

Article

A Data-Driven Approach for Handling Missing Data in a Real Multiple Sclerosis Dataset Based on Machine Learning

Shima Pilehvari ^{1,*}, Wei Peng ^{1,*}, Mohammad Ali Sahraian ² and Sharareh Eskandarieh ^{2,*}¹ Faculty of Engineering and Applied Science, University of Regina, Regina, SK S4S 0A2, Canada² Multiple Sclerosis Research Centers, Neuroscience Institute, Tehran University of Medical Sciences, Tehran 11367-46911, Iran

* Correspondence: wei.peng@uregina.ca (W.P.); sh_eskandarieh@tums.ac.ir (S.E.)

Abstract

Background and Objective: Reliable medical research depends on data integrity, yet clinical datasets often contain real and systematically missing values. This study aimed to develop a robust, clinically realistic imputation framework for a raw Multiple Sclerosis (MS) dataset affected by real-world missingness. **Methods:** We propose an innovative data-driven approach called the Sequential Multiple Imputation Bootstrapping (SMIB) model, which orders imputation based on feature correlation and incorporates bootstrapping to enhance stability and generalizability. Relevant features were identified using RF importance and Mutual Information scores and imputed using a hybrid machine learning framework that combines Random Forest (RF), Multilayer Perceptron (MLP), k-Nearest Neighbors (kNN), and a multiple imputation (MI) algorithm. The proposed method was evaluated using 15-fold cross-validation and a masking-based evaluation strategy. Model performance was assessed using accuracy, precision, recall, specificity, F1 score, Mean Squared Error (MSE), Mean Absolute Error (MAE), and R^2 . **Results:** RF-based SMIB achieved superior performance, with final imputation accuracy reaching up to 97% for categorical outcomes and strong numerical performance (R^2 up to 0.999; MSE as low as 2.48×10^{-5}). Sequential ordering and weighted bootstrapping improved stability in imbalanced clinical data under real-world missingness. Compared with the widely adopted Multiple Imputation by Chained Equations (MICE) approach, the proposed SMIB framework consistently demonstrated superior predictive performance across all evaluated categorical and numerical outcomes. **Conclusions:** The SMIB framework provides a robust and clinically aligned strategy for handling real-world missing values in MS datasets, improving imputation accuracy while preserving feature dependencies and feature relationships. The method supports reliable predictive analytics in healthcare contexts.

Keywords: multiple sclerosis; machine learning; missing data; sequential imputation; bootstrapping; data-driven



Academic Editor: Francesca Luisa Conforti

Received: 18 April 2026

Revised: 20 July 2026

Accepted: 29 July 2026

Published: 31 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Incomplete medical data poses a challenge in research. Traditional techniques such as mean imputation, linear interpolation, and extrapolation are commonly used to handle missing values in medical datasets, but they often produce biased or inaccurate estimates. These conventional imputation methods attempt to estimate missing values as entries before model training, potentially resulting in the development of a self-fulfilling model [1]. Some studies in Multiple Sclerosis (MS) research remove missing values instead of imputing

them, simplifying preprocessing but risking the loss of valuable information and reduced predictive accuracy [2]. However, excluding missing values may compromise predictive accuracy. In addition, high-dimensional clinical datasets require effective feature selection to reduce redundancy and improve model performance [3,4].

To address these limitations, more advanced imputation strategies have been developed in recent years. Single imputation methods often fail to reflect uncertainty, leading to biased estimates and reduced reliability [5]. Baker et al. [6] demonstrated that multiple imputation frameworks can reduce bias and improve true signal detection in high-dimensional biomedical data. These approaches outperform single imputation and complete-case analyses by reducing bias, controlling false discoveries, and improving detection of true signals. These limitations of single imputation highlight the need for more advanced approaches, motivating the sequential MI-bootstrapping method used in this study. Pan et al. [7] introduced a multiple imputation framework for addressing missing data in multireader multicase (MRMC) study designs, demonstrating that this approach can provide almost unbiased estimates of diagnostic capability and robust confidence intervals even in small-sample scenarios, thereby enhancing the reliability of statistical conclusions in clinical datasets. Chén et al. used this MI technique during model development. However, the lack of imputation in testing limits the accuracy of prediction results. They recommended exploring advanced imputation methods to ensure reliable test data without artificially inflating performance [8]. Recent advances in machine learning-based imputation have demonstrated improved bias reduction, predictive accuracy, and robustness compared to traditional statistical approaches. For example, Samad et al. [9] combined clustering with deep neural networks for multiple imputation, outperforming traditional regressors. More recently, Ou et al. [10] proposed a hybrid RF+ Generative Adversarial Network imputer tailored for complex missing-data patterns. Although promising, these methods often face limitations, such as dependence on large datasets, potential sensitivity to the mechanism of missingness, or limited external validation, which underlines the motivation for a more structured, robust, and domain-adapted imputation framework like ours.

Valenzuela et al. [11] tackled missing values in their study by using the Amelia R package to impute 10 complete datasets, with 76% of patients originally having missing measurements. This approach involves drawing random values from inferred distributions for imputation and applying “Rubin’s rules” [12] for calculating standard errors. The model’s AUC of 0.80 reflects effective preservation of data relationships, validating the imputation process and its contribution to enhancing prognostic accuracy in clinical outcomes. El Badisy et al. [5] compared machine learning-based imputation methods, such as RF and CART, on a breast cancer survival dataset. ML-based multiple imputation reduced bias, improved confidence interval coverage, and enhanced predictive accuracy, whereas single imputation was more prone to bias. This demonstrates the strong impact of machine learning on preserving data integrity in clinical datasets. In Aghili’s study [13], kNN, Matrix Factorization, Singular Value Decomposition (SVD), Soft Impute, and Gradient Boosting (GB) were applied and achieved the highest performance with accuracies up to 93.40%, a precision up to 92.37%, and a recall up to 91.44% for the imputation. However, these approaches often depend on complex architectures or require extensive supplementary data, limiting their applicability in practical settings. Ouyang et al. [14] developed an ImputeRNN approach that used a mask matrix to address missing-data bias. Tested extensively on two EMR datasets, ImputeRNN outperformed traditional methods like Feature Regression (FR) and Mask GRU (MG), achieving an RMSE of 0.53 and an AUC of 0.89 on Physionet, and a 0.47 RMSE with a 0.91 AUC on MIMIC-III. However, its reliance on a complex RNN architecture may hinder its use in resource-limited settings.

For the MS databases, Pellegrini et al. noted that 44% and 48% of MS patients from the DEFINE and CONFIRM groups, respectively, had missing baseline MRI data. To address this, they used a Monte Carlo Markov chain imputation strategy to create 10 imputed datasets for building prognostic models. They then consolidated the rankings of prognostic factors across these datasets by calculating their median values. Despite these efforts, the overall model performance remained modest (all C-indices ≤ 0.65), highlighting the need for exploring alternative predictors and definitions to enhance prognostic accuracy for MS progression [15].

Despite advances in machine learning-based imputation, existing approaches often rely on assumptions regarding missingness mechanisms or require large, well-structured datasets, limiting their applicability in real-world clinical settings. In this study, we propose a correlation-guided Sequential Multiple Imputation Bootstrapping (SMIB) framework designed to handle heterogeneous and non-random missing data in Multiple Sclerosis datasets. The proposed framework integrates MI, weighted bootstrapping, and structured sequential ordering to enhance stability, preserve clinically meaningful feature relationships, and improve imputation accuracy for both categorical and numerical variables. To comprehensively evaluate the proposed framework, two complementary comparative analyses were performed. First, SMIB was compared with the baseline Multiple Imputation Bootstrapping (MIB) framework to isolate the contribution of the proposed correlation-guided sequential imputation strategy. Second, SMIB was compared with the widely adopted Multiple Imputation by Chained Equations (MICE) approach [16] to evaluate its performance relative to an established MI benchmark commonly used in biomedical research. Together, these comparisons provide both an internal ablation study and an external benchmark evaluation of the proposed methodology.

