UHASSELT
KNOWLEDGE IN ACTION

Maastricht University

Faculty of Sciences
School for Information Technology
Master of Statistics and Data Science

Master's thesis

Improving the calibration of machine-learning models for predicting disease progression of Multiple Sclerosis patients

Javier Rodriguez Soto

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics and Data Science, specialization Biostatistics

SUPERVISOR :

dr. Thijs BECKER

De heer Axel-Jan ROUSSEAU

2020
2021

Transnational University Limburg is a unique collaboration of two universities in two countries: the University of Hasselt and Maastricht University.

UHASSELT
KNOWLEDGE IN ACTION

www.uhasselt.be
Universiteit Hasselt
Campus Hasselt:
Martelarenlaan 42 | 3500 Hasselt
Campus Diepenbeek:
Agoralaan Gebouw D | 3590 Diepenbeek

Maastricht University

Faculty of Sciences
School for Information Technology
Master of Statistics and Data Science

Master's thesis

Improving the calibration of machine-learning models for predicting disease progression of Multiple Sclerosis patients

Javier Rodriguez Soto

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics and Data Science, specialization Biostatistics

SUPERVISOR :

dr. Thijs BECKER

De heer Axel-Jan ROUSSEAU

Contents

List of Abbreviations

ADA Adaptative Boosting. 1, 14, 18, 20–27, 29

AH Abductor hallucis. 4–6

APL Abductor pollicis longus. 4–6

AUC Area Under the Curve. 1, 10, 18, 20–22, 24, 28–30

AUROC Area Under the Receiver Operating Characteristics. 8–10

BS Brier Score. 8, 10, 24, 29

ECE Expected Calibration Error. 12, 22, 29

EDSS Expanded Disability Score Scale. 2, 4–6, 8, 14, 28

EP Evoked Potential. 1–3, 14

FS Functional System. 2

GBDT Gradient Boosted Decision Trees. 1, 14, 18, 20, 22–27, 29

LR Logistic Regression. 1, 14, 15, 18–22, 24, 28, 29

MEP Motor Evoked Potential. 3, 4, 6, 8, 14, 18, 28, 30

ML Machine Learning. 3, 9, 14, 17, 18, 28–30

MRI Magnetic Resonance Imaging. 2

MS Multiple Sclerosis. 1, 3, 28, 30

RF Random Forest. 1, 14–16, 18–22, 24, 28–30

ROC Receiver Operating Characteristics. 8, 9, 13

SVM Support Vector Machines. 1, 15, 18, 20, 23, 29

List of Figures

List of Tables

Abstract

Background: Evoked potentials (EPs) are electrical signals that are produced by the nervous system in response to an external stimulus.They are used to monitor disease progression of Multiple Sclerosis (MS) patients. Previous studies have used several machine learning algorithms to prove this relationship, but until now the calibration properties of these models have not been sufficiently investigated. This research performs a machine learning analysis on latencies of motor EP and investigates how well the probabilistic outputs of the model are calibrated.

Methods: In this study, motor EP data from 514 multiple sclerosis patients was analyzed. Modeling was performed using logistic regression (LR), random forest (RF), gradient boosted decision trees (GBDT), adaptative boosting (ADA) and support vector machines (SVM). Probabilistic outcomes were calibrated by using Platt Scaling and Isotonic regression.

Results: The results show a satisfactory performance of the RF model (AUC 0.655 ± 0.046). RF probabilistic outcomes had good calibration levels, with an expected calibration error of only 5%. The use of class weights to correct for the class imbalance in RF and LR model resulted in a notorious miscalibration of the the probabilistic outcomes, but the use of both Isotonic regression and Platt scaling were able to correct it.

Conclusions: Using machine learning methods on motor EP showed favorable results and a calibrated predictive performance. The use of class weight methods to correct for imbalance data should be combined with calibration techniques.

Key Words: Calibration; Multiple sclerosis; Motor evoked potential; Expected calibration error; Machine learning

---

1 Introduction

1.1 General Background

Multiple Sclerosis (MS) is an inflammatory disorder that affects the central nervous systems, i.e. the brain and spinal cord. It is characterized by lymphocytic infiltration which leads to the demyelination of the nerve cell's axon. MS is a complex disease that can lead to substantial disabilities of the patient's motor, autonomic, and neurocognitive function. Although the disease is usually not life shortening, its manifestations are multiple and may include weakness, imbalance, sensory loss, cognitive dysfunction, bowel/bladder dysfunction, and vision impairment. MS is triggered by environment factors in individuals with complex genetic risk profiles [1]. The clinical evolution of the disease shows considerable inter-patient variability, with about 10-20% of patients developing only minimal (motor) disability after 20 years [26].

A great variety of advances in pharmacological treatments of MS have been made in the last decades, but the development of reliable predictors for individual disease evolution remains as a major challenge. Development of a satisfactory tool that quantifies the disability

progression in patients with MS has proved challenging. Diagnostic techniques such as conventional magnetic resonance imaging (MRI) have emerged historically as outcome measures in clinical trials in MS and it is considered the best technique to identify MS lesions [24]. More recently, a considerable number of tools have been developed that measure the severity of the disability functions. The most popular and most universally accepted metric tool is the Expanded Disability Status Scale (EDSS), first developed by Kurztke in 1983 [16].

EDSS is a comprehensive tool that allows a neurological evaluation based in eight individual functional systems (FSs)– vision, brainstem, pyramidal, cerebellar sensory, bowel/bladder, cerebral, and ambulation. All these individual evaluations are combined into a single score that reflect the overall neurological disability. EDSS provides an ordinal rating system (Fig. 1) that ranges from 0 (normal neuronal exam status) to 10 (death due to MS). Steps 1.0 to 4.5 refer to people with MS who are able to walk (fully ambulatory), EDSS steps 5.0 to 9.5 are defined by the walking impairment.

In addition to MRI and EDSS, several research groups had suggested the use of Evoked Potentials (EP) measures to predict the development of clinical disabilities [25][18]. EPs are electrical signals (or potentials) that are produced by the nervous system (e.g. cerebral cortex) in response to an external stimulus. The resulting signal is measured by special equipment at several body locations along the neural tracts involved. Besides the potentials in the cerebral cortex, signals can be recorded from the spinal cord, peripheral nerves and muscles. Features of EP were demonstrated to correspond well with clinical symptoms of MS and the degree of disability. EP have remained as a useful electrophysiological biomarker of the degree and dynamics of neurological deficit in the course of MS.

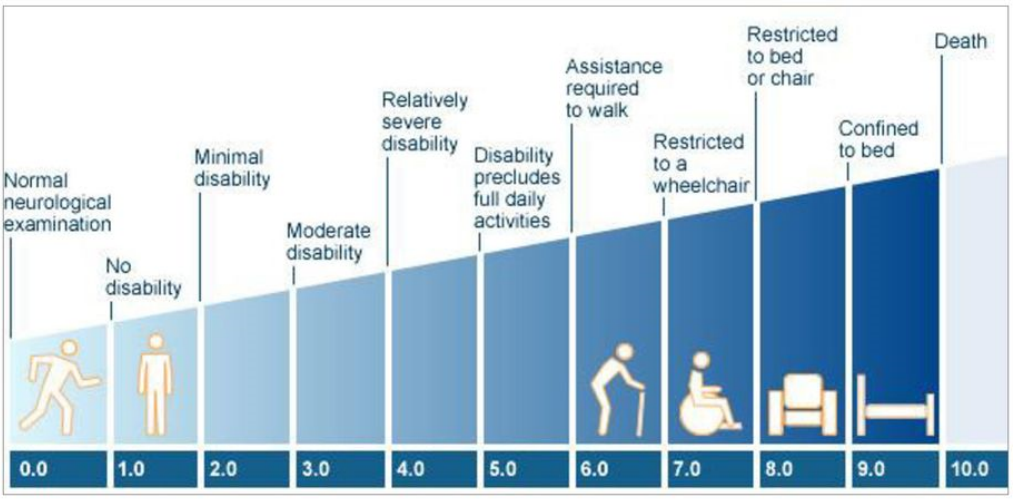


Figure 1: Stages of the EDSS score

1.2 Research Questions

The current research paper focuses on the study of several machine learning (ML) algorithms that uses motor EP measurements for predicting whether the MS patient is getting worse or not (binary classification) after two years. The proposed models use MEP latency data, in addiction to patient age and gender. Latency is a feature extracted from the EP and measures the time delay of the signal between the initial input and the beginning of the response. The data was obtained from a cohort of patients following treatment from the Rehabilitation & MS Center in Overpelt, Belgium. In addition, this research paper studies how well the probabilistic outputs of the models are calibrated, and how calibration can be improved.

To summarize, our research questions are:

1. Investigate the predictive performance of Motor Evoked Potentials (namely latency) and basic clinical variables for several machines learning algorithms.

2. The second research question aims to investigate and improve the calibration of machine learning models for predicting disease progression in Multiple Sclerosis patients.

2 Data

2.1 Data Description

In this study, the analysis was conducted on longitudinal data obtained from the the Rehabilitation & MS Center in Overpelt, Belgium, which is a recognised centre for care for people with multiple sclerosis, Parkinson's disease, Cerebrovascular accident patients (stroke) and coma patients, among others chronic diseases.

Two datasets were available for the research, containing *clinical* evaluations and *motor evoked* data. The clinical dataset included data over repeated patient visits. On each of the visits, spanning between 1982 and 2016, EDSS was quantified. All patients had at least one available visit in the dataset, and a maximum of 51 visits. The motor evoked dataset contained MEP measurements for each patient visit, spanning between 1992 and 2002. Latency (i.e. the time for the signal to arrive as illustrated in Fig. 2) measurements were included. Each patient visit had four different MEP measurements, two in muscles located in both hands (Abductor pollicis brevis - APL muscle), and two in both feet (in the Abductor Hallucis - AH muscle), with a total number of visits spanning between 1 and 18. Furthermore, information about the team of nurses that performed the measurement and type of machine was provided. Basic patient demographic information, such as age, and gender, were also included.

