Advances in Nonlinear Variational Inequalities
ISSN: 1092-910X
Vol 28 No. 3s (2025)

A Comparative Analysis of Machine Learning Imputation Techniques for MAR Missingness

Shweta Tiwaskar[1], Sandip Thite[1], Rashid Mamoon[1]

[1]Department of Computer Engineering, Faculty of Science and Technology, Vishwakarma University, Pune, India.

Article History:

Received: *22-09-2024*

Revised: *17-11-2024*

Accepted: *26-11-2024*

Abstract:

Electronic health records (EHR) are essential for making informed patient care decisions., but missing data can hinder decision-making. This study addresses the issue of missing data, specifically under the Missing at Random (MAR) mechanism, which is common in real-world datasets. While statistical methods are traditionally used for data imputation, machine learning (ML) approaches offer greater flexibility and can capture complex relationships within the data. The paper evaluates three prominent ML-based imputation techniques—K Nearest Neighbor Imputation (KNNI), Multivariate Imputation by Chained Equations (MICE), and MissForest—focusing on their performance in handling MAR missingness in multivariate configurations. The study simulates MAR missingness (5%-30% of the dataset) across multiple variables and imputes the missing values using these methods. The imputed datasets are evaluated against a complete subset of the original data using several performance metrics e.g. (accuracy, F1 score, MAE, RMSE, R-squared, Pearson correlation, and BIC etc.). particularly examining correlations between missing and observed values. To calculate these performance metrics, eighteen imputed datasets are compared with one complete subset of original dataset. As compared to KNNI and MICE, MissForest imputation method demonstrated reduced SD, MAE, and RMSE in 83.33% of MR cases, and higher R-squared values in all (100%) MR cases. MissForest performs better in 100% of MR cases in all the five performance metrics of model performance. This suggests that MissForest is a superior imputation method for handling MAR missingness in multivariate settings.

Keywords: Multivariate Imputation, Machine Learning, MissForest, Data Quality, MAR Missingness.

1. Introduction

Complete and accurate electronic health records (EHR) are essential for making informed patient care decisions and reducing medical errors, but the handling of missing data and overall data quality are equally important factors. Missing data can occur due to human error, unrecorded attributes, and other factors. Large hospital departments, especially those monitoring chronic diseases like diabetes, frequently encounter datasets with significant missing data. To address this issue, researchers use methods like Listwise Deletion, Pairwise Deletion, and Imputation [1]. Listwise Deletion removes entire cases with missing data, while Pairwise Deletion only removes missing values for specific variables in an observation unit [2]. Deletion methods can significantly reduce sample sizes, making

them less recommended, especially in healthcare domain as the observed data is too important to be discarded. Imputation, considered the best method, fills missing values with estimated ones based on available data. Imputation methods are divided into statistical-based and ML-based methods. Statistical based replace missing data with values like mean, median, mode. ML-based method uses predictive models to estimate and fill missing values. ML methods are more flexible than statistical and can capture higher order interaction between data [3]. This study will compare three ML based methods KNNI, MICE and MissForest for handling missing data in medical dataset. Missing data has three types: Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR). In MCAR, the likelihood of a missing value for a variable is unrelated to any other variable in the dataset. In MAR, missingness is not related to the variable itself but may be influenced by other variables in the dataset. In MNAR, the likelihood of data being missing is influenced by the values of the missing data itself [4]. Most existing imputation algorithms are based on the MCAR assumption, but real-life missingness is generally of MAR or MNAR type. However, research on MAR and MNAR imputation is limited. Simulating the MAR mechanism involves selecting the most correlated complete feature to determine missingness, but this selection is complicated by the distribution of missing values and feature correlations, which vary across datasets. When many features have missing values, fewer complete features are available, reducing the likelihood of finding a highly correlated one. This variability and complexity make MAR simulation challenging and highlights the need for further research on the different correlations between missing and observed values in MAR cases [5]. Addressing MAR missingness is more critical as it reflects more realistic data conditions in healthcare settings [6]. A significant disparity exists as the majority of these studies concentrate on the more common MCAR scenario, with fewer addressing the more complex cases of MAR and MNAR [7]. MissForest imputation is a tree-based method that employs random forests as the prediction model. The stopping criterion for MissForest is reached when the average difference between the newly imputed data and the previously imputed data increases for the first time. A more in-depth discussion is available in [2]. MICE is a widely used technique for addressing missing data, also referred to as fully conditional specification. MICE estimates missing values multiple times (m times), It does this in steps: for each variable with missing data, it uses information from other variables to guess the missing values. This process is repeated m times generating m imputed datasets. After analyzing each complete dataset, MICE combine the results into one complete imputed dataset, considering the types of attributes with missing values in the dataset [8]. The MICE process involves a series of estimations where each variable with missing data is modeled based on its distribution. Detailed explanation of MICE can be found in [9]. KNNI identifies the k nearest neighbors for a missing value from all complete instances in a dataset. For categorical features, it imputes the missing value with the most frequent value among the neighbors. For numerical features, it uses the mean of the neighbors (mean rule). Known for its simplicity, understandability, and relatively high accuracy, KNNI is widely used in real-world data processing [10]. For additional insights of KNNI, see [11]. Given the challenges and importance of accurately addressing MAR missingness, this study aims to provide a thorough analysis of three ML-based imputation methods for MAR missingness. Results on the complete and imputed datasets are evaluated using twelve evaluation criteria across four categories, emphasizing the need for robust methods to handle MAR missingness effectively. In author's previously published paper [12], an empirical evaluation of three ML-based

imputation methods is performed, for MCAR missingness on a diabetes dataset using eleven evaluation criteria. The results demonstrated that MissForest outperformed the other two methods. However, there is a need to extend this work to understand the impact of MAR missingness on the performance of these three ML methods. This extended study, aim to evaluate these methods under MAR missingness using 12 evaluation criteria, with the addition of standard deviation as an extra criterion. This extended analysis will provide deeper insights into the robustness and effectiveness of the imputation methods under MAR missing data mechanisms. The contributions of this research paper are as follows:

