Imputation Techiques for Clinical Data of Ischemic Stroke Patients

Filipa Matos Marques
filipa.m.marques@ist.utl.pt

Instituto Superior Técnico, Lisboa, Portugal

December 2020

Abstract

In the 21st century, every year, approximately 880 thousand people living in Europe suffer an ischemic stroke. Predicting the patient's outcome is key to choosing the course of treatment. In this master thesis, it was predicted the functional outcome, by the binary version, of the modified Rankin Scale at two points in time: three months and one year after the stroke took place. Often, data provided by health organisations to conduct these studies is incomplete which can impair the results. Thus the need arises to choose a proper way to handle the missing data. Here missing values were imputed with six different methods and the classifiers were then trained with seven distinct machine learning models. It was shown the area under the receiver operating characteristic curve for the best classifiers, at the three months and one-year marks, are 0.8217 and 0.7537, respectively. Moreover, it was not found a statistically significant difference between the performance of the distinct imputation methods for each machine learning model.

Keywords: Ischemic Stroke, Missing Data, Imputation Techniques, Machine Learning

1. Introduction

A cerebrovascular accident (CVA), or stroke, results from ischemia caused by thrombosis, malformation, stenosis, or a haemorrhage from a ruptured aneurysm [1]. Roughly, 1.1 million people living in Europe suffer a stroke yearly in the 21st century, ischemic strokes accounting for approximately 80% of cases and this number is expected to rise to 1.5 million because of the ageing population [2, 3]. In the year 2013, 11.8% of all deaths were attributed to stroke making it the second main cause of death in the world (half of these were from ischemic strokes). Furthermore, in 2013, a CVA is also the third most common cause of disability (4.5%) being responsible for 113 million disability-adjusted life-years globally [4].

The course of treatment is highly dependable on the predicted outcome of the patient meaning that any tool created to help predict the patients' functional outcome are immensely useful. Moreover, it is common for both the patient and the family to ask for a long term prognosis which is an answer that is neither immediate nor straightforward [5].

Over the last decade, the medical community has been searching for the best scores to predict the patients' functional outcome using data available at admission, making it possible to have a more informed treatment decision. Among them, the Acute Stroke Registry and Analysis of Lausanne (ASTRAL) [6], the DRAGON [7] and the Totaled HealthRisks in Vascular Events (THRIVE) [8] scores.

Currently, the modified Rankin Scale (mRS) is the gold standard used scale for "measuring the degree of disability or dependence in the daily activities of people who have suffered a stroke or other causes of neurological disability" [9, 10]. The scale goes as follows (the physician should choose the best fit of the patients' ability) [10]: i Score 0: No symptoms. ii Score 1: No significant disability. Able to carry out all usual activities, despite some symptoms. iii Score 2: Slight disability. Able to look after own affairs without assistance, but unable to carry out all previous activities. iv Score 3: Moderate disability. Requires some help, but able to walk unassisted. v Score 4: Moderately severe disability. Unable to attend to own bodily needs without assistance, and unable to walk unassisted. vi Score 5: Severe disability. Requires constant nursing care and attention, bedridden, incontinent. vii Score 6: Dead.

Some of this scale's major strengths are: i) it covers the full range of functional outcomes, from no symptoms to death [11] ii) its categorization is intuitive and easily understood by clinicians and patients [11] iii) its concurrent validity is demonstrated by strong correlation with measures of stroke pathology and agreement with other stroke scales [12]. The main criticisms from the scientific community to mRS has been its subjectivity when determining between categories and its reproducibility by examiners and patients [12].

In order to predict the mRS artificial intelligence (AI) was used. The term AI was used for the first time in a conference in 1956 at Dartmouth [13]. Nowadays, AI is comprised, for example, of machine learning (ML) methods able to identify patterns and account for complex interactions within the data [14].

The amount of large-scale annotated clinical data is increasing due to the adoption of electronic health record (EHR) systems, ML methods are also getting better every year and are readily available in opensource packages. These along with the rapidly growing computational power and cloud storage have contributed to the current growth in AI which, in turn, is expected to alter the landscape of medical practice in the close future [15, 16].

Nowadays, AI systems have already specialist-level performance in a wide range of medical tasks [17, 18]. Furthermore, they also allow physicians to be in contact with areas they haven't been able to before, as AI enables remote healthcare services for rural and low-income zones [19].

Missing data is a problem present in EHR and, given that a lot of ML models only work on complete datasets, it is a problem that needs to be dealt with either by deleting incomplete observations or by imputing it, *i.e.*, replacing any values that are missing with a value estimated by the remaining available information [20].

Here we use machine learning techniques to predict the functional outcome of a patient using the binary version of the modified Rankin Scale: good outcome for scores 0 to 2 and poor outcome for scores 3 to 6. This is done at two points in time: three months and one year after the initial stroke. In addiction, we are also interested in studying the impact of missing data and the choice of the data imputation technique in machine learning models.

We start by performing the imputation using six distinct approaches: mode/median according to the quantitative/qualitative nature of the variables, mode/median according to the quantitative/qualitative nature of the variables and taking into account the dependence of a few variables, hotdeck, k-nearest neighbours, decision trees and multiple imputation with posterior predictive distribution/conditional mean imputation once again according to the quantitative/qualitative nature of the variables. We then compare the impact of the different imputation methods in each machine learning model (L1 regression, Support Vector Machines, Random Forest, Xgboost, Neural Networks, Classification And Regression Trees and k-Nearest Neighbours).

To our knowledge, similar work has only been done by *Woźnica et al.* [21] who analysed different imputation methods for a collection of datasets and a collection of machine learning algorithms. Similarly, *Jadhav et al.* [22] and *Kyureghian et al.* [23] have evaluated some of the existing imputation techniques, yet not in the same way. They focused on the quality of imputed data, by assessing the accuracy of predicting the missing values to fully known simulated data.

2. Background

2.1. Missing Data Mechanisms

It is important to understand the mechanisms by which the data is missing before addressing the issue of imputation since it will have an impact on some of the assumptions made. *Little et al.* [24, 25] formulated three possible missing data mechanisms taking into account the relation between the missing (unobserved) and the available (observed) data.

