scientific reports

Check for updates

OPEN

Multiple imputation integrated to machine learning: predicting post-stroke recovery of ambulation after intensive inpatient rehabilitation

Alice Finocchi[1], Silvia Campagnini[1]✉, Andrea Mannini[1], Stefano Doronzio[1,2], Marco Baccini[1], Bahia Hakiki[1,2], Donata Bardi[1], Antonello Grippo[1,3], Claudio Macchi[1,2], Jorge Navarro Solano[4], Michela Baccini[5,6] & Francesca Cecchi[1,2,6]

Good data quality is vital for personalising plans in rehabilitation. Machine learning (ML) improves prognostics but integrating it with Multiple Imputation (MImp) for dealing missingness is an unexplored field. This work aims to provide post-stroke ambulation prognosis, integrating MImp with ML, and identify the prognostic influential factors. Stroke survivors in intensive rehabilitation were enrolled. Data on demographics, events, clinical, physiotherapy, and psycho-social assessment were collected. An independent ambulation at discharge, using the Functional Ambulation Category scale, was the outcome. After handling missingness using MImp, ML models were optimised, cross-validated, and tested. Interpretability techniques analysed predictor contributions. Pre-MImp, the dataset included 54.1% women, 79.2% ischaemic patients, median age 80.0 (interquartile range: 15.0). Post-MImp, 368 non-ambulatory patients on 10 imputed datasets were used for training, 80 for testing. The random forest (the validation best-performing algorithm) obtained 75.5% aggregated balanced accuracy on the test set. The main predictors included modified Barthel index, Fugl-Meyer assessment/motricity index, short physical performance battery, age, Charlson comorbidity index/cumulative illness rating scale, and trunk control test. This is among the first studies applying ML, together with MImp, to predict ambulation recovery in post-stroke rehabilitation. This pipeline reliably exploits the potential of incomplete datasets for healthcare prognosis, identifying relevant predictors.

Keywords Ambulation, Machine learning, Multiple imputation, Prediction, Rehabilitation, Stroke

Personalised medicine represents a new frontier in healthcare. Data-driven approaches are crucial in optimising individualised rehabilitation pathways by providing reliable, interpretable, and patient-centric predictions[1]. Moreover, there is a pressing demand for trustworthy prognostic solutions, enabling users to understand and interpret automatic decisions[2]. However, while tools for personalised treatment decisions are becoming more prevalent in healthcare, their clinical validation and impact on treatment improvement remain largely underexplored[3].

Treatment personalisation is particularly relevant in rehabilitative medicine[4], where the goal is to adapt the rehabilitation plan to the unique needs of each patient, given the evidence of its positive effects on recovery[5]. According to Kokkotis et al.[6], machine learning (ML) tools can be applied to predict long-term recovery rates from the earliest hours of hospitalisation after a stroke. This suggests ML can assist medical practitioners in deploying novel, individualised rehabilitation approaches, to enhance the quality of life for survivors and the overall quality of care. However, examples of technological tools that support personalised post-stroke rehabilitation treatments are still scarce[7].

The use of ML technologies in healthcare presents pitfalls such as prediction inaccuracy, privacy vulnerabilities, and data scarcity that can hinder the attainment of real-life comparable results[8]. A critical challenge is collecting high-dimension and high-quality data for reliable and reproducible predictions, due to limitations in sample size

[1]IRCCS Fondazione Don Carlo Gnocchi onlus, Firenze, Italy. [2]Department of Experimental and Clinical Medicine, University of Florence, Firenze, Italy. [3]Azienda Ospedaliera Universitaria Careggi (AOUC), Firenze, Italy. [4]IRCCS Fondazione Don Carlo Gnocchi onlus, Milano, Italy. [5]Department of Statistics, Computer Science, Applications, University of Florence, Firenze, Italy. [6]Michela Baccini and Francesca Cecchi contributed equally to this work. ✉email: scampagnin@ricres.org

and data quality in real-world scenarios[8]. In this context, the presence of missing data may represent a significant technical problem[9], because it can result in a loss of information, reduced sample size, bias in the results, and underestimated uncertainty[9–11]. When it is not possible to avoid missing values by optimising data collection, Multiple Imputation (MImp) is a suitable method for obtaining unbiased results while appropriately considering variability[11,12]. While in single imputation only one value is imputed for each missing entry, causing statistical analyses to overlook the uncertainty around the values which are not observed[10], MImp is a statistical technique involving the generation of multiple plausible estimates for missing values, allowing a correct quantification of the uncertainty associated with missing observations in the data[13].

In ML, the presence of missing data is often resolved by simply removing or exclusively filling the entries with a single imputation procedure[13]. However, the integration of MImp techniques with ML methods is possible, despite being rarely addressed, and may lead to superior results, enhancing prediction performance[14]. Pioneering contributions currently exist specifically addressing the use of MImp techniques in ML, exploring alternative procedures and their feasibility[15,16]. Rios et al.[15] conducted an evaluation of the impact of missing values on the accuracy estimates of ML models, employing seven distinct methods for missing data management, such as the MImp method, cluster-based imputation or regression-based imputation. In this work, MImp emerged as a promising compromise between feasibility and accuracy, in predicting patient-specific risk of adverse cardiac events.

Despite the increasing prevalence of ML methods applied to stroke and ambulation recovery studies[17], in accordance with current information no attention has been given to integrating advanced missing data management techniques with ML ones. Therefore, it becomes urgent to explore and evaluate methods that ensure the robustness and reliability of missing data handling without compromising the overall effectiveness of the analytical process.

This study focuses on the development of predictive models for the prognosis of stroke rehabilitation outcomes, based on the datasets of two multisite observational studies, prospectively and systematically enrolling all adults addressing intensive inpatient rehabilitation within 30 days after stroke[18,19]. The recovery of independent ambulation is a key stroke rehabilitation outcome, directly related to community mobility and participation[20], and improved quality of life in the chronic stage of stroke, as well as a determinant of caregiver's burden[21]. Further independent walking is a well-known top priority of stroke patients and their families, having a relevant impact on the patients' social destination after discharge, and mobility. For these reasons, the recovery of independent ambulation can be considered one of the most relevant patient-centred outcomes, as also reported in the International Standard Set of Patient-Centered Outcome Measures After Stroke[22]. Thus, we focused on the recovery of independent ambulation at discharge from rehabilitation in the subset of stroke survivors, who ambulated independently before stroke but lost the ability after stroke. After an accurate phase of data pre-processing, this study integrated MImp techniques with a cross-validated ML-based predictive model. Then, influential predictors of ambulation outcomes were identified, by using explainable Artificial Intelligence (AI) techniques.

Methods

Study design and sample

This work was based on data prospectively collected on a cohort of 448 post-stroke survivors derived from two observational multi-site databases: RIPS[18] and STRATEGY[19] studies, aimed at identifying predictors of rehabilitation outcome within 30 days from stroke. Both studies included a core of clinical and functional variables assessing the domains of body structure and function, activity, and participation as potential predictors of post-acute stroke rehabilitation outcomes. The systematic recruitment of the patients involved all individuals admitted to the Intensive Rehabilitation Units (IRUs) of Fondazione Don Gnocchi (FDG) centres, in Italy. Specifically, RIPS enrolled 234 patients from Florence, La Spezia, Massa and Fivizzano between December 2019 and December 2020. The STRATEGY project is still ongoing, with the enrolment of patients in 13 FDG centres across Italy from June 2021. This work considered 214 STRATEGY patients enrolled within July 26, 2023, in Florence and Milan (IRCCS "S. Maria Nascente").

The study protocols were a-priori registered on ClinicalTrials.gov (registration number RIPS: NCT03866057, registration number STRATEGY: NCT05389878) and were submitted and approved by the local ethical committees (RIPS: Florence, 14513; La Spezia, 294/2019; Massa and Fivizzano, 68013/2019; STRATEGY: Florence, 19779_oss; Milan, 04_13/10/2021). All medical research involving human subjects were conducted in accordance with the Declaration of Helsinki.

The studies shared the same inclusion criteria:

- First-ever or recurrent ischaemic or haemorrhagic stroke diagnosed clinically and with brain imaging occurred within 30 days from recruitment;
- First-ever admission to the IRU for the considered stroke;
- Age > 18 years old;
- Written informed consent.

Patients were excluded in the presence of transitory ischaemic attack and if addressed to the severe brain injury high-complexity rehabilitation wards due to a severe neurological condition (severe stroke with a coma lasting at least 24 h[18,19]).

