benign/malignant course identified in classification problem 7. The darker the shade of blue, the closer to the diagnosis is the year when the feature was retrieved and the darker the shade of green the closer to the diagnosis is the accumulation of the years (non-starting on onset) where the feature was identified.

The results obtained for the classification problem 7 are presented in Table 6.8. It can be seen that a good performance was obtained in all year models. Overall, the third-year model can be selected as the best performing year, once it was obtained the highest values of G-mean, AUC and Sensitivity among all year models and the other metrics hadn't a great difference when compared to other year models. To visualize such result, it can be noted that using an SVM linear and three years of features, an AUC of 0.91 ± 0.04, a geometric mean of 0.83 ± 0.05, a specificity of 0.80 ± 0.04, a sensitivity of 0.86 ± 0.11 and an F1-Score of 0.52 ± 0.06 was obtained. Furthermore, it is important to highlight that the LDA was the best classifier for the first-year model, the KNN-3 was the best for the fifth year model and the SVM linear was selected in the remaining year models.

In Figure 6.9 it can be noted that both Score Pyramidal and EDSS are predictive in all year models. Among the other features identified as predictive in the first year, only the Score Cerebellar and the Score Brainstem are also predictive in other year models. It is also interesting to visualize other cases such as the number of relapses, the CNS Bowel Bladder and the Score Bowel that only started to be predictive in the second year model. Lastly, it can be noted that the CNS Neuropsycho Functions and the Score cerebellar were exclusively predictive in the third and fourth-year model, respectively.

6.2.8 Classification problem 8 - Increase EDSS < 1.5 after 5 years

Table 6.9: Results of the performance obtained for the classification problem 8.

	Classifier	KNN-3	LDA	SVM Linear	Linear Regression
1 year model	AUC	0.53 ± 0.08	**0.61 ± 0.09**	0.58 ± 0.10	0.55 ± 0.09
	G-Mean	0.49 ± 0.08	**0.56 ± 0.07**	0.53 ± 0.07	0.56 ± 0.07
	Specificity	0.36 ± 0.09	**0.41 ± 0.09**	0.38 ± 0.09	0.42 ± 0.09
	Sensitivity	0.70 ± 0.17	**0.78 ± 0.14**	0.77 ± 0.14	0.77 ± 0.14
	F1-Score	0.10 ± 0.02	**0.12 ± 0.02**	0.12 ± 0.02	0.12 ± 0.02
2 year model	AUC	0.61 ± 0.09	**0.69 ± 0.08**	0.68 ± 0.08	0.62 ± 0.12
	G-Mean	0.58 ± 0.10	**0.64 ± 0.08**	0.63 ± 0.09	0.64 ± 0.08
	Specificity	0.49 ± 0.10	**0.56 ± 0.09**	0.55 ± 0.10	0.56 ± 0.09
	Sensitivity	0.71 ± 0.20	**0.76 ± 0.15**	0.75 ± 0.18	0.74 ± 0.15
	F1-Score	0.13 ± 0.03	**0.15 ± 0.03**	0.15 ± 0.03	0.15 ± 0.03
3 year model	AUC	0.60 ± 0.08	**0.67 ± 0.08**	0.66 ± 0.07	0.58 ± 0.10
	G-Mean	0.57 ± 0.08	**0.63 ± 0.09**	0.61 ± 0.08	0.62 ± 0.09
	Specificity	0.50 ± 0.13	**0.57 ± 0.11**	0.55 ± 0.12	0.57 ± 0.11
	Sensitivity	0.69 ± 0.17	**0.72 ± 0.18**	0.72 ± 0.17	0.70 ± 0.18
	F1-Score	0.13 ± 0.03	**0.15 ± 0.04**	0.14 ± 0.03	0.15 ± 0.04
4 year model	AUC	0.61 ± 0.08	**0.69 ± 0.06**	0.65 ± 0.06	0.60 ± 0.09
	G-Mean	0.57 ± 0.07	**0.62 ± 0.07**	0.58 ± 0.08	0.62 ± 0.07
	Specificity	0.53 ± 0.10	**0.62 ± 0.08**	0.62 ± 0.09	0.62 ± 0.08
	Sensitivity	0.64 ± 0.15	**0.64 ± 0.12**	0.56 ± 0.16	0.63 ± 0.12
	F1-Score	0.12 ± 0.03	**0.15 ± 0.03**	0.13 ± 0.03	0.15 ± 0.03

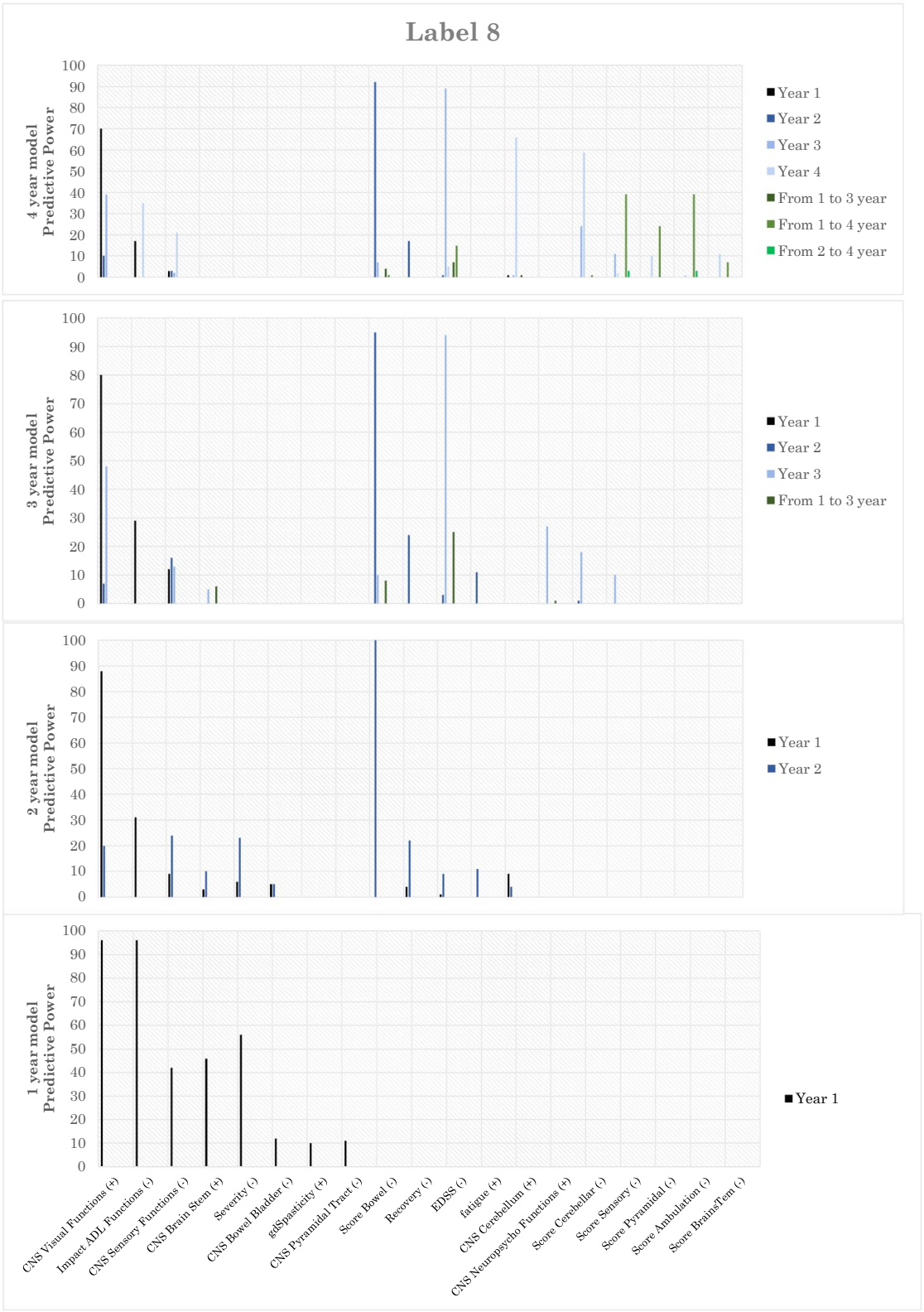

Figure 6.10: Features with highest predictive power of a benign/malignant course identified in classification problem 8. The darker the shade of blue, the closer to the diagnosis is the year when the feature was retrieved and the darker the shade of green the closer to the diagnosis is the accumulation of the years (non-starting on onset) where the feature was identified.

The results obtained for the classification problem 8 are presented in Table 6.9. It can be noted that the LDA was the best performing classifier for all year models. Although the model presents an acceptable value of sensitivity, it was obtained a low value of F1-Score in every year model. The best performing year was the fourth, with an AUC of 0.69 ± 0.06, a geometric mean of 0.62 ± 0.07, a specificity of 0.62 ± 0.08, a sensitivity of 0.64 ± 0.12 and an F1-Score of 0.15 ± 0.03.

As can be seen in Figure 6.10, it was identified several distinct features. The most predictive features in the first year are the CNS Visual Functions, the Impact ADL Functions and the Severity. On the second year model, the CNS Visual Functions remain as one of the most predictive features alongside with Score Bowel. On the remaining year models, those two features present also a high value of predictive power together with EDSS. From the remaining features, it was identified 4 features (Score Sensory, Score Pyramidal, Score Ambulation and Score Brainstem) that were only predictive on the fourth year model.

6.2.9 Classification problem 9 - Progression index < 0,2 after a duration of 5 years

Table 6.10: Results of the performance obtained for the classification problem 9.

Best results	Classifier	KNN-3	LDA	SVM Linear	Linear Regression
1 year model	AUC:	0.58 ± 0.09	**0.68 ± 0.10**	0.66 ± 0.09	0.65 ± 0.09
	G-Mean:	0.50 ± 0.08	**0.60 ± 0.10**	0.57 ± 0.09	0.61 ± 0.10
	Specificity:	0.35 ± 0.11	**0.55 ± 0.13**	0.51 ± 0.15	0.56 ± 0.12
	Sensitivity:	0.76 ± 0.14	**0.70 ± 0.19**	0.70 ± 0.19	0.69 ± 0.19
	F1-Score:	0.19 ± 0.04	**0.24 ± 0.06**	0.22 ± 0.05	0.24 ± 0.06
2 year model	AUC:	0.61 ± 0.08	**0.70 ± 0.07**	0.68 ± 0.07	0.65 ± 0.07
	G-Mean:	0.55 ± 0.07	**0.63 ± 0.07**	0.59 ± 0.08	0.63 ± 0.07
	Specificity:	0.43 ± 0.12	**0.58 ± 0.10**	0.53 ± 0.15	0.58 ± 0.10
	Sensitivity:	0.74 ± 0.16	**0.70 ± 0.15**	0.70 ± 0.16	0.69 ± 0.15
	F1-Score:	0.21 ± 0.04	**0.25 ± 0.05**	0.24 ± 0.05	0.25 ± 0.05
3 year model	AUC:	**0.69 ± 0.07**	0.73 ± 0.08	0.71 ± 0.08	0.69 ± 0.07
	G-Mean:	**0.66 ± 0.07**	0.64 ± 0.09	0.64 ± 0.09	0.64 ± 0.08
	Specificity:	**0.65 ± 0.07**	0.72 ± 0.07	0.68 ± 0.08	0.72 ± 0.07
	Sensitivity:	**0.69 ± 0.13**	0.58 ± 0.15	0.62 ± 0.17	0.57 ± 0.15
	F1-Score:	**0.28 ± 0.05**	0.28 ± 0.07	0.27 ± 0.07	0.28 ± 0.07
4 year model	AUC:	**0.68 ± 0.08**	0.70 ± 0.07	0.71 ± 0.08	0.67 ± 0.08
	G-Mean:	**0.67 ± 0.08**	0.58 ± 0.10	0.64 ± 0.08	0.58 ± 0.10
	Specificity:	**0.72 ± 0.07**	0.76 ± 0.07	0.67 ± 0.07	0.77 ± 0.07
	Sensitivity:	**0.63 ± 0.14**	0.46 ± 0.15	0.62 ± 0.14	0.46 ± 0.15
	F1-Score:	**0.30 ± 0.07**	0.25 ± 0.07	0.27 ± 0.06	0.25 ± 0.07