The three possible missing data mechanisms are then defined as follows [20]:

i Missing Completely at Random (MCAR): the probability of an observation being missing depends only on itself. MCAR is the highest level of random given that the missingness does not depend on any information in the dataset. In a medical setting, this might correspond to a doctor forgetting to record the gender of every seventh patient that comes in the emergency room - there is no hidden mechanism related to any variable and it does not depend on any characteristic of the patients.

ii Missing at Random (MAR): the probability of a value being missing is related only to the observable information, *i.e.*, some statistical relationship exists between the observed and the missing variables meaning the missing data may be traceable from the observed values in the dataset. As a medical example, let's assume elderly people are less probable to notify the physician they have had pneumonia before, the response rate of the variable "pneumonia" will be correlated to the variable "age".

iii Not Missing at Random (NMAR): the probability of a value being missing depends both on missing and observed values. It refers to the case when neither MCAR nor MAR holds, the pattern of missing data is not random and non predictable from available values.

NMAR is usually regarded as the worst type as it might lead to bias whereas MCAR and MAR may lead to loss of statistical power [26, 27]. Determining the missing mechanism is usually impossible, as it depends on unseen data. A t-test comparing the characteristics of the groups' missing values and observed values on a certain variable will yield different characteristics if the data is not MCAR yet the result is merely indicative since it always depends on the sample size of the data. Additionally, there is no method for distinguishing between MAR or NMAR data [25]. Given this impossibility we must rely on sensitivity analyses and testing how the inference holds under different conditions, *e.g.* diabetic patients will have their blood sugar measured more often than non diabetic patients meaning the variable "blood sugar" depends on the variable "diabetic" [22].

2.2. Handling Missing Values

The goal of the various imputation methods is the accurate estimation of population parameters in order to keep the power of the following data analysis and data mining techniques. There is no rule as to what method should be chosen to handle the missing values of a given dataset yet there is a common agreement that imputation should be used with care in datasets with over 25% of the data missing [22].

The easiest way to handle missing data is to omit the observations or cases that have missing values. Although this is often the standard method, it reduces the dataset. Therefore should only be used when a small amount of missing values is present [22]. Moreover, usually, deletion methods lead to valid inferences only for MCAR data [28]. There are two general approaches:

- Complete-Case Analysis / Listwise Deletion: observations with one or more missing values are discarded. It is assumed that the sample is representative of the whole population meaning the analysis will not be biased towards a subgroup.
- Available-Case Analysis/ Pairwise Deletion: observations with one or more missing values are only discarded if they are being analysed. Consequently, sample sizes will be different making it impossible to make a statistical comparison of the results [20].

Another way is to use single imputation which fills missing values with a predicted value, all the while ignoring uncertainty, resulting often in the underestimation of variance [20]. Similarly to the deletion meth-

ods, several approaches can be taken. Below is a non-comprehensive list:

- Imputation with a constant: the missing values are replaced with a constant. When dealing with a categorical variable one might replace it with "Missing" or a value of no significance, *e.g.*, "999".
- Mode, Mean and Median Imputation: the categorical and numerical missing values are replaced by the variables mode or mean/median, respectively. The mean should only be used for populations which have a normal distribution, otherwise the median should be used [20]. The latter is also more robust to outliers. There are disadvantages [29]: i) The new variance understates the true variance. ii) The new distribution has more values under the category containing mean/median/mode than the true population. iii) The correlations between variables are diminished.

  A special case can be used, conditional mode/mean/median. Here a variable is grouped according to a second variable and the mode/mean/median is computed for every unique value of the second variable. It might be useful when a known relation exists.
- Hot Deck Imputation: the missing values are replaced with a value from the known data's estimated distribution. The implementation is done in two steps, first the data is grouped in clusters and each missing value is attributed to a cluster. Then a distribution for the variable with the missing values is created for each cluster and the missing value is filled. This simple approach allows for the variable distribution's preservation however it underestimates the variability [30]. Moreover, the definition of "similar" for the creation of clusters is not straightforward, several metrics can be used which will result in different imputations [29].
- Model-Based Imputation: the missing values are replaced by values estimated by a predictive model. The complete data will be used to create a model, *e.g.*, regression, logistic regression, neural networks or other (non) parametric modelling techniques. Due to its characteristics, the model won't have high accuracy when the data is MCAR. When rightly applied, its estimated values are usually more well-behaved than the true values [20].
- Regression Imputation: the missing values are replaced by values estimated by a regression model (a particular case of a Model-Based Imputation). This imputation method, like the Hot Deck, is able to preserve the distribution shape however it might produce biased results when applied to NMAR and MAR data.

  There are disadvantages since it does not take into account the uncertainty in the missing data [29]: i) It assumes the estimated variable correlates with the remaining variables in the dataset. ii) It reinforces relationships already existent in the dataset reducing its generalization capability. iii) It understates the distribution's variance. iv) The estimated value is not constrained and may consequently be outside predetermined boundaries for set variable thus requiring additional adjustment.
- k-Nearest Neighbours Imputation: the missing values are replaced by the mean of the k values coming from the k most similar complete observations. There are several ways to compute this similarity (distance functions, *e.g.*, Euclidean, Manhattan, Mahalanobis, Pearson, etc) notwithstanding it is very time consuming for a large dataset. Moreover, the value given to k should be thoroughly investigated, the value should be large enough to encompass all significant attributes yet not as large that would include attributes which significantly differ from our target observation [20]. Its main advantage is the fact that the correlation structure of the data is taken into consideration. Additionally, it can handle both discrete and continuous variables [20].

2.2.1 Multiple Imputation

Single imputation tends to underestimate the variance and ignores uncertainty [20] while multiple imputation incorporates uncertainty into its methods [31]. *Rubin* [32] created a method which takes the average of the outcome across multiple imputed datasets. The imputation of multiple plausible values allows the model to account for uncertainty. This Monte Carlo technique consists of three steps: i Imputation: missing values are replaced, using a method of choice, M times (5–10 is generally sufficient) [28]. ii Analysis: every M completed dataset is analysed (*e.g.* it is built a logistic regression classifier for outcome prediction), resulting in M analyses [20]. iii Pooling: the M analysis and results are consolidated into one final one, *e.g.*, by computing the mean and the 95 % CI of the M analyses [20].

The above three steps make multiple imputation a very time consuming step, which is why many analyst opt not to choose set imputation.

3. Implementation

3.1. Database: Precise Stroke

The dataset used in this study results from a collaboration, between investigators from Instituto de Engenharia de Sistemas e Computadores, Investigação e Desenvolvimento in Lisbon (INESC-ID) and from the Santa Maria Hospital in Lisbon, in the project Precise. Our original dataset was comprised of 536 patients however the mRS three months and one year after the event was only recorded for 243 and 234 patients, respectively. This database has data collected at admission, follow-up data, data collected on discharge, three months and one year after the initial stroke. Before data pre-processing 93% of the dataset features had more than 30% of its data missing, 90% of the dataset features had more than 50% of its data missing and 64% of the dataset features had more than 70% of its data missing. Further research was not done given the elevated number of features in the dataset, 393 and 466 for the mRS three months and one year, respectively. For more information regarding missing data exploration in R the work by *Ghazali et al.* should be consulted [33].