Both in RIPS and STRATEGY, the integrated rehabilitation pathway (IRP) was developed based on the AHA/ASA Stroke rehabilitation guidelines[23] and on the SPREAD (Italian Stroke Guidelines) 2011[24]. The IRP was defined by an interdisciplinary team, coordinated by a physiatrist. Each patient was involved in at least three hours of rehabilitation per day and received clinical observation and management, nurse management, and

physiotherapy. Speech/swallowing training, neuropsychological treatment, occupational therapy, psychological support, and aid advice were prescribed by the physiatrist according to the team assessment. Discharge was generally determined by the achievement of the previous level of independence or a functional level adequate to prosecute outpatient rehabilitation (most often implying the recovery of ambulation), or when the functional improvement reached a plateau, and no further improvement was expected[25].

In the STRATEGY study, patients were assessed at admission, discharge from the IRUs, and 3 and 6-month follow-up after the stroke event through a telephonic interview. In for the RIPS study, they were assessed at admission, discharge from the IRU, and 6-month follow-up after the stroke event through in-person visit and telephonic interview. On both studies, the assessment addressed different domains, namely demographics, clinical and nursing complexity, neurological profile, functional evaluation, and neuropsychological evaluation. Additionally, in RIPS study, both neurophysiological and genetic assessment were included. Further details on the rehabilitation intervention on the patients, as well as the time points and content of the assessments, can be found elsewhere[18,26].

Of the above-mentioned samples, only enrolled and non-ambulatory patients at baseline were retained for the analysis. The selection of non-ambulatory patients was performed using the Functional Ambulation Categories (FAC)[27]. The ambulation item of the modified Barthel Index (mBI) was considered when the FAC was missing[28]. Hence, patients with FAC < 4 (or mBI < 15 when missing) were considered as non-ambulating in class 0, and vice versa in class 1.

Measures

Despite involving multiple assessments, the timepoints considered in this study were the baseline and discharge from the rehabilitation stay. The selected outcome was the recovery of independent ambulation at discharge. The selection of non-ambulatory patients was performed with the same modalities of the sample definition, i.e., using the FAC scale in conjunction with the ambulation item of mBI at discharge (when the first one was missing). As it will be better explained in the next section, the ML model was developed considering as candidate predictors of the ambulation outcome a subset of independent categorical and numerical variables measured at admission. The predictors can be categorised as follows: demographics, description of the event, clinical assessment, functional profile, and psycho-social assessment. The detailed list is reported in supplementary material, SM, (Supplementary Table SM.1).

- Demographic predictors were age, sex, educational level, and cohabitation.
- Predictors describing the event were aetiology (ischaemic, haemorrhagic), time from the event (days), area of the lesion (-, supra-tentorial, sub-tentorial, both), recurrence event and the side of the lesion.
- Predictors from the cognitive and psycho-social assessment were: modified functional walking categories (mFWC), a worldwide used, reliable and valid tool to assess community ambulation in stroke survivors; the Frenchay activities index (FAI), a widely used tool to measure participation after stroke, cross-culturally adapted on an Italian stroke survivor population; the Montreal cognitive assessment (MoCA)[29], a rapid screening instrument that assesses different cognitive domains, or, when missing, the mini mental state examination (MMSE)[30], a brief screening tool that provides a quantitative assessment of cognitive impairment.
- Predictors from the physiotherapy assessment were: mBI, a clinical tool used to assess the independence level across various medical condition; the FAC[31], a scale to assesses the level of functional ambulation in patients with gait limitation; the Trunk control test (TCT), used to assess four axial trunk movements in patients that have suffered from a neurological disorder; the short physical performance battery (SPPB), a clinical tests battery (i.e., tandem, 4 m walking test, five times sit-to stand) for evaluating lower extremity functioning in older persons; the modified Rankin scale (mRS), a six points-scale to measure of the degree of disability in daily life activities in stroke survivors or other causes of neurological disability; modified Ashworth scale-lower limbs (mAS_LL), the most widely accepted clinical test used to measure muscle tone on a five-points scoring; the lower limb sections of the motricity index or, when missing, the Fugl-Meyer assessment (MI/FMA_LL)[32,33], two widely adopted clinical test to assess and monitor over time the performance of persons with motor disability of neurologic origin.
- Predictors from the clinical assessment were: the National Institute of Health Stoke Scale (NIHSS), used as quantitative measure of severity of symptoms associated with stroke; the communication disability scale (CDS), a rating scale that stratifies the patient's communicative disability into five levels; the Aphasia item of NIHSS; the Charlson comorbidity index (CCI)[34], an assessment tool designed specifically to predict long-term mortality, or, when missing, the cumulative illness rating scale (CIRS)[35], a validated multidimensional test commonly used as part of the comprehensive geriatric assessment; the presence of one of the following conditions: reduced vigilance or coma, clinical instability, acute infection, delirium, depression, dysphagia, malnutrition, pressure ulcers, bladder catheter, central venous catheter, pain, nose-gastric tube or percutaneous endoscopic gastrostomy.

Data analysis

For what concerns the model development, the analysis pipeline can be summarised in three main steps (Fig. 1): dataset conversion, data pre-processing, multiple imputation of missing data, and development of the ML prediction model.

The first step of the analysis pipeline concerned the merging and selecting process of the datasets, which was carried out using The MathWorks, Inc. (2023). *MATLAB version: 9.14.0 (R2023a)*, Accessed: September 04, 2023. https://www.mathworks.com, accessed: September 04, 2023).

Specifically, the analysis was conducted by considering both RIPS and STRATEGY within a single dataset. Since these were originally two separate studies, despite having in common most of the assessment domains,

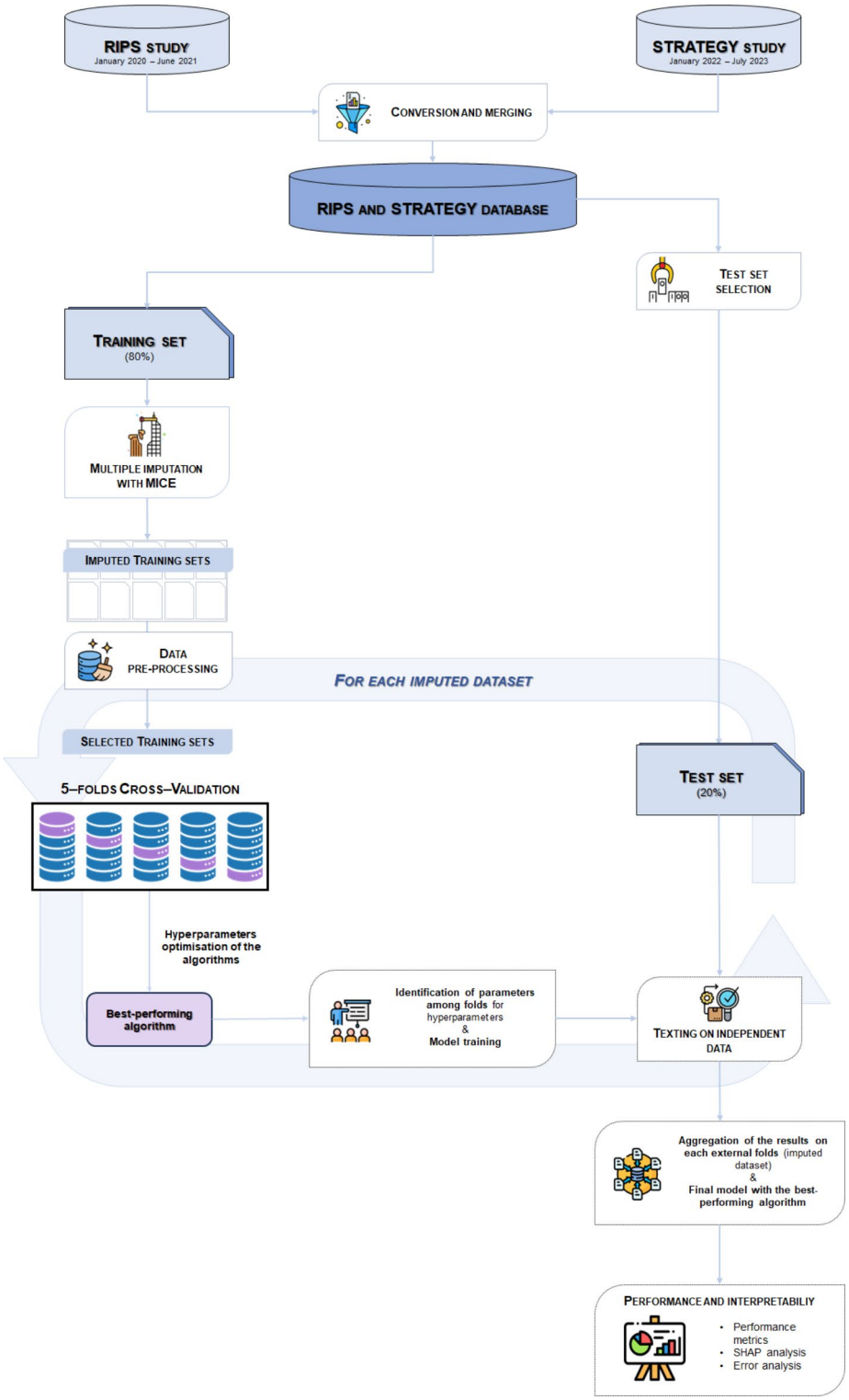

Fig. 1. Analysis pipeline.

they entail different types of variables, such as coding or nomenclature. Consequently, a data pre-processing merging was carried out on the two datasets. Identical variables were simply overlaid, while, for non-matching variables, conversions were applied based on accepted standards or clinical experience. Regarding the latter process, the main differences encountered between RIPS and STRATEGY content concerned the motor, the cognitive, and the comorbidity assessment.