2. Materials and Methods

2.1. Data Collection and Initial Cleaning

The dataset was sourced from the Sina Hospital in Tehran and was derived from the National Multiple Sclerosis Registry of Iran (NMSRI) [17], ensuring a representative and diverse sample. Initially comprising 8347 records and 62 attributes—including patient demographics (age, gender), symptoms, treatments, risk factors, and health conditions—the data required thorough preprocessing due to its original documentation in Persian. To enhance relevance and quality, domain experts (an epidemiologist and a neurologist) refined the dataset by eliminating non-critical attributes, such as surgery history and residential city.

2.2. Data Preprocessing

Data preprocessing encompassed three critical steps: addressing missing values, standardizing features, and performing feature screening to reduce noise and redundancy. Unlike synthetic simulation studies, the missingness in this dataset was real and not artificially induced, reflecting authentic clinical data collection processes. Following Rubin's framework [12], the missingness mechanisms are discussed conceptually in relation to Missing Completely at Random (MCAR), Missing at Random (MAR), and Not Missing at Random (NMAR) patterns. However, no synthetic MCAR, MAR, or NMAR mechanisms were imposed. All missing values originated from real clinical workflows. Masking was used strictly as an evaluation technique and not to generate artificial missingness. Although these mechanisms were not explicitly modeled, the proposed method was evaluated under real-world missingness conditions rather than idealized simulation-based assumptions. Table 1 summarizes the prevalence of missing data across crucial attributes.

Table 1. Percentages of missing values.

Attribute	% Missing Values
Age of onset of first symptoms of MS	15.75%
Age of diagnosis of MS by a doctor	15.75%
Age of converting to secondary progressive MS	12.44%
Type of MS	16.59%
Expanded Disability Status Scale (EDSS) score	50.98%

Data Scaling and Encoding

To optimize performance in gradient-sensitive models such as MLP, feature scaling was implemented using Min–Max normalization. This approach was selected to preserve bounded feature ranges and maintain consistency across models during training. Categorical features, meanwhile, were preserved through label encoding without scaling to retain their inherent structure. Although RF is insensitive to scaling, normalization was necessary for models like MLP; thus, preprocessing ensured comparability across models [18].

2.3. Model-Based Imputation

2.3.1. Initial Model-Based Approaches

Model-based imputation methods were used to replace missing values across both categorical and numerical variables more accurately than traditional approaches. The following machine learning models were tested:

- *Random Forest (RF)*: An ensemble method effective for high-dimensional data [19].
- *Multilayer Perceptron (MLP)*: A neural network-based method with strong function approximation capabilities [20,21].
- *k-Nearest Neighbors (kNN)*: A distance-based method that estimates missing values based on neighboring data points [22]. Although kNN often shows lower predictive accuracy than RF or MLP, it was intentionally included for its interpretability and sensitivity to local patterns, allowing us to compare whether models with different learning biases preserve relationships after imputation. The performance of these models in imputing missing values was assessed, leading to further refinement of the methodology [23].

2.3.2. Multiple Imputation (MI) and Bootstrapping

To enhance the robustness of missing-data handling and model stability, multiple imputation (MI) and bootstrapping were jointly applied:

- *Multiple Imputation (MI)*: By replacing missing values with reasonable estimates, multiple complete datasets are generated, thereby reducing bias and maintaining statistical variability, which is in line with the established principles of multiple imputation proposed by Donald Rubin [24,25]. The specific number of imputations was not fixed at this stage and was optimized later as part of the proposed methodology.
- *Bootstrapping*: A resampling-based technique was used to generate multiple training datasets for assessing classifier stability and robustness [26]. The number of bootstrap samples was treated as a tunable parameter and finalized during the model-optimization phase.

RF and MLP performed better when combined with bootstrapping, leading to their selection for further refinement, while kNN served as a baseline to evaluate interpretability and local sensitivity.

2.4. Sequential MI Bootstrapping (SMIB)—A Novel Approach

We propose an integration of the MI method with bootstrapping to impute missing data, systematically determining the optimal number of imputations ($m = 5$) and bootstrap samples ($B = 100$) [27,28]. As illustrated in Figure 1, while error metrics (MSE, MAE, RMSE) exhibit minimal sensitivity to parameter variations, the R^2 values distinctly identify this configuration as optimal. Here, “sequential” refers to attribute-by-attribute imputation, ordered by correlation strength, which distinguishes our method from iterative but unordered techniques. The correlation analysis was performed only to determine the sequential order of the target attributes containing naturally occurring missing values. The target attributes were not selected arbitrarily. They correspond to the clinically important outcome variables in the MS registry dataset [17] that also contained substantial naturally occurring missing values. Their clinical relevance was confirmed through consultation with a neurologist and an epidemiologist during dataset preparation, and they were therefore selected as representative targets for evaluating the proposed sequential imputation framework. For each target attribute, Pearson correlation coefficients were computed between each target attribute and all available encoded predictor variables, including both numerical variables and categorical variables after label encoding. Categorical variables were transformed into numerical representations using label encoding solely to enable computation of Pearson correlation coefficients for determining the sequential imputation order. The encoded values were not interpreted as quantitative measurements but served as computational representations for estimating dependency strength. Since correlation analysis was used only to establish the relative ordering of target attributes, while the subsequent imputation was performed using Random Forest, the influence of label encoding on the overall imputation process was limited. We acknowledge that Pearson correlation may not fully capture associations among purely nominal variables; however, it provided a simple and consistent criterion for establishing the sequential imputation order in the present study. The cumulative correlation strength was defined as the sum of the absolute Pearson correlation coefficients between each target attribute and all available encoded predictor variables. Target attributes were subsequently ranked in descending order according to this cumulative score, such that attributes exhibiting stronger overall dependency relationships entered the sequential imputation process earlier. No graph-based traversal algorithm was employed. Instead, each target attribute was ranked independently according to its cumulative correlation strength before sequential imputation. This criterion was selected because it provides a straightforward measure of the overall dependency of each target attribute on the remaining available variables, allowing attributes with stronger overall relationships to be imputed first.

Although the proposed framework shares conceptual similarities with iterative imputation approaches such as MICE and Fully Conditional Specification (FCS) [16,29], the proposed SMIB framework differs in three important aspects. First, the imputation sequence is explicitly guided by correlation-based dependency ranking rather than unordered iterative cycling. Second, weighted bootstrapping is integrated directly into the imputation pipeline to enhance stability under heterogeneous and imbalanced clinical missingness patterns. Third, the framework combines sequential multiple imputation with ensemble-based RF prediction to preserve nonlinear clinical relationships while maintaining robustness across mixed numerical and categorical variables.

Unlike the baseline MIB framework, which applies MI and weighted bootstrapping without considering the order of missing attributes, the proposed SMIB framework performs imputation sequentially according to the correlation strength between variables. This strategy enables previously imputed variables to contribute to subsequent imputations while preserving clinically meaningful dependencies among patient attributes.