3.2. Data Cleaning and Manipulation

Data cleaning was performed for both predicted outcomes in the same manner by deleting features that

contained more than 90% of missing values and features which were meta-data, *e.g.* record number. Variables that record times were converted into time differences between variables, *e.g.* time of the initial event and time of arrival at the hospital becomes time between the event and arrival. Variables consisting of a true/false list were used to create a column for each list entry. Furthermore, observations having a field "Unknown" or "Untested" were set to "NA" whereas the field "Not Applicable" was kept. Patients for which the mRS was not recorded or were dead by the time of its assessment were removed. Moreover, using the caret package [34], features with zero variance and near-zero variance (the feature must have a ratio of the most common value to the second most common value lower than 95:5) were removed as well as features with a correlation higher than 85%. These last features were removed in order to enable the use of multiple imputation. After cleaning the data, the resulting dataset had 138 features and 243 patients for the mRS three months and 192 features and 234 patients for the mRS one year.

The target variable was the mRS three months and one year after the event. To turn the problem into a binary classification problem the mRS was discretized into two classes: i Good outcome: defined by mRS $\leq$ 2 ii Poor outcome: defined by mRS $>$ 2 This particular discretization is of medical relevance because it separates the patients who will be able to live a rather normal independent life from the ones who will require significant assistance.

3.3. Data Imputation

The database Precise Stroke has a number of dependent fields, *i.e.* fields which can only be filled when a third field has a pre-determined answer, as is shown in Fig. 1: "Idade" can only be filled when the previous field's ("Hipertensão Arterial" or "Diabetes Mellitus") answer is "Sim". In this cases primal data imputation was performed, single value imputation was used with the value "9999" or "0" depending on the variable's quantitative or qualitative nature, respectively.

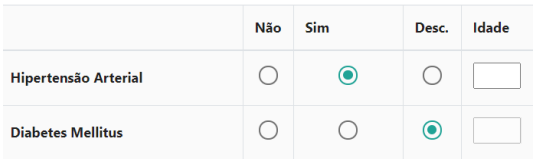

Figure 1: Example of two dependent fields on the database Precise Stroke.

After data cleaning, manipulation and a primal data imputation, 6 experiments were designed that aimed to assess with what precision we could predict the patient's mRS three months and the mRS one year after admission. Each experiment corresponds to a different imputation method:

- Mode/Median Imputation (Imp. 1), according to the quantitative/qualitative nature of the variables (imputeTS package in R [35]).
- Mode/Median Imputation (Imp. 2), according to the quantitative/qualitative nature of the variable and taking into account the dependence of a few variables (imputeTS package in R for each pair dependent/independent variable [35]).
- Hotdeck Imputation (Imp. 3) (VIM package in R [36]).
- k-Nearest Neighbours Imputation (Imp. 4). Done by using the default settings (k=5 and a variation of the Gower Distance) of the VIM package in R [36].
- Decision Trees Imputation (Imp. 5) (missForest package in R [37, 38]).
- Multiple Imputation (Imp. 6) with posterior predictive distribution or conditional mean-imputation, according to the quantitative or qualitative nature of the variables. Done by using the default settings (maximum number of iterations=30 and number of chains=4) of the mi package in R [39].

3.4. Machine Learning Models

For each experiment we used the following classifiers:

- Logistic Regression L1-regularised; caret method "regLogistic" from R package LiblineaR [40, 41].
- Support Vector Machines: caret method "svmPoly" from R package kernlab [42, 43].
- Random Forest: caret method "rf" from R package randomForest [44].
- Extreme Gradient Boosting: caret method "xgbLinear" from R package xgboost [45, 46, 47].
- Neural Network: caret method "nnet" from R package nnet [48, 49].
- Classification And Regression Trees: caret method "rpart" from R package rpart [50].
- k-Nearest Neighbours: caret method "knn" from R itself [48, 49].

3.5. Evaluation and Training

To measure the performance of the models it was used the AUC. To train and validate the model it was used 10-fold cross validation, using the caret package [34]. ROC and PR curves were created, using the mLeval R package [51], to further compare the models. In order to determine if the differences observed between the different classifiers' AUC were statistically significant the DeLong's test was applied, using the pROC R package, and a p-value threshold of 0.05 was chosen. To determine the best parameterization for each model a grid search was performed over a set of reasonable values.

4. Results

4.1. Modified Rankin Scale at Three Months

From table 1 it can be seen how there is no one better imputation method, it greatly depends on the model being used to train the classifier. *Woźnica et al.* [21] arrived at the same conclusion, more complex methods aren't always the best option. Here the combination which achieved the best results was performing hotdeck imputation and using neural networks as the classification model with an AUC of 0.8217.

In Figure 2 can be found the twenty most important variables and its relative importance (scale of 100%) for the best modified Rankin Scale classifier at three months. 13 out of the 20 variables are all known predictors that are used by traditional scores (National

Institutes of Health Stroke Scale/Score (NIHSS) [52], Hospital Anxiety and Depression Scale (HADS) [53], Mini-Mental State Examination (MMSE) [54] and the Montreal Cognitive Assessment (MoCA) [55]) also appear in the most important features list of the classifiers. The major presence of the NIHSS score comes as no surprise given its metrics measure the symptoms' severity and there is a direct correlation between the severity of the symptoms and the patient's likelihood to recover [5]. Interestingly, features related to recovery are also represented, and ranked fourth and fifth nonetheless, highlighting the importance of physical therapy and speech therapy.

In a medical context, a classifier which has an 80% sensitivity, *i.e.* is able to predict eight in every ten patients who will require significant assistance (positive class), is considered a good model [56]. Looking at table 2, the partial AUC values for an 80% sensitivity were computed and the correction by McClish was applied. The best imputation method is the same as when the total AUC is computed, table 1, as well as the best pair imputation method/classification model.

Furthermore, it was previously discussed how the metric AUC could be misleading when computed for an imbalanced dataset as a small variation in the number of correct and incorrect predictions resulted in a large change in the ROC curve and, consequently, in the AUC score providing an excessively optimistic value for performance [57]. *Fernández et al.* advise the reader to, in this situation, use the precision-recall curve and AUC-PR. The different metrics AUC and AUC-PR in tables 1 and 3, respectively, are not in agreement when electing the best imputation method and classification model pair. For AUC-PR the best imputation method would be decision trees paired with a random forest classifier. These results show how important it is the choice of the evaluation metric.