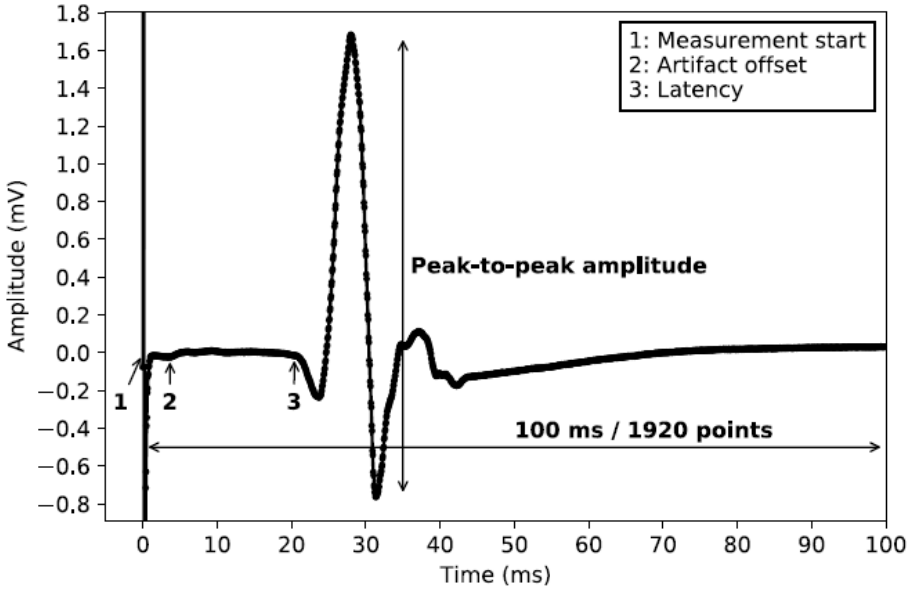
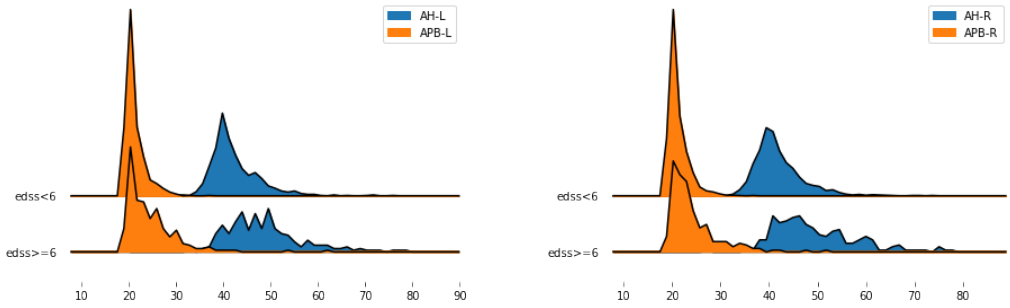


Figure 2: Example of a single motor evoked potential (MEP) measurement.

2.2 Characteristics of Study Sample

The *clinical* dataset included 582 patients, with a total of 7414 visits. On each of the visits, EDSS was recorded. First visit occurred in 1982, and the last in 2016. Most of the patients had multiple visits over a broad period of time, ranging from 1 day to 16 years. The average time a patient remained in the study was 5 years. The average number of visits per patient was 12 (range 1-50). The median EDSS score was 3.0 (range 0-9.5, mean 3.652) at study entry, and 3.0 (range 0-9.5, mean 3.431) at the end of the study. 130 patients had a significant improvement of their EDSS score at the end of the study, 236 did not have either improvement nor deterioration, 216 patients got worse.

On the other hand, the *motor evoked* dataset included information over 520 patients, from which 373 were females and 141 males. Patients date of birth ranged from 1909 and 1994. There were a total of 3326 visits, on each of the visits, four different motor evoked potential latency measures were recorded. The mean AH value for patients with an EDSS <6 was 40.96 (right hand muscle) and 41.19 (left hand muscle), and 48.89 (right) and 47.73 (left) for patients with EDSS ≥ 6. For APL muscles the mean for patients with an EDSS < 6 was 20.58 (right foot muscle) and 20.41 (left foot muscle), and 24.10 (right) and 23.88 (left) for patients with EDSS ≥ 6. See Fig. 3 for visual representation.



(a) Left side body

(b) Right side body

Figure 3: MEP latency distribution for patients with EDSS<6 and EDSS$\geq$ 6 (assistance required to walk). Divided between left and right side of the body.

Other variables that were included in the dataset were; *machine*, indicating which of the two available machines the measurement was performed on; *team*, indicating which of the two teams available at the centre performed the measurements; *measurement id*, unique identifier for a measurement; and *test id*, a unique identifier for a test. None of the previous variables were used for the study as it was considered out of scope for this research.

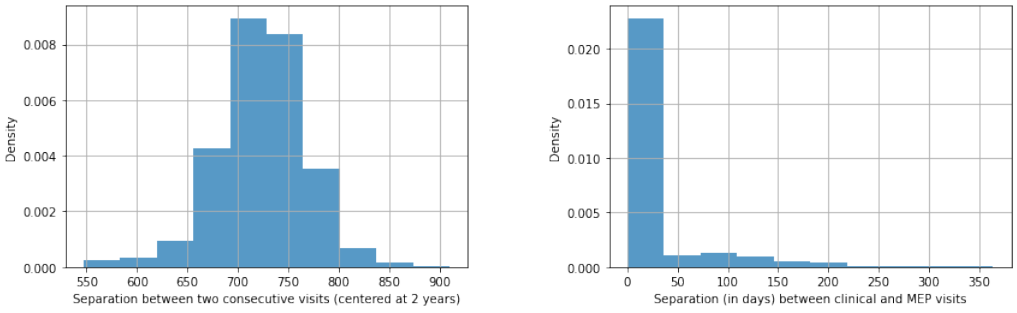
2.3 Data Preprocessing

From the *motor evoked* dataset, four new features were generated, aiming to count for individual changes in MEP measurements. The new feature calculated the increase (decrease) of the APL and AH latencies between two consecutive visits.

A common binary target was defined for each visit (patient is getting worse or not after two years). We used the standard definition of disability progression [13], see below in equation 1 for details. Where $EDSS_{T_0}$ is the EDSS value at the time of the first visit, Δ_{EDSS} is the disability progression after two years (± 0.5 years), where the closest visit to the two years mark was selected in case of multiple matching. We allowed patients to have multiple positive outcome events. Visits without a two-year mark follow-up were discarded.

$$w_i = \begin{cases} 1 & \text{if } \Delta_{EDSS} \geq 1.5 \ \& \ EDSS_{T0} = 0 \\ 1 & \text{if } \Delta_{EDSS} > 1.5 \ \& \ EDSS_{T0} \leq 5.5 \\ 1 & \text{if } \Delta_{EDSS} \geq 0.5 \ \& \ EDSS_{T0} > 5.5 \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

After these conditions were processed, 3368 visits (45%) did not have a two year (± 6 months) follow up visit and were removed from the data. Histogram of the difference in time between the time of the first visit (T_0) and the two years follow up (T_1) for the target calculation is showed in Fig. 4 (left). Roughly 12% of the patient visits in the *clinical* dataset were targeted as a disease progression.



(a) EDSS:T_0 and T_1

(b) MEP visit and T_0

Figure 4: Time difference between visits.

At least 68 patients from the *clinical* data were removed as they did not have a matching patient id in the *motor evoked* dataset. The remaining 514 patients were combined with those from the motor evoked dataset into a single one based on their unique patient identifier and the time between clinical evaluation visits (EDSS T_0) and motor evoked visit, with an

allowed maximal separation of one year. This final dataset contained a total set of 2164 patient visits. Another additional 25 patients were removed as both clinical and motor evoked potential visit dates were separated by more than a year. From the remaining set of visits (2139), 250 were classified as disease progression, approximately 12%. 144 patients had at least one positive disease occurrence. Time difference between the clinical evaluation and motor evoked visits is showed in Fig. 4 (right).

Standard Pearson correlation was obtained for all the features in the final dataset. As it can be observed in the bottom row in Fig. 5, no significant associations were found between any of the variables and the disability progression (w).

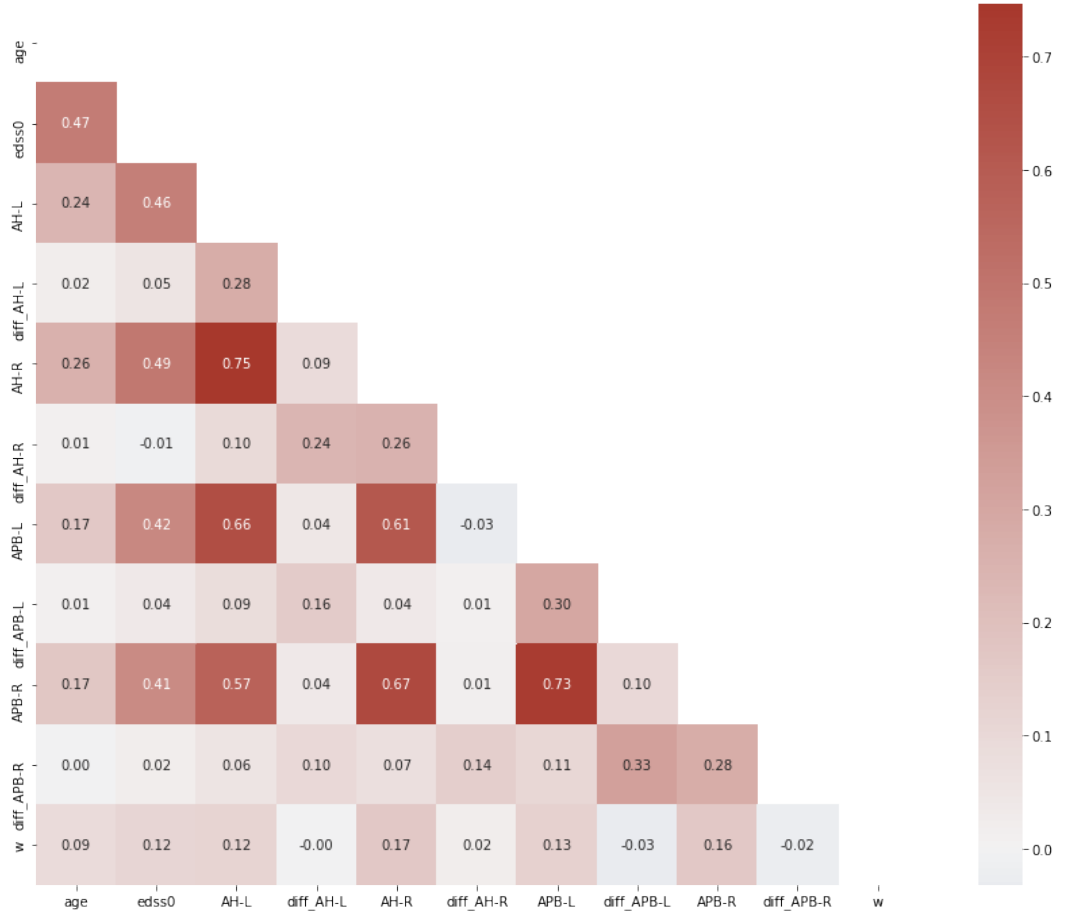


Figure 5: Pairwise correlations among Age of the patient, EDSS evaluation at first visit (edss0), AH and APL latency, AH and APL latency difference between two visits (diff_AH and diff_APB), and disability progression (w).