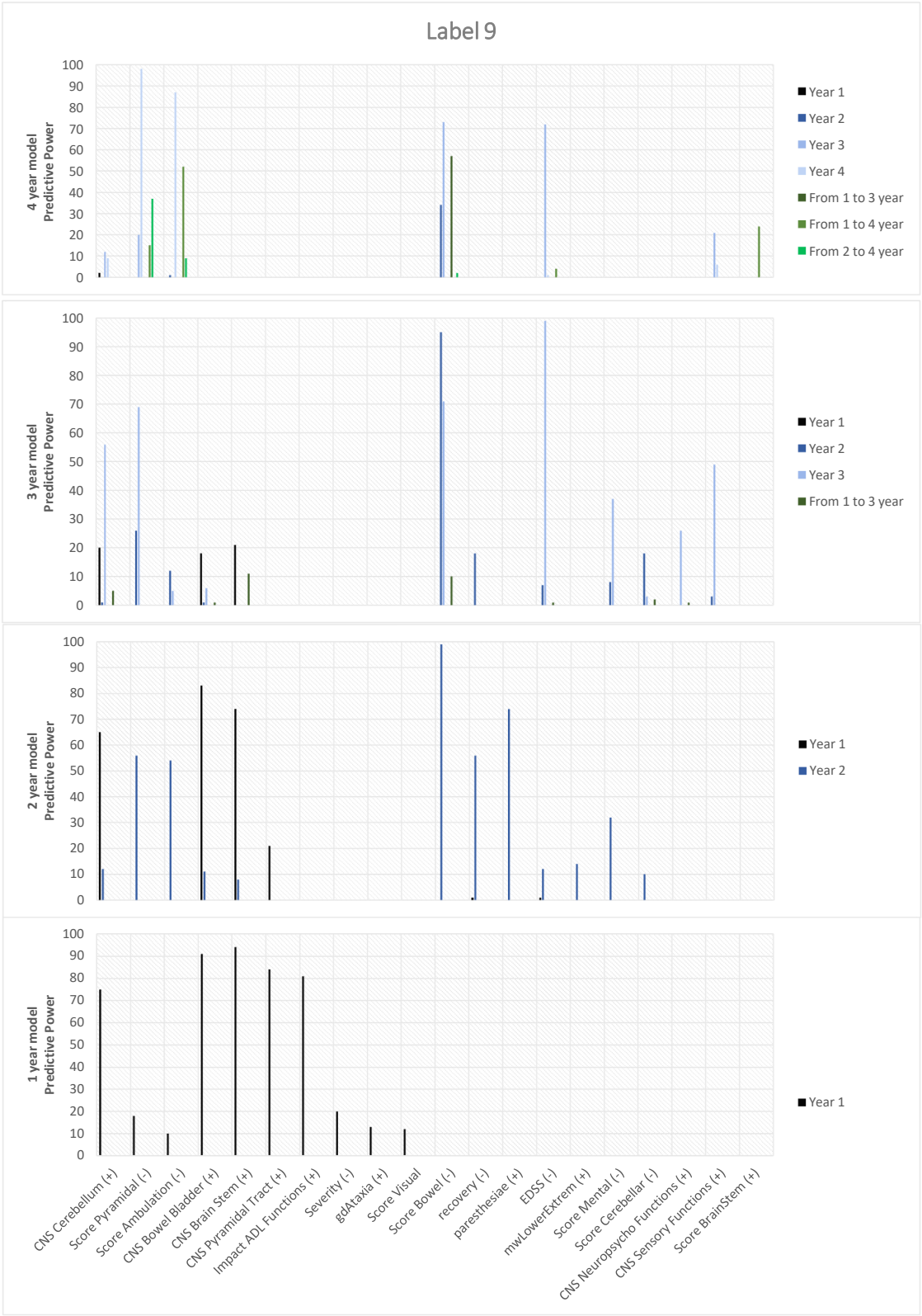

Figure 6.11: Features with highest predictive power of a benign/malignant course identified in classification problem 9. The darker the shade of blue, the closer to the diagnosis is the year when the feature was retrieved and the darker the shade of green the closer to the diagnosis is the accumulation of the years (non-starting on onset) where the feature was identified.

The results obtained for the classification problem 9 are presented in Table 6.10. Overall, the LDA was the best classifier in the two initial year models while the KNN-3 was the best performing in the two last year models. The F1-Score metric exhibits low results for in all classifiers for all year models. The third-year with KNN-3 as classifier it is the best performing case with an AUC of 0.69 $\pm$ 0.07, a geometric mean of 0.66 $\pm$ 0.07, a specificity of 0.65 $\pm$ 0.07, a sensitivity of 0.69 $\pm$ 0.13 and an F1-Score of 0.28 $\pm$ 0.05.

In Figure 6.11 it is possible to visualize that half of the features were identified as predictive in the first year model and the other half were only predictive over the following years. From the features identified in the first year, only the CNS Cerebellum, the Score Pyramidal and the Score ambulation were predictive in all year models. Considering the features not predictive in the first year, the Score Brainstem was the unique feature that only started to be predictive in the fourth year model, once the other features were predominantly predictive in the second and third year models.

6.2.10 Comparison of the results of benign/malignant classification problems

Table 6.11: Best results obtained in disease severity classification problems using 1 year of features.

Best results 1 year model					
	AUC	**G-Mean**	**Specificity**	**Sensitivity**	**F1-Score**
Classification problem 1 (EDSS $\leq$ 2 after 10th year using baseline visits)	0.62 ± 0.08	0.56 ± 0.08	0.61 ± 0.14	0.52 ± 0.11	0.61 ± 0.09
Classification problem 2 (EDSS $\leq$ 3 after 10th year using baseline visits)	0.67 ± 0.09	0.60 ± 0.11	0.65 ± 0.12	0.58 ± 0.17	0.48 ± 0.11
Classification problem 3 (EDSS $\leq$ 3 in the 10 year)	**0.81 ± 0.13**	**0.68 ± 0.17**	**0.81 ± 0.11**	**0.62 ± 0.25**	**0.58 ± 0.18**
Classification problem 4 (EDSS $\leq$ 3 during the initial 10 years)	**0.82 ± 0.06**	**0.70 ± 0.07**	**0.84 ± 0.13**	**0.60 ± 0.10**	**0.73 ± 0.07**
Classification problem 5 (EDSS $\leq$ 2 after 5 years and $\leq$ 3 after 10 years)	0.74 ± 0.08	0.65 ± 0.09	0.78 ± 0.17	0.56 ± 0.12	0.66 ± 0.10
Classification problem 6 (EDSS $\leq$ 4 after 10 years using baseline visits)	0.68 ± 0.12	0.47 ± 0.24	0.85 ± 0.06	0.34 ± 0.22	0.31 ± 0.10
Classification problem 7 (EDSS $\leq$ 3 in the 6th year)	**0.78 ± 0.05**	**0.73 ± 0.06**	**0.74 ± 0.05**	**0.73 ± 0.11**	**0.41 ± 0.06**
Classification problem 8 (Increase EDSS $<$ 1.5 after 5 years)	0.61 ± 0.09	0.56 ± 0.07	0.41 ± 0.09	0.78 ± 0.14	0.12 ± 0.02
Classification problem 9 (Progression index $<$0.2 after a duration of 5 years)	0.68 ± 0.10	0.60 ± 0.10	0.55 ± 0.13	0.70 ± 0.19	0.24 ± 0.06

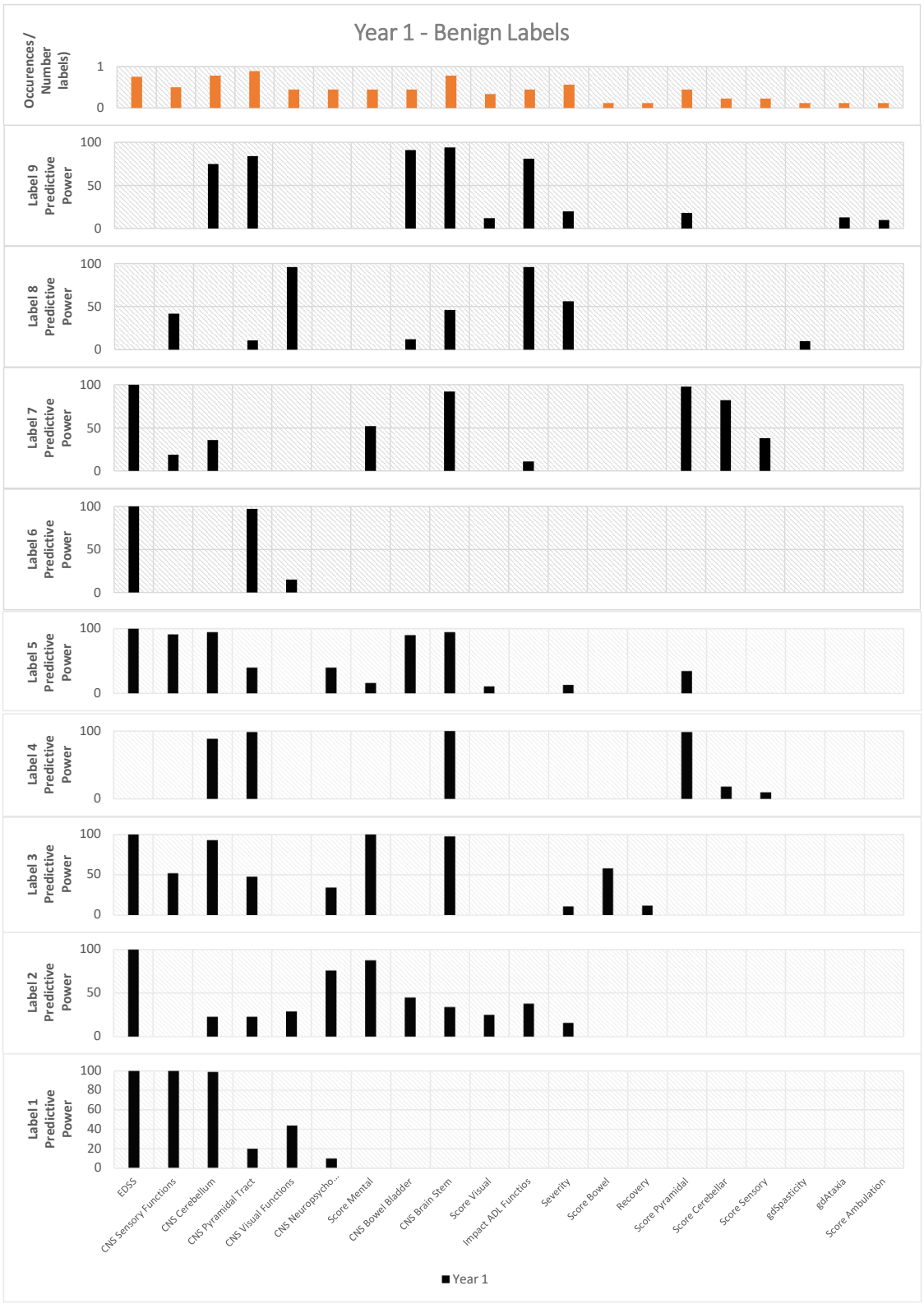

Figure 6.12: Features with highest predictive power of a benign/malignant course identified in the first year.

The best results for every classification problem using 1 year of features are presented in Table 6.11. Overall, it is possible to note several differences between the results obtained for different classification problems. Firstly, the classification problems 6, 8 and 9 are the worst-performing labels and all have in common the quite imbalanced dataset and the consideration of 10 years of follow-up. From the remaining labels, the classification problem 3, 4 and 7 were identified as the best-performing labels. All those classification problems use a value of EDSS of 3 as a threshold, although the follow-up period is different (10 years for classification problem 3 and 4, and 6 years for classification problem 7). Moreover, those 3 classification problems consider baseline and non-baseline information. Once the classification problem 1 and classification problem 2 present slightly worse results than the three best labels, can indicate that the presence of non-baseline information on the first year contributes to better results. Lastly, classification problem 5 exhibits good performance, similar to the best labels, although it was not chosen as one of the best-performing labels once it contains a slightly worse sensitivity.

Regarding the predictive features, the Figure 6.12 presents an overview of all features identified in each classification problem using only 1 year of features. The top part of the figure, represented in orange, indicates the relative frequency of the occurrence of the features among all labels in order to facilitate the visualization of the most frequent features. The relative frequency was given by the number of features where the feature was identified as predictive divided by the total number of labels. It is important to note that in the case of the classification problem 4 the EDSS was not considered as a feature and for such reason, the total number of labels considered was only 8 in that case.

Among the totality of the features identified as predictive in the first year, the CNS Pyramidal Tract was the most commonly identified, being present in 8 out of 9 labels. Moreover, the CNS Cerebellum and CNS Brainstem were also important predictors, once they were present in 7 out of 9 labels. The EDSS was identified only in 6 out of 8 labels, although it presented a predictive power of 100 which indicates that is an essential predictor for such labels.

The reason to identify the most common features among the classification problems is due to the fact that if one feature is identified regardless of the classification problem considered, it is more probable that the feature can be assumed as a known predictor of a benign case.

Table 6.12: Best results obtained in disease severity classification problems using 2 years of features.