Concerning the motor assessment, the scores of FMA[33] for RIPS and of the MI[32] for STRATEGY were converted in percentage of the maximum score, focusing on the motor components of FMA since the MI test

only evaluates motor function. Likewise, for comorbidities, the CIRS[35], in RIPS, and the CCI[34], in STRATEGY, were converted by calculating a percentage value. In the case of CIRS, the total score was used. In the case of CCI, a non-weighted total score was calculated (i.e., by assigning only one point to diabetes, liver disease, and cancer/tumour items) and then converted. Concerning the cognitive assessment, the MoCA[29], in RIPS, was converted into the MMSE[30] according to the conversion method on the raw scores of Aiello et al.[36]. Subsequently, the MMSE raw scores were adjusted by age and education level using the version from Carpinelli Mazzi et al.[37].

A total number of 35 features out of the 154 available, involving the different areas of the comprehensive assessment of the patients (see previous section), were selected based on clinical criteria and successively used to define the ML algorithms. Once the merging of the two datasets was completed, the test set was extracted for the final assessment of methods, including only non-ambulatory patients upon admission without missing entries on the 35 predictors and the outcome, so that no imputation was needed on them. The test set included about 20% of the non-ambulatory patients upon admission and the training/validation set encompassed the remaining 80% of the data.

Missing values in the training set were imputed through a Multiple Imputation by Chained Equations (MICE) procedure, under the assumption of Missing At Random (MAR)[13,38]. Note that the MICE was performed on the entire set of the collected variables, not only on the 36 selected predictors, to account for all information available to fill in missing data. It was implemented in the R Core Team (2023). R: A Language and Environment for Statistical. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/, using the mice function of the MICE package[39]. MICE requires the specification of predictive models on each variable with missing values. It was defined on each of them a Random Forest (RF) model conditional on all the other features. The number of MICE iterations was set to 30 and 10 imputed datasets were created. To maintain consistency in the imputation of variables that derived from or depended on others, steps of passive imputation or post-processing were applied. After excluding from the 10 imputed training sets non-ambulatory patients at the admission and patients with missing outcomes before MImp, a ML-based prediction was performed on them.

Descriptive statistics were obtained both for the merging dataset and separately for the test set and the training set before imputation, and on the patients excluded from the analysis, to better understand the differences between the included and the excluded observations. More specifically, descriptive (before MImp) and univariate analyses (after MImp, on each imputed dataset) were performed. Categorical variables were described in terms of counts and relative frequencies; continuous variables in terms of mean and standard deviation (std) or median and interquartile range (IQR), depending on the Kolmogorov-Smirnov test. Given the single split of the test set a comparison between the distributions of the selected predictors in the training and test sets was performed to assess whether there was a selection bias. Specifically, t-tests for continuous variables following a normal distribution, Kruskal-Wallis tests for numerical variables with a generic distribution, and chi-squared tests for categorical/binary variables were employed. Further, on the training set, univariate statistical analyses were conducted using Generalised Linear Regression models to investigate the association between the predictors and the outcome.

The model implementation was performed on Python, using the Scikit-learn library[40]. For each imputed dataset, after normalising the features, four different algorithms were trained and cross-validated: Elastic-Net Regularised Logistic Regression (LogReg), k-Nearest Neighbours (kNN), RF, and Support Vector Machine (SVM). The hyper-parameters of the models (Table 1) were optimised within imputation via a five-fold cross-validation (CV), using Balanced Accuracy (BA) as a performance metric, to account for the fact that the outcome was unbalanced.

The best-performing algorithm, chosen based on the mean BA across the 5 test sets of the CV, was trained on the entire training set and then tested on the test set. The overall performance on the test set was assessed by calculating both the BA and F1 scores on the aggregate solution obtained by averaging the classification scores of each individual across the models selected as the best ones in the 10 imputed datasets. The performance metrics were also used to evaluate the reliability of ML analysis. Note that, while the 5-fold cross-validation represents the inner loop of the validation scheme, the analysis of the 10 imputed data sets can be considered, in a sense, an outer loop.

On each of the 10 models selected (one for each imputation), interpretability analysis was applied, using the Shap library[41]. The results of the interpretability analysis, in terms of predictor contributions, were at last aggregated by computing the mean of the Shap values estimated on each imputed dataset. On the results of the aggregated solution, an error analysis was also conducted to obtain a better understanding of the prediction mechanism of the model and evaluate the reliability of ML analysis. Specifically, a categorical variable with four categories was created to identify true positive, true negative, false positive, and false negative records. To gain a comprehensive understanding of the error evolution, a descriptive analysis of the main baseline clinical characteristics was conducted. Variable descriptions were provided in terms of counts and relative frequencies for categorical variables, while for numerical variables, the mean or median was presented along with the respective standard deviation (std) and interquartile range (IQR) depending on the Shapiro–Wilk test (for the reduced sample size). Since this variable consists of four categories, an analysis of variance (ANOVA) was employed for continuous variables following a normal distribution, Kruskal–Wallis tests for numerical variables with a generic distribution, and chi-squared tests for categorical/binary variables. The error analysis also involved the evaluation of the performance of the aggregated solution assuming different classification thresholds on the classification probability. Thresholds were defined between 25% and 75%, with a step of 5%. For each diverse classification F1 score, BA, sensitivity, and specificity were evaluated with respect to the actual labels.

Finally, due to the fact that MImp is not as widely used as single imputation methods in ML applications, the same algorithms, parameter optimisation ranges and performance metrics were used after a single imputation of missing values performed with kNN imputer implemented in the Scikit-learn library[40] of Python. This analysis

Classifiers	Parameters	Values range
Elastic-Net Regularized Logistic Regression		
	C class_weight iterations penalty solver	[1e-2;1e2], log = True "balanced" 1e4 "l1" "saga"
k-Nearest Neighbours		
	Algorithm leaf_size n_neighbours p weights	"auto" [1;1e2] [3;1e1] 2 ["uniform"]
Random Forest		
	Bootstrap class_weight criterion max_features min_samples_leaf min_samples_split n_estimators	["True, "False"] ["balanced", "balanced_subsample"] ["gini", "entropy"] None [3; 14], log = True [5; 22], log = True [3; 20], log = True
Support Vector Machine		
	C class_weight gamma kernel probability	[1e-6; 1e2], log = True ["balanced"] [1e-5, 1e6] ["rbf", "linear"] True

Table 1. Range of hyper-parameters optimisation for each trained algorithm.

is not intended as a comparison of methods, which is beyond the scope of this paper, but is intended to show how much the results would vary using more commonly used imputation technique.

Results

A total number of 234 and 214 patients were included in the RIPS and STRATEGY studies, respectively, enrolled between December 2019 and July 2023. During the pre-processing phase, the two datasets were merged into a single dataset of 448 patients.

After the described conversion between the RIPS and STRATEGY studies, 154 common variables were obtained. Subsequently, 80 non-ambulatory patients upon admission (approximately 20% of the total) and without missing values on any predictors and on the outcome variable, were randomly selected. A selection of 35 predictors was also performed on the test set, involving the different areas of the comprehensive assessment of the patients. A total of 368 patients with the full set of 154 variables, constituting the training set, underwent MImp.

Descriptive statistics of the predictors for the overall, training, and test samples can be found in the in Supplementary Table SM.2. To provide an overall view of the solely patients entering the analyses, all ambulatory individuals upon admission with missing outcomes were excluded from the descriptive analysis (of both training and test sets), resulting in a sample size of 277, 80, and 357 observations for training, test, and merged sets, respectively.

After MImp, 10 imputed datasets were obtained, and the 35 predictors were selected. Then, the records in the training set were extracted, in terms of ambulation at admission and non-missing outcome. The dataset was unbalanced, with only 23% of non-walking patients at admission (with available outcomes) who regained ambulation upon discharge.