3 Methodology

The following section includes a brief review of some methods used to evaluated model performance. It is followed by study of some available methods to correct for outcome class imbalance that is present in this research. Further, we test several machine learning models to predict EDSS progression from MEP. Finally, some calibration techniques are presented.

3.1 Evaluation of Predictive Models

There exist different methods to evaluate model performance of the predicted probability estimations in classification problems. Common metrics for binary classifiers are: Brier score (BS), Log Loss, and area under the receiver operating characteristics AUROC.

3.1.1 Brier Score

Originally introduced in 1950 by Glenn Brier to evaluate weather forecasts [6]. The most common formulation of the score is:

$$BS = \frac{1}{N}\sum_{t=1}^{N}(\hat{p}_i - Y_i)^2 \tag{2}$$

BS measures the mean squared difference between the predicted probability $\hat{p}_i$ and the observed Y_i outcome. BS always takes values between 0 and 1, since this is the largest possible difference between a predicted probability and the actual outcome. The smaller the Brier score, the better, so the metric is also referred as Brier Score Loss.

BS is an appropriate evaluation metric for binary and categorical outcomes that can be structured as true or false, but inappropriate for ordinal outcomes with more than 3 possible outcomes and also bad for rare events [4].

This metric should be used with care because a lower Brier score does not always mean a better calibrated model. This is because the BS can be decomposed as the sum of the calibration and refinement loss [3]. Calibration loss is defined as the mean squared deviation from empirical probabilities derived from the slope of ROC segments. On the other hand, refinement loss is defined as the expected optimal loss as measured by the area under the optimal cost curve. Refinement loss can change independently from calibration loss, therefore a lower BS does not necessarily mean a better calibrated model.

3.1.2 Log loss

Log loss, also called logistic regression loss or cross entropy loss, is an indicative of how close the prediction probability is to the correspondent actual or true value (0 or 1 for

binary classification). For a single sample with true label $y \in 0, 1$ and a probability estimated $\hat{p}_i = Pr(Y_i = 1)$ the log loss is defined as:

$$L_{log}(Y_i - \hat{p}_i) = -(Y_i \log(\hat{p}_i) + (1 - Y_i) \log(1 - \hat{p}_i)) \tag{3}$$

The more the predicted probability diverges from from the actual value, the higher the log-loss value. While training a classification model, we would expect that the predicted probability to be as close as possible to the actual value (of 0 or 1), whereas prediction probabilities farther away from its true value are penalized. Log-loss is a good metric to evaluate classifier calibration [15] especially if cost sensitive decisions are made based on the classifier outcome. A smaller value of logarithmic loss means better calibration.

3.1.3 Area Under the Receiver Operating Characteristics

The Receiver Operator Characteristic curve (ROC curve) is frequently used in evaluating binary classifiers in ML applications [19]. The ROC curve typically features the true positive rate (or sensitivity) on the Y axis, against the false positive rate (1 - specificity) on the X axis (Fig. 6). The overall performance of a classifier, summarized over all possible thresholds, is given by the area under the receiver operating characteristics (AUROC). An ideal curve will hug the top left corner of the ROC plot, indicating a high true positive rate and a low false positive rate. The larger the AUROC the better the classifier, and it will equal to 1 in a perfect model. If the model is not better than random classification, then the AUROC will be 0.5.

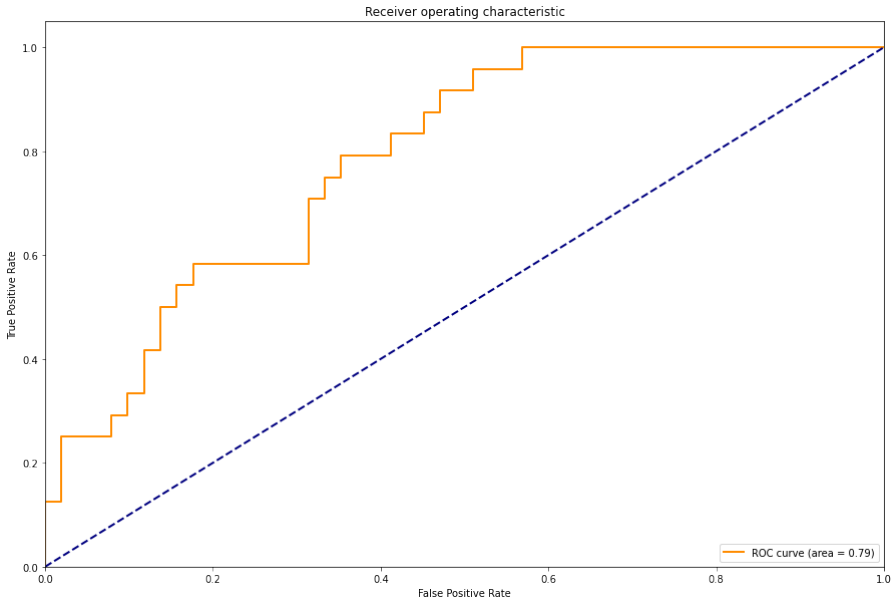


Figure 6: Example AUROC.

3.2 Calibration

An accurate estimation of the outcome probabilities for individuals is important in medical practice. For example, for a given model, probabilities may be systematically high for all patients, irrespective of whether they experienced the event or not. These predictions may support critical clinical decision making and better inform patients. For example, a strong overestimation of a positive disease progression in a patient in early phase of MS could lead the patient to abandon any personal future plans, or the physician to recommend the patient to start with a more aggressive treatment against MS.

However, even very accurate classifiers may output class probabilities of rather poor quality, specially in cases with imbalanced data. Algorithms should not only give higher probability estimates for patients with the event than for patients without the event ('discrimination'), but also should have well estimated probabilities, relating to the agreement between the estimate and observed number of events [29]. A classifier that more often predicts the correct class and also generates better rankings than some other classifier may still produce badly calibrated probability estimates. Typically, discrimination is quantified by means of using the AUROC. Models with good discrimination ability will have high sensitivity and specificity simultaneously, leading to a high values of AUROC, but their estimated probabilities can be unreliable.

Probability calibration maps original class probabilities to more accurate ones (Fig. 7). Recent systematic reviews have found that calibration is assessed far less than discrimination ([2], [5], [17]). A properly calibrated model may result in a lower AUC, making the algorithm less preferable than a competitor algorithm, but it will not have overestimated the risk. Poor calibration can make predictions misleading, leading patients and physicians to make decisions in anticipations of an event, or the absent thereof.

In this research, the two most popular methods of calibrating machine learning models are used, namely *isotonic* regression and *Platt* scaling methods.

3.3 Assessing Calibration

Besides BS and log-loss to calculate the difference between the calculate probabilities and the true value, there are some common visual representations and metrics to evaluate the model calibration of a certain model. The most used methods are presented below.

3.3.1 Reliability Diagrams (Calibration curves)

In a binary classification problem, a model is trained to estimate the probability of an example to be classified as a positive class, i.e. $f(x_i) = p(y_i = 1|x_i)$. Once the probabilities for a test set are obtained, these are partitioned into M partitions (each of size $\frac{1}{M}$), in which each

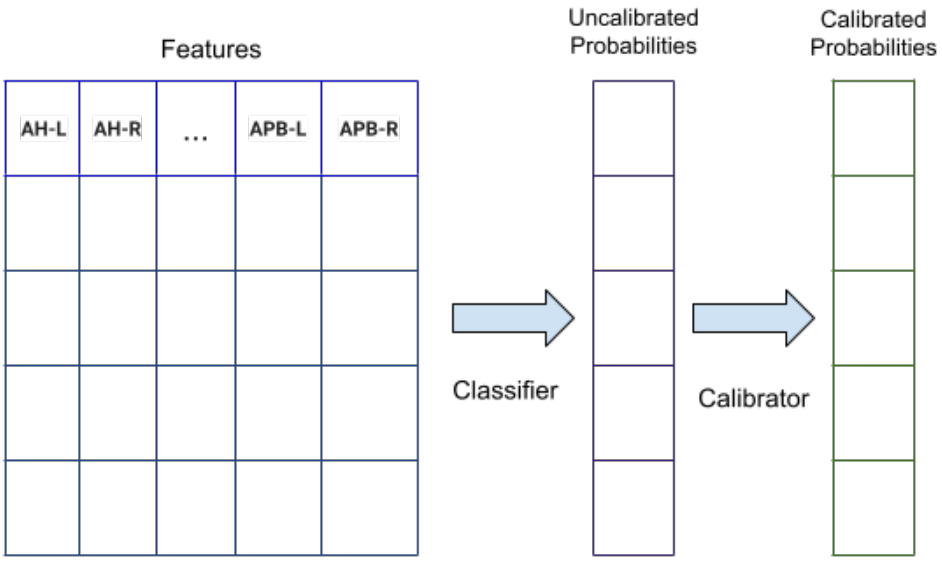


Figure 7: two-step calibration approach.

represents a disjoint interval of probabilities between 0 and 1. Then for each partition, the mean predicted probability, or average confidence within the bin, is calculated as:

$$conf(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \hat{p}_i, \tag{4}$$

Where B_m is the set of indices of samples whose prediction confidence falls into the interval $I_m = (\frac{m-1}{M}, \frac{m}{N})$, and $\hat{p}$ is the predicted probability for sample i. Finally, the relative frequency of positive examples, or accuracy, is calculated as:

$$acc(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} 1(\hat{y}_i = y_i), \tag{5}$$

Where $\hat{y}_i$ and y_i and the predicted and true class labels for sample i. The reliability diagram plots *accuracy* against *confidence* as shown in Fig. 8. If the model is perfectly calibrated, the diagram should plot the identity function. Any deviation from a perfect diagonal represents miscalibration. The most common forms of miscalibration are:

- *Systematic Overestimation*. Compared to the true distribution, the predicted probability distribution is shifted towards the right. This is common phenomenon when the data for the model is imbalanced and has very few positives cases.
- *Systematic underestimation*. Compared to the true distribution, the predicted probability distribution is shifted towards the left.
- *Center of the distribution is too heavy*. This phenomenon occurs when the algorithm used tend to push the predicts probabilities away from 0 and 1. An example of such algorithms are support vector machines and boosted trees [23].