Best results 2 year model					
	AUC	**G-Mean**	**Specificity**	**Sensitivity**	**F1-Score**
Classification problem 1 (EDSS $\leq$ 2 after 10th year using baseline visits)	0.62 ± 0.09	0.59 ± 0.09	0.54 ± 0.15	0.66 ± 0.12	0.71 ± 0.08
Classification problem 2 (EDSS $\leq$ 3 after 10th year using baseline visits)	0.69 ± 0.09	0.62 ± 0.10	0.61 ± 0.09	0.66 ± 0.17	0.51 ± 0.11
Classification problem 3 (EDSS $\leq$ 3 in the 10 year)	0.78 ± 0.11	0.65 ± 0.19	0.80 ± 0.12	0.59 ± 0.25	0.56 ± 0.15
Classification problem 4 (EDSS $\leq$ 3 during the initial 10 years)	**0.80 ± 0.08**	**0.70 ± 0.08**	**0.83 ± 0.15**	**0.61 ± 0.11**	**0.74 ± 0.09**
Classification problem 5 (EDSS $\leq$ 2 after 5 years and $\leq$ 3 after 10 years)	**0.75 ± 0.05**	**0.70± 0.07**	**0.76 ± 0.13**	**0.65 ± 0.11**	**0.72 ± 0.07**
Classification problem 6 (EDSS $\leq$ 4 after 10 years using baseline visits)	0.68 ± 0.11	0.51 ± 0.20	0.84 ± 0.07	0.37 ± 0.21	0.30 ± 0.10
Classification problem 7 (EDSS $\leq$ 3 in the 6th year)	**0.88 ± 0.04**	**0.80 ± 0.05**	**0.82 ± 0.05**	**0.79 ± 0.11**	**0.51 ± 0.06**
Classification problem 8 (Increase EDSS $<$ 1.5 after 5 years)	0.69 ± 0.08	0.64 ± 0.08	0.56 ± 0.09	0.76 ± 0.15	0.15 ± 0.03
Classification problem 9 (Progression index $<$0.2 after a duration of 5 years)	0.70 ± 0.07	0.63 ± 0.07	0.58 ± 0.10	0.70 ± 0.15	0.25 ± 0.05

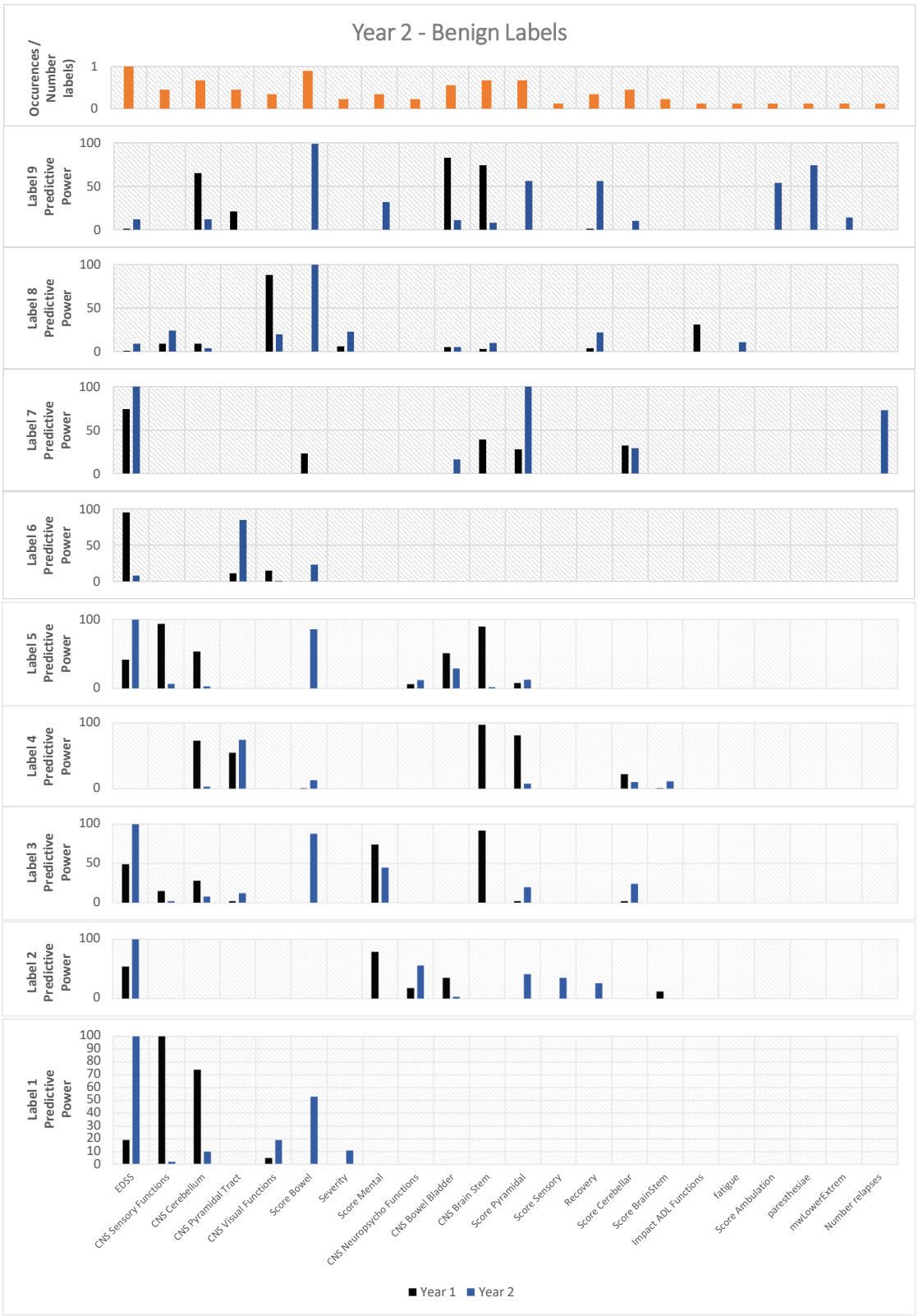

Figure 6.13: Features with highest predictive power of a benign/malignant course identified in the second year. The darker the shade of blue, the closer to the diagnosis is the year when the feature was retrieved.

The Table 6.12 summarizes the best results achieved in each classification problem using 2 years of features. Similarly to the previous case, the classification problems 6, 8 and 9 are the worst-performing labels. Although the classification problems 8 and 9 present a good result of sensitivity, they also present a quite low value of F1-Score. Regarding the other classification problems, the best-performing labels, in this case, are the label 4, 5 and 7. Those 3 labels have in common the fact that they consider baseline and non-baseline information. It is important to note that the classification problems 1 and 2 were also performing well, containing a value of sensitivity higher than 2 of the best labels, although they present lower values of AUC, G-Mean and Specificity. The classification problem 3 exhibited a lower value of sensitivity when compared to the three best labels. Overall, it is possible to identify a similar scenario when compared to the previous case where only 1 year of features was used. In both cases, the best-performing labels used non-baseline information and the worst performing labels correspond to imbalanced datasets.

Lastly, it is important to note that the performance of the model increased for the majority of the labels by the addition of the second year of features.

Regarding the features most commonly identified as predictive, the EDSS was the unique feature present in all possible classification problems. Moreover, the Score Bowel was also an important predictor in 8 out of 9 classification problems and the CNS Cerebellum, the CNS Brainstem and Score Pyramidal were the other important predictors identified in 6 out of 9 classification problems.

Table 6.13: Best results obtained in disease severity classification problems using 3 years of features.

Best results 3 year model					
	AUC	**G-Mean**	**Specificity**	**Sensitivity**	**F1-Score**
Classification problem 1 (EDSS $\leq$ 2 after 10th year using baseline visits)	**0.68 ± 0.06**	**0.63 ± 0.06**	**0.66 ± 0.14**	**0.61 ± 0.10**	**0.69 ± 0.07**
Classification problem 2 (EDSS $\leq$ 3 after 10th year using baseline visits)	0.69 ± 0.11	0.62 ± 0.12	0.66 ± 0.11	0.61 ± 0.18	0.51 ± 0.12
Classification problem 3 (EDSS $\leq$ 3 in the 10 year)	0.74 ± 0.11	0.64 ± 0.20	0.78 ± 0.12	0.58 ± 0.27	0.54 ± 0.16
Classification problem 4 (EDSS $\leq$ 3 during the initial 10 years)	0.78 ± 0.09	0.69 ± 0.09	0.89 ± 0.15	0.55 ± 0.10	0.70 ± 0.09
Classification problem 5 (EDSS $\leq$ 2 after 5 years and $\leq$ 3 after 10 years)	**0.82 ± 0.05**	**0.75± 0.07**	**0.76 ± 0.12**	**0.75 ± 0.10**	**0.79 ± 0.06**
Classification problem 6 (EDSS $\leq$ 4 after 10 years using baseline visits)	0.64 ± 0.12	0.47 ± 0.23	0.81 ± 0.09	0.33 ± 0.20	0.30 ± 0.14
Classification problem 7 (EDSS $\leq$ 3 in the 6th year)	**0.91 ± 0.04**	**0.83 ± 0.05**	**0.80 ± 0.04**	**0.86 ± 0.11**	**0.52 ± 0.06**
Classification problem 8 (Increase EDSS $<$ 1.5 after 5 years)	0.67 ± 0.08	0.63 ± 0.09	0.57 ± 0.11	0.72 ± 0.18	0.15 ± 0.04
Classification problem 9 (Progression index $<$0.2 after a duration of 5 years)	0.69 ± 0.08	0.67 ± 0.08	0.72 ± 0.07	0.63 ± 0.14	0.30 ± 0.07

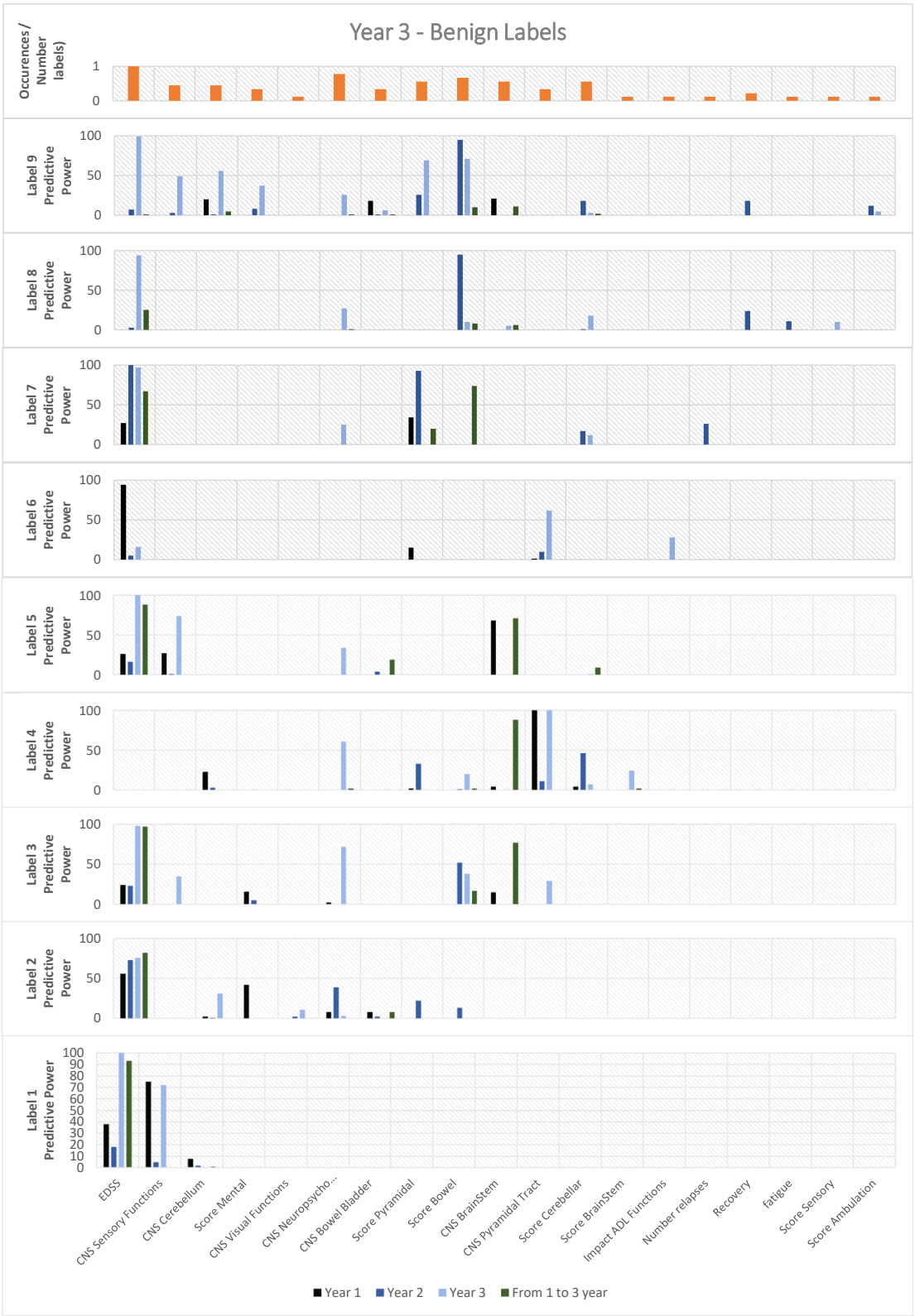

Figure 6.14: Features with highest predictive power of a benign/malignant course identified in the third year. The darker the shade of blue, the closer to the diagnosis is the year when the feature was retrieved and the darker the shade of green the closer to the diagnosis is the accumulation of the years (non-starting on onset) where the feature was identified.