The variable entitled for the selection of non-ambulatory patients at admission contained 14 missing observations, the imputations of which ultimately led to different numbers of ambulatory and non-ambulatory patients in each imputed replica. The 14 patients were always imputed as non-ambulatory, except for the dataset4 and dataset8, with 13 non-ambulatory and one ambulatory, and 12 non-ambulatory and two ambulatory patients, respectively.

The consequence of the selection of non-ambulatory patients with non-missing outcomes resulted in 9 datasets of 277 observations, and 1 dataset of 276 observations (Fig. 2). The description of the included and the excluded patients, and the results of the univariate analyses, examining each imputed data set the relationships between the chosen predictor variables and the outcome are displayed in Supplementary Table SM.3 and Supplementary Table SM.4.

After the standardisation of the features and the optimisation of the hyperparameters, the results on the ML procedure reported the RF as the best-performing algorithm on the validation set (mean BA: 81.2%). Table 2 displays the performances on the test set of each imputed dataset using the RF algorithm. Additionally, confusion matrices for all 10 datasets are presented in Supplementary Figure SM.1. Lastly, the aggregated solution obtained a BA and an F1 score on the test set of 75.7% and 69.0%, respectively (Table 2).

The features that were consistently involved in the prediction with the biggest contributions were mBI, MI/FMA_LL and SPPB, among those that contributed transversally across the 10 imputations (Fig. 3). Supplementary Figure SM.2 shows the list of the 10 beeswarm plots for each imputed dataset. The results (both in aggregated,

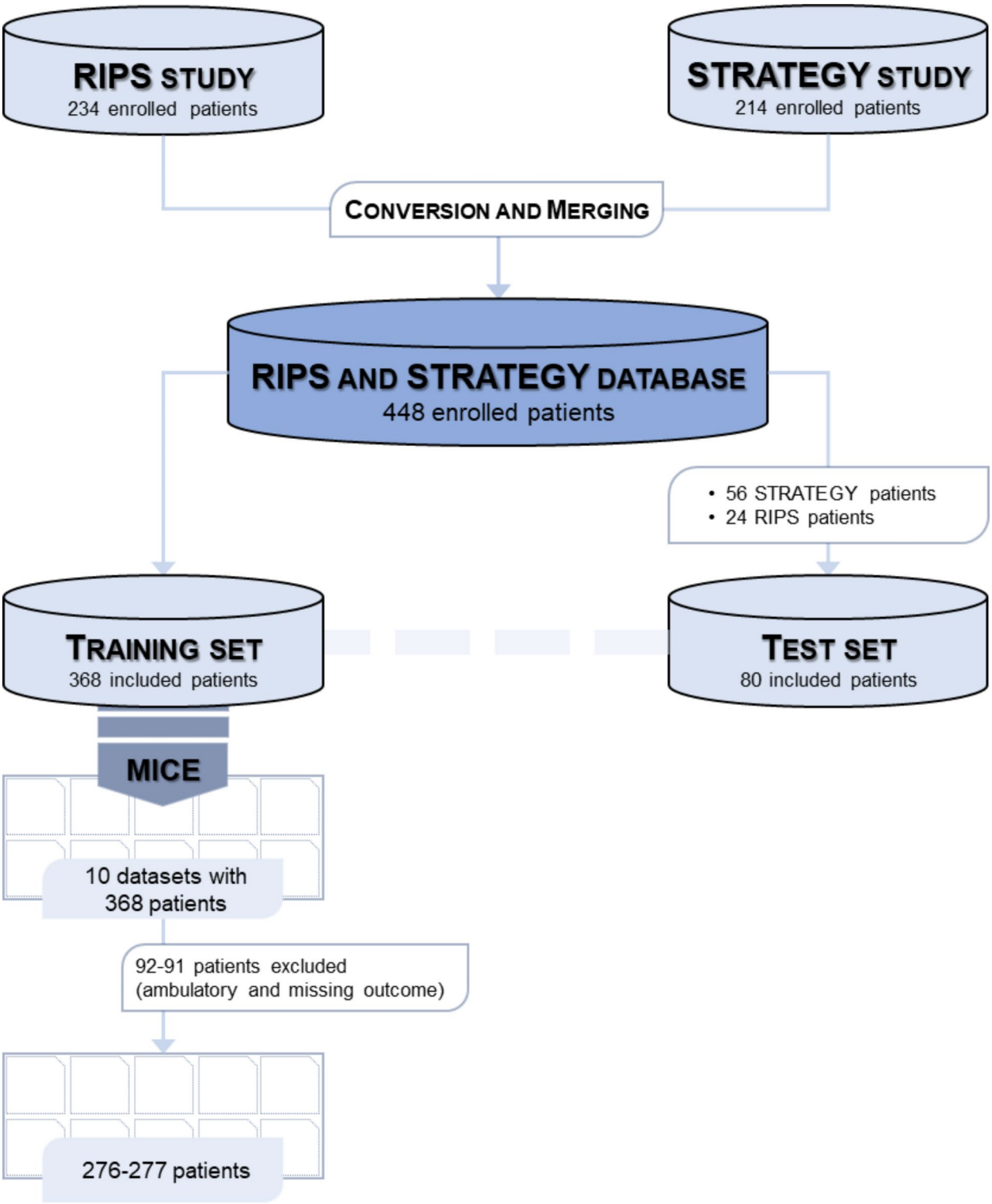

Fig. 2. Flow chart of the study.

Fig. 3, and in each imputed dataset, Supplementary Table SM.2) confirm what has been indicated by the previous statistical analyses, revealing that the features with the biggest contributions at admission are also associated with the outcome at a p-value level below 1% (Supplementary Table SM.4). Consistently identified predictors of good ambulation recovery included high modified Barthel Index scores, good motor function (MI/FMA_LL and SPPB), lower age, limited clinical complexities and good trunk control. The above-mentioned results found on the aggregated solution, concerning the contributions of predictors, were consistently found also across the results of each imputed dataset.

Datasets	F1 score %	BA %	specificity %	sensitivity %
Dataset 1	67.7	74.2	72.6	75.9
Dataset 2	70.0	76.4	80.4	72.4
Dataset 3	66.7	73.7	78.4	68.1
Dataset 4	62.7	69.5	66.7	72.4
Dataset 5	59.0	67.3	72.6	62.1
Dataset 6	63.3	71.0	76.5	65.5
Dataset 7	65.6	72.7	76.5	69.0
Dataset 8	61.5	68.8	68.6	69.0
Dataset 9	66.7	74.2	86.3	62.1
Dataset 10	67.7	74.2	72.6	72.6
Aggregated solution	69.0	75.7	82.4	69.0

Table 2. F1 score, balanced accuracy (BA), specificity, and sensitivity on each imputed dataset and the aggregated solution by averaging the classification scores on the best-performing algorithm (RF). Significant values are given in bold.

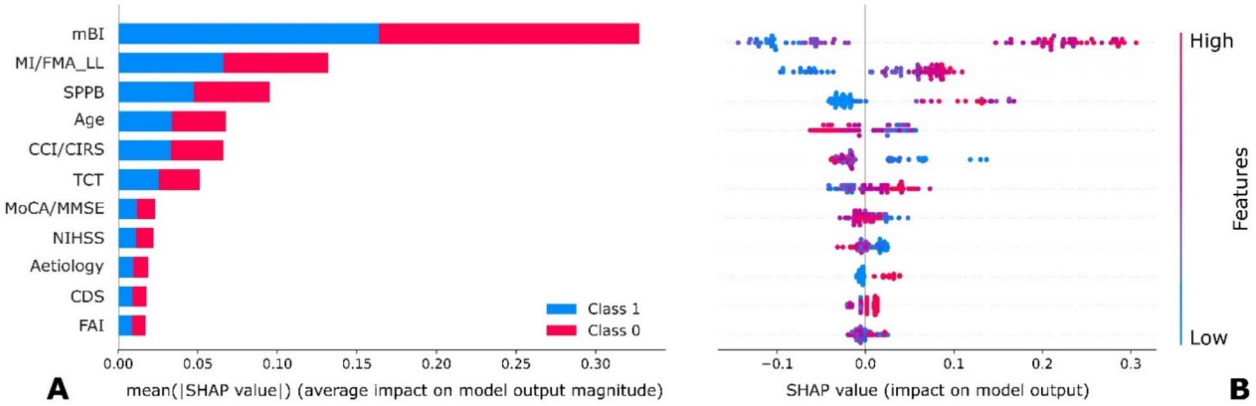

Fig. 3. Contributions of the predictors aggregated over the imputed datasets. In panel (**A**), a bar plot of the global contributions is presented, whilst in panel (**B**) a beeswarm plot with patients-wise contributions is presented. The results are presented for the RF algorithm. *Abbreviations: CCI/CIRS* Charlson comorbidity index or cumulative illness rating scale, *CDS* communication disability scale, *FAI* Frenchay activities index, *mBI* modified Barthel index, *MI/FMA_LL* lower limb score at the motricity index or the Fugl-Meyer assessment, *MoCA/MMSE* Montreal cognitive assessment or mini-mental state examination, *NIHSS* National Institutes of Health Stroke Scale, *SPPB* short physical performance, *TCT* Trunk control test.