- *Tails of the distribution are too heavy*. This is the opposite case. Other methods such as naive bayes push predictions closer to 0 and 1 [23].

The reliability diagram can help to understand the relative calibration of the predictive probabilities from different classification models.

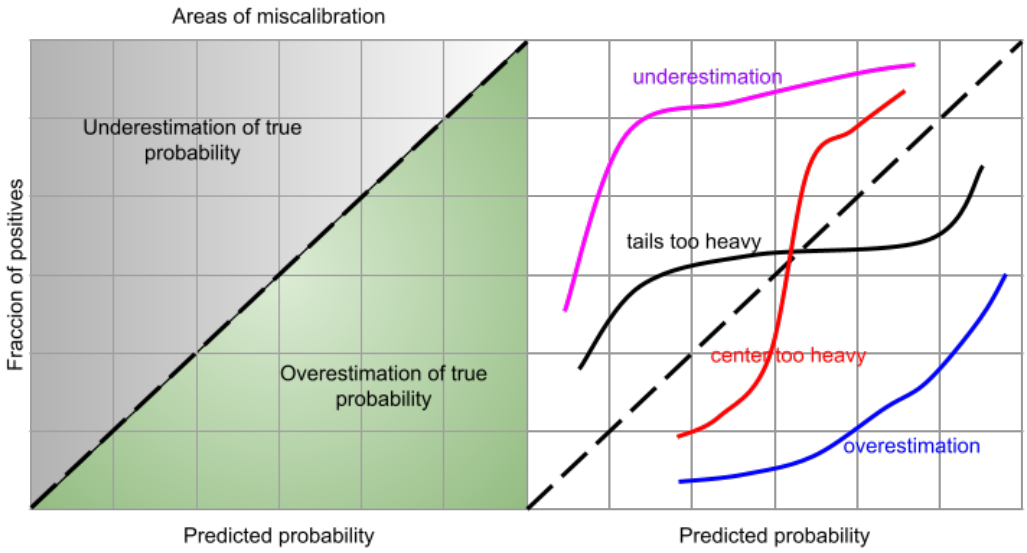


Figure 8: Example of miscalibration.

3.3.2 Expected Calibration Error

In order to be able to not just evaluate a model with a diagram, sometimes it is more convenient to have a scalar summary statistic of calibration. The Expected Calibration Error (ECE) [22] partitions each prediction probability into M equally spaced bins, similarly to reliability diagrams, and then it simply takes a weighted average over the absolute accuracy/confidence difference.

$$ECE = \sum_{m=1}^{M} \frac{|BM|}{n} |acc(B_m) - conf(B_m)| \tag{6}$$

Difference between confidence and accuracy for a given bin represents the calibration gap. ECE is used as a primary empirical metric in this thesis.

3.4 Calibration Methods

In this research, two approaches for performing calibration of probabilistic predictions are used: a parametric approach based on Platt sigmoid model and a non-parametric approach based on isotonic regression.

It should be noted that probability calibration should be done using a separate test/validation data that is not used for model fitting. Two methods can be done: the use of an independent

calibration set, and using cross validation techniques to generate scores from the training set. In the cross validation approach to generate calibration data, the training data is divided into k folds, the model is trained on the remaining k-1 folds, and then make predictions on the left-out fold. After iterating this process k times with a different fold, we will have a set of predictions. This set of predictions serve as an appropriate calibration set.

3.4.1 Platt scaling

This method was invented by John Platt [27] in the context of calibrating support vector machines. Platt scaling is often called sigmoid scaling, both terms will be used interchangeably in this research. Platt scaling works by fitting a logistic regression model to the classifiers predicted probability scores. The regressor has the following form:

$$p(Y_i = 1|f_i) = \frac{1}{1 + exp(Af_i + B)} \tag{7}$$

Where Y_i is the true label of sample i, and f_i is the output of the un-calibrated classifier for the same sample i. Parameters A and $B \in \mathbb{R}$ are two scalar values that are learned by the algorithm and are to be determined when fitting the regressor via maximum likelihood.

This method assumes that the calibration curve can be corrected by applying a sigmoid function to the raw predictions. This method works best if the calibration error is symmetrical, meaning the classifier output for each binary class is normally distributed with the same variance [14], but it can be a problem for highly imbalanced classification problems, where outputs do not have equal variance.

3.4.2 Isotonic Regression

Isotonic regression [32] is a non-parametric method to calibrate scoring classifiers in binary classification. It uses the convex closure of the ROC curve to discretise the scores into bins; the slope of each segment can be interpreted as an empirical likelihood ratio, from which a calibrated posterior probability for the corresponding bin can be derived. The resulting calibration outcome consists of a step-wise non-decreasing function, which can be interpolated if a continuous calibration map is needed.

Given the output un-calibrated prediction f_i from the classifier, and the true label Y_i, the isotonic regression is simply $Y_i = m(f_i) + \varepsilon_i$,where m is the non-decreasing function.

Isotonic regression is more general when compared to *sigmoid*, its only restriction is that the mapping function is monotonically increasing. It is more powerful as it can correct any monotonic distortion of the non-calibrated model. However it is more prone to overfitting, especially in small datasets ([20]).

3.5 Machine Learning Implementation

Different ML algorithms were used to learn the relation between the MEP, and related patient demographic variables with the clinical outcomes (disability progression):

- Logistic Regression (LR).
- Random Forest (RF).
- Gradient Boosted Decision Trees (GBDT) .
- AdaBoost Classifier (ADA).
- Non-linear SVM.

In the literature, the main modelling technique for modelling the relationship between EP with EDSS is Logistic Regression. Three additional ensemble methods were used for the study, namely Random Forest (RF), Gradient Tree Boosting (GBDT) and AdaBoost (ADA) Classifier.

We opted for these ensemble methods because they are non-linear classifiers and are robust against overfitting ([12]). Ensemble methods has been proved to be effective in imbalanced data problems [10]. Their main goal is to combine the predictions of several base estimators (known as weak learners) with a given learning algorithm in order to improve generalizability and robustness over a single estimator. Two families are usually distinguished: average and boosting methods. Averaging methods (e.g. Random Forest) consists in building several independent estimators and then average their predictions, each of the independent estimators has high variance, but low bias. Averaging these estimators reduces the variance. Boosting (e.g. ADA and GBDT) methods works in a similar way, except that the estimators are built sequentially, where one tries to reduce the bias of the previous estimators.

Additionally, a Vector Machines (SVM) model, with Non-linear decision boundaries using Radial Basis Function (rbf) kernel was implemented. This model allows to address the problem of possible non-linear boundaries between classes by enlarging the feature space using polynomial or other functions to calculate the bounds between classes.

All ML algorithms were implemented with Scikit-learn (or sklearn) library for the python programming language.

3.6 Implementation Pipeline

The use of more complex models may lead to a situation of overfitting the data. Overfitting occurs when the model fits the noise instead of the signal. This is an undesirable situation

because the fit obtained will not yield accurate estimates of the response on new observations that were not part of the original training dataset. In order to avoid this, it is usual to split the processed data into two randomized sets, frequently named train and test sets. We ensure a correct separation of patients between the training and test set by using their unique identifier. Then, features were standardized by removing the mean and scaling to unit variance. This scaling is a common requirement for many machine learning estimators: they might behave badly if the individual features are not of similar scale.

In order to address the issue with the high imbalance in the data (low occurrence of positive outcomes), balanced class weights were used and compared to the non-weighted model, unfortunately this approach is only possible for some of selected classifiers (such as LR, RF and SVM). This approach incorporates weights (e.g. giving higher weight to the minority class) to the classes into the algorithm cost function. For our research, we aimed for balance weights between classes; the values of the disability progression are used to automatically adjust weights inversely proportional to class frequencies. The following weights were obtained:

$$W_j = \frac{Total\_Number\_Samples}{Number\_Classes * Number\_Samples\_Class_j} = [w_1 = 0.564, w_0 = 4.382] \quad (8)$$

Then, these weights are introduced into the classifier cost function. For example, a common cost function for LR model is the *log loss*. The weight adapted cost function has then the following form:

$$log\ loss = \frac{1}{N}\sum_{i=1}^{N}[-(w_0(y_i * \log \hat{y}_i) + w_1(1-y_i) * \log(1-\hat{y}_i))] \quad (9)$$

We can infer from this process that majority classes will obtain smaller weights, and smaller updates to the model coefficients, while minority classes will have larger weights that will have larger impact to the coefficients.

Next, cross-validation was used for hyperparameter tuning in order to improve the model performance scores. A hyperparameter is a parameter whose value is used to control the learning process, and is not directly derived via training of the model. The model is trained in k-1 of the folds as training data and the resulting model is validated on the remaining fold/test set, then the average across the five performances is kept as final result. For cross validation is again important that patients do not appear in two different folds during the grouped 4-fold cross-validation. A broad search within all possible hyperparameter combinations was done by using a *randomized search on hyperparameters* class of the Model Selection library in Sklearn. In contrast with the exhaustive search (performed by *Grid Search* class), randomized search does not try all parameter combinations, but rather a fixed number of parameter settings is sampled from the specified distributions. The number of parameter settings that are tried is given by *number of iterations* parameter, which was fixed to 100.

Decision over the best hyperparameters was done with Brier scoring.

Due the small size of the dataset and the relative small number of patients, the performance of the algorithms is greatly influenced by the choice of the training and set sizes. To have an understanding about the influence of those limitations and model performance behaviour, a simulation was performed; we run the model fitting and calibration process for different choices of train/test splits, for a total of 1000 times each, ensuring that the same patient is not included in both the training and in the test set. To illustrate this, for some of the splits we obtained AUC values of 0.831, while others obtain values of 0.378. We chose four different sizes for the training set, composed of 20, 30, 50 and 80 percent of the dataset. This method also gives information on the necessary dataset size to achieve a certain performance[21][8], and on how much room for improvement the algorithm has when given more data [33], [28].

Then, the estimated probabilities were calibrated with *isotonic regression* and *Platt scaling* techniques. To the limited amount of observations, cross-validation technique was used ($K = 4$) on the training set to estimate the parameters of the calibration function. For each cross-validation split, the model fits the base estimators to the training subset, and then it is calibrated. Finally, calibration metrics and plots were obtained for the model and compared with the uncalibrated model. An schematic of the process is depicted in Fig. 9.