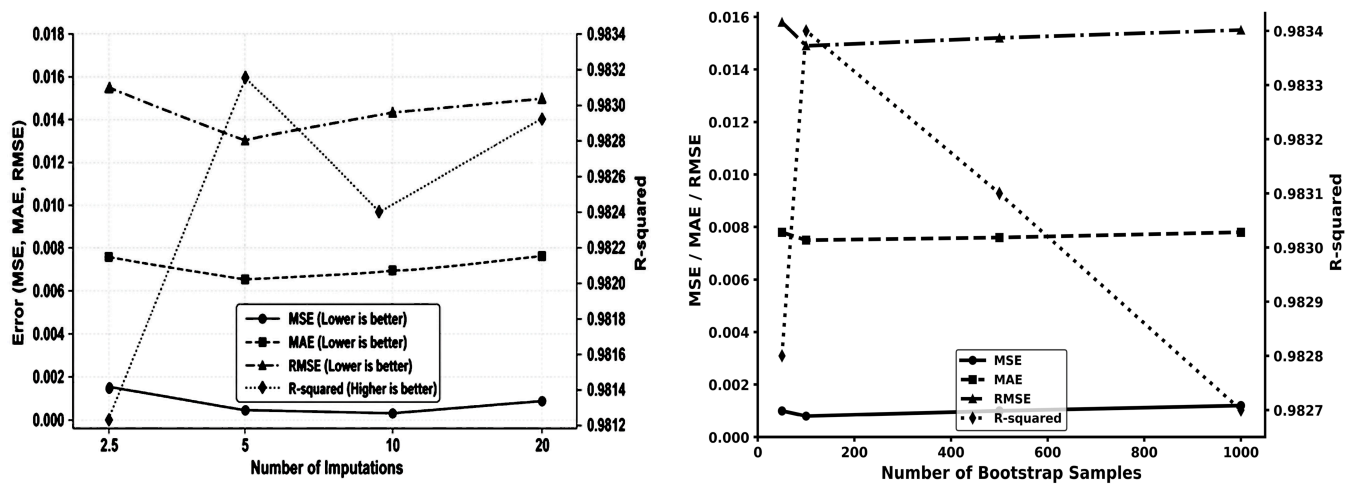


Figure 1. Optimized imputation and bootstrapping numbers.

2.4.1. SMIB Method Steps

To further improve imputation, an innovative approach, sequential MI bootstrapping (SMIB) [30], was developed. The SMIB methodology is described step by step as follows:

- Arrange attributes containing missing values according to their Pearson correlation coefficients to establish a stable and interpretable sequential imputation order. For each target attribute, Pearson correlation was calculated with all available encoded predictor variables. Categorical variables were first transformed into numerical representations through label encoding, allowing them to participate in the dependency analysis together with numerical variables. The resulting correlation scores were used only for ranking the imputation sequence and not for prediction itself.
- Perform feature selection using a hybrid approach (RF importance combined with Mutual Information scores). Random Forest importance captures nonlinear predictive relationships, whereas Mutual Information quantifies the statistical dependency between the candidate predictor variables (input features) and the current target attribute being imputed. Predictor importance was evaluated independently by both methods, and variables identified as important by both approaches were retained as input features for Random Forest imputation [31].
- Apply RF [32] within SMIB to impute the highest-priority attribute in the sequential order.
- Repeat the process until all missing values are imputed.
- Consolidate the results for downstream analysis.

Feature selection plays a vital role in optimizing the imputation process. We employed a combination of RF importance and Mutual Information, which evaluates attribute relevance by capturing both nonlinear relationships and dependencies across categorical and numerical features. By prioritizing impactful attributes and eliminating redundant or irrelevant ones, this approach enhances the precision of imputation results. Unlike the correlation analysis, which determines only the sequential ordering of target attributes, feature selection identifies the most informative predictor variables used by the Random Forest model for imputing the current target attribute. Thus, the ordering stage and predictor-selection stage serve two distinct purposes within the proposed framework. This process was carefully aligned with clinical interpretability, particularly ensuring that EDSS and MS type were never artificially masked but only imputed when naturally missing. A flowchart of the SMIB method is illustrated in Figure 2.

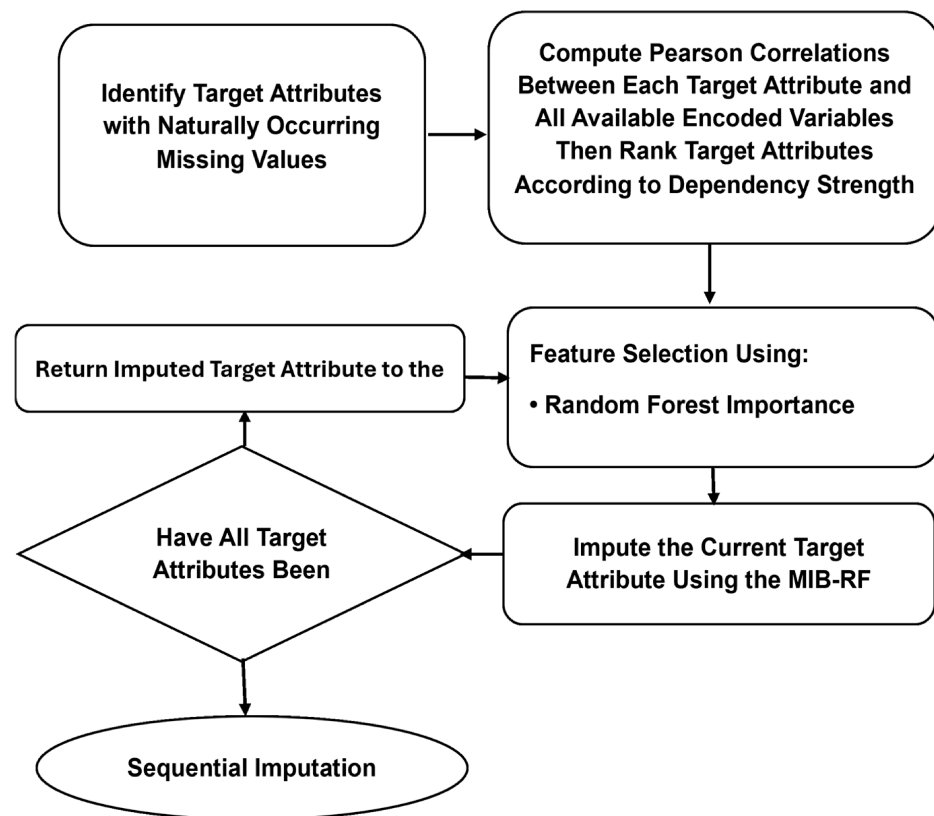


Figure 2. Workflow of the proposed SMIB framework.

The Pearson correlation coefficient [33] is calculated to quantify associations between pairs of variables using the following formula:

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{\sum (x_i - \bar{x}) \sum (y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2} \sqrt{\sum (y_i - \bar{y})^2}} \quad (1)$$

where $\bar{x} = \frac{1}{n} \sum_{i=1}^N x_i$, $\bar{y} = \frac{1}{n} \sum_{i=1}^N y_i$ are means and “cov” represents the covariance between two random variables, X and Y , while σ denotes the standard deviation; these coefficients, which range from -1 to $+1$, indicate the extent to which two variables are linearly related, with positive values suggesting a direct relationship and negative values indicating an inverse relationship. These coefficients aid in optimizing imputation by identifying attribute dependencies. Although Pearson correlation is a linear dependency measure, it was not used as the prediction model. Instead, it provided a practical mechanism for prioritizing the sequence of target attributes. Once the sequential order was established, nonlinear relationships among predictors were learned by the Random Forest model during imputation. In this framework, Pearson correlation was used specifically to guide the sequential ordering of target attributes, while nonlinear dependency modeling was performed during the RF-based imputation and bootstrapping stage.

Figure 2 illustrates the proposed sequential imputation process. Pearson correlation is used only to determine the order in which target attributes containing naturally occurring missing values are imputed. For each target attribute, informative predictors are then selected using Random Forest importance and Mutual Information before Random Forest-based imputation. After imputation, the completed attribute is reintegrated into the dataset and becomes available as a predictor for subsequent imputation steps. The process is repeated until all target attributes with missing values have been imputed.

2.4.2. Weighted Bootstrapping

To enhance imputation accuracy, a weighted bootstrapping technique was implemented. This method assigns higher sampling probabilities to underrepresented data patterns, improving model generalization by prioritizing rare attribute combinations. The weighting scheme is calculated as follows:

$$w_i = \frac{r_i + m_i}{\sum_{j=1}^N (r_j + m_j)} \quad (2)$$

where r_i denotes the rarity score, defined as the reciprocal frequency of the attribute combination in the dataset, and m_i represents the missingness score, defined as the proportion of missing values within record i [34,35].