On the results of the aggregated solution, differences in the main baseline clinical characteristics among groups of true negative, true positive, false negative, and false positive were analysed. Statistically significant differences were found for the recurrence of stroke, FAC, ambulation item of the mBI, and NIHSS, mBI and MoCA/MMSE scores (Supplementary Table SM.5). Specifically concerning the mBI, which resulted among the strongest predictor with higher values contributing to the recovery of ambulation, it can be visible how false positive have a significantly higher median value (58.0[17.0]) if compared to the false negative ones (37.0[25.0]).

The error analysis involved the evaluation of the performance of the aggregated solution considering different classification thresholds on the classification probability. These analyses allowed for the identification of the model performance, in terms of F1 score, BA, sensitivity, and specificity, when removing the most uncertain cases (Fig. 4; Table 3). It is visible how performances are increasing till the threshold between 30% and 70%, where the retained number of observations is reducing to almost 50%.

Lastly, the analysis pipeline with the single imputation method obtained a BA of 70.0% and an F1 score of 62.3%, with a decrease of 5.7% points in BA and 6.7 in F1 score compared to MImp.

Discussion

This study developed and cross-validated ML-based prognostic models integrated with MImp techniques for the prediction of independent ambulation recovery in post-stroke survivors at discharge from rehabilitation.

Missing data may represent a significant problem in health data analysis, and it is crucial to employ methodologies that appropriately address it[42]. The recognition of missing data as a potential source of uncertainty underscores the importance of strong methodologies, such as MImp, in healthcare research. In the

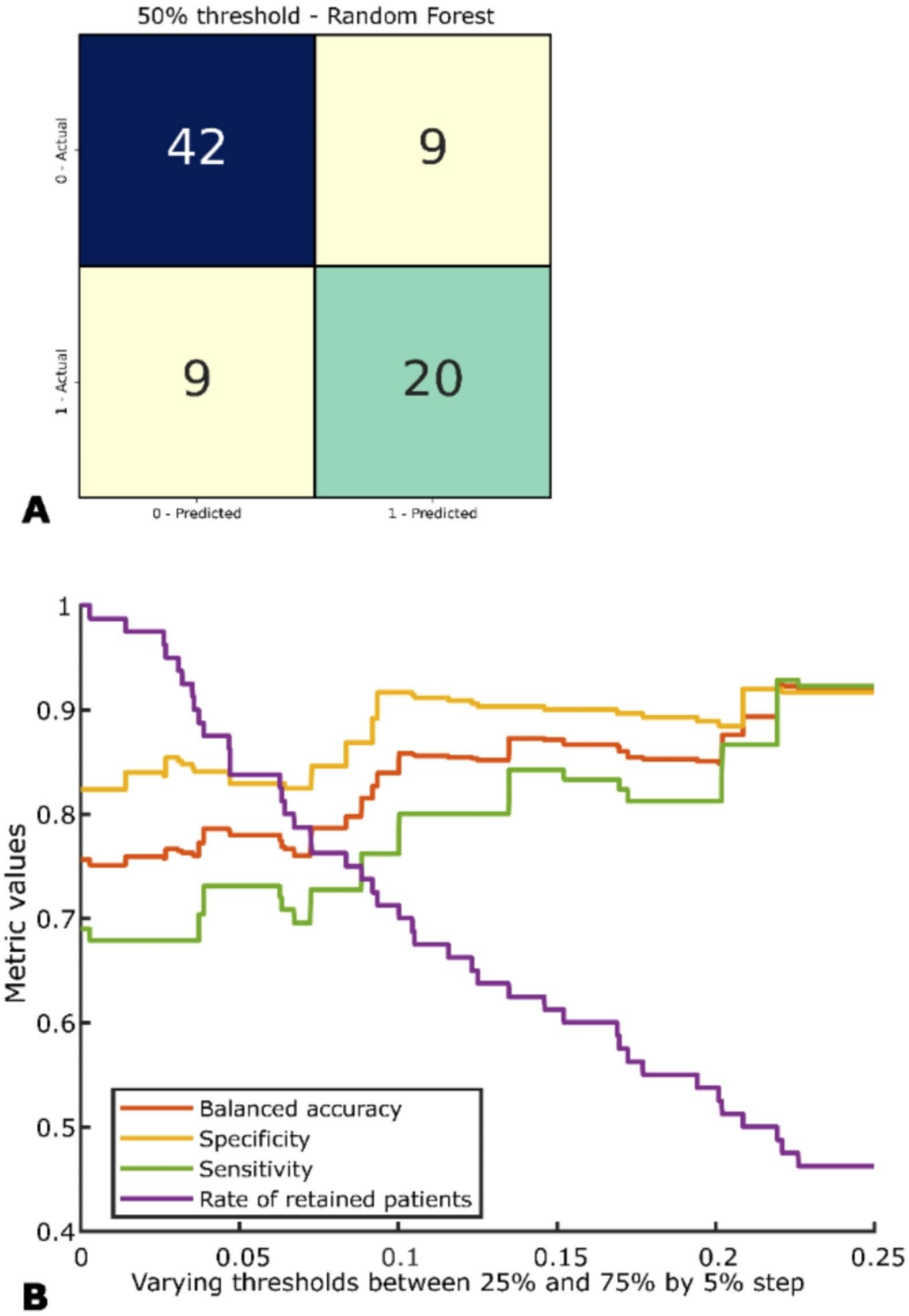

Fig. 4. Confusion matrix of the aggregated solution with threshold of 50% (**A**). In (**B**), BA, sensitivity, specificity, and rate of retained patients on the test set are proposed for different classification thresholds.

absence of comprehensive data collection during the study's design phase[43], the application of MImp becomes essential. Although this work does not include evaluations of MImp performance or simulations aimed at verifying it, the proposal of this study of combining MImp with ML relies on solid bases. First of all, MImp is highly effective for dealing with missing entries, and unanimously considered the gold standard in the statistical domain. Additionally, even if the issue of missing data in the ML domain is much less explored, the existing results seem to indicate that MImp is promising. For example, Tran et al. show that integrating MImp with ML

Thresholds	F1 score %	BA %	Specificity %	Sensitivity %	*N*
50%	69.0	75.7	82.4	69.0	80
45%	73.1	78.0	82.9	73.1	67
40%	80.0	83.9	91.7	76.2	57
35%	84.2	87.1	90.0	84.2	49
30%	81.3	85.1	88.9	81.3	43
25%	88.9	92.0	91.7	92.3	37

Table 3. F1 score, balanced accuracy (BA), sensitivity, specificity and numerosity of the retained set on the test set for the aggregated solution with thresholds varying from 25 to 75%, with a step of 5%.

achieves significantly better classification accuracy than other methods that employ more common imputation procedure. Aleryani et al. also find a similar result. A relevant issue in combining MImp with ML techniques concerns the pooling method, that is the method used to derive one overall result from many imputed data sets. Compare different approaches for pooling the multiple imputed data sets after multivariate imputation by chained equation (MICE), finding that combining the predictions (bagging approach), as in this work, is the best pooling method. It outperforms, for example, the simpler approach consisting in pooling the imputed data sets before applying the ML algorithms.

For both MImp and ML methods, a larger sample size enhances the precision in imputation and predictions. Consequently, the joint clinical and automatic strategy allowed the merge of two temporally non-overlapping datasets (RIPS[18] and STRATEGY[19]), increasing the sample size, and the subsequent model interpretation analysis. Merging two different, though similar, datasets led to the utilisation of converted scales according to the minimal stroke protocol (PMIC2020)[44] and possibly the introduction of a potential bias in prediction results. To reduce potential bias, the conversion was performed utilising normative data[36], when possible, and with clinical support for appropriate information fitting. The application of automatic solutions after a first clinical approach guaranteed the interpretability of the solution developed and its results, promoting an increasing trust, usability, and acceptance of these solutions[45]. Then, an independent and complete test set enabled estimates of data that were not influenced by imputation, providing a clean reference for the assessment of the results.