Analysing the Table 6.13 the three best-performing labels using 3 years of features are the classification problem 1, 5 and 7. There is still a significant difference between the results obtained with the classification problems 5 and 7 and the classification problem 1, once the classification problem 1 has a lower performance than the other two labels. This can be resultant from the fact that besides the threshold at the tenth year, the classification problem 5 also considers a threshold in the fifth year, and the classification problem 7 considers a threshold in the sixth year, which occurs earlier than the classification problem 1. For this reason, with three years of features, it appears that the labels predicting earlier scenarios perform better. Even for the classification problem 9, the results were better, even though the F1-Score value remained low. The classification problem 8 is the most imbalanced scenario, with the lower number of malignant cases, which explains the poor performance even for a prediction in the fifth year. The overall performance is also increasing for the majority of the labels by the addition of the third year of features.

Regarding the predictive features, the EDSS was identified as predictive in all classification problems. Moreover, the CNS Neuropsycho Functions was identified in 7 out of 9 classification problems and the Score Bowel in 6 out of 9 classification problems. The remaining features were only coincident in 5 or fewer classification problems.

Table 6.14: Best results obtained in disease severity classification problems using 4 years of features.

Best results 4 year model					
	AUC	**G-Mean**	**Specificity**	**Sensitivity**	**F1-Score**
Classification problem 1 (EDSS $\leq$ 2 after 10th year using baseline visits)	0.76 ± 0.08	0.70 ± 0.08	0.76 ± 0.14	0.66 ± 0.12	0.74 ± 0.08
Classification problem 2 (EDSS $\leq$ 3 after 10th year using baseline visits)	0.74 ± 0.09	0.66 ± 0.11	0.76 ± 0.13	0.59 ± 0.19	0.55 ± 0.13
Classification problem 3 (EDSS $\leq$ 3 in the 10 year)	**0.81 ± 0.12**	**0.70 ± 0.19**	**0.78 ± 0.11**	**0.68 ± 0.27**	**0.60 ± 0.17**
Classification problem 4 (EDSS $\leq$ 3 during the initial 10 years)	0.81 ± 0.06	0.72 ± 0.08	0.91 ± 0.12	0.57 ± 0.10	0.72 ± 0.08
Classification problem 5 (EDSS $\leq$ 2 after 5 years and $\leq$ 3 after 10 years)	**0.84 ± 0.05**	**0.77 ± 0.06**	**0.79 ± 0.09**	**0.75 ± 0.09**	**0.79 ± 0.06**
Classification problem 6 (EDSS $\leq$ 4 after 10 years using baseline visits)	0.60 ± 0.12	0.41 ± 0.26	0.82 ± 0.10	0.30 ± 0.23	0.29 ± 0.10
Classification problem 7 (EDSS $\leq$ 3 in the 6th year)	**0.90 ± 0.04**	**0.82 ± 0.05**	**0.83 ± 0.05**	**0.82 ± 0.10**	**0.54 ± 0.06**
Classification problem 8 (Increase EDSS $<$ 1.5 after 5 years)	0.69 ± 0.06	0.62 ± 0.07	0.62 ± 0.08	0.64 ± 0.12	0.15 ± 0.04
Classification problem 9 (Progression index $<$0.2 after a duration of 5 years)	0.69 ± 0.08	0.67 ± 0.08	0.72 ± 0.07	0.63 ± 0.14	0.30 ± 0.07

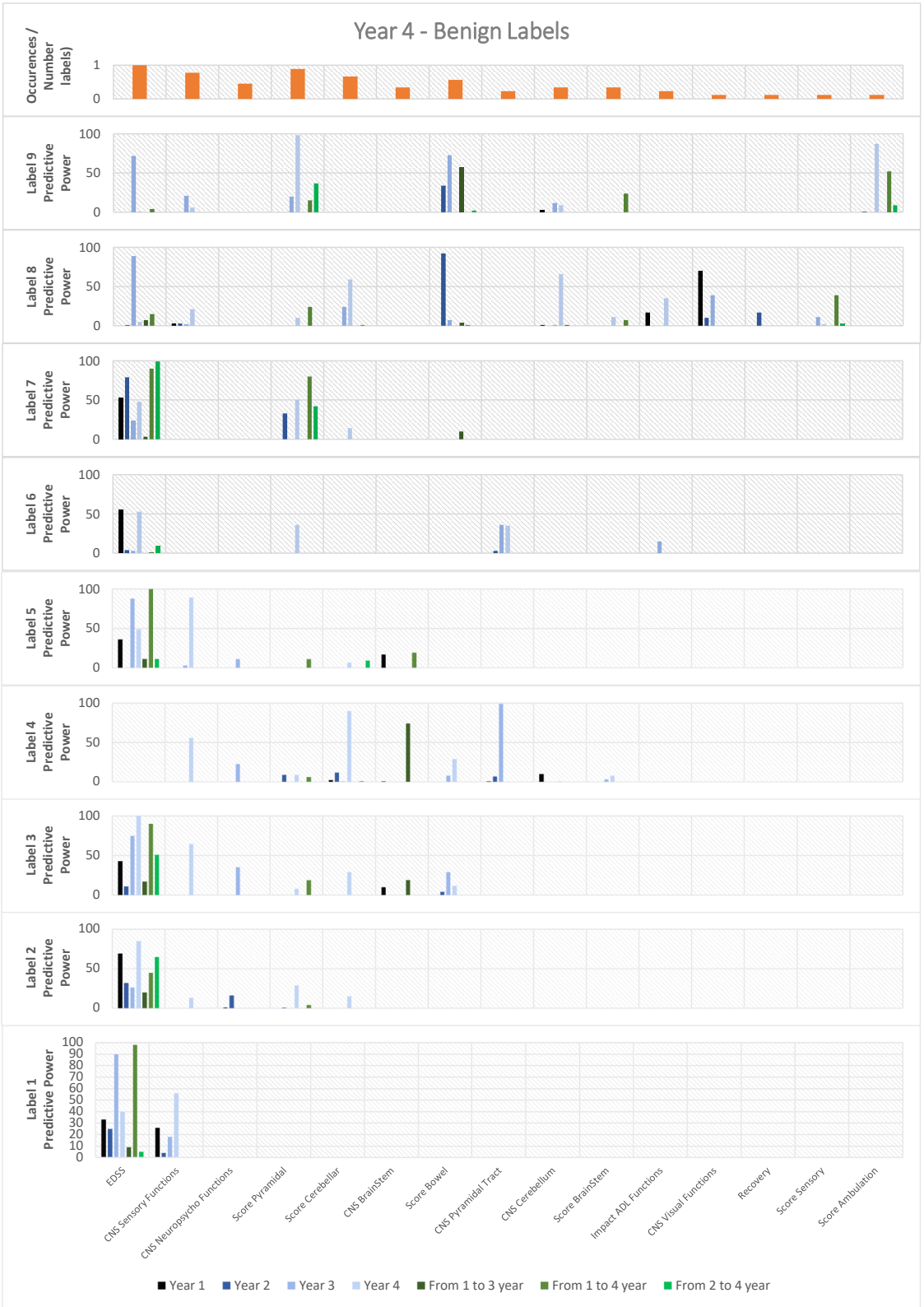

Figure 6.15: Features with highest predictive power of a benign/malignant course identified in the fourth year. The darker the shade of blue, the closer to the diagnosis is the year when the feature was retrieved and the darker the shade of green the closer to the diagnosis is the accumulation of the years (non-starting on onset) where the feature was identified.

The three best-performing labels using 4 years of features are the classification problems 3, 5 and 7 as demonstrated in the Table 6.14. Those 3 labels use non-baseline information, which one more time appear to be a significant factor. Moreover, the classification problem 6 is also performing poorly in this case. The classification problem 8 and 9 are performing well in all metrics except for the F1-Score, which may be caused by the highly imbalanced dataset, once the model is predicting correctly the largest part of the samples of the majority class, although it incorrectly classify several samples of the minority class. Moreover, the classification problem 1 also increase the performance with the addition of the fourth year of features.

Regarding the predictive features, the EDSS was identified in all classification problems again. The Score Pyramidal was present in 8 out of 9 classification problems and the CNS Sensory Functions in 7 out of 9 classification problems. Furthermore, the CNS Brainstem was predictive in 6 out of 9 classification problems, while the remaining features were predictive in 5 or fewer classification problems.

Table 6.15: Best results obtained in disease severity classification problems using 5 years of features.

Best results 5 year model					
	AUC	**G-Mean**	**Specificity**	**Sensitivity**	**F1-Score**
Classification problem 1 (EDSS $\leq$ 2 after 10th year using baseline visits)	0.76 ± 0.08	0.71 ± 0.09	0.86 ± 0.12	0.58 ± 0.09	0.71 ± 0.08
Classification problem 2 (EDSS $\leq$ 3 after 10th year using baseline visits)	0.81 ± 0.09	0.70 ± 0.11	0.84 ± 0.12	0.61 ± 0.17	0.61 ± 0.13
Classification problem 3 (EDSS $\leq$ 3 in the 10 year)	**0.89 ± 0.10**	**0.78 ± 0.18**	**0.88 ± 0.10**	**0.73 ± 0.25**	**0.71 ± 0.17**
Classification problem 4 (EDSS $\leq$ 3 during the initial 10 years)	0.84 ± 0.06	0.72 ± 0.07	0.95 ± 0.08	0.55 ± 0.10	0.70 ± 0.09
Classification problem 5 (EDSS $\leq$ 2 after 5 years and $\leq$ 3 after 10 years)	-	-	-	-	-
Classification problem 6 (EDSS $\leq$ 4 after 10 years using baseline visits)	**0.89 ± 0.07**	**0.80 ± 0.08**	**0.93 ± 0.03**	**0.69 ± 0.14**	**0.63 ± 0.09**
Classification problem 7 (EDSS $\leq$ 3 in the 6th year)	**0.91 ± 0.04**	**0.82 ± 0.06**	**0.90 ± 0.04**	**0.76 ± 0.13**	**0.61 ± 0.07**
Classification problem 8 (Increase EDSS $<$ 1.5 after 5 years)	-	-	-	-	-
Classification problem 9 (Progression index $<$0.2 after a duration of 5 years)	-	-	-	-	-