The attribute combinations were constructed using the predictor variables selected during the feature-selection stage for the current target attribute. Consequently, the rarity score reflects the frequency of clinically relevant predictor combinations rather than all variables in the dataset, preventing high-dimensional sparsity while preserving meaningful patient profiles.

The computed weights are converted into sampling probabilities to direct the resampling process. Next, resampling with replacement is performed, generating diverse datasets while preserving dataset integrity. Finally, models such as MLP and RF are trained and evaluated on the resampled datasets to quantify the imputation's impact. The probability of selecting record i is

$$P_i = \frac{w_i}{\sum_{j=1}^N w_j} \quad (3)$$

2.4.3. SMIB Formula

The imputation estimate θ_{MIB} is derived by combining the results that across all imputed datasets and bootstrap samples, and it is calculated as follows:

$$\theta_{MIB} = \frac{1}{m} \sum_{i=1}^m \left(\frac{1}{B} \sum_{j=1}^B \theta_b^{(j)} \right) \quad (4)$$

where m is the number of imputations (fixed at 5 after preliminary stability analysis using R^2 metrics); B is the number of bootstraps (e.g., 100); and $\theta_b^{(j)}$ represents the prediction obtained from bootstrap j within imputation i [36]. Categorical attributes were numerically encoded and normalized during the RF bootstrapping process to maintain computational consistency. However, for categorical variables such as "Type of MS," the final imputed category was determined using majority voting (mode aggregation) across bootstrap predictions rather than direct arithmetic averaging. Algorithm 1 presents the pseudocode of the proposed Sequential Multiple Imputation Bootstrapping Random Forest (SMIB-RF) framework, illustrating the sequential ordering mechanism, reintegration of previously imputed attributes, and RF-based bootstrapping process used for robust missing-value estimation. This sequential MI bootstrapping approach was designed to balance computational efficiency and predictive robustness, with evaluation focused primarily on predictive accuracy rather than variance-based inference due to clinical dataset constraints.

Algorithm 1. Sequential MI Bootstrapping with Random Forest (SMIB-RF)

```

1:   Input: Dataset  $D$  with missing values
2:   Output: Imputed dataset  $D'$ 
3:   Step 1: Identify Attributes with Missing Values
4:       Identify features  $F$  in  $D$  that contain missing values.
5:   Step 2: Rank Attributes by Correlation
6:       Compute Pearson Correlation Coefficient for determining the sequential
7:       ordering
8:       of missing attributes (in descending order).
9:   Step 3: Feature Selection
10:      Apply Random Forest importance and Mutual Information scores for
11:      feature
12:      selection.
13:      Remove irrelevant features to improve model efficiency.
14:   Step 4: Initialize Bootstrapping and Multiple Imputation
15:      Set number of imputation  $m$  and bootstrap samples  $B$ .
16:   Step 5: Sequential Imputation Process
17:      for each feature  $X_i$  in ranked missing value list do
18:          Identify the top  $k$  most informative available features, including
19:          previously
20:          imputed attributes.
21:          for each bootstrap sample  $b = 1$  to  $B$  do
22:              Create bootstrap sample  $D_b$  by resampling from training
23:              partitions of  $D$ .
24:              Train a Random Forest model  $RF_b$  using complete and
25:              previously imputed
26:              cases in  $D_b$ .
27:              Predict missing values for  $X_i$  using  $RF_b$ .
28:          end for
29:          if  $X_i$  is numerical then
30:              Compute final imputed value as the mean of all  $B$ 
31:              predictions:
32:              
$$X_{imputed} = \frac{1}{B} \sum_{b=1}^B X_b$$

33:          else if  $X_i$  is categorical then
34:              Compute final imputed value using majority voting
35:              across all  $B$  predictions.
36:          else if
37:              Update dataset  $D$  with imputed feature  $X_i$ 
38:          end for
39:   Step 6: Reincorporate Imputed Attributes
40:      Reincorporate newly imputed attributes into  $D$  for downstream
41:      imputations.
42:   Step 7: Termination
43:      if all missing values are filled up then
44:          Return final imputed dataset  $D'$ 
45:      else
46:          Repeat process until convergence.
47:      end if

```

2.5. Model Evaluation and Experimental Design

2.5.1. Hyperparameter Optimization

Hyperparameter tuning was performed for RF and MLP models using cross-validation to optimize predictive performance [37,38]. The optimized hyperparameters are presented in Table 2. The k-Nearest Neighbors (kNN) model was included as a benchmark due to its interpretability and sensitivity to local neighborhood patterns, while RF and MLP demonstrated superior predictive stability.

Table 2. Hyperparameters for the RF and MLP.

Hyperparameter	RF Value	Hyperparameter	MLP Value
<i>n_estimators</i>	200	<i>hidden_layer_sizes</i>	100
<i>max_depth</i>	20	<i>activation</i>	'relu'
<i>min_samples_split</i>	4	<i>solver</i>	'adam'
<i>min_samples_leaf</i>	3	<i>alpha</i>	0.0001
<i>max_features</i>	'sqrt'	<i>learning_rate</i>	'adaptive'

2.5.2. Evaluation Strategy

The imputation methods were evaluated using a masking-based approach combined with 15-fold cross-validation to simulate real-world missing-data conditions. Observed values were deliberately masked during validation to enable objective comparison between imputed and true values under realistic missingness conditions. In this study, accuracy is defined as the agreement between imputed values and the original (masked) values across both categorical and numerical attributes. All correlation ranking, feature selection, bootstrapping, and imputation model fitting procedures were performed exclusively within the training partitions during cross-validation to prevent information leakage into validation or testing data.

2.5.3. Dataset Partitioning

The dataset was partitioned into 80% training data and 20% testing data to ensure fair performance evaluation. The training data were further divided into training and validation subsets using a 75–25 split to support model tuning and validation.

2.5.4. Performance Metrics

Attributes were categorized into categorical and numerical types. EDSS was analyzed as a categorized variable because it represents an ordinal clinical disability scale with discrete score intervals rather than a fully continuous measurement. Categorical attributes were evaluated using classification metrics, including accuracy, precision, recall, F1 score, and confusion matrices. Numerical attributes were assessed using regression metrics, including Mean Squared Error (MSE), Mean Absolute Error (MAE), and the coefficient of determination (R^2). Performance results were reported as averages across multiple imputations to ensure stability. Inferential statistical measures such as confidence intervals were not included in this study and remain an important direction for future work.

2.6. Comparative Evaluation

To evaluate the proposed methodology, two complementary comparative analyses were conducted. First, the proposed SMIB framework was compared with the baseline MIB framework. Since both frameworks employ identical multiple imputation, weighted bootstrapping, feature selection, and prediction models, differing only in the inclusion of

the proposed correlation-guided sequential imputation strategy, this comparison serves as an ablation study that isolates the contribution of the sequential component.

Second, the proposed SMIB framework was compared with the widely adopted Multiple Imputation by Chained Equations (MICE) framework, one of the most commonly used MI methods in biomedical research. Both approaches were evaluated using identical datasets, prediction models, and performance metrics to provide a fair and objective comparison of the proposed methodology relative to an established benchmark.

3. Results

This section presents the key quantitative findings from the proposed imputation methods for predicting clinical factors in MS.

3.1. Numerical Imputation Performance Results

A comparative analysis of numerical imputation methods is detailed in Table 3. All reported performance values represent averages across 5 imputations and 100 bootstrap runs to ensure stability and reduce randomness.

Table 3 summarizes the performance of KNN, RF, MLP, and techniques such as MLP and RF with MI bootstrapping, focusing on MAE, MSE, and R^2 metrics. Averaged performance metrics were used to quantify overall model effectiveness.