The outcome of interest was selected on the FAC scale, given the need to measure the recovery of independent walking. This scale is a tool that measures ambulation ability, considering both indoor and outdoor settings, based on the human assistance required during walking, regardless of the use of assistive devices[46]. The FAC assesses the full spectrum of ambulation, providing essential insight for predicting the patient's future independence. While in a previous study[47] a fine-grained prediction of overall functional recovery in stroke patients was searched, choosing the modified Barthel Index score as the primary outcome, rather than using a disability level cut-off, in this study the focus is on independent walking, as a top priority of patients and families[48]. Indeed, the choice of a dichotomised outcomes lies in the relevance of the information that needs to be shared with health professionals, patients, and caregivers, of whether the patient would not achieve independence in ambulation, rather than attempting to establish to what extent he/she will recover ambulation[22]. Thus, in analogy with previous literature, the outcome was dichotomised based on the FAC cut-off suggested by Mehrholz et al.[49] in FAC ≥ 4 (independent walking) and FAC < 4 (supervised or assisted walking). The outcome of this study paves the way to the development and external validation of ever more accurate models predicting the recovery of independent ambulation at discharge from intensive inpatient rehabilitation after stroke; this will enable clinicians to provide comprehensive responses to families and patients inquiring about the potential return to independence[49], allowing them to anticipate the level and duration of support required after discharge from the hospital[50].

One of the aims of this study was to achieve an analysis pipeline and prototypal solution for an effective clinical decision-support system for rehabilitation, aiding clinicians in delivering confident responses to patients' families regarding the potential for independent walking recovery. This aligns with a similar objective seen in the study from Smith et al.[50], where the time to walking independently after stroke (TWIST) algorithm was employed to predict independent walking post-stroke defined as FAC ≥ 4. The meticulous handling of missing data is an often-overlooked aspect[50] and its omission is associated with distortion in results and underestimation of variability. This research aimed to address issues related to poor data management, ensuring more accurate predictions. This study employed advanced ML models, contributing to enhancing the accuracy of the outcomes[13].

In comparison to the earlier study of Campagnini et al.[51] on post-stroke functional recovery, targeting a more generic recovery outcome on a different dataset, this work focuses specifically on predicting independent walking ability, using the FAC with the mBI as a major predictor. Although a direct comparison is not possible, the achieved accuracy is slightly increased in this study with respect to the earlier one. It should be considered that key differences among the studies exist in terms of data content and methodology. This study results from a prospective data acquisition on a wider set of information, enabling a more comprehensive assessment of patients. The selected outcome, which is the recovery of a specific activity—ambulation—among the many assessed by mBI, along with a larger sample size and a greater number of independent variables encompassed in the current investigation, could have had a positive impact on the analytical results.

In addition to the development and cross-validation of the classifiers, an analysis of the interpretability of the best-performing model was also performed. More specifically, the analysis was conducted through the application of the Shap library[41], which evaluates the weight of the feature on the prediction in a global and

patient-specific manner. The analysis of the weights of factors showed great importance on walking ability, including the mBI level, the trunk control, the communication level, and daily activity domains. Also, the lower limb motor functioning (MI or FMA scale) and the level of physical performance (SPPB) were significant. The cognitive function (MoCA combined with MMSE), age, presence of comorbidity, severity, and type of stroke were confirmed as predictors. This different selection of predictors underscores the significance of conducting a thorough and broad evaluation of patients, particularly within the rehabilitation domain. This assessment should encompass various domains, such as functional, motor, cognitive, clinical history, and clinical complexity components (PMIC2020[19,20,47]). Among these factors, specifically, mBI, MI/FMA, SPPB, age, CCI/CIRS, and TCT at baseline contributed the most to the prediction of the model. These results are in line with previous studies. In fact, in two retrospective studies, global functioning, age, trunk control, along with lower-limb motor functions were also found to be influential in the recovery of ambulation at discharge from rehabilitation[52,53]. Indeed, in another retrospective study that analysed predictors of ambulation recovery in a similar intensive inpatient rehabilitation setting, but in a different population (patients with hip fracture), older age, higher comorbidity, impaired trunk control, and lower functional status upon admission were associated with unfavourable outcomes[54]; this suggests that these factors are relevant to the recovery of ambulation in intensive rehabilitation after a catastrophic event, regardless of the specific clinical condition involved. Also, a review[55] concluded that TCT and lower limb motricity seem to be the best predictors of gait recovery at six months after stroke. Further, the study from Guralnik et al.[56] demonstrated how the SPPB performance measure can validly characterise older individuals for lower extremity functions, thus providing useful insights into the individual's functional status. Consistently with existing literature, the primary predictors identified were usually the modified Barthel Index at admission, holding information on functional independence and basic daily activities among stroke survivors[28,57]. While these findings align with existing literature employing traditional methodologies for predicting walking ability in stroke survivors, for study repeatability it is necessary to report that the timing of assessments can significantly influence the performance of the prediction models on ambulation ability, as previously reported[58].

Beyond studies applying traditional statistical methods to the prediction of the recovery of ambulation after a stroke, a recent narrative review investigated the role of ML in predicting central nervous systems rehabilitation outcomes[17]. Three papers included in the review, all by the same research group, explored ML methods for the prediction of the recovery of ambulation after stroke[59–61]. All considered the recovery of ambulation at six months after stroke, by dichotomising the FAC score, with the same cut-off suggested by Mehrholtz et al.[49] and used in this study. None of them, however, included imputation methods in the pipeline. Kim et al.[59] performed a retrospective study using only Magnetic Resonance Imaging (MRI) (30 days within stroke) of 221 patients with a corona radiata infarct undergoing post-stroke rehabilitation and used a Convolutional Neural Network (CNN) to predict the FAC at six months after onset. Similarly in Shin et al.[60] MRI data from 1233 post-stroke patients were processed with the same purpose. These studies achieved 79.1%[63] and 76.1% accuracy[64]. In both cases, results are reported on the validation set, lacking an independent test subset. Compared with these two studies, a limitation of this work might be the absence of imaging variables; however, this study still achieved comparable accuracies using only clinical variables, which are easily obtainable in most rehabilitation settings. In this way, a potentially more deployable and reproducible model than the one based on MRI scans has been proposed. Actually the first work published by Kim et al.[61] retrospectively considered 833 consecutive stroke survivors, and applied deep neural network (DNN), random forest (RF) and logistic regression models (LRM) to the prediction of the recovery of ambulation at 6 months, based on clinical variables collected within 30 days from stroke: age, sex, type of stroke (ischaemic/haemorrhagic), modified Brunnstrom Classification (mBC), FAC, and Medical Research Council (MRC) score for muscle strength of hip flexor, knee extensor, and ankle dorsiflexor of the affected side. They achieved 69.3% accuracy using random forests; the availability of a larger pool of participants enabled them to use deep learning methods, reaching 78.7% accuracy with DNN. Despite the complexity of the proposed solution (three layers, 256 neurons per layer), an analysis of the interpretability was not reported.

With respect to these studies, the decision to evaluate rehabilitation outcomes at discharge rather than at a specific time point may be considered a limitation. Indeed, although ambulation is one of the most relevant goals of intensive rehabilitation, discharge may be influenced by economic and psychosocial factors that are at least partially independent of the motor recovery trajectory and potential[56]. On the other hand, when the outcome is collected six months after the stroke, as in the previous three studies proposing ML application for ambulation recovery post-stroke[59-61], it may be influenced by numerous factors unrelated to rehabilitation. None of these studies reported information about how rehabilitation was conducted or its duration. In the ongoing STRATEGY study, data are being collected at a follow-up six months post-stroke, including ambulation and global functioning but also recording information on rehabilitation and adverse events during the same period.

According to current knowledge, this study is the first application of ML models to predict the recovery of ambulation in stroke patients based on a prospective dataset. This study systematically considered all post-acute stroke patients accessing inpatient rehabilitation, enabling predictions that are not limited to a specific type of stroke (excluding severe vascular brain lesions). Additionally, this work considered a comprehensive set of clinical variables, including comorbidity, which is often reported in stroke patients and may influence stroke rehabilitation outcomes[62].

This study proposes an analysis pipeline, integrated of both MImp and ML, that could be transversally applied in diverse settings for more robust management of missing data, and thus more reliable predictions. Concerning the pipeline, some limitations, in the selection of the test set should be highlighted: in this work, the test set was by design excluded from imputation, thus its extraction was performed a priori, through single split, randomly selecting complete observations of non-ambulatory patients. By doing this, only a specific portion of the data was selected as a candidate for the test set, introducing a potential selection bias. Despite a random selection was performed, the test set showed statistically significant differences with respect to the training set. In particular,

the test set sample showed a slightly reduced stroke severity and higher functional level. This translated into a lower sensitivity of this work solution with respect to its specificity in identifying patients that will recover independent ambulation. This aspect should be considered when interpreting the ML-model prediction on new patients. A possible solution in future work could be to redesign the method pipeline, applying MImp within the cross-validation cycle.

Another interesting aspect that could be considered in future research is the introduction of automatic processes for data cleaning, such as the integration of multivariate cell-wise outlier detection with MImp or possibly the inclusion of other instrumental data among predictors, which could be beneficial for predicting motor functions[17,50].