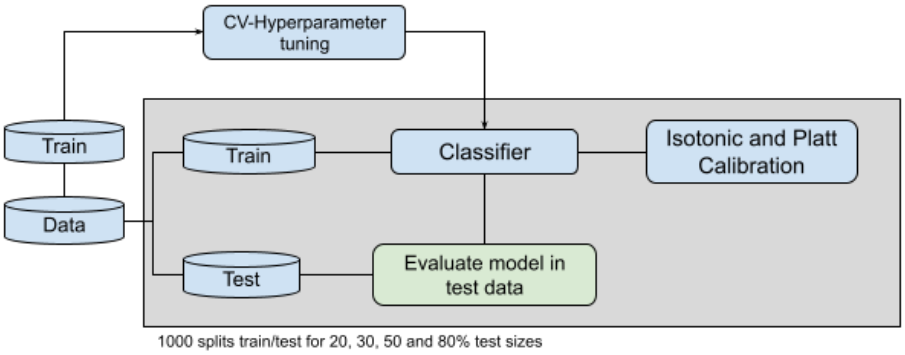


Figure 9: process pipeline.

From the process pipeline (Fig. 9), it can be observed that the hyperparameter tuning process was excluded from the iteration process. The reason was merely due memory limitations of the virtual environment were models were running. As an example, RF model tuning alone would take approximately between 10 and 20 minutes for a single execution, while simulation (with the 1000 repeated executions) of the model fitting and calibration for the same model takes around 2-3 hours. Repeating the process of model tuning jointly

with fitting and calibration for the specified number of iterations would easily take days of computational time with the current specifications. In order to circumvent this, it was decided to leave the tuning process outside of the main loop. Some data leakage is likely to happen, because the tuning process optimizes the hyperparameter in a set of patients, included in the train set, that will be, during the iteration step, in the test set too. In order to minimize this issue, the number of hyperparameters included in the process was kept small, with a maximum of four or five, depending of the ML algorithm approach.

4 Results

The following section describes the results for both of the research questions. First, outcome and interpretation of the predictive performance of MEP and patient demographic information for the different ML approaches, and secondly, the results of the calibration phase.

4.1 Predictive performance of Motor Evoked Potential

The parameter tuning procedure was performed using the randomized search class as described in section 3.6. For logistic regression model, the randomized search of hyperparameter was performed over two penalization norms, *l1* and *l2*, and the (inverse) of the regularization strength (C). The set of parameters giving the best score were: C: 0.16 and C:1×10^5 for the non-weighted and weighted LR models respectively, and *L1* (Lasso penalty) for both models.

For RF, the search was performed over the following hyperparameters: number of trees in the forest, proportion of minimum number of samples required to split an internal node, proportion of minimum number of samples required to be at a leaf node, and the maximum depth of the resulting trees averages. The set of parameters giving the best score were: Number of trees: 235, minimum samples split: 20%, minimum samples leaf: 10%, and maximum depth: 30, for the non-weighted model. For the class weighted version the best hyperparameters were: Number of trees: 1185, minimum samples split: 10%, minimum samples leaf: 10%, and maximum depth: 31.

The next implemented algorithm was GBDT. The hyperparameter search was conducted over learning rate, which shrinks the contribution of each of the individual trees, number of boosting stages (or estimators) to perform, proportion of minimum number of samples required to split an internal node, proportion of minimum number of samples required to be at a leaf node, and maximum depth of the individual regression estimators. Best found parameters were: learning rate: 0.01, number boosting stages: 235, minimum samples split: 80%, minimum samples leaf: 40%, and maximum depth: 25.

The hyperparameter search for ADA was performed over the maximum number of estimators at which boosting is terminated (number estimators), and for weight applied to each classifier at each boosting iteration (learning rate). The hyperparameters were: number estimators: 100 and learning rate: 0.01. Finally, for SVM the search was performed only for parameter ν, which limits the fraction of training errors and support vectors, the optimized value was $\nu = 0.5$.

All AUC results for the different test sizes are summarized numerically in table 1 and graphically in Fig. 10.

	Mean AUC			
% Train	LR	LR(Weighted)	RF	RF(Weighted)
20	0.460 ± 0.031	0.570 ± 0.037	0.620 ± 0.024	0.626 ± 0.022
50	0.515 ± 0.048	0.604 ± 0.028	0.645 ± 0.025	0.639 ± 0.027
70	0.565 ± 0.045	0.613 ± 0.039	0.652 ± 0.036	0.641 ± 0.034
80	0.575 ± 0.055	0.612 ± 0.054	0.655 ± 0.046	0.644 ± 0.049
	AUC			
% Train	Gradient Boosting	Ada Boost	SVM*	SVM* (Weighted)
20	0.623 ± 0.021	0.607 ± 0.030	0.500 ± 0.047	0.483 ± 0.069
50	0.637 ± 0.024	0.625 ± 0.027	0.487 ± 0.054	0.452 ± 0.044
70	0.637 ± 0.037	0.629 ± 0.040	0.518 ± 0.064	0.492 ± 0.084
80	0.637 ± 0.052	0.628 ± 0.053	0.520 ± 0.060	0.493 ± 0.107

Table 1: The leftmost column indicates what percentage of the dataset was used for training. LR (Logistic Regression) indicate the classifier that was used. The values after $\pm$ indicate the standard deviations. (SVM results were obtained after only 10 iterations, instead of the default 10^3).

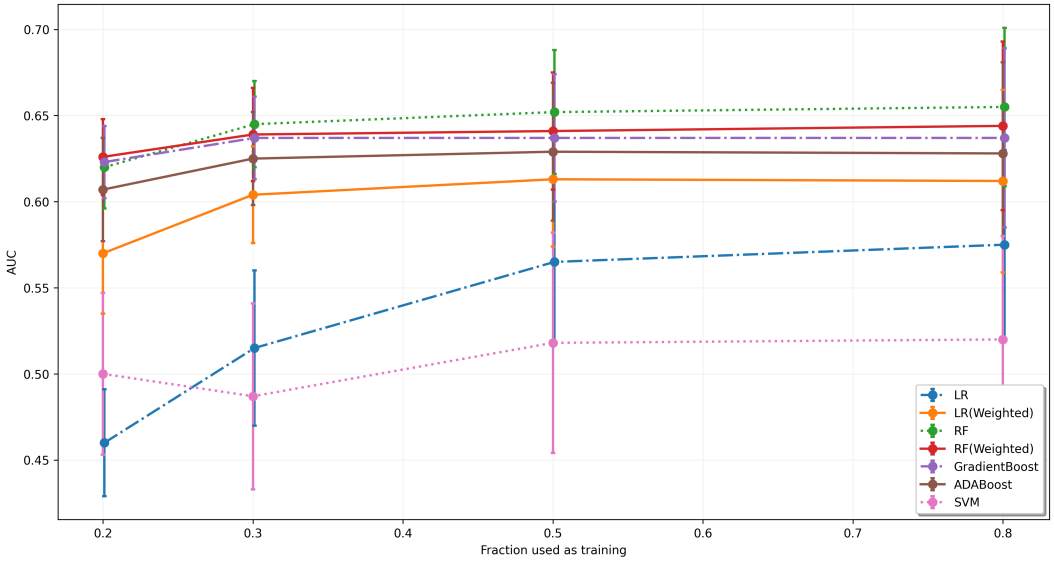


Figure 10: Results of the disability progression classification. Results are shown for different sizes of the training set, with the error bar indicating the standard deviation.

As expected, the performance of the model fit increases as the size of the training set increases, while the variance of the results also increases due the smaller sample size of the test set. RF outperform LR model, except for the class weighted model. The use of a balanced class weight in order to correct for the dataset imbalance has a positive effect over AUC scores. On the other hand, the use of balanced class weights in the RF model does not translate into a better performance, as there is no difference with the unweighted model, showing the good performance of ensemble methods in handling imbalanced data.

Other ensemble methods (GBDT and ADA) perform similarly to RF, these methods are also capable of quickly having steady AUC scores with less training data; their AUC does not significantly increases when using beyond 30% as a training data. Ensemble methods have the capacity of combine predictions in order to improve generalizability and robustness over single estimators such as LR. On the other side, SVM algorithm (both unweighted and weighted versions) had the worse performance, with scores below the rest of the models.

Densities of the averaged AUC scores for all algorithms (except for SVM) and 80% training set sizes are presented in Fig. 11. We can see that there is considerable overlap between the curves (with the only exception of LR). Best mean estimates were for RF, while the worse for LR.

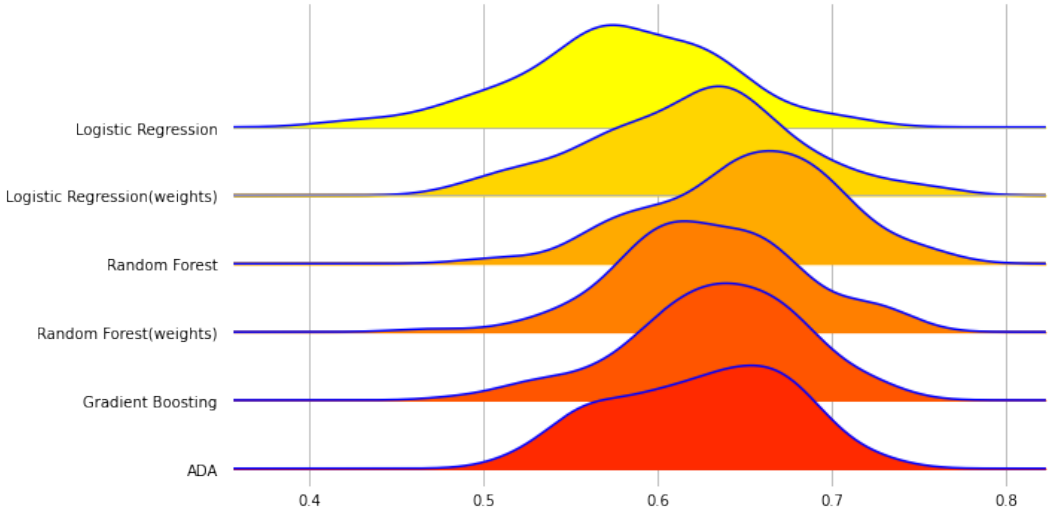


Figure 11: AUC distribution and mean scores for 80% Train.