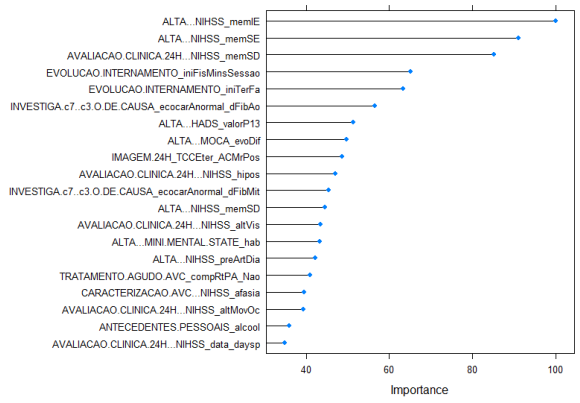

Figure 2: The twenty most important variables and its relative importance (scale of 100%) for the best model predicting the modified Rankin Scale at three months: hotdeck imputation and neural network classification.

Given that the observed performances were so close to each other, paired DeLong's tests using a p-value of 0.05 were performed to determine whether the observed differences were statistically significant. Table 4 and 5 show the results among each classification model and imputation method, respectively. For readability reasons, the p-value was omitted and a check-mark was placed instead when the difference between the classifiers are statistically significant. Models which did not show any significant difference were also omitted.

Taking a closer look at each model in table 4 it can be concluded that the great majority of imputation methods are statistically equivalent. To the best of our knowledge, this can be explained by two facts: the elevated amount of missing values (93% of the dataset features had more than 30% of its data missing) and the small number of patients used to conduct this study. Previous studies [21, 22, 23] have shown the difference between the performance of the several imputation methods is small and small sample sizes only allow large differences to be detected [58]. Moreover, it is known imputation should be used carefully in datasets with over 25% of the data missing [22], the high proportion of missing data may introduce considerable bias resulting in too similar imputations. On the other hand, in table 5, it can be concluded that the great majority of classification models are statistically different.

ROC curves for the six imputation methods and seven different models predicting the modified Rankin Scale at three months are not shown given that the curves overlap several times and there is no clear distinction between them which is expected considering the close AUC values in table 1 and the statistical tests performed, table 4.

Coherently, the precision-recall curves are zigzagged, reason why they are also not shown. It is common to have noisy curves for small recall values however when this tendency persists for higher recalls, curves for different classifiers crossing each other very frequently, it makes it hard to choose the best classifier by analyzing set curves.

4.2. Modified Rankin Scale at One Year

5. Modified Rankin Scale at One Year

Contrarily to the results for the models predicting the modified Rankin Scale at three months, from table 6 the best imputation model can be chosen, hotdeck, which yields the best results for four out of the seven models supporting the previous conclusion that more complex methods aren't always the best option [21]. Here, similarly to the results at three months, the combination which achieved the best results was performing hotdeck imputation and using neural networks as the classification model with an AUC of 0.7537.

It was expected an increase in performance for the models predicting the modified Rankin Scale at one year when compared to predicting the modified Rankin Scale at three months since stroke symptoms are maximal on onset and decrease in severity with time. Moreover, as time goes by, the patients' state is less likely to significantly change, meaning, it was expected to be easier to predict the patients' functional outcome at one year based on their state at three months than their functional outcome at three months based on the patients' state a few days after stroke (when the symptoms are more likely to vary greatly on a daily basis)

	Logistic Regression	Support Vector Machine	Random Forest	Extreme Gradient Boosting	Neural Network	CART	k-Nearest Neighbours
Imp. 1	0.7665 + 0.0218	0.7615 + 0.0219	**0.8004 + 0.0209**	0.7754 + 0.0216	0.7979 + 0.0208	0.6250 + 0.0230	0.7110 + 0.0219
Imp. 2	0.7705 + 0.0217	0.7589 + 0.0219	0.8001 + 0.0209	0.7566 + 0.0220	0.7984 + 0.0207	0.6433 + 0.0235	0.6919 + 0.0218
Imp. 3	0.7815 + 0.0214	0.7730 + 0.0216	0.7998 + 0.0209	0.7737 + 0.0216	**0.8217 + 0.0198**	0.6456 + 0.0225	0.7061 + 0.0218
Imp. 4	0.7760 + 0.0216	**0.7811 + 0.0214**	0.7916 + 0.0212	**0.7818 + 0.0214**	0.7945 + 0.0209	0.6218 + 0.0233	0.7071 + 0.0220
Imp. 5	0.7891 + 0.0212	0.7672 + 0.0217	0.7876 + 0.0213	-	0.7365 + 0.0224	**0.7087 + 0.0230**	0.7303 + 0.0220
Imp. 6	**0.8014 + 0.0209**	0.7791 + 0.0214	0.7972 + 0.0210	-	0.7768 + 0.0215	0.6768 + 0.0228	**0.7314 + 0.0220**

Table 1: AUC results for six imputation methods and seven different models predicting the modified Rankin Scale at three months. The best imputation method for each model is highlighted.

	Logistic Regression	Support Vector Machine	Random Forest	Extreme Gradient Boosting	Neural Network	CART	k-Nearest Neighbours
Imp. 1	0.8708	0.8670	**0.9114**	0.8819	0.8979	0.6819	0.8045
Imp. 2	0.8759	0.8640	0.9104	0.8590	0.8967	0.7065	0.7783
Imp. 3	0.8885	0.8805	0.9110	0.8791	**0.9151**	0.7119	0.7981
Imp. 4	0.8825	**0.8918**	0.9027	**0.8895**	0.8930	0.6769	0.7989
Imp. 5	0.9006	0.8739	0.8978	-	0.8320	**0.7948**	0.8295
Imp. 6	**0.9129**	0.8896	0.9086	-	0.8794	0.7548	**0.8311**

Table 2: Partial AUC results, 80% sensitivity, for six imputation methods and seven different models predicting the modified Rankin Scale at three months. The best imputation method for each model is highlighted.