On the other hand, the analyses, while providing an interpretable, predictive model of ambulation recovery, resulted in clinical insights identifying influential factors in the recovery. Finally, according to current knowledge, no earlier studies applied ML models to predict ambulation recovery in post-stroke patients undergoing intensive rehabilitation, based on a prospective data acquisition trial. Thus, this work showed promising results that could support clinical decisions by assisting in the design of optimal rehabilitation programs based on realistic therapeutic goals.

Conclusions

This work focused on the integration of MImp techniques with ML-based predictive models to propose an integrated pipeline for data-driven prognostic modelling. The pipeline efficacy was verified on a clinically relevant problem of rehabilitation medicine which is ambulation recovery prediction in post-stroke survivors at discharge from inpatient rehabilitation. The method was validated using data from two prospective observational studies systematically recruiting patients addressing inpatient rehabilitation within 30 days from stroke onset. The achieved performances and identification of key predictors among a set of comprehensive easily collected clinical variables, applied to clinical research databases, enhance the likelihood prediction of post-stroke patients regaining walking ability. This provides ground for the development of a tool to support clinical decisions during rehabilitation. The interpretability analysis underscores the need for a comprehensive assessment of patients undergoing rehabilitation, using standardised and validated measures. The heightened accuracy and reliability of estimates provide a foundation for more informed clinical decision-making, particularly vital in the context of post-stroke survivor care. Lastly, this study reveals insights that should be considered in future research endeavours aiming to maximize information in the face of incomplete data, thereby advancing the methodology and reliability of prognostic studies in healthcare.

Data availability

Data will be available upon request to the corresponding author for research purposes.

Received: 15 May 2024; Accepted: 26 September 2024
Published online: 24 October 2024

References

1. Rinott, R. et al. Prognostic data-driven clinical decision support - formulation and implications. *Stud. Health Technol. Inf.***169**, 140–144 (2011).
2. Baptista, M., Goebel, K. & Henriques, E. Relation between Prognostics Predictor Evaluation Metrics and local interpretability SHAP values. *Artif. Intell.* **306**, 103667 (2022).
3. Ostropolets, A., Zhang, L. & Hripcsak, G. A scoping review of clinical decision support tools that generate new knowledge to support decision making in real time. *J. Am. Med. Inf. Assoc. JAMIA*. **27**, 1968–1976 (2020).
4. Rincé, G. et al. Impact of an individual personalised rehabilitation program on mobility performance in older-old people. *Aging Clin. Exp. Res.* **33**, 2821–2830 (2021).
5. Dür, M., Wenzel, C., Simon, P. & Tucek, G. Patients' and professionals' perspectives on the consideration of patients' convenient therapy periods as part of personalised rehabilitation: a focus group study with patients and therapists from inpatient neurological rehabilitation. *BMC Health Serv. Res.* **22**, 372 (2022).
6. Kokkotis, C. et al. Machine learning techniques for the prediction of functional outcomes in the Rehabilitation of Post-stroke patients: a scoping review. *BioMed*. **3**, 1–20 (2023).
7. Campagnini, S. et al. Machine learning methods for functional recovery prediction and prognosis in post-stroke rehabilitation: a systematic review. *J. Neuroeng. Rehabil.* **19**, 54 (2022).
8. Habehh, H., Gohel, S. Machine Learning in Healthcare. *Curr Genomics.* **22**(4), 291–300. https://doi.org/10.2174/1389202922666210705124359 (2021).
9. Madley-Dowd, P., Hughes, R., Tilling, K. & Heron, J. The proportion of missing data should not be used to guide decisions on multiple imputation. *J. Clin. Epidemiol.* **110**, 63–73 (2019).
10. Zhang, N. Journal of Hospital Medicine Leadership Team. Methodolgical Progress Note: Handling Missing Data in Clinical Research. *J Hosp Med.* **14**(4), 237–239. https://doi.org/10.12788/jhm.3330 (2020).
11. Kang, H. The prevention and handling of the missing data. *Korean J. Anesthesiol.* **64**, 402–406 (2013).
12. Shao, X. et al. Development and validation of risk prediction models for stroke and mortality among patients with type 2 diabetes in northern China. *J. Endocrinol. Invest.* **46**, 271–283 (2023).
13. Rubin, D. B. Inference and Missing Data. *Biometrika*. **63**, 581–592 (1976).
14. Navarro, C. L. A. et al. Risk of bias in studies on prediction models developed using supervised machine learning techniques: systematic review. *BMJ*. **375**, n2281 (2021).
15. Rios, R. et al. Handling missing values in machine learning to predict patient-specific risk of adverse cardiac events: insights from REFINE SPECT registry. *Comput. Biol. Med.* **145**, 105449 (2022).
16. Hoogland, J. et al. Handling missing predictor values when validating and applying a prediction model to new patients. *Stat. Med.* **39**, 3591–3607 (2020).
17. Chang, M. C. et al. The Use of Artificial Intelligence to predict the prognosis of patients undergoing Central Nervous System Rehabilitation: a narrative review. *Healthcare*. **11**, 2687 (2023).
18. Hakiki, B. et al. Predictors of function, activity, and participation of Stroke patients undergoing Intensive Rehabilitation: a Multicenter prospective observational study protocol. *Front. Neurol.* **12**, 632672 (2021).