The model with the best average AUC score was the RF, but in order to see if scores were significantly different from each other and not obtained by chance, we run Friedman's test. This test computes the average rank of each classifier and then compares with a critical value from a χ^2 distribution. If there were no differences between all algorithm scores, we would

expect the ranks to be evenly spread among all iterations, so the null hypothesis is that all classifiers are the same. Results for Friedman's test (χ_f^2=72, $p-value < 0.001$) show that we can reject the hypothesis that all the classifiers are the same. This test does not differentiate were the difference resides, and a post-hoc test is needed to pinpoint where the difference resides.

To determine exactly which models are different, we calculate the critical difference by using the Nemenyi [9] post-hoc test. Nemenyi test is similar in calculations to Friedman test. For any two classifiers f_{j1} and f_{j2}, the q statistic is simply calculated as:

$$q = \frac{\bar{R}_{.j1} - \bar{R}_{.j2}}{\sqrt{\frac{k(k+1)}{6n}}} \tag{10}$$

Where k ($k = 6$) is the total number of classifiers, and n ($n = 100$) the number of iterations.

Results are presented in Fig. 12. Although the best scoring model was RF, there is no statistical difference between any of the ensemble models according to Nemenyi (only borderline difference between RF and ADA). The use of class weights does significantly improve AUC performance of the LR model, while for RF there was no difference between the weighted and non-weighted model.

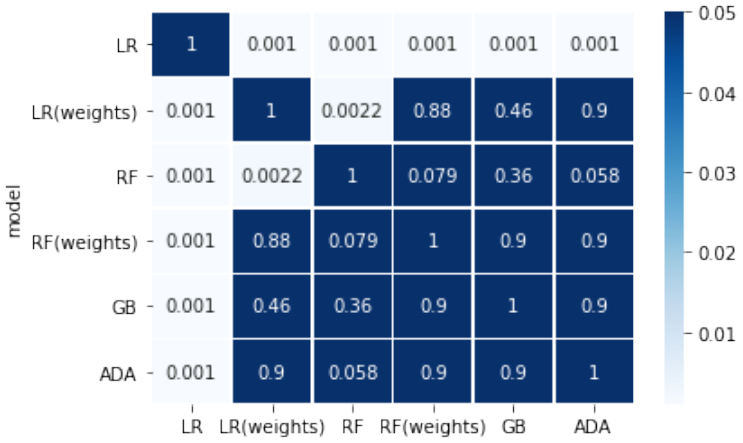


Figure 12: Nemenyi results for the different ML classifiers

4.2 Calibration

After the model parameter optimization was finished, calibration methods were implemented over the predicted probabilities of the previously fitted algorithm. As described in section section 3.6, two different calibration techniques were implemented (Isotonic and Platt scaling), using cross-validation ($k = 4$) over the training set to ensure unbiased data is always used to

fit the calibrator. Three different methods were used to asset the accuracy of the probabilistic predictions, namely *Brier* score, *Log-loss*, and *ECE* (with $M = 10$ bins).

Table 2 displays model ECE score results for all implementations. It can be observed that LR is already a well calibrated model, with results showing as little as $3\% - 5\%$ ECE ranges, while results after calibration are not significantly improved. The combination of using a simple model such as logistic regression and the use of penalized regression techniques, such as Lasso regularization, usually return well calibrated estimates, specially for small datasets [30].

Next results are for the LR model with class weights. It should be noted that the use of class weights was introduced to correct the class imbalance of the data. The introduction of weights in the model fitting had a significant impact in the model AUC scores, with a $\sim 7\%$ improvement when compared to the unweighted model, but the introduction of class weights in the cost function did result in an inflation of the model predicted probabilities.

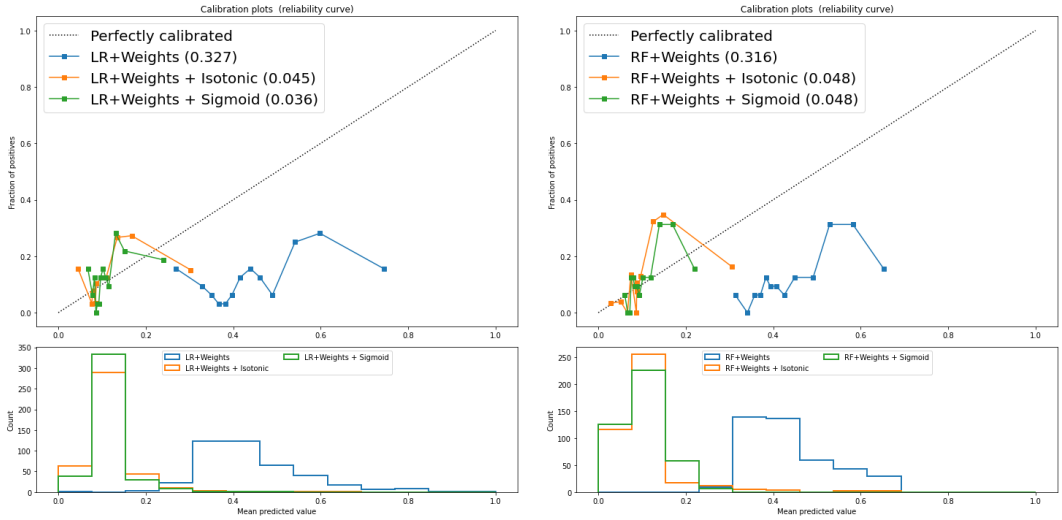
As it can be seem in Fig. 13a, the calibration line (in blue) for the class weighted LR model is shifted towards the right of the main diagonal. This shifting means that the model is showing a systematic overestimation when compared to the true distribution of the test dataset. Fortunately, both of the calibration techniques applied are able to correct the miscalibration and bring the probabilities closer to the identity line, although the form and distribution of the probability outcomes still indicate that the tails of the distribution are too heavy (predictions closer to 0).

Similar behaviour is observed for the use of the RF classifier in table 2. RF probability estimates had already a good level of calibration, but the use of class weight translates in a strong miscalibration of the probability estimates. Again, as it was in the LR weighted case, both isotonic and Platt's scaling techniques are able to correct it (Fig. 13b). The GBDT model had similar response to RF. Both methods combine decision trees, where the difference resides on how results are combined. In our study case, results are similar before, and after calibration.

Another model that shown some level of miscalibration was ADA. As briefly explained in section 3.5, ensemble techniques such as ADA has been proven to be a very efficient classification method, ADA model creates accurate predictions by combining many weak and inaccurate rules. On the other hand, this method tends to provide inaccurate estimates [7]. This conduct can be observed in table 2. ECE scores for the uncalibrated model show scores ranges between 10% (with a 20% train set) and 9% (with a 80% train set). The use of calibrators is able to correct the score values.

	Mean ECE			
Model	20% Train	50% Train	70% Train	80% Train
LR	0.041 ± 0.015	0.034 ± 0.013	0.037 ± 0.014	0.056 ± 0.014
LR + Isotonic	0.042 ± 0.015	0.036 ± 0.013	0.035 ± 0.014	0.049 ± 0.017
LR + Platt	0.042 ± 0.016	0.036 ± 0.013	0.029 ± 0.012	0.031 ± 0.015
LR(Weighted)	0.355 ± 0.023	0.355 ± 0.020	0.352 ± 0.017	0.334 ± 0.028
LR(Weighted)+Isotonic	0.041 ± 0.014	0.035 ± 0.013	0.033 ± 0.012	0.042 ± 0.017
LR(Weighted)+Platt	0.040 ± 0.014	0.033 ± 0.012	0.028 ± 0.011	0.030 ± 0.014
RF	0.045 ± 0.012	0.037 ± 0.010	0.030 ± 0.012	0.033 ± 0.012
RF + Isotonic	0.043 ± 0.014	0.037 ± 0.012	0.034 ± 0.012	0.042 ± 0.016
RF + Platt	0.047 ± 0.012	0.040 ± 0.010	0.033 ± 0.009	0.033 ± 0.013
RF(Weighted)	0.350 ± 0.022	0.345 ± 0.018	0.332 ± 0.017	0.304 ± 0.020
RF + Isotonic(Weighted)	0.042 ± 0.012	0.036 ± 0.013	0.034 ± 0.012	0.040 ± 0.015
RF + Platt (Weighted)	0.045 ± 0.011	0.036 ± 0.011	0.033 ± 0.010	0.031 ± 0.011
GBDT	0.048 ± 0.012	0.041 ± 0.011	0.034 ± 0.009	0.032 ± 0.011
GBDT + Isotonic	0.046 ± 0.013	0.039 ± 0.012	0.035 ± 0.011	0.043 ± 0.016
GBDT + Platt	0.048 ± 0.012	0.041 ± 0.010	0.034 ± 0.010	0.035 ± 0.013
ADA	0.101 ± 0.023	0.098 ± 0.020	0.093 ± 0.021	0.089 ± 0.025
ADA + Isotonic	0.043 ± 0.015	0.036 ± 0.012	0.033 ± 0.012	0.038 ± 0.015
ADA + Platt	0.046 ± 0.014	0.038 ± 0.011	0.033 ± 0.011	0.034 ± 0.013
SVM*	0.037 ± 0.015	0.038 ± 0.015	0.030 ± 0.011	0.035 ± 0.015
SVM* + Isotonic	0.037 ± 0.011	0.037 ± 0.008	0.027 ± 0.006	0.036 ± 0.014
SVM* + Platt	0.036 ± 0.009	0.036 ± 0.012	0.028 ± 0.010	0.033 ± 0.012
SVM* (Weighted)	0.051 ± 0.015	0.043 ± 0.011	0.032 ± 0.014	0.039 ± 0.016
SVM* + Isotonic (Weighted)	0.039 ± 0.009	0.034 ± 0.014	0.023 ± 0.010	0.036 ± 0.017
SVM* + Platt (Weighted)	0.040 ± 0.013	0.038 ± 0.012	0.021 ± 0.009	0.035 ± 0.017

Table 2: Mean ECE scores. The values after $\pm$ indicate the standard deviations. (SVM* results were obtained after only 10 iterations, instead of the default 10^3).



(a) LR class weights

(b) RF class weights

Figure 13: Reliability diagram and histogram plot with the distribution of probabilities of the classifier. The top diagram shows the calibration lines for the respective model (blue), and the results from the two Isotonic (orange) and Platt's (green) calibration methods.

Mean log-loss and BS scores are both presented in tables 3 and 4 respectively. Model miscalibration for class weighted versions of LR and RF model is well captured by both evaluation metrics.