Table 3. Features comparison of different imputation methods.

Imputation Method	Measured Method	Age of MSFS	Age of DMSD	Age of SPMS
SMIB with MLP	MSE	0.00425	0.00088	0.00042
	MAE	0.04746	0.01645	0.00704
	R^2	0.7025	0.9339	0.9845
	RMSE	0.06518	0.02964	0.02039
SMIB with RF	MSE	0.00288	0.00018	0.00003
	MAE	0.03796	0.00615	0.00101
	R^2	0.7997	0.9871	0.9990
	RMSE:	0.05364	0.01346	0.00516
RF	MSE	0.00525	0.00129	0.00017
	MAE	0.05573	0.02011	0.00255
	R^2	0.6704	0.9122	0.9927
	RMSE	0.07248	0.03589	0.01285
MLP	MSE	0.00530	0.00126	0.00022
	MAE	0.05716	0.02036	0.00769
	R^2	0.6675	0.9141	0.9901
	RMSE	0.07279	0.03550	0.01496
KNN	MSE	0.00531	0.00141	0.00134
	MAE	0.05673	0.02692	0.01424
	R^2	0.6665	0.9041	0.9406
	RMSE	0.07290	0.03751	0.03657

Abbreviations: SPMS = secondary progressive MS; MSFS = MS first symptoms; DMSD = diagnosis of MS by a doctor.

3.2. Categorical Imputation Performance Results

Table 4 analyzes the accuracy, precision, recall, and F1 score of five imputation techniques for two categorical attributes, Categorized EDSS and Type of MS, focusing on their

effectiveness. Observed values were masked only for evaluation purposes. All naturally missing values were preserved to maintain clinical integrity.

Table 4. Various imputation techniques for categorical features.

Imputation Methods	Categorized EDSS	Type of MS
MI Bootstrapping with MLP	Accuracy: 0.91, Precision: 0.89, Recall: 0.91, F1 Score: 0.90	Accuracy: 0.92, Precision: 0.90, Recall: 0.92, F1 Score: 0.90
MI Bootstrapping with RF	Accuracy: 0.93, Precision: 0.92, Recall: 0.93, F1 Score: 0.92	Accuracy: 0.96, Precision: 0.96, Recall: 0.96, F1 Score: 0.96
RF	Accuracy: 0.93, Precision: 0.91, Recall: 0.93, F1 Score: 0.91	Accuracy: 0.91, Precision: 0.90, Recall: 0.91, F1 Score: 0.90
MLP	Accuracy: 0.92, Precision: 0.91, Recall: 0.92, F1 Score: 0.91	Accuracy: 0.91, Precision: 0.90, Recall: 0.91, F1 Score: 0.90
KNN	Accuracy: 0.91, Precision: 0.90, Recall: 0.91, F1 Score: 0.90	Accuracy: 0.91, Precision: 0.89, Recall: 0.91, F1 Score: 0.90

3.3. Model Comparison and Visualization

Figure 3 presents heatmaps of confusion matrices for MIB-RF and MIB-MLP imputation methods, comparing their classification performance on two categorical attributes: Categorized EDSS and Type of MS. Both matrices show higher classification accuracy for the predominant class (Class 2), while rare classes (Classes 0 and 1) exhibit higher misclassification rates.



Figure 3. Confusion matrices comparing RF- and MLP-based imputation performance.

Figure 4 presents scatter plots of imputed versus actual values for the RF- and MLP-based bootstrapping methods, illustrating the agreement between the predicted and observed values.

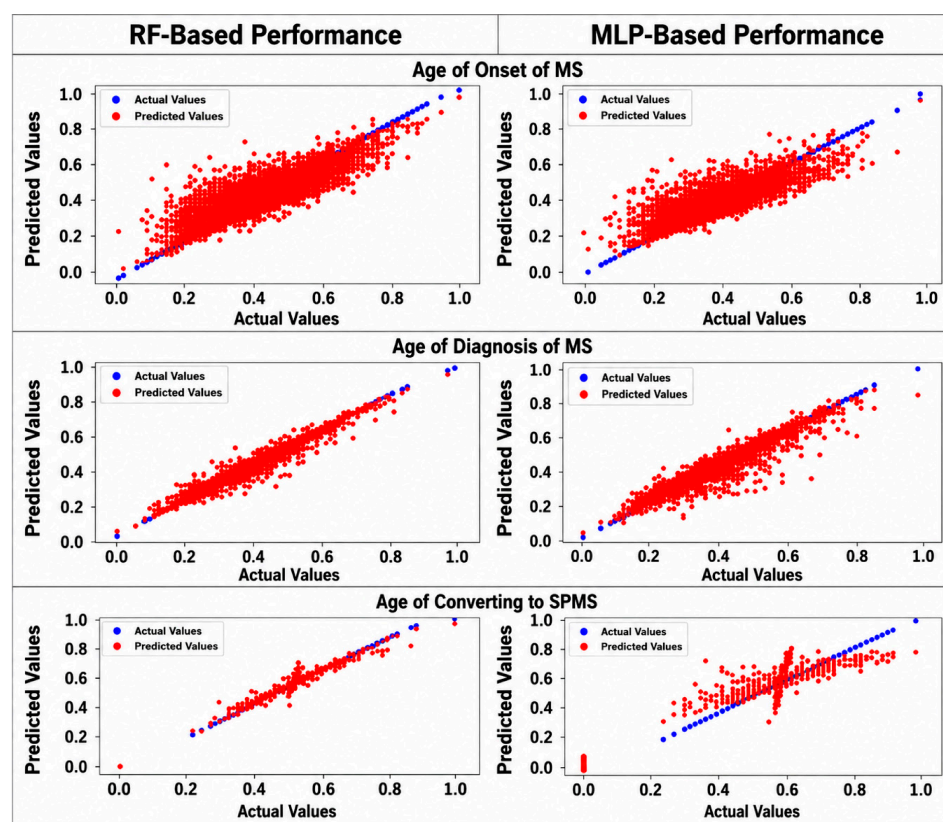


Figure 4. Scatter plots of imputed versus actual values using RF- and MLP-based performance.

Figure 5 illustrates the correlation-guided sequential imputation order used within our framework. The diagram presents the sequence in which attributes containing missing values were imputed according to their dependency ranking, enabling progressively refined imputations throughout the SMIB process.

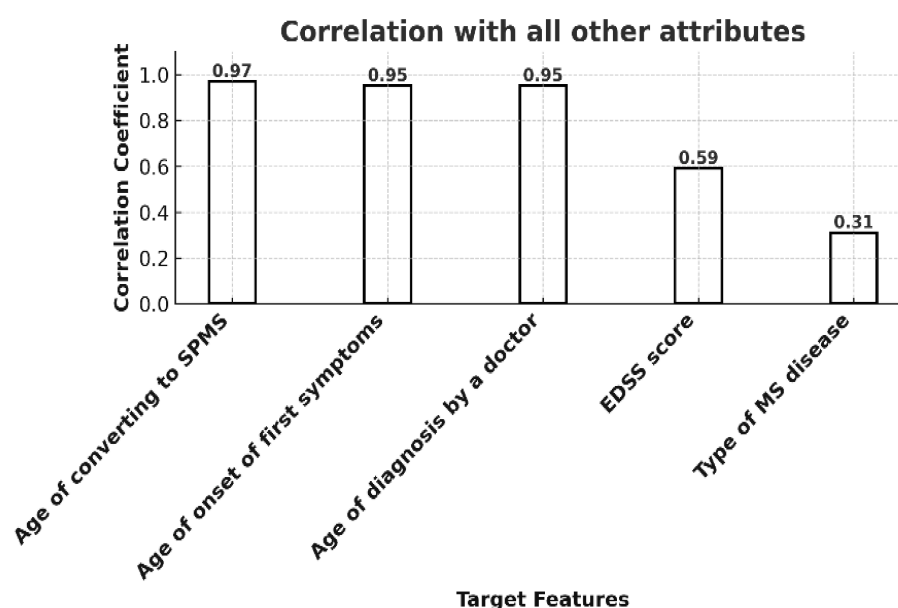


Figure 5. Correlation-guided sequential imputation order within the SMIB framework.