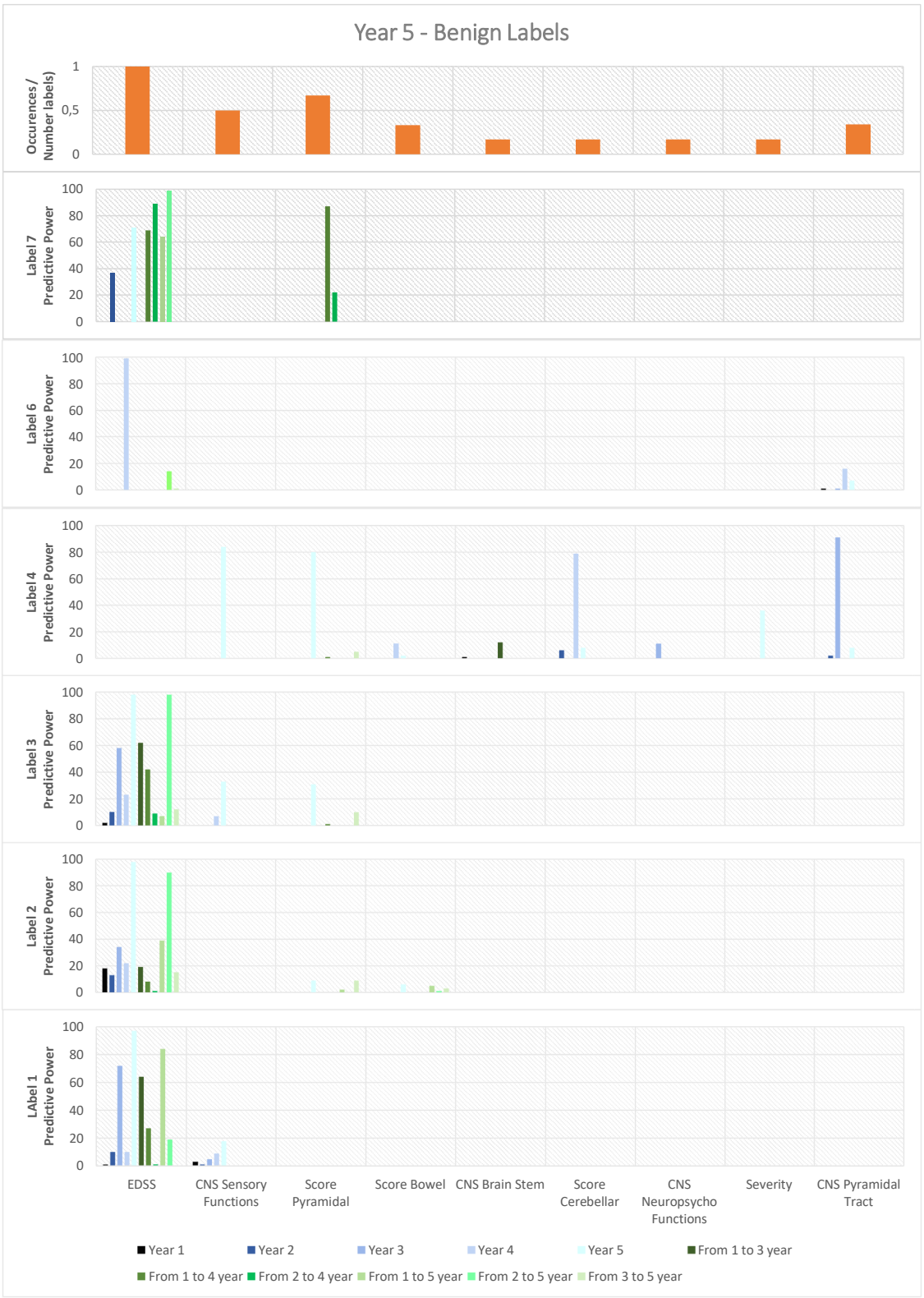

Figure 6.16: Features with highest predictive power of a benign/malignant course identified in the fifth year. The darker the shade of blue, the closer to the diagnosis is the year when the feature was retrieved and the darker the shade of green the closer to the diagnosis is the accumulation of the years (non-starting on onset) where the feature was identified.

The Table 6.15 contains the results of the classification problems where it was used 5 years of features for prediction. In this case, the classification problems 5, 8 and 9 don't have results once they represent a prediction in the fifth year and for such reason, there was not included 5 years of features. From the remaining classification problems, the three with the best performance are the classification problem 3, 6 and 7. In this case, the classification problem 6 increased significantly the performance. The majority of the classification problems did not increase the performance with the addition of the fifth year of features.

Concerning the predictive features, the EDSS was identified one more time as predictive in all classification problems. The other significant predictor was the Score Pyramidal that was present in 4 out of 6 classification problems. The remaining features were present in 3 or fewer classification problems.

7

Discussion

In this chapter, several aspects related to the objective and procedure used in this master thesis are discussed. First of all, the objective of this master thesis was to predict the progression of MS integrating clinical data from different sources, although several aspects were not considered. The epidemiology would be an interesting factor to consider, as the disease is highly affected by geographic factors, but once the database is constituted by only Caucasian patients, it had to be disregarded. Similarly, data relative to the family history, CSF and evoked potentials was also not considered because there was plenty missing data, although there is evidence that genetic factors are related with a higher risk of MS, and that CSF and evoked potentials are key factors used in other studies.

The consideration of reports of MRI examinations could also have been an interesting aspect to use as clinical data once the majority of the studies use MRI examinations to retrieve potential features predictive of disease progression. However, in this master thesis, the reports of MRI examinations could not be considered once it only contained information from a reduced number of patients.

In terms of therapy, the data was not considered because it could introduce a bias to the results, once the medication chosen may introduce a medical opinion of the course of the disease.

Furthermore, the data sets for all predictions are small, once the number of patients used was quite reduced when compared to other studies in the literature [59] [63] [74] [77] [79] [82]. Even from the selected patients, it was necessary to proceed to the imputation of missing data, once the data set contained several missing values.

With regards to the EDSS, this scale was used in every benign/malignant definition, once the definitions are composed by a threshold value of EDSS during a certain amount of time. Some of the values used were retrieved from the literature while some others were used as a test for comparisons with existing labels. From the labels extracted from the literature, some had to be slightly changed due to the

missing values, highly imbalanced data set and reduced number of patients.

The initial years of the disease were highlighted once it was desirable to identify the early clinical factors that are responsible for the transition to an SP course and the ones that lead to a benign/malignant case.

From these factors, one may conclude that a study containing a higher number of patients could be a vital aspect to obtain more reliable results and conclusions. In order to confirm our findings, it is necessary to use other data sets with less missing data, and evaluate the labels of benign retrieved from literature without any modification. Lastly, integration of clinical data from more sources could also be an interesting approach to identify more diverse predictive features of a benign course and an SP development.

7.1 Comparison with the State of the art

The state of the art was essential for the selection of approaches for this master thesis. Although the initial objective was to predict the evolution of MS, the concept of comparing different definitions of benign MS arises from the literature.

The literature itself is reduced in terms of studies involving ML procedures, although some approaches already exist and in the future it is expected that there will be even more. The majority of the papers consist of an analysis of MRI examinations, inferential statistics or clinical reviews of the current predictive features. Although the number of studies is small, it was possible to identify among the existing studies several incoherences in the definitions used that lead to a difficulty in the comparison of the results obtained.

The benign form of MS is described with several differences from author to author, varying in terms of the years of follow-up admitted and in the EDSS value considered. Such modifications result in an enormous variation in the percentage of patients identified as benign, which can be a reason to the different conclusions regarding the clinical factors that can lead to a better or worse prognosis of the patients. This inconsistency verified was the reason to explore the benign MS and develop a model capable of predicting the benign/malignant course of MS, using definitions from the literature and new definitions in order to verify which models are easier to predict and which features are consistent among all definitions.

The other aspect identified in the literature was the development of SP course, that was identified as a problem due to the difficulty of identifying the patients that will evolve to such course. Even from the patients considered in this study, there is a possibility that a patient may be considered RR type, but will still develop SP

course. The studies regarding SP development were also quite reduced and they have identified few clinical conditions resulting in the development of that course.

Besides the few existing studies, a comparison with the literature was performed for both SP development and disease severity. In terms of model performance of SP development prediction, Adrian Ion. Margineanu [80], obtained an F1-score of 0.87 after training Fisher LDA and SVM-RBF classifiers on clinical, lesion loads and metabolic features. Once the other metrics are not assessed in Adrian Ion. Margineanu study, it can only be seen that the model developed in this master thesis is underperforming in terms of F1-score.

Regarding the features predictive of an SP development the *polisyntomatic onset, late age at onset, female sex, motor and sphincter relapses* were identified by Bergamaschi [82]. In the present study, only the CNS Bowel Bladder (related with sphincter relapses) and CNS Pyramidal, Score Pyramidal, Impact ADL Functions (all related with motor functions) were also identified in common with literature. The majority of the features predictive of an SP development identified in the literature were demographic factors, although only dynamic information related to EDSS, impact ADL functions and functional systems were predictive in the present study. The fact that there is no consensus of the predictive features from study to study lead to non-existence of features that could validate the present model, although the non-presence of static features may be a limiting factor once they were identified in other studies [82][81][80].

Regarding the performance of disease severity model, the Table 7.1 summarizes the results obtained by Zhao et al. [77] and the results obtained in each classification problem.

Table 7.1: Comparison of the results obtained with literature.

Performance of disease severity model		
	Specificity	**Sensitivity**
Zhao et al. [77] - Increase EDSS $\leq$ 1.5 after 5 years	0.68	0.71
EDSS $\leq$ 2 after 10th year using baseline visits	0.76	0.66
EDSS $\leq$ 3 after 10th year using baseline visits	0.84	0.61
EDSS $\leq$ 3 in the 10 year	0.88	0.73
EDSS $\leq$ 3 during the initial 10 years	0.84	0.60
EDSS $\leq$ 2 after 5 years and $\leq$ 3 after 10 years	0.79	0.75
EDSS $\leq$ 4 after 10 years using baseline visits	0.93	0.69
EDSS $\leq$ 3 in the 6th year	0.80	0.86
Increase EDSS $<$ 1.5 after 5 years	0.62	0.64
Progression index $<$ 0.2 after a duration of 5 years	0.65	0.69

Zhao et al. obtained a sensitivity of 71% and specificity of 68% and in the present study were obtained worst results when the same Benign MS definition than Zhao et al. was used (Increase EDSS $<$ 1.5 after 5 years), even though, for other benign MS definitions (EDSS $\leq$ 3 in the 10 year, EDSS $\leq$ 2 after 5 years and $\leq$ 3 after 10 years, and EDSS $\leq$ 3 in the 6th year) were achieved better results. Although a comparison of the sensitivity and specificity obtained was performed, the absence of the other metrics limit the comparison of results. Furthermore, in terms of the predictive features of disease severity, only the EDSS and features related to functional systems were identified in the present study. All those features were also present in literature, which reinforces the confidence in the results obtained.

Finally, the nonexistence of several studies with similar procedures for both RR/SP and benign/malignant prediction limits the comparison of the results, as well as the incoherence of the predictive features identified in the existing studies that lead to difficulties to validate the obtained results.

7.2 Dataset Description

One of the major limitations of this master thesis was the dataset used. Although the initial database contained 1134 patients, from these only 181 could be used. The great part of the patients contained only information several years after the diagnosis which leads to their exclusion once it was impossible to retrieve information from the initial years of the disease.

Furthermore, although the dataset contained clinical information of several aspects such as MRI, concomitant diseases, evoked potentials, among others, these factors had to be ignored because they contained plenty of missing data and were related to a small number of patients. For this reason, only information from the identification, visits and relapses was considered, although more data could lead to better results.

Moreover, among the considered patients it was possible to note that the number of information from patient to patient is quite variant and there are plenty of missing data in some fields. The diagnosis date is an example of a field that was missing for some patients, and in such cases was used the onset date for select the patients included in each dataset. For the RR/SP selection was also included patients with the SP diagnosis date after the fifth year of tracking. The assumptions made may lead to incorrect identification of the years where the features were predictive if the diagnosis date was earlier than the onset date, and if the SP diagnosis date was several years after diagnosis date. Even though, those considerations were essential to guarantee an acceptable number of patients included.

Furthermore, the routine visits are usually scheduled every 3 or every 6 months depending on the patient condition, whereby was expected to exist data of at least 2 visits in each year to every patient, which does not occur. The majority of the patients does not contain information of 2 visits in every year they are followed and there are several cases of patients that do not contain data of any visit in some years. For this reason, the temporal segmentation was annual, which contributed to deal with the missing data. However, the nonexistence of information in a certain year might also affect the drawing conclusions regarding the years in which the features are predictive.

The number of important features that had to be disregarded due to the missing data present and the reduced number of patients that were used are negative aspects of the data set used, that may limit the conclusions drawn from this master thesis. Nevertheless, the data set used is a real data set from a hospital, representing a real scenario, which demanded that the methodology used was capable of deal with missing values, leading also to an advantage.

7.3 Experimental Procedure

A process of feature engineering was performed in order to increase the number of features by creating features using temporal segmentation and applying statistical operators to the raw data. The process of feature engineering was a complex

procedure that involved a consideration of accumulative windows starting in different years and non-accumulative windows in order to retrieve the most information. This procedure was also used as a strategy to deal with missing data, once the temporal segmentation performed annually and the accumulative windows considered allowed to increase the amount of data used for predictions. Furthermore, some other features identified in the state-of-the-art such as the number of functional systems involved at the onset, the EDSS at onset and the number of visits and relapses at each year were also introduced.

The dataset obtained with the feature engineering process was used to explore different labels and compare the differences in the results achieved with them.

Regarding the benign labels, some of the labels were retrieved from literature. Several limitations arise to test those labels due to the dataset used. Among the labels identified in the literature, there were several cases that were defined by and EDSS ≤ 3 after 15 years [70] [72] [73] [75] and other defined as EDSS ≤ 3 after 20 years [76] that were not tested once the dataset contains information of a lower amount of years for the majority of the patients, which made unviable to test these labels for such periods. Furthermore, the consideration of the same EDSS value with only 10 years of follow-up was already considered as a label to test due to some studies addressing such definition, which led to the exclusion of those labels with a value of EDSS of 3 and follow-up periods that lasted for more than 10 years.