	Logistic Regression	Support Vector Machine	Random Forest	Extreme Gradient Boosting	Neural Network	CART	k-Nearest Neighbours
Imp. 1	0.470	**0.640**	0.630	**0.650**	0.550	0.140	0.500
Imp. 2	0.440	0.630	0.630	0.600	0.550	0.100	0.520
Imp. 3	0.400	0.600	0.610	0.600	0.550	**0.400**	0.400
Imp. 4	0.480	0.600	0.600	0.600	0.530	0.130	0.480
Imp. 5	**0.600**	**0.640**	**0.660**	-	0.530	0.130	**0.600**
Imp. 6	**0.600**	0.620	0.590	-	**0.570**	0.390	**0.600**

Table 3: AUC-PR results for six imputation methods and seven different models predicting the modified Rankin Scale at three months. The best imputation method for each model is highlighted.

	Imp. 1	Imp. 2	Imp. 3	Imp. 4	Imp. 5	Imp. 6
Imp. 1						✓
Imp. 2						✓
Imp. 3						
Imp. 4						
Imp. 5						
Imp. 6						

((a)) Logistic Regression L1-regularised.

	Imp. 1	Imp. 2	Imp. 3	Imp. 4	Imp. 5	Imp. 6
Imp. 1					✓	
Imp. 2					✓	
Imp. 3					✓	✓
Imp. 4					✓	
Imp. 5						✓
Imp. 6						

((b)) Neural Network.

	Imp. 1	Imp. 2	Imp. 3	Imp. 4	Imp. 5	Imp. 6
Imp. 1					✓	✓
Imp. 2					✓	✓
Imp. 3					✓	
Imp. 4					✓	✓
Imp. 5						
Imp. 6						

((c)) Classification And Regression Trees.

	Imp. 1	Imp. 2	Imp. 3	Imp. 4	Imp. 5	Imp. 6
Imp. 1						
Imp. 2					✓	✓
Imp. 3						
Imp. 4						
Imp. 5						
Imp. 6						

((d)) k-Nearest Neighbours.

Table 4: Paired DeLong's test results for each model - prediction of modified Rankin Scale at three months. Note: The following models are not presented since there were no significant differences present: Support Vector Machine, Random Forest, Extreme Gradient Boosting.

[5]. Comparing the AUC for both three months and one year, tables 1 and 6 respectively, it is seen this is not true. A possible explanation is that the added features have a portion of missing data too big that, when completed with the different imputation techniques, add noise rather than any relevant information resulting in the worsening of the AUC.

In Figure 3 can be found the twenty most important features and its relative importance (scale of 100%) for the best modified Rankin Scale classifier at one year. The tendency seen at three months of variables corresponding to known predictors that are used by traditional scores is still present as anticipated. Only 4 of the 20 features were recorded at three months which is a lower number than expected given that, as mentioned above, the patients' situation is less likely to significantly change the more time has passed since the stroke occurred [5]. This corroborates the above hypothesis, data recorded at three months added more noise than information. Although the majority of the most important variables at one year were present at three months, the overlap between the variables at the two dates is very small.

Looking at table 7, the partial AUC values for an 80% sensitivity were computed and the correction by McClish was applied. As for the three months mark, the best imputation method is the same as when the total AUC is computed, table 6, as well as the best

	KNN	CART	NN	Xgboost	RF	SVM	LR
KNN		✓	✓	✓	✓	✓	✓
CART			✓	✓	✓	✓	✓
NN						✓	✓
Xgboost							
RF						✓	✓
SVM							
LR							

((a)) Mode/Median Imputation (Imp. 1).

	KNN	CART	NN	Xgboost	RF	SVM	LR
KNN		✓	✓	✓	✓	✓	✓
CART			✓	✓	✓	✓	✓
NN				✓		✓	
Xgboost					✓		
RF						✓	
SVM							
LR							

((b)) Mode/Median Imputation taking into account the dependence of a few variables (Imp. 2).

	KNN	CART	NN	Xgboost	RF	SVM	LR
KNN		✓	✓	✓	✓	✓	✓
CART			✓	✓	✓	✓	✓
NN				✓		✓	✓
Xgboost							
RF							
SVM							
LR							

((c)) Hotdeck Imputation (Imp. 3).

	KNN	CART	NN	Xgboost	RF	SVM	LR
KNN		✓	✓	✓	✓	✓	✓
CART			✓	✓	✓	✓	✓
NN							
Xgboost							
RF							
SVM							
LR							

((d)) K-Nearest Neighbours Imputation (Imp. 4).

	KNN	CART	NN	RF	SVM	LR
KNN				✓	✓	✓
CART				✓	✓	✓
NN				✓		✓
RF						
SVM						
LR						

((e)) Decision Trees Imputation (Imp. 5).

	KNN	CART	NN	RF	SVM	LR
KNN		✓	✓	✓	✓	✓
CART			✓	✓	✓	✓
NN						
RF						
SVM						
LR						

((f)) Multiple Imputation (Imp. 6).

Table 5: Paired DeLong's test results for each imputation method - prediction of modified Rankin Scale at three months.

pair imputation method/classification model.

Furthermore, the different metrics AUC and AUC-PR in tables 6 and 8, respectively, are again not in agreement when electing the best imputation method and classification model pair. Here, for the AUC-PR evaluation metric, there is a tie for the best pair: multiple imputation/logistic regression and mode/median imputation with dependent variables/extreme gradient boosting.

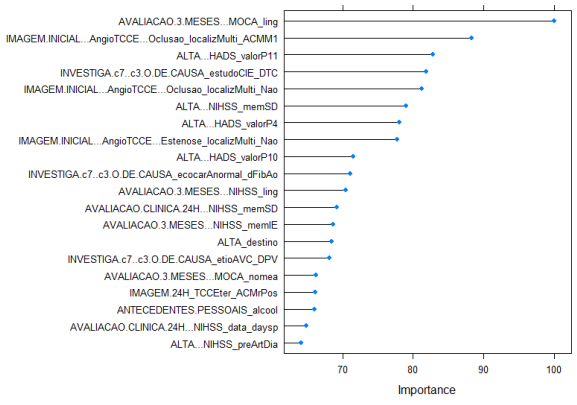

Figure 3: The twenty most important variables and its relative importance (scale of 100%) for the best model predicting the modified Rankin Scale at one year: hotdeck imputation and neural network classification.

Table 9 and 10 show the results of the paired DeLong's tests, using a p-value of 0.05, among each classification model and imputation method, respectively.

Taking a closer look at each model in table 9 it can be concluded that the great majority of imputation methods are statistically equivalent. The same happened for the results at three months and both outcomes can be explained by the facts enumerated before: the elevated amount of missing values and the small number of patients used to conduct this study. On the other hand, in table 10, contrarily to what happens for the results at three months, it is seen greater statistical equivalence between classification models. Again, a possible explanation lies in the added features: adding it, with its big portion of missing data might have resulted in the addiction of noise rather than any relevant information.