3.4. Comparative Analysis of RF-Based Strategies

To evaluate the contribution of the proposed correlation-guided sequential imputation strategy, a comparative analysis was conducted between the baseline Multiple Imputation Bootstrapping Random Forest (MIB-RF) framework and the proposed SMIB-RF framework. Since both frameworks employ identical MI, weighted bootstrapping, feature selection, and Random Forest prediction procedures, differing only in the proposed sequential feature dependency imputation strategy, this comparison isolates the contribution of the sequential component and therefore serves as an ablation analysis of the proposed methodology.

A comprehensive assessment of the effectiveness of two RF strategies, MIB-RF and sequential MI bootstrapping RF (SMIB-RF), is presented in Table 5 across categorical and numerical attributes. Table 5 presents the performance comparison across categorical and numerical MS-related variables.

Table 5. Ablation analysis comparing MIB-RF and the proposed SMIB-RF framework.

Methods	Categ. EDSS (MIB-RF)	Type of MS (MIB-RF)	Age of Onset (MIB-RF)	Age of Dx. (MIB-RF)	Age of Conv. (MIB-RF)	Categ. EDSS (SMIB-RF)	Type of MS (SMIB-RF)	Age of Onset (SMIB-RF)	Age of Dx. (SMIB-RF)	Age of Conv. (SMIB-RF)
Best Accu.	0.929	0.955	-	-	-	0.928	0.971	-	-	-
Best Preci.	0.918	0.955	-	-	-	0.9189	0.970	-	-	-
Best Recall	0.929	0.955	-	-	-	0.928	0.971	-	-	-
Best F1	0.917	0.955	-	-	-	0.9217	0.970	-	-	-
Best MSE	-	-	0.00288	0.00018	2.67×10^{-5}	-	-	0.00306	0.00024	2.48×10^{-5}
Best MAE	-	-	0.03912	0.00615	0.00101	-	-	0.03796	0.0075	0.00096
Best R ²	-	-	0.7842	0.9871	0.999	-	-	0.7997	0.9824	0.9992

Abbreviations: Categ. = categorized; Dx. = diagnosis; Conv. = conversion; Accu. = accuracy; Preci. = precision.

3.5. Comparison with an Established Multiple Imputation Method

To assess the practical effectiveness of the proposed methodology relative to an established MI approach, the proposed Sequential Multiple Imputation Bootstrapping (SMIB) framework was compared with MICE, one of the most widely adopted MI techniques in biomedical research. Since both methods were evaluated using identical datasets, prediction models, and performance metrics, this comparison provides a fair and objective assessment of the practical effectiveness of the proposed framework relative to a widely accepted benchmark. The comparative results are summarized in Table 6.

The comparison presented in Table 6 demonstrates that the proposed SMIB framework achieved superior predictive performance relative to the conventional MICE framework across the evaluated categorical and numerical prediction tasks. The observed improvements are consistent with the methodological characteristics of the proposed SMIB framework described in Section 2.4. Specifically, the combination of correlation-guided sequential ordering and weighted bootstrapping enables clinically related variables to be imputed earlier while improving robustness under heterogeneous missingness conditions. As shown in Table 6, these design characteristics resulted in higher predictive accuracy for the categorical variables together with lower prediction errors and higher R² values for the numerical variables. Although the magnitude of improvement varied among the evaluated variables, the proposed SMIB framework consistently achieved better performance than the conventional MICE approach across all reported evaluation metrics. These findings indicate that the observed performance gains arise not from the use of multiple imputation alone, but from its integration with correlation-guided sequential ordering and weighted bootstrapping within the proposed SMIB framework.

Table 6. Performance comparison between the conventional MICE framework and the proposed SMIB framework.

Categorical Variables						
Metric	Categorized EDSS		Type of MS			
	MICE	SMIB	MICE	SMIB		
Accuracy	0.9145	0.9280	0.8923	0.9710		
Precision	0.9040	0.9189	0.8743	0.9700		
Recall	0.9145	0.9280	0.8923	0.9710		
F1 Score	0.9078	0.9217	0.8797	0.9700		
Numerical Variables						
Metric	Age of Onset		Age of Diagnosis		Age of Converting to SPMS	
	MICE	SMIB	MICE	SMIB	MICE	SMIB
MAE	0.0608	0.0391	0.0226	0.0075	0.0028	0.0010
RMSE	0.0816	0.0553	0.0409	0.0155	0.0150	0.0050
R ²	0.5530	0.7842	0.8797	0.9824	0.9894	0.9992

4. Discussion

This study demonstrates that integrating correlation-guided sequential ordering with multiple imputation and bootstrapping improves both numerical and categorical imputation performance in real-world MS datasets. Unlike simulation-based studies, all evaluations were conducted on authentic clinical data with naturally occurring missingness, enhancing the practical relevance and robustness of the findings. The proposed SMIB framework provides a structured and clinically interpretable mechanism for preserving inter-feature dependencies while improving stability under real-world missingness conditions.

4.1. Numerical Imputation Performance Discussion

As shown in Table 3, SMIB-RF consistently achieved lower MSE and MAE and higher R² values across all continuous outcomes, indicating improved numerical stability and accuracy. These results highlight the effectiveness of combining multiple imputation with ensemble-based bootstrapping in capturing nonlinear dependencies within clinical data. The observed improvement in R² further suggests that feature dependency sequential ordering enhances predictive consistency beyond conventional imputation pipelines. This is supported by the scatter plots in Figure 4, where RF-based SMIB demonstrates tighter alignment between imputed and observed values compared with MLP-based approaches.

4.2. Categorical Imputation Performance Discussion

For categorical variables, Table 4 shows that combining MI with bootstrapping also improves classification outcomes relative to standalone RF, MLP, or kNN. MIB-RF yielded the strongest accuracy, precision, recall, and F1 scores for both Categorized EDSS and Type of MS, demonstrating the robustness of ensemble learners in capturing complex, nonlinear relationships within mixed clinical attributes.

The confusion matrices in Figure 3 further illustrate that RF-based imputers maintain higher classification performance across dominant classes while exhibiting expected challenges in minority classes. This behavior reflects inherent class imbalance within clinical datasets rather than model instability. The controlled masking strategy ensured that evaluation remained unbiased by preserving the original missingness structure while enabling reliable performance assessment.

4.3. Role of Correlation-Guided Sequential Ordering

Figure 5 illustrates the correlation-guided sequential imputation order used within the SMIB framework. This hierarchical strategy ensured that variables with stronger dependency relationships were imputed first, enabling downstream targets to benefit from progressively refined prior imputations. By incorporating sequential dependency ranking into the imputation process, the proposed framework improves coherence and stability across both categorical and numerical attributes. This structured imputation strategy differentiates SMIB from traditional iterative methods by explicitly leveraging feature dependency relationships to support clinically consistent imputations.

4.4. SMIB vs. MIB-RF

Table 5 compares the proposed SMIB-RF framework with the non-sequential MIB-RF configuration. Overall, incorporating correlation-guided sequential ordering provided incremental performance gains, with the greatest improvements observed for Type of MS and selected numerical prediction tasks, while maintaining competitive performance for the remaining variables. These findings demonstrate the added value of incorporating correlation-guided sequential ordering into the proposed framework.