Regarding the tested labels, there were cases defined by a value of EDSS in a specific year, even though the database contained no data of visits or relapses in that year. To overcome this limitation, and maintain enough patients to test, it was considered the EDSS value in the consecutive years in those labels. Among the literature labels, some other modifications were performed. The label 5 should consider only the baseline EDSS although both the baseline and non-baseline EDSS were admitted to obtain a higher number of patients with EDSS value in the years considered or consecutive years. Label 6 should be evaluated after a follow-up of 15 years, but only 10 years were considered once there were few patients that contained enough information after 15 years. In the labels 8 and 9, it was also considered the EDSS of baseline and non-baseline and assumed that the limit cases (increase EDSS = 1.5 and progression index = 0.2 respectively) were malignant, once both these modifications allow obtaining a more balanced scenario.

7.4 Results

The results of RR/SP prediction and benign/malignant classification problems are discussed separately once they constitute two distinct prediction scenarios although in both cases a comparison with the literature was performed.

7.4.1 RR/SP classification problem

The prediction of SP development was performed using 5 different models that considered different past data. As previously mentioned, to the year 1 only features from the first year were used, while for year 2 features from both year 1 and year 2 were considered. By analyzing the Table 6.1 it can be seen the results obtained for the prediction of SP development in each one of those year models.

The best results were achieved with the SVM classifier in the two first years and with a KNN-3 in the remaining years. Overall, it is possible to note an increase in all five metrics over the years, with the best results for all of them being achieved using 5 years of features. This seems to demonstrate that the higher the number of features considered, the better the performance of the model, although for the fourth-year a decrease in the performance was verified, once the specificity increases but the sensitivity decreases. In fact, once the SP cases correspond to the minority and is more important to identify such cases due to the necessity of administering a more aggressive medication, it is desirable to obtain higher sensitivity.

It is important to highlight that for F1-Score metric the results obtained were quite low in all years. This metric, that considers precision (true positive cases among all cases classified as positive) presents a poor performance that is probably caused by the highly imbalanced data set. Although the performance is not good for F1-Score, this metric is important because it provides a more realistic performance evaluation.

Overall, the developed model presents acceptable results in terms of AUC, specificity, sensitivity and G-mean although the low value obtained for F1-Score metric presents a major limitation of the model. One may conclude that once the number of SP samples is reduced, although the majority of RR samples are correctly classified, the number of RR miss-classified samples is still superior to the number of correctly classified SP samples which lead to poor F1-Score results. For such reason, the model is limited and should be used with caution once the incorrect identification of RR samples leads to the administration of more aggressive medication for those patients which can cause negative effects on patients.

7.4.2 Feature analysis - RR/SP classification problem

In Figure 6.1 the features predictive of an SP development case are identified. The figure represents the features and their predictive power, calculated by the number of the times that the features appeared among the 100 iterations.

Although several statistical operators were used to the creation of the features, and a division into routine/non-routine visits was performed, those are not mentioned in the results because heterogeneity was verified. There wasn't a predominance of any statistical operator, and by excluding such information the results are easier to present and interpret. There wasn't also any relevant superiority between routine/non-routine features whereby such information is not presented. Figure 6.1 allows to verify the evolution of the features over the years.

Firstly, it is possible to note that all identified features correspond to dynamic features, and the majority of them are related to the functional systems. It can be noted the presence of 11 predictive features, retrieved from functional systems information from both visits and relapses. Additionally, only the EDSS and the Impact ADL Functions were identified as predictive features. From this fact, one may conclude that to evaluate the SP development, both EDSS and functional systems data are the most important information, as both features were identified in distinct year-models, as opposed to the Impact ADL functions that was only predictive in the first year. Moreover, it is also important to highlight the complete absence of features related to demographic information and symptoms.

An initial conclusion that can be retrieved is that the most recent features are usually the most predictive ones. This means that when two years of features are considered, the features of the second year are the most predictive ones, while when three years are admitted the features of the third year are the most common. Although this does not occur in every case, the majority of the features identified follow this logic. Moreover, it can be seen that the features from the initial years remain important once they are also identified in the later years.

Additionally, it was possible to note a presence of both features retrieved from both accumulative windows starting and not starting on the onset, although the features from accumulative windows starting on onset appear to be more predominant.

Among the 13 features identified, it is possible to note that each feature presents a particular pattern, in which the EDSS and the Score Brainstem are identified in all 5 models, the CNS Sensory functions was identified only in the 4 initial years and the remaining features in fewer year-models. The predictive power also varied from feature to feature and even within the same feature from year-model to year-model.

Table 7.2 summarizes the predictive features of an RR/SP case, indicating the name of the feature, the years where the feature was identified as an important predictor and whether the feature was predictive of an RR or SP course. It can be noted that the majority of the features identified as predictive are indicative of an SP course. The identified features provide valuable information to physicians, as they inform which patient characteristics should be given attention in each year. For example, it can be seen that the CNS Brainstem was a predictor of a RR course during the two initial years.

Table 7.2: Most commonly predictive features of an RR/SP course chronologically ordered by year of occurrence.

Features	RR Course	SP Course	Years
CNS Cerebellum	x	x	1
CNS Pyramidal Tract	x	x	1
Impact ADL Functions		x	1
CNS Brainstem	x		1,2
Score Sensory		x	1,2
EDSS		x	1,2,3,4,5
Score Brainstem	x		1,2,3,4,5
CNS Sensory Functions		x	1,2,3,4
CNS Neuropsycho Functions		x	1,3
CNS Bowel Bladder	x		2,3,4
Score Pyramidal		x	3
Score Cerebellar		x	3,4,5
Score Mental		x	3,4,5

7.4.3 Benign/malignant classification problems

In terms of benign/malignant MS prediction, several classification problems were used to explore the existing definitions used by several authors in literature and advance with some others that could be interesting approaches.

The first conclusion that can be retrieved from the model performances obtained for each classification problem is that the definition of benign MS used affect significantly the results obtained, once the performance of the model and the best performing year varied significantly for each classification problem. Although for the

majority of the classification problems the higher the number of years of features used, the better the model performed, this does not occur to every case once classification problem 4 present worst performance over time and there are cases whose better performance is identified on the third year. Furthermore, the best classifiers identified in each year-model also varied from classification problem to classification problem. Such facts lead to the impossibility to retrieve a common best classifier and year-model to all classification problems, once both depend on the definition considered.

The fact that some definitions are similar, and present only small variations in terms of the information considered (EDSS threshold used and the number of year of follow-up considered) enable a comparison of the performance of some of those labels, which allows concluding about the factors that lead to such diversity in the performances obtained.

From the analyses of the Tables 6.2 until 6.15 it is possible to retrieve information of the performance of the model for each classification problem over years and also a comparison between the classification problems, essential to retrieve conclusions.

The classification problems 1, 2 and 6 considered an EDSS threshold of 2, 3 and 4, respectively, after 10 years of follow-up, using only baseline information, and those labels can be compared, once the unique variation is the EDSS value considered. By comparing the performance of those classification problems, it can be seen if the value of EDSS selected can lead to variations in the efficiency of the model. Overall, the results in the label 1 and 2 are similar. The label 6 presents a poor performance compared to the remaining 2 labels, which is caused by the more imbalanced scenario of the dataset used for the label 6. This idea suggests that labels with higher values of EDSS as threshold can lead to a reduced number of patients identified as malignant, which impacts negatively the performance of the models due to the imbalanced datasets.

Classification problem 3 was defined by the EDSS of both visits and relapses in the tenth year, which is different from the label 2 where only the baseline EDSS was admitted. This is an important factor once in the relapses the EDSS suffers an increase and for such reason, it is expected that more patients are identified as malignant. From the comparison of the performance of those labels, it can be seen that the classification problem 3 presents slightly better results than the classification problem 2 in the majority of the cases, which indicates that the model performs better for prediction of an EDSS that includes relapses values.

For the classification problem 4, an EDSS <=3 during the initial 10th years

was admitted as opposed to the label 3 where the EDSS <=3 was verified only on the 10th year. The comparison of those labels suggests that for the initial years those labels perform similarly, with slightly better performance for the classification problem 4 until the third-year model. After that, the classification problem 3 present significant better results which indicate that with more years of information, the classification problem 4 perform worst.

The classification problem 5, can also be compared with the classification problem 3 once the unique difference between them is that the classification problem 5 also consider and EDSS value 2 after 5 years as a condition. It can be seen that the classification problem 5 presents better results for every year, with exception to the first, which is an indicator that the disease severity is easier predicted when diminished follow-up periods are also admitted. The same conclusion is obtained for classification problem 7, where a follow-up of 6 years is admitted and better results are achieved when compared to the classification problem 3, even with a more imbalanced dataset.

For labels 8 and 9, where 5 years of follow up were considered, it is possible to note that the model was underperforming in both cases. Although it was obtained acceptable results of sensitivity, both labels present low results in the F1-Score metric. The poor results can be due to the consideration of an increase on the EDSS and progression index for label 8 and 9, respectively, which appear to be harder to predict than simply an EDSS value as a threshold. Moreover, those two labels are also the most imbalanced ones which might also be one of the reasons for the poor performance verified.

In general, it is possible to conclude that the label 5 and label 7 are the better-performing labels among all other labels. The label 7 consider only 6 years of follow-up and the label 5 consider 10 years of follow-up, although it also admits an threshold after 5 years, which appears to lead to the good results. From this fact, one may conclude that the models perform better when EDSS thresholds are established for lower follow-up periods. Moreover, it was also possible to verify, as expected, that the model performs better for balanced datasets than for imbalanced datasets. The consideration of the non-baseline information appears to be also a factor that affects the results once the labels considering that information present better results. Furthermore, the EDSS value selected can also be a factor that influences the results, once higher EDSS thresholds can lead to higher imbalanced datasets which cause poorer results.

From these facts, one may conclude that a label presenting a low number of years of follow-up, defined by a lower EDSS threshold, and using both baseline and

non-baseline information is the option that would present the better performance. Lastly, it is important to remember that the low number of patients, distinct to different classification problems, as well as the modifications introduced to definitions retrieved from literature are limiting aspects of this thesis.

7.4.4 Features analysis - benign/malignant classification problems

From the analysis of the Figures 6.3 until 6.16 it is possible to retrieve several conclusions regarding the features identified as predictive on the different classification problems considered.

Similarly to the RR/SP classification problem, the information regarding the several statistical operators used and the division into routine/non-routine visits performed was not mentioned in the results due to the nonexistence of a predominance of any statistical operator or routine/non-routine features.

First of all, it can be noted that the features identified as predictive are distinct for distinct labels. Table 7.3 summarizes the totality of features that were identified in at least one classification problem. By observing this table and compare it with the literature it is possible to take some initial conclusions. In the results obtained there is not a single static feature, although in literature features such as age at onset [63] [59] [26] [61], sex [72] [26], the course of the disease [66] [63] [78] [67] [56], and the number of function systems involved at onset [55] [56] [67] were identified in some studies.

Table 7.3: Features identified as predictive of disease severity among all classification problems.

Features identified as predictive of disease severity				
EDSS	CNS Neuropsycho Function	Impact ADL Functions	Score Cerebellar	Score Brainstem
CNS Sensory Functions	Score Mental	Severity	Score Sensory	Fatigue
CNS Cerebellum	CNS Bowel Bladder	Score Bowel	GdSpasticity	Paresthesiae
CNS Pyramidal Tract	CNS Brainstem	Recovery	GdAtaxia	MwLowerExtrem
CNS Visual Functions	Score Visual	Score Pyramidal	Score Ambulation	Number relapses

The features identified belong exclusively to visits and relapses information, without being observed a predominance of one of them. Furthermore, from the 25 identified features, just some of them are identified in each label and for such reason, an analysis to identify the more common features among all classification problems were carried out. With such analyses, the confidence in the features identified as predictive is higher.

An initial conclusion retrieved from the analysis of the figures regarding the features predictive of a benign scenario is that each classification problem presented a different number of features identified, predictive on different years, with different predictive powers. This fact demonstrates that different classification problems lead to different features considered which is according to the literature where different authors identify some features as predictive while others don't.

Moreover, it is possible to note for the majority of the classification problems, the most recent information is the most important, although the past information remains predictive. For example, it was demonstrated that the features from the third year are more predictive in the third year model, although some information from previous years is also predictive. Furthermore, features retrieved from both accumulative windows starting on the onset and non starting on onset were identified in all labels.

In addition, in the current study, several features commonly identified among the labels were summarized in Table 7.4. The information was retrieved from the analysis of the Figures 6.12 until 6.16, in which the features appearing in the majority of the classification problems were included in the table, once they appear to be reliable predictors.

Table 7.4: Most commonly predictive features among all benign/malignant classification problems chronologically ordered by year of occurrence.