The ROC curves for the six imputation methods and seven different models predicting the modified Rankin Scale at one year are not shown given that the same conclusion can be drawn as at three months: there is no clear distinction between them, the different classifier for each classification method are equivalent which is supported by the AUC values in table 6 and the statistical tests performed, table 9.

As at the three-month timeline, the precision-recall curves are zigzagged, reason why they are also not shown.

6. Conclusions

It was possible to conclude that machine learning can indeed effectively predict the functional outcome of an ischemic stroke patient. The AUC for the three months

	Logistic Regression	Support Vector Machine	Random Forest	Extreme Gradient Boosting	Neural Network	CART	k-Nearest Neighbours
Imp. 1	0.7140 + 0.0244	0.7166 + 0.0239	0.6756 + 0.0231	0.7092 + 0.0244	0.7254 + 0.0241	0.6680 + 0.0247	0.6982 + 0.0235
Imp. 2	0.7103 + 0.0244	0.7188 + 0.0239	0.6632 + 0.0230	0.7067 + 0.0243	0.7488 + 0.0235	0.6341 + 0.0242	**0.6989 + 0.0235**
Imp. 3	**0.7295 + 0.0242**	**0.7296 + 0.0239**	0.6683 + 0.0230	**0.7320 + 0.0242**	**0.7537 + 0.0232**	0.6739 + 0.0247	0.6966 + 0.0233
Imp. 4	0.7236 + 0.0243	0.7236 + 0.0239	0.6802 + 0.0233	0.7075 + 0.0244	0.7358 + 0.0238	0.6297 + 0.0241	**0.6989 + 0.0235**
Imp. 5	0.7278 + 0.0242	0.7233 + 0.0241	**0.6826 + 0.0232**	-	0.6857 + 0.0248	0.6782 + 0.0247	0.6885 + 0.0238
Imp. 6	0.7171 + 0.0242	0.6959 + 0.0239	0.6687 + 0.0231	-	0.7146 + 0.0244	**0.6861 + 0.0245**	0.6950 + 0.0240

Table 6: AUC results for six imputation methods and seven different models predicting the modified Rankin Scale at one year. The six different imputations used: Imp. 1 - Mode/Median Imputation, Imp. 2 - Mode/Median Imputation taking into account the dependence of a few variables, Imp. 3 - Hotdeck Imputation, Imp. 4 - K-Nearest Neighbours Imputation, Imp. 5 - Decision Trees Imputation, Imp. 6 - Multiple Imputation. The best imputation method for each model is highlighted.

	Logistic Regression	Support Vector Machine	Random Forest	Extreme Gradient Boosting	Neural Network	CART	k-Nearest Neighbours
Imp. 1	0.8062	0.8117	0.7564	0.7999	0.8078	0.7427	0.7877
Imp. 2	0.8013	0.8147	0.7388	0.7970	0.8357	0.6957	**0.7888**
Imp. 3	**0.8264**	**0.8288**	0.7460	**0.8298**	**0.8374**	0.7508	0.7857
Imp. 4	0.8187	0.8211	0.7626	0.7973	0.8186	0.6894	**0.7888**
Imp. 5	0.8247	0.8198	**0.7662**	-	0.7663	0.7569	0.7735
Imp. 6	0.8108	0.7837	0.7465	-	0.8065	**0.7679**	0.7822

Table 7: Partial AUC results, 80% sensitivity, for six imputation methods and seven different models predicting the modified Rankin Scale at one year. The six different imputations used: Imp. 1 - Mode/Median Imputation, Imp. 2 - Mode/Median Imputation taking into account the dependence of a few variables, Imp. 3 - Hotdeck Imputation, Imp. 4 - K-Nearest Neighbours Imputation, Imp. 5 - Decision Trees Imputation, Imp. 6 - Multiple Imputation. The best imputation method for each model is highlighted.

	Logistic Regression	Support Vector Machine	Random Forest	Extreme Gradient Boosting	Neural Network	CART	k-Nearest Neighbours
Imp. 1	0.560	0.610	0.670	0.590	0.080	0.340	0.590
Imp. 2	0.610	0.600	0.690	**0.710**	0.100	0.340	0.580
Imp. 3	0.680	0.540	0.580	0.620	0.120	**0.430**	0.460
Imp. 4	0.630	0.630	**0.700**	0.600	0.250	0.340	**0.650**
Imp. 5	0.520	0.520	0.610	-	0.480	0.250	0.530
Imp. 6	**0.710**	**0.640**	0.670	-	**0.550**	0.380	0.370

Table 8: AUC-PR results for six imputation methods and seven different models predicting the modified Rankin Scale at one year. The six different imputations used: Imp. 1 - Mode/Median Imputation, Imp. 2 - Mode/Median Imputation taking into account the dependence of a few variables, Imp. 3 - Hotdeck Imputation, Imp. 4 - K-Nearest Neighbours Imputation, Imp. 5 - Decision Trees Imputation, Imp. 6 - Multiple Imputation. The best imputation method for each model is highlighted.

	Imp. 1	Imp. 2	Imp. 3	Imp. 4	Imp. 5	Imp. 6
Imp. 1					✓	
Imp. 2					✓	✓
Imp. 3					✓	✓
Imp. 4					✓	
Imp. 5						
Imp. 6						

((a)) Neural Network

	Imp. 1	Imp. 2	Imp. 3	Imp. 4	Imp. 5	Imp. 6
Imp. 1				✓		
Imp. 2			✓		✓	✓
Imp. 3				✓		
Imp. 4					✓	✓
Imp. 5						
Imp. 6						

((b)) Classification And Regression Trees

Table 9: Paired DeLong's test results for each model - prediction of modified Rankin Scale at one year. Note: The following models are not presented since there were no significant differences present: Logistic Regression, Support Vector Machine, Random Forest, Extreme Gradient Boosting and k-Nearest Neighbours. The six different imputations used: Imp. 1 - Mode/Median Imputation, Imp. 2 - Mode/Median Imputation taking into account the dependence of a few variables, Imp. 3 - Hotdeck Imputation, Imp. 4 - K-Nearest Neighbours Imputation, Imp. 5 - Decision Trees Imputation, Imp. 6 - Multiple Imputation.

and one-year mark are 0.8217 and 0.7537, respectively meaning more data does not necessarily imply better results. Moreover, although there is a clear distinction between classifiers trained with only different machine learning methods, the same cannot be said for classifiers trained with only different imputation methods. Finally, it was highlighted how important it is to choose the right evaluation metric according to the problem's specificity. In the future we wish to improve our performances by using more and richer records from the Precise Stroke Database. By adding more patients and with less missing data we hope to be able to answer the question: which imputation method results better for electronic health records and does this answer depend on the machine learning method being used for training.