Since both MIB-RF and SMIB-RF employ identical multiple imputation, weighted bootstrapping, feature selection, and Random Forest prediction procedures, differing only in the inclusion of the proposed correlation-guided sequential imputation strategy, the comparison presented in Table 5 serves as an ablation analysis that isolates the contribution of the proposed sequential component. Overall, the findings indicate that embedding RF within a multiple imputation and bootstrapping framework yields stable and accurate imputations. More importantly, the integration of correlation-guided sequential ordering enhances structural coherence and improves performance consistency across both numerical and categorical variables, reinforcing the practical value of the proposed SMIB framework in real-world clinical applications.

The comparison with the conventional MICE framework (Table 6) further confirms that the observed performance improvements are attributable to the proposed methodology rather than to the use of MI alone. While MICE represents one of the most widely adopted benchmark methods for multiple imputation, the proposed SMIB framework consistently achieved superior performance across all evaluated categorical and numerical outcomes. These improvements can be attributed to the combination of correlation-guided sequential ordering and weighted bootstrapping, which together preserve clinically meaningful feature dependencies while improving model robustness under complex clinical missingness.

Despite these promising results, several limitations should be acknowledged. The study was primarily conducted on a single registry dataset, which may limit direct generalizability to other clinical populations. However, the dataset reflects real-world clinical conditions with heterogeneous and non-random missingness, supporting the practical relevance of the findings. Future work will extend the proposed methodology to additional datasets and multi-center studies to further validate its robustness and generalizability.

Although the present study compares the proposed framework with the widely adopted MICE approach, additional comparisons with other established imputation methods, including missForest and other established machine learning-based imputation methods [32], together with broader statistical analyses such as confidence intervals, remain valuable directions for future research.

5. Conclusions

This study introduced a sequential MI bootstrapping (SMIB) framework for handling missing values in MS datasets by integrating MI with structured bootstrapping and feature dependency sequencing. The RF-based implementation demonstrated strong performance for both categorical and numerical variables, reflecting improved stability, coherence, and reliability under complex and heterogeneous missingness patterns. Ensemble learning and weighted bootstrapping further enhanced generalizability, emphasized underrepresented clinical patterns, and reduced overfitting, while the sequential ordering preserved clinically meaningful feature relationships throughout the imputation process.

Although kNN and MLP produced competitive results, the findings underscored the advantages of tree-based ensembles in capturing nonlinear dependencies and maintaining robust performance across diverse clinical attributes. The novelty of the SMIB framework lies in its systematic integration of established methods into a unified, domain-specific pipeline that enhances imputation quality while preserving clinically meaningful feature relationships. The comparative analyses further demonstrated that incorporating the proposed correlation-guided sequential imputation strategy contributes directly to the observed performance improvements over the baseline MIB framework. Moreover, comparison with the widely adopted MICE approach demonstrated the practical effectiveness and competitive performance of the proposed SMIB framework for handling heterogeneous missing data in real-world clinical MS datasets.

Future work will extend this framework through variance-based inference, multi-center validation, and additional comparisons with other established imputation methods, including missForest, to further evaluate the robustness and generalizability of the SMIB across diverse clinical datasets.

Author Contributions: Conceptualization, S.P. and W.P.; methodology S.P.; software, S.P.; validation, S.P., W.P., M.A.S. and S.E.; formal analysis, S.P.; investigation, S.P., W.P., M.A.S. and S.E.; resources, M.A.S. and S.E.; data curation, S.P.; writing—original draft preparation, S.P.; writing—review and editing, W.P., M.A.S. and S.E.; visualization, S.P., W.P., M.A.S. and S.E.; supervision, W.P.; project administration, W.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: The source registry (NMSRI) obtained written informed consent from all patients. As described in detail by Ayoubi et al. [39], all patients enrolled in the registry provided written informed consent and were informed that their anonymized data would be used for research purposes. This study falls within the scope of that consent.

Data Availability Statement: Restrictions apply to the availability of these data. The data were obtained from the National Multiple Sclerosis Registry of Iran (NMSRI) and are not publicly available due to privacy and ethical restrictions. Access to the data may be granted upon reasonable request to the corresponding author and with permission from the data-providing institution.

Acknowledgments: The authors would like to thank Shamim Pilehvari for assistance in facilitating collaboration with the national MS registry of Iran.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Zhao, Y.; Berretta, M.; Wang, T.; Chitnis, T. GRU-DF: A Temporal Model with Dynamic Imputation for Missing Target Values in Longitudinal Patient Data. In Proceedings of the 2020 IEEE International Conference on Healthcare Informatics (ICHI), Oldenburg, Germany, 30 November–3 December 2020; pp. 1–7.