Finally, the impact of the calibration methods over AUC scores is presented in table 5. For all cases were the baseline model was already calibrated (LR, RF and GBDT), as well as in models were miscalibration was present at some level (LR + class weights, RF + class weights, and marginally in ADA), results indicate that the calibration use does not have impact over the AUC scores.

	Mean Log-loss			
Model	20% Train	50% Train	70% Train	80% Train
LR	0.389 ± 0.018	0.368 ± 0.026	0.361 ± 0.034	0.360 ± 0.044
LR + Isotonic	0.409 ± 0.081	0.378 ± 0.039	0.376 ± 0.049	0.372 ± 0.055
LR + Platt	0.364 ± 0.012	0.361 ± 0.023	0.362 ± 0.034	0.362 ± 0.044
LR(Weighted)	0.681 ± 0.043	0.666 ± 0.023	0.665 ± 0.024	0.665 ± 0.026
LR(Weighted)+Isotonic	0.399 ± 0.069	0.367 ± 0.032	0.360 ± 0.041	0.358 ± 0.048
LR(Weighted)+Platt	0.363 ± 0.013	0.356 ± 0.021	0.353 ± 0.032	0.352 ± 0.043
RF	0.358 ± 0.015	0.351 ± 0.022	0.348 ± 0.031	0.348 ± 0.041
RF + Isotonic	0.397 ± 0.070	0.366 ± 0.034	0.362 ± 0.044	0.361 ± 0.052
RF + Platt	0.359 ± 0.014	0.352 ± 0.021	0.349 ± 0.031	0.349 ± 0.042
RF(Weighted)	0.588 ± 0.018	0.624 ± 0.016	0.635 ± 0.018	0.640 ± 0.019
RF + Isotonic(Weighted)	0.410 ± 0.106	0.377 ± 0.052	0.363 ± 0.046	0.352 ± 0.048
RF + Platt (Weighted)	0.356 ± 0.013	0.356 ± 0.022	0.346 ± 0.030	0.344 ± 0.041
GBDT	0.354 ± 0.012	0.352 ± 0.022	0.351 ± 0.032	0.349 ± 0.042
GBDT + Isotonic	0.386 ± 0.063	0.361 ± 0.032	0.356 ± 0.038	0.352 ± 0.047
GBDT + Platt	0.357 ± 0.013	0.352 ± 0.022	0.350 ± 0.033	0.349 ± 0.042
ADA	0.382 ± 0.007	0.384 ± 0.009	0.383 ± 0.017	0.384 ± 0.026
ADA + Isotonic	0.366 ± 0.035	0.355 ± 0.024	0.352 ± 0.032	0.352 ± 0.045
ADA + Platt	0.358 ± 0.012	0.354 ± 0.023	0.351 ± 0.031	0.351 ± 0.044

Table 3: Mean Log-loss scores. The values after $\pm$ indicate the standard deviations.

	Mean Brier			
Model	20% Train	50% Train	70% Train	80% Train
LR	0.109 ± 0.004	0.104 ± 0.009	0.103 ± 0.012	0.103 ± 0.016
LR + Isotonic	0.107 ± 0.004	0.105 ± 0.009	0.105 ± 0.013	0.104 ± 0.017
LR + Platt	0.104 ± 0.004	0.103 ± 0.008	0.104 ± 0.013	0.104 ± 0.017
LR(Weighted)	0.235 ± 0.016	0.233 ± 0.010	0.233 ± 0.010	0.233 ± 0.012
LR(Weighted)+Isotonic	0.106 ± 0.005	0.103 ± 0.008	0.102 ± 0.012	0.102 ± 0.015
LR(Weighted)+Platt	0.104 ± 0.004	0.102 ± 0.008	0.101 ± 0.012	0.101 ± 0.016
RF	0.103 ± 0.004	0.101 ± 0.008	0.100 ± 0.011	0.101 ± 0.015
RF + Isotonic	0.105 ± 0.005	0.103 ± 0.008	0.101 ± 0.011	0.102 ± 0.015
RF + Platt	0.103 ± 0.004	0.102 ± 0.008	0.101 ± 0.012	0.101 ± 0.015
RF(Weighted)	0.200 ± 0.009	0.217 ± 0.008	0.221 ± 0.009	0.224 ± 0.009
RF + Isotonic(Weighted)	0.104 ± 0.005	0.104 ± 0.008	0.101 ± 0.011	0.100 ± 0.015
RF + Platt (Weighted)	0.102 ± 0.004	0.103 ± 0.008	0.100 ± 0.011	0.099 ± 0.015
GBDT	0.102 ± 0.004	0.101 ± 0.008	0.101 ± 0.012	0.101 ± 0.016
GBDT + Isotonic	0.104 ± 0.005	0.102 ± 0.008	0.101 ± 0.012	0.101 ± 0.016
GBDT + Platt	0.102 ± 0.004	0.101 ± 0.008	0.101 ± 0.012	0.100 ± 0.016
ADA	0.110 ± 0.002	0.110 ± 0.004	0.110 ± 0.008	0.110 ± 0.012
ADA + Isotonic	0.104 ± 0.004	0.102 ± 0.008	0.101 ± 0.012	0.101 ± 0.016
ADA + Platt	0.103 ± 0.004	0.102 ± 0.008	0.101 ± 0.012	0.101 ± 0.016

Table 4: Mean Brier scores. The values after $\pm$ indicate the standard deviations.

Model	Mean AUC (80% Train)
LR	0.575 ± 0.055
LR + Isotonic	0.550 ± 0.059
LR + Platt	0.553 ± 0.066
LR(Weighted)	0.612 ± 0.054
LR(Weighted) + Isotonic	0.616 ± 0.053
LR(Weighted) + Platt	0.614 ± 0.053
RF	0.655 ± 0.046
RF + Isotonic	0.652 ± 0.047
RF + Platt	0.656 ± 0.046
RF(Weighted)	0.644 ± 0.049
RF(Weighted) + Isotonic	0.643 ± 0.049
RF(Weighted) + Platt	0.646 ± 0.049
GBDT	0.637 ± 0.052
GBDT + Isotonic	0.638 ± 0.052
GBDT + Platt	0.637 ± 0.052
ADA	0.628 ± 0.053
ADA + Isotonic	0.637 ± 0.053
ADA + Platt	0.637 ± 0.053

Table 5: AUC mean scores of the different ML algorithms before and after calibration. The values after $\pm$ indicate the standard deviations.

5 Discussion and limitations

The principal aim of this research was to study whether we can predict disability progression by using MEP latency measures from a cohort of MS patients, while ensuring that the predicted probabilities are well calibrated. In order to study the predictive performance, we make use of different ML algorithms. In addition, correct model calibration of the predictions was evaluated, and corrected with the use of two different calibration techniques.

One of the challenges we had in this research was to deal with the disproportion of positive outcomes, that is, the low occurrence of positive cases of disability progression in the data. Class imbalance is a common issue in classification problems. Multiple workarounds are possible, but finally two solutions were considered to compensate for the class imbalance. Oversampling, or more concretely, minority class oversampling, was first implemented. Unfortunately, this approach had to be abandoned due some limitations of the software package implementation, and later difficulties with the type of cross validation approach that was used in the study: it required a cumbersome coding implementation in order to correctly split patients between train and test set groups during the iteration process. We opted for a second approach of using balanced class weights, which could be easily implemented. Promising results were obtained in the first phases of the research, an improvement of the AUC scores between 6% and 19% of the first implemented model (LR) was achieved.

While we were limited by the relative small dataset size and the number of features to work with, we tried to maximize the model performance and reduce the risk of overfitting by doing hyperparameter optimization for all the models considered in this research. Hyperparameter search was performed within a small set of hyperparameters (to minimized the effects of information leakage), but exhaustively within hyperparameter space. For example, tuning for RF was performed for five hyperparameters, but allowing for a large search range to be able to find an optimal global minimum, which then was combined with the cross-validation procedure, giving a total of 400 model fits. This process required long execution, so it was decided to keep it outside the main process pipeline (see Fig. 3.6). Given the nature of the research question (MEP study), and the limited amount of features, we did not perform feature selection.

Model scores before calibration are presented in Fig. 1. We see that overall best scores are obtained when learning phase is performed with training sizes are between 70% and 80%. RF model ranks as the highest from the implemented algorithms (AUC 0.65 ± 0.05), which is commonly considered as a poor score, but somewhat decent for a model containing only MEP latency features, EDSS at T_0 and age. RF has already been used in combination with multiple time series MEP measures in a previous study [31], with good performance results (AUC 0.75 ± 0.07). This difference ($\Delta AUC = 15\%$) in model performance between

the two studies, could be attributed to the lower number of features available in this study. RF AUC scores are closely followed by its weighted version, the other two ensemble alternatives (GBDT and ADA models), and by the class weighted LR model. On the opposite side, unweighted LR gave the worst performance results, showing that the use of the class weights could be an interesting solution to explore for unbalanced datasets.

Next, model calibration was studied. Mean ECE, log-loss, and BS values are reported in tables 2, 3 and 4 respectively. From these tables, we see that most (but not all) of the algorithms that are used in this study show a relatively good level of standard calibration. Models such as RF and GBDT returned well calibrated probabilities (ECE 3%). Showing that decision tree based classifiers return well calibrated estimations. LR, with the use of Lasso penalty, also returns slightly higher, but still good calibrated results (ECE 5.6%). Here, the use of Platt scaling is able to decrease the calibration error to the same level of the ensemble family models (ECE 3.1%). Most notably, the use of class weights had a negative impact over the calibrated probabilities. The introduction of balanced weights in LR and RF models brings as a result an overestimation of their predicted probabilities, causing a shift in their calibration curves, as seem in Fig. 13a and 13b.

From table 5, we observe that the implementation of the calibration methods has no, or little impact over the AUC scores. Using Platt and Isotonic calibration improves their miscalibration errors, while still keeping the AUC scores.

Finally, the use of the non-linear SVM proved to be infeasible under the selected working environment. We used *Google Colaboratory (Colab)* platform to build ML models, using python and scikit-learn. Colab is a free cloud service that allows implementation of ML models and cooperation between people. But because it is a free platform, it comes with its limitations, for computer intensive processes, execution times become lengthy (+12 hours), and frequent disconnections were common. The issues made that most of the SVM simulations were performed with a smaller number of iterations (10 instead of the standard 10^3 iterations). Even with the smaller number of iterations, some of the executions did not return any results in a reasonable time. In those cases, SVM results are not displayed in the tables or in plots.

6 Conclusion

In this study we have attempted to find models that can predict whether a MS patient is getting worse or not after two years. The prediction task was performed with latency data from MEP measures, and ML techniques were applied.