References

[1] X. Gao, Y. Uchiyama, X. Zhou, T. Hara, T. Asano, and H. Fujita, "A fast and fully automatic method

	KNN	CART	NN	Xgboost	RF	SVM	LR
KNN							
CART			✓	✓		✓	✓
NN					✓		
Xgboost							
RF						✓	✓
SVM							
LR							

((a)) Mode/Median Imputation (Imp. 1).

	KNN	CART	NN	Xgboost	RF	SVM	LR
KNN		✓	✓		✓		
CART			✓	✓		✓	✓
NN				✓	✓		✓
Xgboost					✓		
RF						✓	✓
SVM							
LR							

((b)) Mode/Median Imputation taking into account the dependence of a few variables (Imp. 2).

	KNN	CART	NN	Xgboost	RF	SVM	LR
KNN			✓	✓			
CART			✓	✓		✓	✓
NN					✓		
Xgboost					✓		
RF						✓	✓
SVM							
LR							

((c)) Hotdeck Imputation (Imp. 3).

	KNN	CART	NN	Xgboost	RF	SVM	LR
KNN		✓	✓				
CART			✓	✓	✓	✓	✓
NN					✓		
Xgboost							
RF						✓	✓
SVM							
LR							

((d)) K-Nearest Neighbours Imputation (Imp. 4).

	KNN	CART	NN	RF	SVM	LR
KNN					✓	✓
CART					✓	✓
NN					✓	✓
RF					✓	✓
SVM						
LR						

((e)) Decision Trees Imputation (Imp. 5).

	KNN	CART	NN	RF	SVM	LR
KNN						
CART						
NN				✓		
RF						✓
SVM						
LR						

((f)) Multiple Imputation (Imp. 6).

Table 10: Paired DeLong's test results for each imputation method - prediction of modified Rankin Scale at one year. The seven different methods used: KNN - k-Nearest Neighbours, CART - Classification And Regression Trees, NN - Neural Network, Xgboost - Extreme Gradient Boosting, RF - Random Forest, SVM - Support Vector Machines, LR - Logistic Regression.

for cerebrovascular segmentation on time-of-flight (TOF) MRA image.," *Journal of digital imaging*, vol. 24, pp. 609–25, 8 2011.

[2] Y. Béjot, H. Bailly, J. Durier, and M. Giroud, "Epidemiology of stroke in Europe and trends for the 21st century," *La Presse Médicale*, vol. 45, pp. e391–e398, 12 2016.

[3] T. Truelsen, B. Piechowski-Jozwiak, R. Bonita, C. Mathers, J. Bogousslavsky, and G. Boysen, "Stroke incidence and prevalence in Europe: a review of available data," *European Journal of Neurology*, vol. 13, pp. 581–598, 6 2006.

[4] V. L. Feigin, B. Norrving, and G. A. Mensah, "Global Burden of Stroke," *Circulation Research*, vol. 120, pp. 439–448, 2 2017.

[5] M. Monteiro, A. C. Fonseca, A. T. Freitas, T. Pinho E Melo, A. P. Francisco, J. M. Ferro, and A. L. Oliveira, "Using Machine Learning to Improve the Prediction of Functional Outcome in Ischemic Stroke Patients," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 15, pp. 1953–1959, 11 2018.

[6] G. Ntaios, M. Faouzi, J. Ferrari, W. Lang, K. Vemmos, and P. Michel, "An integer-based score to predict functional outcome in acute ischemic stroke: The ASTRAL score," *Neurology*, vol. 78, pp. 1916–1922, 6 2012.

[7] D. Strbian, A. Meretoja, F. J. Ahlhelm, J. Pitkäniemi, P. Lyrer, M. Kaste, S. Engelter, and T. Tatlisumak, "Predicting outcome of IV thrombolysis - Treated ischemic stroke patients: The DRAGON score," *Neurology*, vol. 78, pp. 427–432, 2 2012.

[8] A. C. Flint, S. P. Cullen, B. S. Faigeles, and V. A. Rao, "Predicting long-term outcome after endovascular stroke treatment: The totaled health risks in vascular events score," *American Journal of Neuroradiology*, vol. 31, pp. 1192–1196, 8 2010.

[9] J. T. Wilson, A. Hareendran, M. Grant, T. Baird, U. G. Schulz, K. W. Muir, and I. Bone, "Improving the assessment of outcomes in stroke: Use of a structured interview to assign grades on the modified Rankin Scale," *Stroke*, vol. 33, pp. 2243–2246, 9 2002.

[10] J. L. Saver, B. Filip, S. Hamilton, A. Yanes, S. Craig, M. Cho, R. Conwit, and S. Starkman, "Improving the reliability of stroke disability grading in clinical trials and clinical practice: The rankin focused assessment (RFA)," *Stroke*, vol. 41, pp. 992–995, 5 2010.

[11] J. P. Broderick, O. Adeoye, and J. Elm, "Evolution of the Modified Rankin Scale and Its Use in Future Stroke Trials," 7 2017.

[12] J. K. Harrison, K. S. McArthur, and T. J. Quinn, "Assessment scales in stroke: Clinimetric and clinical considerations," 2 2013.

[13] S. J. Russell, *Artificial intelligence : a modern approach*. Prentice Hall, 2010.

[14] R. C. Deo, "Machine learning in medicine," *Circulation*, vol. 132, pp. 1920–1930, 11 2015.

[15] K. H. Yu, A. L. Beam, and I. S. Kohane, "Artificial intelligence in healthcare," 10 2018.

[16] S. Jha and E. J. Topol, "Adapting to artificial intelligence: Radiologists and pathologists as information specialists," 12 2016.

[17] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, R. Kim, R. Raman, P. C. Nelson, J. L. Mega, and D. R. Webster, "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," *JAMA - Journal of the American Medical Association*, vol. 316, pp. 2402–2410, 12 2016.

[18] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, pp. 115–118, 2 2017.

[19] A. Qayyum, J. Qadir, M. Bilal, and A. Al-Fuqaha, "Secure and Robust Machine Learning for Healthcare: A Survey," *IEEE Reviews in Biomedical Engineering*, 1 2020.

[20] C. M. Salgado, C. Azevedo, H. Proença, and S. M. Vieira, "Missing data," in *Secondary Analysis of Electronic Health Records*, pp. 143–162, Springer International Publishing, 1 2016.