2. Zhao, Y.; Wang, T.; Bove, R.; Cree, B.; Henry, R.; Lokhande, H.; Polgar-Turcsanyi, M.; Anderson, M.; Bakshi, R.; Weiner, H.L.; et al. Ensemble Learning Predicts Multiple Sclerosis Disease Course in the SUMMIT Study. *npj Digit. Med.* **2020**, *3*, 135. [[CrossRef](#)] [[PubMed](#)]
3. Shadbahr, T.; Roberts, M.; Stanczuk, J.; Gilbey, J.; Teare, P.; Dittmer, S.; Thorpe, M.; Torné, R.V.; Sala, E.; Lió, P. The Impact of Imputation Quality on Machine Learning Classifiers for Datasets with Missing Values. *Commun. Med.* **2023**, *3*, 139. [[CrossRef](#)] [[PubMed](#)]
4. Pudjihartono, N.; Fadason, T.; Kempa-Liehr, A.W.; O'Sullivan, J.M. A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction. *Front. Bioinform.* **2022**, *2*, 927312. [[CrossRef](#)] [[PubMed](#)]
5. El Badisy, I.; Graffeo, N.; Khalis, M.; Giorgi, R. Multi-Metric Comparison of Machine Learning Imputation Methods with Application to Breast Cancer Survival. *BMC Med. Res. Methodol.* **2024**, *24*, 191. [[CrossRef](#)] [[PubMed](#)]
6. Baker, B.H.; Sathyanarayana, S.; Szpiro, A.A.; MacDonald, J.W.; Paquette, A.G. RNAseqCovarImpute: A Multiple Imputation Procedure That Outperforms Complete Case and Single Imputation Differential Expression Analysis. *Genome Biol.* **2024**, *25*, 236. [[CrossRef](#)] [[PubMed](#)]
7. Pan, Z.; Qin, Y.; Bai, W.; He, Q.; Yin, X.; He, J. Implementing Multiple Imputations for Addressing Missing Data in Multireader Multicase Design Studies. *BMC Med. Res. Methodol.* **2024**, *24*, 217. [[CrossRef](#)] [[PubMed](#)]
8. Chén, O.Y.; Lipsmeier, F.; Phan, H.; Dondelinger, F.; Creagh, A.; Gossens, C.; Lindemann, M.; de Vos, M. Personalized Longitudinal Assessment of Multiple Sclerosis Using Smartphones. *IEEE J. Biomed. Health Inform.* **2023**, *27*, 3633–3644. [[CrossRef](#)] [[PubMed](#)]
9. Samad, M.D.; Abrar, S.; Diawara, N. Missing Value Estimation Using Clustering and Deep Learning within Multiple Imputation Framework. *Knowl. Based Syst.* **2022**, *249*, 108968. [[CrossRef](#)] [[PubMed](#)]
10. Ou, H.; Yao, Y.; He, Y. Missing Data Imputation Method Combining Random Forest and Generative Adversarial Imputation Network. *Sensors* **2024**, *24*, 1112. [[CrossRef](#)] [[PubMed](#)]
11. Valenzuela, R.M.; Kaufman, M.; Balashov, K.E.; Ito, K.; Buyske, S.; Dhib-Jalbut, S. Predictive Cytokine Biomarkers of Clinical Response to Glatiramer Acetate Therapy in Multiple Sclerosis. *J. Neuroimmunol.* **2016**, *300*, 59–65. [[CrossRef](#)] [[PubMed](#)]
12. Rubin, D.B. Multiple Imputation. In *Flexible Imputation of Missing Data*, 2nd ed.; Chapman and Hall/CRC: Boca Raton, FL, USA, 2018; pp. 29–62.
13. Aghili, M.; Tabarestani, S.; Adjouadi, M. Addressing the Missing Data Challenge in Multi-Modal Datasets for the Diagnosis of Alzheimer's Disease. *J. Neurosci. Methods* **2022**, *375*, 109582. [[CrossRef](#)] [[PubMed](#)]
14. Ouyang, J.; Zhang, Y.; Cai, X.; Zhang, Y.; Yuan, X. ImputeRNN: Imputing Missing Values in Electronic Medical Records. In *Proceedings of the Database Systems for Advanced Applications*; Jensen, C.S., Lim, E.-P., Yang, D.-N., Lee, W.-C., Tseng, V.S., Kalogeraki, V., Huang, J.-W., Shen, C.-Y., Eds.; Springer International Publishing: Cham, Switzerland, 2021; pp. 413–428.
15. Pellegrini, F.; Copetti, M.; Sormani, M.P.; Bovis, F.; de Moor, C.; Debray, T.P.A.; Kieseier, B.C. Predicting Disability Progression in Multiple Sclerosis: Insights from Advanced Statistical Modeling. *Mult. Scler. J.* **2020**, *26*, 1828–1836. [[CrossRef](#)] [[PubMed](#)]
16. Groothuis-oudshoorn, K. mice: Multivariate Imputation by Chained. *J. Stat. Softw.* **2011**, *45*, 1–67. [[CrossRef](#)]
17. Shahin, S.; Eskandarieh, S.; Moghadasi, A.N.; Razazian, N.; Baghbanian, S.M.; Ashtari, F.; Manouchehrinia, A.; Beiki, O.; Mohebi, F.; Dezfouli, M.M. Multiple Sclerosis National Registry System in Iran: Validity and Reliability of a Minimum Data Set. *Mult. Scler. Relat. Disord.* **2019**, *33*, 158–161. [[CrossRef](#)] [[PubMed](#)]
18. Al Bataineh, A.; Manacek, S. MLP-PSO Hybrid Algorithm for Heart Disease Prediction. *J. Pers. Med.* **2022**, *12*, 1208. [[CrossRef](#)] [[PubMed](#)]
19. Jackson, K.C.; Sun, K.; Barbour, C.; Hernandez, D.; Kosa, P.; Tanigawa, M.; Weideman, A.M.; Bielekova, B. Genetic Model of [MS] Severity Predicts Future Accumulation of Disability. *Ann. Hum. Genet.* **2019**, *84*, 1–10. [[CrossRef](#)] [[PubMed](#)]
20. Seo, W.; Khairulov, T.; Baek, H.-J.; Lee, J. ChainImputer: A Neural Network-Based Iterative Imputation Method Using Cumulative Features. *Symmetry* **2025**, *17*, 869. [[CrossRef](#)]
21. Balazs, A.; Liski, E.; Tuominen, S.; Kangas, A. Comparison of Neural Networks and K-Nearest Neighbors Methods in Forest Stand Variable Estimation Using Airborne Laser Data. *ISPRS Open J. Photogramm. Remote Sens.* **2022**, *4*, 100012. [[CrossRef](#)]
22. Halder, R.K.; Uddin, M.N.; Uddin, M.A.; Aryal, S.; Khraisat, A. Enhancing K-Nearest Neighbor Algorithm: A Comprehensive Review and Performance Analysis of Modifications. *J. Big Data* **2024**, *11*, 113. [[CrossRef](#)]
23. Joel, L.O.; Doorsamy, W.; Paul, B.S. A Comparative Study of Imputation Techniques for Missing Values in Healthcare Diagnostic Datasets. *Int. J. Data Sci. Anal.* **2025**, *20*, 6357–6373. [[CrossRef](#)]
24. Awounvo, S.; Kieser, M.; Feißt, M. Combining Multiple Imputation with Internal Model Validation in Clinical Prediction Modeling: A Systematic Methodological Review. *J. Clin. Epidemiol.* **2025**, *186*, 111916. [[CrossRef](#)] [[PubMed](#)]
25. Fichman, M.; Cummings, J.N. Multiple Imputation for Missing Data: Making the Most of What You Know. *Organ. Res. Methods* **2003**, *6*, 282–308. [[CrossRef](#)]
26. Efron, B. Bootstrap Methods: Another Look at the Jackknife. *Ann. Stat.* **1979**, *7*, 1–26. [[CrossRef](#)]
27. Heymans, M.W.; Van Buuren, S.; Knol, D.L.; Van Mechelen, W.; De Vet, H.C.W. Variable Selection under Multiple Imputation Using the Bootstrap in a Prognostic Study. *BMC Med. Res. Methodol.* **2007**, *7*, 33. [[CrossRef](#)] [[PubMed](#)]

28. Singh, S.; Sedory, S.A. Sufficient Bootstrapping. *Comput. Stat. Data Anal.* **2011**, *55*, 1629–1637. [CrossRef]
29. Demirtas, H. Flexible Imputation of Missing Data. *J. Stat. Softw.* **2018**, *85*, 1–5. [CrossRef]
30. Robbins, M.W.; Burgette, L. Resampling Methods with Multiply Imputed Data. *Biometrika* **2025**, *112*, asaf059. [CrossRef] [PubMed]
31. Alduailij, M.; Khan, Q.W.; Tahir, M.; Sardaraz, M.; Alduailij, M.; Malik, F. Machine-Learning-Based DDoS Attack Detection Using Mutual Information and Random Forest Feature Importance Method. *Symmetry* **2022**, *14*, 1095. [CrossRef]
32. Aracri, F.; Bianco, M.G.; Quattrone, A.; Sarica, A. Imputation of Missing Clinical, Cognitive and Neuroimaging Data of Dementia Using MissForest, a Random Forest Based Algorithm. In Proceedings of the 2023 IEEE 36th International Symposium on Computer-Based Medical Systems (CBMS), L'Aquila, Italy, 22–24 June 2023; pp. 684–688.
33. Zhou, H.; Deng, Z.; Xia, Y.; Fu, M. A New Sampling Method in Particle Filter Based on Pearson Correlation Coefficient. *Neurocomputing* **2016**, *216*, 208–215. [CrossRef]
34. Barbe, P.; Bertail, P. *The Weighted Bootstrap*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2012; Volume 98, ISBN 1461225329.
35. Greifer, N. Estimating Effects After Weighting 2023. Available online: <https://github.com/ngreifer/WeightIt/blob/master/vignettes/estimating-effects.Rmd> (accessed on 15 April 2026).
36. Schomaker, M.; Heumann, C. Bootstrap Inference When Using Multiple Imputation. *Stat. Med.* **2018**, *37*, 2252–2266. [CrossRef] [PubMed]
37. Hidayaturohman, Q.A.; Hanada, E. A Comparative Analysis of Hyper-Parameter Optimization Methods for Predicting Heart Failure Outcomes. *Appl. Sci.* **2025**, *15*, 3393. [CrossRef]
38. Ali, H.; Muthudoss, P.; Chauhan, C.; Kaliappan, I.; Kumar, D.; Paudel, A.; Ramasamy, G. Machine Learning-Enabled NIR Spectroscopy. Part 3: Hyperparameter by Design (HyD) Based ANN-MLP Optimization, Model Generalizability, and Model Transferability. *AAPS PharmSciTech* **2023**, *24*, 254. [CrossRef] [PubMed]
39. Ayoubi, S.; Asadigandomani, H.; Bafrani, M.A.; Shirkoohi, A.; Nasiri, M.; Sahraian, M.A.; Eskandarieh, S. The National Multiple Sclerosis Registry System of Iran (NMSRI): Aspects and methodological. *Mult. Scler. Relat. Disord.* **2023**, *72*, 104610. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.