Features	Benign MS	Malignant MS	Years	Identified in literature:
CNS Pyramidal Tract	x	x	1	Both benign and malignant
CNS Cerebellum	x	x	1, 2	Both benign and malignant
CNS Brainstem	x		1, 2	Only benign
EDSS		x	1, 2, 3, 4, 5	Both benign and malignant
Score Bowel		x	2, 3	Both benign and malignant
Score Pyramidal		x	2, 4, 5	Both benign and malignant
CNS Neuropsycho Functions	x		3	Only benign
CNS Sensory Functions	x	x	4	Only benign
Score Cerebellar		x	4	Both benign and malignant

In this table, there is an indication of features that the physicians should pay attention to, once they were identified as the most common features among the majority of the classification problems. Moreover, it is also exhibited the year models where each feature was predictive as well as an indication whether they were predictors of a benign, malignant or both scenarios. Furthermore, the Table 7.4 also contains an comparison with the literature. The fact that all features identified were also identified in the literature is an important aspect that contributes to a validation of the results and the developed model.

Moreover, it can also be noted that there were features predicting only benign or malignant scenarios, which occurred when all classification problems agreed, but also features indicating both benign and malignant scenario, which occurred when different conclusions were retrieved from different classification problems. Furthermore, was observed a clear predominance of the features predictive of malignant scenarios.

Additionally, all features identified as reliable predictors of benign/malignant scenario were functional systems, with exception to the EDSS. Although the EDSS is not a functional system, it is a scale that measures the impairment of patients using grades attributed to the 8 different functional systems, and for such reason, both EDSS and functional systems are related. From this fact, one may conclude that any symptom or demographic feature was commonly identified among the classification problems, as opposed to the scores attributed to each functional systems (in both visits and relapses) and the EDSS value that constituted the important features identified.

Lastly, the information of the Table 7.4 can be valuable to assist in the prognosis of a patient, once it is easily identified the factors that might lead to a severe scenario. For example, it can be seen that the Central Nervous System Brainstem appear to be a predictor of a benign scenario during the two initial years, although it does not remain as an important predictor during the remaining years.

8

Conclusion

The overall goal of this Master Thesis was to evaluate and predict the progression of patients with MS using a database from the Neurology Department of Centro Hospitalar e Universitario de Coimbra. To achieve such objective an ML model was developed, where the clinical data from the initial years after the diagnosis was used to predict later stages.

The ML algorithm created, robust to real-life scenarios, was used to classify each patient into RR/SP subtypes and benign/malignant scenario, and extract the clinical information that was predictive of each case, as well as the corresponding years in which the features were identified. Furthermore, it was also possible to compare the performances of the different comparative problems considered.

Overall, the best results were achieved in the fifth year for the RR/SP classification problem with an AUC of 0.84 ± 0.08, a G-mean of 0.78 ± 0.11, an specificity of 0.78 ± 0.06, an sensitivity of 0.78 ± 0.18 and an F1-Score of 0.23 ± 0.07 obtained. The classification problem 7, that stands for an EDSS ≤ 3 in the 6th year, was the best performing classification problem evaluating the disease severity, with an AUC of 0.91 ± 0.04, a geometric mean of 0.83 ± 0.05, a specificity of 0.80 ± 0.04, a sensitivity of 0.86 ± 0.11 and an F1-Score of 0.52 ± 0.06 obtained on the third year. Furthermore, the classification problem 5, defined as an EDSS ≤ 2 in the 5th year and an EDSS ≤ 3 in the 10th year, was the second-best performing classification problem, with a slightly worst performance when compared to the previously mentioned case.

From the results obtained, one may conclude that the models perform better when more years of information are considered. It was also noted that it is easier to predict benign/malignant scenarios when an EDSS threshold is defined after 5/6 years than when it is only defined after 10 years. Besides, predicting the disease severity using lower EDSS threshold values is simpler, once high EDSS values tend to lead to more imbalanced data sets.

Regarding the predictive features identified, the EDSS and functional systems

related features appear to be the most significant ones, and the most recent information of those features appear to be the most predictive, although the information from the initial years remains important.

The modification introduced in some benign definitions retrieved from literature, the reduced number of patients considered, the different amount of patients admitted for different classification problems constitute some of the limitations of this master thesis. From those facts, it is necessary to test this procedure in different data sets, in order to validate the drawing conclusions.

In a future study, an integration of more data, related to MRI reports, CSF analysis and evoked potentials could be an important addition to identify other important predictors. Moreover, it could also be interesting to add more samples to verify if those predictors remain the most important ones. Furthermore, performing this comparative study in different data sets, using an equal amount of patients to every classification problem, would be an important factor that could lead to more fair conclusions.

Bibliography

[1] M. T. Wallin, W. J. Culpepper, E. Nichols, Z. A. Bhutta, T. T. Gebrehiwot, S. I. Hay, I. A. Khalil, K. J. Krohn, X. Liang, M. Naghavi, *et al.*, "Global, regional, and national burden of multiple sclerosis 1990–2016: a systematic analysis for the global burden of disease study 2016," *The Lancet Neurology*, vol. 18, no. 3, pp. 269–285, 2019.

[2] G. Giovannoni, H. Butzkueven, S. Dhib-Jalbut, J. Hobart, G. Kobelt, G. Pepper, M. P. Sormani, C. Thalheim, A. Traboulsee, and T. Vollmer, "Brain health: time matters in multiple sclerosis," *Multiple sclerosis and related disorders*, vol. 9, pp. S5–S48, 2016.

[3] J. D. Haines, M. Inglese, and P. Casaccia, "Axonal damage in multiple sclerosis," *Mount Sinai Journal of Medicine: A Journal of Translational and Personalized Medicine*, vol. 78, no. 2, pp. 231–243, 2011.

[4] A. Ochoa-Morales, T. Hernández-Mojica, F. Paz-Rodríguez, A. Jara-Prado, Z. Trujillo-De Los Santos, M. Sánchez-Guzmán, J. Guerrero-Camacho, T. Corona-Vázquez, J. Flores, A. Camacho-Molina, *et al.*, "Quality of life in patients with multiple sclerosis and its association with depressive symptoms and physical disability," *Multiple sclerosis and related disorders*, vol. 36, p. 101386, 2019.

[5] H. Inojosa, U. Proschmann, K. Akgün, and T. Ziemssen, "A focus on secondary progressive multiple sclerosis (spms): challenges in diagnosis and definition," *Journal of neurology*, pp. 1–12, 2019.

[6] G. Ramsaransing and J. De Keyser, "Benign course in multiple sclerosis: a review," *Acta Neurologica Scandinavica*, vol. 113, no. 6, pp. 359–369, 2006.

[7] D. Bzdok, N. Altman, and M. Krzywinski, "Points of significance: statistics versus machine learning," 2018.

[8] C. Caravagna, "What is multiple sclerosis," *Front. Young Minds*, vol. 7, no. 7, 2019.

[9] D. A. Dyment, A. D. Sadnovich, and G. C. Ebers, "Genetics of multiple sclerosis," *Human Molecular Genetics*, vol. 6, no. 10, pp. 1693–1698, 1997.

[10] A. Ascherio, K. L. Munger, R. White, K. Köchert, K. C. Simon, C. H. Polman, M. S. Freedman, H.-P. Hartung, D. H. Miller, X. Montalbán, *et al.*, "Vitamin d as an early predictor of multiple sclerosis activity and progression," *JAMA neurology*, vol. 71, no. 3, pp. 306–314, 2014.

[11] L. Alfredsson and T. Olsson, "Lifestyle and environmental factors in multiple sclerosis," *Cold Spring Harbor perspectives in medicine*, vol. 9, no. 4, p. a028944, 2019.

[12] L. Alfredsson and T. Olsson, "Lifestyle and environmental factors in multiple sclerosis," *Cold Spring Harbor perspectives in medicine*, vol. 9, no. 4, p. a028944, 2019.

[13] M. Pugliatti, G. Rosati, H. Carton, T. Riise, J. Drulovic, L. Vécsei, and I. Milanov, "The epidemiology of multiple sclerosis in europe," *European journal of Neurology*, vol. 13, no. 7, pp. 700–722, 2006.

[14] A. J. Thompson, B. L. Banwell, F. Barkhof, W. M. Carroll, T. Coetzee, G. Comi, J. Correale, F. Fazekas, M. Filippi, M. S. Freedman, *et al.*, "Diagnosis of multiple sclerosis: 2017 revisions of the mcdonald criteria," *The Lancet Neurology*, vol. 17, no. 2, pp. 162–173, 2018.

[15] P. Schwenkenbecher, U. Wurster, F. F. Konen, S. Gingele, K.-W. Sühs, M. P. Wattjes, M. Stangel, and T. Skripuletz, "Impact of the mcdonald criteria 2017 on early diagnosis of relapsing-remitting multiple sclerosis," *Frontiers in Neurology*, vol. 10, 2019.

[16] J. F. Kurtzke, "Rating neurologic impairment in multiple sclerosis: an expanded disability status scale (edss)," *Neurology*, vol. 33, no. 11, pp. 1444–1444, 1983.

[17] M. Gaspari, G. Roveda, C. Scandellari, and S. Stecchi, "An expert system for the evaluation of edss in multiple sclerosis," *Artificial intelligence in medicine*, vol. 25, no. 2, pp. 187–210, 2002.

[18] M. M. Goldenberg, "Multiple sclerosis review," *Pharmacy and Therapeutics*, vol. 37, no. 3, p. 175, 2012.

[19] S. Vukusic and C. Confavreux, "Primary and secondary progressive multiple

sclerosis," *Journal of the neurological sciences*, vol. 206, no. 2, pp. 153–155, 2003.

[20] A. Thompson, X. Montalban, F. Barkhof, B. Brochet, M. Filippi, D. Miller, C. Polman, V. Stevenson, and W. McDonald, "Diagnostic criteria for primary progressive multiple sclerosis: a position paper," *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society*, vol. 47, no. 6, pp. 831–835, 2000.

[21] D. H. Miller and S. M. Leary, "Primary-progressive multiple sclerosis," *The Lancet Neurology*, vol. 6, no. 10, pp. 903–912, 2007.

[22] M. Tullman, R. Oshinsky, F. Lublin, and G. Cutter, "Clinical characteristics of progressive relapsing multiple sclerosis," *Multiple Sclerosis Journal*, vol. 10, no. 4, pp. 451–454, 2004.

[23] F. D. Lublin, S. C. Reingold, J. A. Cohen, G. R. Cutter, P. S. Sørensen, A. J. Thompson, J. S. Wolinsky, L. J. Balcer, B. Banwell, F. Barkhof, *et al.*, "Defining the clinical course of multiple sclerosis: the 2013 revisions," *Neurology*, vol. 83, no. 3, pp. 278–286, 2014.

[24] F. D. Lublin, "New multiple sclerosis phenotypic classification," *European neurology*, vol. 72, no. Suppl. 1, pp. 1–5, 2014.

[25] S. Klineova and F. D. Lublin, "Clinical course of multiple sclerosis," *Cold Spring Harbor perspectives in medicine*, p. a028928, 2018.

[26] S. Hawkins and G. McDonnell, "Benign multiple sclerosis? clinical course, long term follow up, and assessment of prognostic factors," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 67, no. 2, pp. 148–152, 1999.

[27] C. H. Polman, "Drug treatment of multiple sclerosis," *Bmj*, vol. 321, no. 7259, pp. 490–494, 2000.

[28] K. Kapitanova and S. Son, "Machine learning basics," in *Intelligent Sensor Networks*, pp. 3–29, CRC Press, 2012.

[29] K. Y. Ngiam and W. Khor, "Big data and machine learning algorithms for health-care delivery," *The Lancet Oncology*, vol. 20, no. 5, pp. e262–e273, 2019.

[30] V. Nasteski, "An overview of the supervised machine learning methods," *HORIZONS.B*, vol. 4, pp. 51–62, 12 2017.

[31] S. A. Alasadi and W. S. Bhaya, "Review of data preprocessing techniques in

data mining," *Journal of Engineering and Applied Sciences*, vol. 12, no. 16, pp. 4102–4107, 2017.

[32] A. Zheng and A. Casari, *Feature engineering for machine learning: principles and techniques for data scientists.* " O'Reilly Media, Inc.", 2018.

[33] F. Nargesian, H. Samulowitz, U. Khurana, E. B. Khalil, and D. S. Turaga, "Learning feature engineering for classification.," in *IJCAI*, pp. 2529–2535, 2017.

[34] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Computers & Electrical Engineering*, vol. 40, no. 1, pp. 16–28, 2014.

[35] A. Bommert, X. Sun, B. Bischl, J. Rahnenführer, and M. Lang, "Benchmark for filter methods for feature selection in high-dimensional classification data," *Computational Statistics & Data Analysis*, vol. 143, p. 106839, 2020.

[36] J. Tang, S. Alelyani, and H. Liu, "Feature selection for classification: A review," *Data classification: Algorithms and applications*, p. 37, 2014.