The modelling results showed that the RF gave the best results (AUC 0.65). Performing slightly better, although not significantly, than other competitor algorithms. Calibration results also determined that the RF predicted probabilities had a good level of calibration. In addition, we determine that the use of calibration can correct deviations in the model predicted probabilities, while it has no major impact on model performance.

Code for this thesis can be found in the following GitHub repository: `https://github.com/yacoan81/motorEvoked.git`.

7 Future Research

In this research, several techniques were explored to deal with the class imbalance. Simple oversampling of the minority class was investigated, but abandoned due to implementation issues and limited time. Several interesting avenues for future research were left open. Once such topic is the use of alternative oversampling techniques, such as Synthetic Minority Oversampling Technique (SMOTE) or Adaptive Synthetic Sampling (ADASYN). Those are common algorithm techniques for imbalanced classification that creates synthetic instances from the minority class, instead of just duplicating existing ones.

Due the lack of time, different classification techniques that take the longitudinal characteristic of the dataset were not investigated. In this research, only the latency difference between two consecutive visits was taken into consideration. It would useful to use either feature engineering to create extra variables that contains patient evolution over time, or most preferable, the use of ML methods that take the patient repeated measures into consideration.

Finally, this research focused in the use of only two calibration techniques. There exist a wide range of different calibration solutions to be studied. Specially the use of calibration based on assignment value techniques, such as Beta distribution calibration. This method has been proven to give even better results than both Isotonic and Platt scaling [11].

References

[1] Raymond D Adams and Charles S Kubik. "The morbid anatomy of the demyelinative diseases". In: *The American journal of medicine* 12.5 (1952), pp. 510–546.

[2] Peter C Austin, Jr Frank E Harrell, and Ewout W Steyerberg. "Predictive performance of machine and statistical learning methods: Impact of data-generating processes on external validity in the "large N, small p" setting". In: *Statistical Methods in Medical Research* 0.0 (0). PMID: 33848231, p. 09622802211002867. DOI: 10.1177/09622802211002867. eprint: https://doi.org/10.1177/09622802211002867. URL: https://doi.org/10.1177/09622802211002867.

[3] Antonio Bella et al. *Calibration of Machine Learning Models.*

[4] Riccardo Benedetti. "Scoring Rules for Forecast Verification". In: *Monthly Weather Review* 138.1 (2010), pp. 203–211. DOI: 10.1175/2009MWR2945.1. URL: https://journals.ametsoc.org/view/journals/mwre/138/1/2009mwr2945.1.xml.

[5] Walter Bouwmeester et al. "Reporting and Methods in Clinical Prediction Research: A Systematic Review". In: 9.5 (2012).

[6] GLENN W. BRIER. "VERIFICATION OF FORECASTS EXPRESSED IN TERMS OF PROBABILITY". In: *Monthly Weather Review* 78.1 (1950), pp. 1–3. DOI: 10.1175/1520-0493(1950)078<0001:vofeit>2.0.co;2. URL: https://app.dimensions.ai/details/publication/pub.1063450417.

[7] Róbert Busa-Fekete et al. "Ranking by calibrated AdaBoost". In: *Proceedings of the Learning to Rank Challenge*. Ed. by Olivier Chapelle, Yi Chang, and Tie-Yan Liu. Vol. 14. Proceedings of Machine Learning Research. Haifa, Israel: PMLR, 25 Jun 2011, pp. 37–48. URL: http://proceedings.mlr.press/v14/busa-fekete11a.html.

[8] Junghwan Cho et al. "How much data is needed to train a medical image deep learning system to achieve necessary high accuracy?" In: *arXiv.org* (2015). URL: https://arxiv.org/pdf/1511.06348.pdf.

[9] Janez Demšar. "Statistical Comparisons of Classifiers over Multiple Data Sets". In: *Journal of Machine Learning Research* 7.1 (2006), pp. 1–30. URL: http://jmlr.org/papers/v7/demsar06a.html.

[10] M. Galar et al. "A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches". In: *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 42 (2012), pp. 463–484.

[11] Martin Gebel. "Multivariate calibration of classifier scores into the probability space". In: 2009.

[12] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Second. Springer, 2009.

[13] Tomas Kalincik et al. “Defining reliable disability outcomes in multiple sclerosis”. In: *Brain* 138.11 (Sept. 2015), pp. 3287–3298. ISSN: 0006-8950. DOI: 10.1093/brain/awv258. eprint: https://academic.oup.com/brain/article-pdf/138/11/3287/13798678/awv258.pdf. URL: https://doi.org/10.1093/brain/awv258.

[14] Meelis Kull, Telmo M. Silva Filho, and Peter Flach. “Beyond sigmoids: How to obtain well-calibrated probabilities from binary classifiers with beta calibration”. In: *Electronic Journal of Statistics* 11.2 (2017), pp. 5052–5080. DOI: 10.1214/17-EJS1338SI. URL: https://doi.org/10.1214/17-EJS1338SI.

[15] Meelis Kull and Peter A Flach. “Novel Decompositions of Proper Scoring Rules for Classification: Score Adjustment as Precursor to Calibration”. English. In: *Machine Learning and Knowledge Discovery in Databases*. Ed. by Annalisa Appice et al. Vol. 1. Lecture Notes in Computer Science. European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases - ECML PKDD PhD Consortium ; Conference date: 07-09-2015 Through 11-09-2015. Switzerland: Springer International Publishing AG, 2015, pp. 68–85. ISBN: 9783319235271. DOI: 10.1007/978-3-319-23528-8_5.

[16] J F Kurtzke. “Rating neurologic impairment in multiple sclerosis: an expanded disability status scale (EDSS).” In: *Neurology* 33.11 (Nov. 1983), pp. 1444–52.

[17] Marije Lamain-de Ruiter et al. “External validation of prognostic models to predict risk of gestational diabetes mellitus in one Dutch cohort, prospective multicentre cohort study”. In: *BMJ* 354 (2016). DOI: 10.1136/bmj.i4338. eprint: https://www.bmj.com/content/354/bmj.i4338.full.pdf. URL: https://www.bmj.com/content/354/bmj.i4338.

[18] Frédéric London, Souraya El Sankari, and Vincent van Pesch. “Early disturbances in multimodal evoked potentials as a prognostic factor for long-term disability in relapsing-remitting multiple sclerosis patients”. In: *Clinical neurophysiology : official journal of the International Federation of Clinical Neurophysiology* 128.4 (Apr. 2017), pp. 561–569. ISSN: 1388-2457. DOI: 10.1016/j.clinph.2016.12.029. URL: https://doi.org/10.1016/j.clinph.2016.12.029.

[19] Sofus A. Macskassy, Foster Provost, and Saharon Rosset. “ROC Confidence Bands: An Empirical Evaluation”. In: *Proceedings of the 22nd International Conference on Machine Learning*. ICML ’05. Bonn, Germany: Association for Computing Machinery, 2005, pp. 537–544. ISBN: 1595931805. DOI: 10.1145/1102351.1102419. URL: https://doi.org/10.1145/1102351.1102419.

[20] Aditya Menon et al. “Predicting accurate probabilities with a ranking loss”. In: *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*. Ed. by John Langford and Joelle Pineau. ICML ’12. Edinburgh, Scotland, GB: Omnipress, July 2012, pp. 703–710. ISBN: 978-1-4503-1285-1.

[21] Sayan Mukherjee et al. “Estimating dataset size requirements for classifying DNA microarray data.” In: *Journal of computational biology : a journal of computational molecular cell biology* 10.2 (2003), pp. 119–42.

[22] Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. “Obtaining Well Calibrated Probabilities Using Bayesian Binning”. In: *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*. AAAI’15. Austin, Texas: AAAI Press, 2015, pp. 2901–2907. ISBN: 0262511290.

[23] Alexandru Niculescu-mizil and Rich Caruana. “Predicting Good Probabilities with Supervised Learning”. In: *In Proc. Int. Conf. on Machine Learning (ICML*. 2005, pp. 625–632.

[24] John H. Noseworthy et al. “Multiple sclerosis”. English (US). In: *New England Journal of Medicine* 343.13 (Sept. 2000). Copyright: Copyright 2007 Elsevier B.V., All rights reserved., pp. 938–952. ISSN: 1533-4406. DOI: 10.1056/NEJM200009283431307.

[25] M R Nuwer et al. “Evoked potentials predict the clinical changes in a multiple sclerosis drug study.” In: *Neurology* 37.11 (Nov. 1987), pp. 1754–61.

[26] Sean J Pittock et al. “Clinical implications of benign multiple sclerosis: a 20-year population-based follow-up study”. In: *Annals of neurology* 56.2 (Aug. 2004), pp. 303–306. ISSN: 0364-5134. DOI: 10.1002/ana.20197. URL: https://doi.org/10.1002/ana.20197.

[27] John Platt et al. “Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods”. In: *Advances in large margin classifiers* 10.3 (1999), pp. 61–74.

[28] Chen Sun et al. *Revisiting Unreasonable Effectiveness of Data in Deep Learning Era*. 2017. arXiv: 1707.02968 [cs.CV].

[29] Ben Van Calster et al. “A calibration hierarchy for risk models was defined: from utopia to empirical data”. In: *Journal of clinical epidemiology* 74 (June 2016), pp. 167–176. ISSN: 0895-4356. DOI: 10.1016/j.jclinepi.2015.12.005. URL: https://doi.org/10.1016/j.jclinepi.2015.12.005.

[30] Ben Van Calster et al. “Calibration: The Achilles heel of predictive analytics”. In: *BMC Medicine* 17 (Dec. 2019). DOI: 10.1186/s12916-019-1466-7.

[31] Jan Yperman et al. “Machine Learning Analysis of Motor Evoked Potential Time Series to Predict Disability Progression in Multiple Sclerosis”. In: *bioRxiv* (2019). DOI: 10.1101/772996. eprint: https://www.biorxiv.org/content/early/2019/09/19/772996.full.pdf. URL: https://www.biorxiv.org/content/early/2019/09/19/772996.

[32] Bianca Zadrozny and Charles Elkan. “Transforming Classifier Scores into Accurate Multiclass Probability Estimates”. In: *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD ’02. Edmonton, Alberta, Canada: Association for Computing Machinery, 2002, pp. 694–699. ISBN: 158113567X. DOI: 10.1145/775047.775151. URL: https://doi.org/10.1145/775047.775151.

[33] Xiangxin Zhu et al. “Do we need more training data or better models for object detection”. In: *In BMVC*. 2012.