[21] K. Woźnica and P. Biecek, "Does imputation matter? Benchmark for predictive models," 7 2020.

[22] A. Jadhav, D. Pramod, and K. Ramanathan, "Comparison of Performance of Data Imputation Methods for Numeric Dataset," *Applied Artificial Intelligence*, vol. 33, pp. 913–933, 8 2019.

[23] G. Kyureghian, O. Capps, and R. M. Nayga, "A missing variable imputation methodology with an empirical application," *Advances in Econometrics*, vol. 27 A, pp. 313–337, 2011.

[24] D. B. Rubin, "Inference and missing data," *Biometrika*, vol. 63, pp. 581–592, 12 1976.

[25] R. J. A. Little and D. B. Rubin, *Statistical Analysis with Missing Data*. Hoboken, NJ, USA: John Wiley & Sons, Inc., 8 2002.

[26] J. L. Schafer and J. W. Graham, "Missing data: Our view of the state of the art," *Psychological Methods*, vol. 7, no. 2, pp. 147–177, 2002.

[27] J. W. Graham, "Missing data analysis: Making it work in the real world," 1 2009.

[28] J. L. Schafer, "Multiple imputation: a primer," *Statistical Methods in Medical Research*, vol. 8, pp. 3–15, 2 1999.

[29] M. L. Brown and J. F. Kros, "Data mining and the impact of missing data," *Industrial Management and Data Systems*, vol. 103, no. 8-9, pp. 611–621, 2003.

[30] P. L. ROTH, "MISSING DATA: A CONCEPTUAL REVIEW FOR APPLIED PSYCHOLOGISTS," *Personnel Psychology*, vol. 47, pp. 537–560, 9 1994.

[31] R. J. A. LITTLE and D. B. RUBIN, "The Analysis of Social Science Data with Missing Values," *Sociological Methods & Research*, vol. 18, pp. 292–326, 11 1989.

[32] D. B. Rubin, ed., *Multiple Imputation for Nonresponse in Surveys*. Wiley Series in Probability and Statistics, Hoboken, NJ, USA: John Wiley & Sons, Inc., 6 1987.

[33] S. M. Ghazali, N. Shaadan, and Z. Idrus, "Missing data exploration in air quality data set using r-package data visualisation tools," *Bulletin of Electrical Engineering and Informatics*, vol. 9, pp. 755–763, 4 2020.

[34] M. Kuhn, "Building predictive models in R using the caret package," *Journal of Statistical Software*, vol. 28, pp. 1–26, 11 2008.

[35] S. Moritz and T. Bartz-Beielstein, "imputeTS: Time series missing value imputation in R," *R Journal*, vol. 9, pp. 207–218, 6 2017.

[36] A. Kowarik and M. Templ, "Imputation with the R package VIM," *Journal of Statistical Software*, vol. 74, pp. 1–16, 10 2016.

[37] D. J. Stekhoven and P. Buehlmann, "MissForest - non-parametric missing value imputation for mixed-type data," *Bioinformatics*, vol. 28, no. 1, pp. 112–118, 2012.

[38] D. J. Stekhoven, "missForest: Nonparametric Missing Value Imputation using Random Forest," 2013.

[39] A. Gelman and J. Hill, "Opening Windows to the Black Box," *Journal of Statistical Software*, vol. 40, 2011.

[40] K.-W. Chang, C.-J. Hsieh, and C.-J. Lin, "LIBLINEAR: A Library for Large Linear Classification Rong-En Fan Xiang-Rui Wang," tech. rep., 2008.

[41] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *Journal of Statistical Software*, vol. 33, pp. 1–22, 2 2010.

[42] H. T. Lin, C. J. Lin, and R. C. Weng, "A note on Platt's probabilistic outputs for support vector machines," *Machine Learning*, vol. 68, pp. 267–276, 10 2007.

[43] T.-F. Wu, C.-J. Lin, and R. C. Weng, "Probability Estimates for Multi-class Classification by Pairwise Coupling," tech. rep., 2004.

[44] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, 10 2001.

[45] S. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," *Advances in Neural Information Processing Systems*, vol. 2017-December, pp. 4766–4775, 5 2017.

[46] S. M. Lundberg and S.-I. Lee, "Consistent feature attribution for tree ensembles," 6 2017.

[47] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, vol. 13-17-August-2016, pp. 785–794, Association for Computing Machinery, 8 2016.

[48] B. D. Ripley, *Pattern recognition and neural networks*. Cambridge University Press, 1 2014.

[49] W. Venables and B. Ripley, *Modern Applied Statistics with S*. Springer-Verlag New York, 4 ed., 2002.

[50] A. D. Gordon, L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, "Classification and Regression Trees.," *Biometrics*, vol. 40, p. 874, 9 1984.

[51] C. R. John, "MLeval: Machine Learning Model Evaluation," 2 2020.

[52] P. Lyden, T. Brott, B. Tilley, K. M. Welch, E. J. Mascha, S. Levine, E. C. Haley, J. Grotta, and J. Marler, "Improved reliability of the NIH stroke scale using video training," *Stroke*, vol. 25, no. 11, pp. 2220–2226, 1994.

[53] R. P. Snaith, "The hospital anxiety and depression scale," *Health and Quality of Life Outcomes*, vol. 1, p. 29, 8 2003.

[54] M. F. Folstein, S. E. Folstein, and P. R. McHugh, ""Mini-mental state". A practical method for grading the cognitive state of patients for the clinician," *Journal of Psychiatric Research*, vol. 12, no. 3, pp. 189–198, 1975.

[55] Z. S. Nasreddine, N. A. Phillips, V. Bédirian, S. Charbonneau, V. Whitehead, I. Collin, J. L. Cummings, and H. Chertkow, "The Montreal Cognitive Assessment, MoCA: A brief screening tool for mild cognitive impairment," *Journal of the American Geriatrics Society*, vol. 53, pp. 695–699, 4 2005.

[56] B. A. Drozdowska, S. Singh, and T. J. Quinn, "Thinking About the Future: A Review of Prognostic Scales Used in Acute Stroke," *Frontiers in Neurology*, vol. 10, no. March, 2019.

[57] A. Fernández, S. García, M. Galar, R. C. Prati, B. Krawczyk, and F. Herrera, *Learning from Imbalanced Data Sets*. Springer International Publishing, 1 ed., 2018.

[58] P. D. Ellis, *The Essential Guide to Effect Sizes: Statistical Power, Meta-Analysis, and the Interpretation of Research Results*. 1 ed., 7 2010.