[37] R. Kohavi, G. H. John, *et al.*, "Wrappers for feature subset selection," *Artificial intelligence*, vol. 97, no. 1-2, pp. 273–324, 1997.

[38] P. A. Patrician, "Multiple imputation for missing data," *Research in nursing & health*, vol. 25, no. 1, pp. 76–84, 2002.

[39] A. R. T. Donders, G. J. Van Der Heijden, T. Stijnen, and K. G. Moons, "A gentle introduction to imputation of missing values," *Journal of clinical epidemiology*, vol. 59, no. 10, pp. 1087–1091, 2006.

[40] J. Tian, H. Gu, and W. Liu, "Imbalanced classification using support vector machine ensemble," *Neural Computing and Applications*, vol. 20, no. 2, pp. 203–209, 2011.

[41] V. Ganganwar, "An overview of classification algorithms for imbalanced datasets," *International Journal of Emerging Technology and Advanced Engineering*, vol. 2, no. 4, pp. 42–47, 2012.

[42] H. Liu and M. Cocea, "Semi-random partitioning of data into training and test sets in granular computing context," *Granular Computing*, vol. 2, no. 4, pp. 357–386, 2017.

[43] R. Kohavi *et al.*, "A study of cross-validation and bootstrap for accuracy esti-

mation and model selection," in *Ijcai*, vol. 14, pp. 1137–1145, Montreal, Canada, 1995.

[44] P. Ahmad, S. Qamar, and S. Q. A. Rizvi, "Techniques of data mining in healthcare: a review," *International Journal of Computer Applications*, vol. 120, no. 15, 2015.

[45] H. O. Alanazi, A. H. Abdullah, and K. N. Qureshi, "A critical review for developing accurate and dynamic predictive models using machine learning methods in medicine and health care," *Journal of medical systems*, vol. 41, no. 4, p. 69, 2017.

[46] J. R. Quinlan, "Learning decision tree classifiers," *ACM Computing Surveys (CSUR)*, vol. 28, no. 1, pp. 71–72, 1996.

[47] C.-Y. J. Peng, K. L. Lee, and G. M. Ingersoll, "An introduction to logistic regression analysis and reporting," *The journal of educational research*, vol. 96, no. 1, pp. 3–14, 2002.

[48] S. Balakrishnama and A. Ganapathiraju, "Linear discriminant analysis-a brief tutorial," in *Institute for Signal and information Processing*, vol. 18, pp. 1–8, 1998.

[49] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE transactions on information theory*, vol. 13, no. 1, pp. 21–27, 1967.

[50] A. Ben-Hur and J. Weston, "A user's guide to support vector machines," in *Data mining techniques for the life sciences*, pp. 223–239, Springer, 2010.

[51] C. M. Bishop, *Pattern recognition and machine learning*. springer, 2006.

[52] R. Espíndola and N. Ebecken, "On extending f-measure and g-mean metrics to multi-class problems," *WIT Transactions on Information and Communication Technologies*, vol. 35, 2005.

[53] S. Narkhede, "Understanding auc-roc curve," *Towards Data Science*, vol. 26, 2018.

[54] M. Bekkar, H. K. Djemaa, and T. A. Alitouche, "Evaluation measures for models assessment over imbalanced data sets," *J Inf Eng Appl*, vol. 3, no. 10, 2013.

[55] J. F. Kurtzke, G. W. Beebe, B. Nagler, L. T. Kurland, and T. L. Auth, "Studies on the natural history of multiple sclerosis—8: Early prognostic features of the

later course of the illness," *Journal of chronic diseases*, vol. 30, no. 12, pp. 819–830, 1977.

[56] K. Lauer and W. Firnhaber, "Epidemiological investigations into multiple sclerosis in southern hesse. v. course and prognosis," *Acta neurologica scandinavica*, vol. 76, no. 1, pp. 12–17, 1987.

[57] J. Benedikz, M. Stefánsson, J. Guomundsson, A. Jónasdóttir, and R. Fossdal, "The natural history of untreated multiple sclerosis in iceland. a total population-based 50 year prospective study," *Clinical neurology and neurosurgery*, vol. 104, no. 3, pp. 208–210, 2002.

[58] M. Hutchinson, "Disability due to multiple sclerosis: a community-based study on an irish county.," *Irish medical journal*, vol. 79, no. 2, p. 48, 1986.

[59] A. J. THOMPSON, M. Hutchinson, J. Brazil, C. Feighery, and E. Martin, "A clinical and laboratory study of benign multiple sclerosis," *QJM: An International Journal of Medicine*, vol. 58, no. 1, pp. 69–80, 1986.

[60] P. Cabre, O. Heinzlef, H. Merle, G. Buisson, O. Bera, R. Bellance, J. Vernant, and D. Smadja, "Ms and neuromyelitis optica in martinique (french west indies)," *Neurology*, vol. 56, no. 4, pp. 507–514, 2001.

[61] G. Ramsaransing, N. Maurits, C. Zwanikken, and J. De Keyser, "Early prediction of a benign course of multiple sclerosis on clinical grounds: a systematic review," *Multiple Sclerosis Journal*, vol. 7, no. 5, pp. 345–347, 2001.

[62] J. Mandrioli, P. Sola, R. Bedin, M. Gambini, and E. Merelli, "A multifactorial prognostic index in multiple sclerosis," *Journal of neurology*, vol. 255, no. 7, p. 1023, 2008.

[63] S. Poser, N. Raun, and W. Poser, "Age at onset, initial symptomatology and the course of multiple sclerosis," *Acta Neurologica Scandinavica*, vol. 66, no. 3, pp. 355–362, 1982.

[64] D. McALPINE, "The benign form of multiple sclerosis. a study based on 241 cases seen within three years of onset and followed up until the tenth year or more of the disease.," *Brain: a journal of neurology*, vol. 84, p. 186, 1961.

[65] U. Leibowitz and M. Alter, "Clinical factors associated with increased disability in multiple sclerosis," *Acta neurologica Scandinavica*, vol. 46, no. 1, pp. 53–70, 1970.

[66] D. I. Shepherd, "Clinical features of multiple sclerosis in north-east scotland," *Acta neurologica Scandinavica*, vol. 60, no. 4, pp. 218–230, 1979.

[67] M. P. Amato, G. Ponziani, M. L. Bartolozzi, and G. Siracusa, "A prospective study on the natural history of multiple sclerosis: clues to the conduct and interpretation of clinical trials," *Journal of the neurological sciences*, vol. 168, no. 2, pp. 96–106, 1999.

[68] F. D. Lublin, S. C. Reingold, *et al.*, "Defining the clinical course of multiple sclerosis: results of an international survey," *Neurology*, vol. 46, no. 4, pp. 907–911, 1996.

[69] K. Bencsik, C. Rajda, J. Füvesi, P. Klivényi, T. Járdánházy, M. Török, and L. Vécsei, "The prevalence of multiple sclerosis, distribution of clinical forms of the disease and functional status of patients in csongrad county, hungary," *European neurology*, vol. 46, no. 4, pp. 206–209, 2001.

[70] P. Perini, C. Tagliaferri, M. Belloni, G. Biasi, and P. Gallo, "The hla-dr13 haplotype is associated with "benign" multiple sclerosis in northeast italy," *Neurology*, vol. 57, no. 1, pp. 158–159, 2001.

[71] H. Kalanie, K. Gharagozli, and A. R. Kalanie, "Multiple sclerosis: report on 200 cases from iran," *Multiple Sclerosis Journal*, vol. 9, no. 1, pp. 36–38, 2003.

[72] E. Portaccio, M. Stromillo, B. Goretti, V. Zipoli, G. Siracusa, M. Battaglini, A. Giorgio, M. Bartolozzi, L. Guidi, S. Sorbi, *et al.*, "Neuropsychological and mri measures predict short-term evolution in benign multiple sclerosis," *Neurology*, vol. 73, no. 7, pp. 498–503, 2009.

[73] M. Rovaris, M. A. Rocca, F. Barkhof, M. Calabrese, N. De Stefano, M. Khalil, F. Fazekas, L. Fisniku, P. Gallo, D. H. Miller, *et al.*, "Relationship between brain mri lesion load and short-term disease evolution in non-disabling ms: a large-scale, multicentre study," *Multiple Sclerosis Journal*, vol. 17, no. 3, pp. 319–326, 2011.

[74] L. E. Hviid, B. C. Healy, D. J. Rintell, T. Chitnis, H. L. Weiner, and B. I. Glanz, "Patient reported outcomes in benign multiple sclerosis," *Multiple Sclerosis Journal*, vol. 17, no. 7, pp. 876–884, 2011.

[75] M. Calabrese, A. Favaretto, V. Poretto, C. Romualdi, F. Rinaldi, I. Mattisi, A. Morra, P. Perini, and P. Gallo, "Low degree of cortical pathology is associated with benign course of multiple sclerosis," *Multiple Sclerosis Journal*, vol. 19, no. 7, pp. 904–911, 2013.

[76] A.-M. Bueno, A.-L. Sayao, M. Yousefi, V. Devonshire, A. Traboulsee, and H. Tremlett, "Health-related quality of life in patients with longstanding 'benign multiple sclerosis'," *Multiple sclerosis and related disorders*, vol. 4, no. 1, pp. 31–38, 2015.

[77] Y. Zhao, B. C. Healy, D. Rotstein, C. R. Guttmann, R. Bakshi, H. L. Weiner, C. E. Brodley, and T. Chitnis, "Exploration of machine learning techniques in predicting multiple sclerosis disease course," *PLoS One*, vol. 12, no. 4, 2017.

[78] T. Reynders, M. D'haeseleer, J. De Keyser, G. Nagels, and M. B. D'hooghe, "Definition, prevalence and predictive factors of benign multiple sclerosis," *eNeurologicalSci*, vol. 7, pp. 37–43, 2017.

[79] Y. Zhao, T. Chitnis, and T. Doan, "Ensemble learning for predicting multiple sclerosis disease course," in *MULTIPLE SCLEROSIS JOURNAL*, vol. 25, pp. 160–161, SAGE PUBLICATIONS LTD 1 OLIVERS YARD, 55 CITY ROAD, LONDON EC1Y 1SP, ENGLAND, 2019.

[80] A. Ion-Mărgineanu, G. Kocevar, C. Stamile, D. M. Sima, F. Durand-Dubief, S. Van Huffel, and D. Sappey-Marinier, "Machine learning approach for classifying multiple sclerosis courses by combining clinical data with lesion loads and magnetic resonance metabolic features," *Frontiers in neuroscience*, vol. 11, p. 398, 2017.

[81] A. Boiko, G. Vorobeychik, D. Paty, V. Devonshire, D. Sadovnick, *et al.*, "Early onset multiple sclerosis: a longitudinal study," *Neurology*, vol. 59, no. 7, pp. 1006–1010, 2002.

[82] R. Bergamaschi, S. Quaglini, M. Trojano, M. P. Amato, E. Tavazzi, D. Paolicelli, V. Zipoli, A. Romani, A. Fuiani, E. Portaccio, *et al.*, "Early prediction of the long term evolution of multiple sclerosis: the bayesian risk estimate for multiple sclerosis (brems) score," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 78, no. 7, pp. 757–759, 2007.

[83] L. A. Arani, A. Hosseini, F. Asadi, S. A. Masoud, and E. Nazemi, "Intelligent computer systems for multiple sclerosis diagnosis: a systematic review of reasoning techniques and methods," *Acta Informatica Medica*, vol. 26, no. 4, p. 258, 2018.

[84] M. F. d. S. Pinto, *Evaluation and prediction of Multiple Sclerosis Disease Progression*. PhD thesis, Universidade de Coimbra, 2018.

[85] P. Paunov, "Data standardization. MATLAB Central File

https://www.mathworks.com/matlabcentral/fileexchange/50857-data-standardization," 2020.

[86] J. Santarcangelo, "jsantarc/Imputation-of-missing-values-Matlab-. GitHub https://github.com/jsantarc/Imputation-of-missing-values-Matlab-/issues/1," 2020.

[87] S. Seaman, J. Galati, D. Jackson, and J. Carlin, "What is meant by" missing at random"?," *Statistical Science*, pp. 257–268, 2013.

[88] V. Bolón-Canedo, N. Sánchez-Maroño, and A. Alonso-Betanzos, "A review of feature selection methods on synthetic data," *Knowledge and information systems*, vol. 34, no. 3, pp. 483–519, 2013.

[89] V. Fonti and E. Belitser, "Feature selection using lasso," *VU Amsterdam Research Paper in Business Analytics*, vol. 30, pp. 1–25, 2017.

[90] A. K. Jain, R. P. W. Duin, and J. Mao, "Statistical pattern recognition: A review," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, no. 1, pp. 4–37, 2000.