19. Chiavilli, M. et al. Design and implementation of a Stroke Rehabilitation Registry for the systematic assessment of processes and outcomes and the development of data-driven prediction models: the STRATEGY study protocol. *Front. Neurol.* **13**, 919353 (2022).
20. Bijleveld-Uitman, M., van de Port, I. & Kwakkel, G. Is gait speed or walking distance a better predictor for community walking after stroke? *J. Rehabil Med.* **45**, 535–540 (2013).
21. Muren, M. A., Hütler, M. & Hooper, J. Functional capacity and health-related quality of life in individuals post stroke. *Top. Stroke Rehabil.* **15**, 51–58 (2008).
22. Salinas, J. et al. An International Standard Set of patient-centered outcome measures after stroke. *Stroke.* **47**, 180–186 (2016).
23. Winstein, C.J., Stein, J., Arena, R., Bates, B., Cherney, L.R., Cramer, S.C., Deruyter, F., Eng, J.J., Fisher, B., Harvey, R.L., Lang, C.E., MacKay-Lyons, M., Ottenbacher, K.J., Pugh, S., Reeves, M.J., Richards, L.G., Stiers, W., Zorowitz, RD. American Heart Association Stroke Council, Council on Cardiovascular and Stroke Nursing, Council on Clinical Cardiology, and Council on Quality of Care and Outcomes Research. Guidelines for Adult Stroke Rehabilitation and Recovery: A Guideline for Healthcare Professionals From the American Heart Association/American Stroke Association. Stroke. 2016 Jun; **47**(6), e98-e169. https://doi.org/10.1161/STR.0000000000000098. [Erratum in: Stroke. 2017 Feb; 48(2):e78. https://doi.org/10.1161/STR.0000000000000120] [Erratum in: Stroke. 2017 Dec; 48(12):e369. https://doi.org/10.1161/STR.0000000000000156]
24. Inzitari, D. & Carlucci, G. Italian stroke guidelines (SPREAD): evidence and clinical practice. *Neurol. Sci.* **27**, s225–s227 (2006).
25. Cecchi, F., Diverio, M., Arienti, C., Corbella, E., Marrazzo, F., Speranza, G., Del Zotto, E., Poggianti, G., Gigliotti, F., Polcaro, P., Zingoni, M., Antonioli, D., Avila, L., Barilli, M., Romano, E., Landucci Pellegrini, L., Gambini, M., Verdesca, S., Bertolucci, F., Mosca, I., Gemignani, P., Paperini, A., Castagnoli, C., Hochleitner, I., Luisi, M.L., Lucidi, G., Hakiki, B., Gabrielli, M.A., Fruzzetti, M., Bruzzi, A., Bacci Bonotti, E., Pancani, S., Galeri, S., Macchi, C., Aprile, I. Development and implementation of a stroke rehabilitation integrated care pathway in an Italian no profit institution: an observational study. *Eur J Phys Rehabil Med.* **56**(6), 713–724. https://doi.org/10.23736/S1973-9087.20.06195-X (2020).
26. Chao, Y.S., Wu, C.J., Wu, H.C., McGolrick, D., Chen, WC. Interpretable Trials: Is Interpretability a Reason Why Clinical Trials Fail? Front Med (Lausanne). **8**, 541405. https://doi.org/10.3389/fmed.2021.541405 (2021).
27. Geyh, S. et al. Identifying the concepts contained in outcome measures of clinical trials on stroke using the International classification of Functioning, Disability and Health as a reference. *J. Rehabil Med.* https://doi.org/10.1080/16501960410015399 (2004).
28. Shah, S., Vanclay, F. & Cooper, B. Improving the sensitivity of the Barthel Index for stroke rehabilitation. *J. Clin. Epidemiol.* **42**, 703–709 (1989).
29. Nasreddine, Z. S. et al. The Montreal Cognitive Assessment, MoCA: a brief Screening Tool for mild cognitive impairment. *J. Am. Geriatr. Soc.* **53**, 695–699 (2005).
30. Folstein, M. F., Folstein, S. E. & McHugh, P. R. Mini-mental state'. A practical method for grading the cognitive state of patients for the clinician. *J. Psychiatr Res.* **12**, 189–198 (1975).
31. Collen, F. M., Wade, D. T. & Bradshaw, C. M. Mobility after stroke: reliability of measures of impairment and disability. *Int. Disabil. Stud.* **12**, 6–9 (1990).
32. Demeurisse, G., Demol, O. & Robaye, E. Motor evaluation in vascular hemiplegia. *Eur. Neurol.* **19**, 382–389 (1980).
33. Cecchi, F. et al. Transcultural translation and validation of Fugl-Meyer assessment to Italian. *Disabil. Rehabil.* **43**, 3717–3722 (2021).
34. Charlson, M. E., Carrozzino, D., Guidi, J. & Patierno, C. Charlson Comorbidity Index: a critical review of Clinimetric Properties. *Psychother. Psychosom.* **91**, 8–35 (2022).
35. Parmelee, P. A., Thuras, P. D., Katz, I. R. & Lawton, M. P. Validation of the cumulative illness rating scale in a geriatric residential population. *J. Am. Geriatr. Soc.* **43**, 130–137 (1995).
36. Aiello, E. N., Pasotti, F., Appollonio, I. & Bolognini, N. Equating Mini-mental State Examination (MMSE) and Montreal Cognitive Assessment (MoCA) scores: conversion norms from a healthy Italian population sample. *Aging Clin. Exp. Res.* **34**, 1721–1724 (2022).
37. Carpinelli Mazzi, M. et al. Mini-mental state examination: new normative values on subjects in Southern Italy. *Aging Clin. Exp. Res.* **32**, 699–702 (2020).
38. Little, R. J. A. & Rubin, D. B. *Statistical Analysis with Missing Data* (Wiley, 2019).
39. van Buuren, S. & Groothuis-Oudshoorn, K. Mice: multivariate imputation by chained equations in R. *J. Stat. Softw.* **45**, 1–67 (2011).
40. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
41. Shapley, L. S. *A Value for N-Person Games.* https://www.rand.org/pubs/papers/P295.html (1952).
42. Marino, M., Lucas, J., Latour, E. & Heintzman, J. D. Missing data in primary care research: importance, implications and approaches. *Fam Pract.* **38**, 199–202 (2021).
43. Di, J. et al. Considerations to address missing data when deriving clinical trial endpoints from digital health technologies. *Contemp. Clin. Trials.* **113**, 106661 (2022).
44. Cecchi, F. et al. Redefining a minimal assessment protocol for stroke rehabilitation: the new 'Protocollo Di Minima per l'ICtus' (PMIC2020). *Eur. J. Phys. Rehabil Med.* **57**, 669–676 (2021).
45. Moss, L., Corsar, D., Shaw, M., Piper, I. & Hawthorne, C. Demystifying the Black Box: the importance of interpretability of Predictive models in Neurocritical Care. *Neurocrit Care.* **37**, 185–191 (2022).
46. Holden, M. K., Gill, K. M., Magliozzi, M. R., Nathan, J. & Piehl-Baker, L. Clinical gait assessment in the neurologically impaired. Reliability and meaningfulness. *Phys. Ther.* **64**, 35–40 (1984).
47. Sodero, A. et al. Predicting the functional outcome of intensive inpatient rehabilitation after stroke: results from the RIPS Study. *Eur. J. Phys. Rehabil Med.* **60**, 1–12 (2024).
48. Moore, S. A., Boyne, P., Fulk, G., Verheyden, G. & Fini, N. A. Walk the talk: current evidence for walking Recovery after Stroke, Future pathways and a Mission for Research and Clinical Practice. *Stroke.* **53**, 3494–3505 (2022).
49. Mehrholz, J., Wagner, K., Rutte, K., Meißner, D. & Pohl, M. Predictive validity and responsiveness of the functional ambulation category in Hemiparetic patients after Stroke. *Arch. Phys. Med. Rehabil.* **88**, 1314–1319 (2007).
50. Smith, M. C., Barber, P. A. & Stinear, C. M. The TWIST Algorithm Predicts Time to walking independently after stroke. *Neurorehabil Neural Repair.* **31**, 955–964 (2017).
51. Campagnini, S. et al. Cross-validation of predictive models for functional recovery after post-stroke rehabilitation. *J. Neuroeng. Rehabil.* **19**, 96 (2022).
52. Hirano, Y. et al. Prediction of independent walking ability for severely hemiplegic stroke patients at Discharge from a Rehabilitation Hospital. *J. Stroke Cerebrovasc. Dis. Off J. Natl. Stroke Assoc.* **25**, 1878–1881 (2016).
53. Ishiwatari, M. et al. Prediction of gait independence using the trunk impairment scale in patients with acute stroke. *Ther. Adv. Neurol. Disord.* **15**, 17562864221140180 (2022).
54. Cecchi, F. et al. Predictors of recovering ambulation after hip fracture inpatient rehabilitation. *BMC Geriatr.* **18**, 201 (2018).
55. Selves, C., Stoquart, G. & Lejeune, T. Gait rehabilitation after stroke: review of the evidence of predictors, clinical outcomes and timing for interventions. *Acta Neurol. Belg.* **120**, 783–790 (2020).
56. Guralnik, J. et al. A short physical performance battery assessing lower extremity function: Association with Self-reported disability and prediction of mortality and nursing home admission. *J. Gerontol.* **49**, M85–94 (1994).

57. Harrison, J. K., McArthur, K. S. & Quinn, T. J. Assessment scales in stroke: clinimetric and clinical considerations. *Clin. Interv. Aging.* **8**, 201–211 (2013).
58. Veerbeek, J.M., Pohl, J., Held, J.P.O., Luft, A.R. External Validation of the Early Prediction of Functional Outcome After Stroke Prediction Model for Independent Gait at 3 Months After Stroke. *Front Neurol.* **13**, 797791. https://doi.org/10.3389/fneur.2022.797791 (2022).
59. Kim, J. K., Choo, Y. J., Shin, H., Choi, G. S. & Chang, M. C. Prediction of ambulatory outcome in patients with corona radiata infarction using deep learning. *Sci. Rep.* **11**, 7989 (2021).
60. Shin, H., Kim, J. K., Choo, Y. J., Choi, G. S. & Chang, M. C. Prediction of Motor Outcome of Stroke patients using a deep learning algorithm with Brain MRI as Input Data. *Eur. Neurol.* **85**, 460–466 (2022).
61. Kim, J. K., Lv, Z., Park, D. & Chang, M. C. Practical machine learning model to predict the Recovery of Motor Function in patients with stroke. *Eur. Neurol.* **85**, 273–279 (2022).
62. Ruksakulpiwat, S. et al. Associations between diagnosis with stroke, comorbidities, and activity of daily living among older adults in the United States. *Chronic Dis. Transl. Med.* **9**, 164–176 (2023).

Author contributions

The contributions of authors of this work are the following:- Conceptualisation: F.C., Mi. B., A.M.- Data analysis: A.F., S.C.- Writing: A.F.- Clinical assessment protocol design: F.C., B.H., A.G., C.M., J.N.S.- Physiotherapy assessment protocol design: S.D., Ma. B.- Cognitive and psycho-social assessment protocol design: D.B.- Manuscript reviewing and editing and approving the final version: A.F., S.C., A.M., S.D., Ma. B, B.H., A.G., C.M., D.B., J.N.S., Mi. B., F.C.

Funding

This work was supported by the Italian Ministry of Health with the "Ricerca Corrente" programs and under the complementary actions to the NRRP "Fit4MedRob – Fit for Medical Robotics".

Declarations

Competing interests

The authors declare no competing interests.

Ethical committee

These studies were a-priori registered on ClinicalTrials.gov (registration number RIPS: NCT03866057, registration number STRATEGY: NCT05389878) and were submitted and approved by the local ethical committees (RIPS: Florence, 14513; La Spezia, 294/2019; Massa and Fivizzano, 68013/2019; STRATEGY: Florence, 19779_oss; Milan, 04_13/10/2021).

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-74537-8.

Correspondence and requests for materials should be addressed to S.C.