- A comparative evaluation of ML-based imputation methods—KNNI, MICE, and MissForest—was conducted for handling MAR missingness, generated by a proposed MAR algorithm in a multivariate configuration.
- A systematic empirical investigation was performed on these three ML-based imputation methods across six levels of missingness, ranging from 5% to 30%, under the MAR mechanism in a multivariate context.
- The three imputation methods were evaluated on one complete and eighteen imputed datasets using twelve evaluation criteria: accuracy, precision, recall, F1 score, MCoff score, MAE, RMSE, R-squared, standard deviation, Pearson correlation, AIC, and BIC scores.

This paper is organized in the following manner. Section 2 presents a review of related work on different imputation methods. The methodology employed in this study is described in Section 3. Section 4 encompasses the experimental setup, results, and analysis of the findings. Section 5 concludes the study with an analysis of the results and final conclusions.

2. Literature Survey

Through this extensive body of work, literature collectively highlights the critical need for sophisticated imputation methods to manage missing data effectively and improve the accuracy of research findings. Authors [13] provide a comprehensive overview of different missing data mechanisms, handling methods, and their impact on statistical inference, setting a foundational understanding for researchers. Building on this foundation, Authors [14] delve into imputation methods such as mean imputation, regression imputation, and multiple imputation, illustrating practical implementation using R and the MICE package. Complementing these insights, Authors [15] discuss various methods including listwise deletion, imputation, and maximum likelihood estimation, highlighting the advantages and disadvantages of each approach. Further advancing the discussion, Authors [16] provide an in-depth look at MICE, addressing common issues and offering best practices to enhance its application. Authors [17] further this by guiding the implementation of MICE in R, discussing the handling of various data types and advanced imputation topics, thereby making the method more accessible for practical use. Finally, Authors [18] offer a practical guide for addressing missingness in research settings, emphasizing the importance of careful planning, thorough reporting, and the implementation of robust methods like multiple imputations and maximum likelihood estimation. Authors [19] evaluate the impact of various imputation methods on healthcare research accuracy and reliability. Their discussion encompasses the occurrence of absent values in healthcare studies and the implications of different imputation strategies. The authors advocate standardized guidelines to handle missing data and suggest future research to enhance imputation methods in

healthcare. Further, the review by [6] examines approaches to generating synthetic missing data based on MCAR, MAR, and MNAR mechanisms. It guides researchers in selecting appropriate methods for generating and handling synthetic missing data to improve analysis robustness. The paper identifies gaps and suggests areas for further research in missing data management. Authors [20] review 111 journal papers on missing value imputation from 2006 to 2017, focusing on experimental design issues. They highlight the limitations in the diversity of methods and datasets used in MVI studies. The authors suggest future research directions to improve MVI practices in data mining and big data analysis. Authors [21] tested various imputation strategies on real datasets simulated with MNAR, MCAR, and MAR missingness. They evaluated nine imputation methods, including RF and KNN, across four different levels of missing values from 5%- 30%. However, their performance evaluation was based on only one criterion: NRMSE. In the literature, various imputation methods have been explored and evaluated for handling missing data in different contexts. A class-weighted iterative KNN imputation method was introduced, exclusively handling MCAR missingness [22]. Performance comparisons include seven imputation methods using only NRMSE as the evaluation criterion [23], without specifying the missing data mechanism, six methods addressing only MCAR missingness with four evaluation criteria [24], and another six methods on a medical dataset using ROC (AUC) without mentioning the missing data mechanism [3]. A comprehensive study evaluated eight methods on a real-world cardiovascular cohort study, considering MAR missing mechanism and multiple criteria including MAE, RMSE, AUC, and up to 20% missingness was handled [25]. Additionally, seven methods were assessed on three healthcare datasets, handling MCAR missingness at four MR levels (10%, 15%, 20%, and 25%), evaluating only MAE and RMSE [26]. Another study compared four methods focusing on MCAR missingness, using model prediction accuracy and imputation error [27], and seven methods were evaluated on five datasets, handling MCAR mechanism with missing rates up to 20%, using only NRMSE as the criterion [28]. Table 1 presents a summary of the literature review. Most studies compare multiple imputation methods, highlighting their advantages but often focusing on specific criteria or missing data mechanisms, while emphasizing the need to address MAR and MNAR mechanisms. However, there is a need for broader evaluation metrics and more forward-looking research directions.

Table 1 Literature Survey

Paper	Advantages	Limitations
[3]	Performance comparison of six data imputation method is performed on medical dataset	Only ROC (AUC) is used as evaluation criteria. Missing mechanism is not mentioned.
[15]	Tests imputation strategies in real datasets; evaluates nine methods (including RF, KNN) at various missing value proportions.	Performance evaluation based solely on NRMSE.
[22]	A class-weighted iterative KNN imputation method is implemented	Only MCAR missingness is handled
[23]	Performance comparison of seven data imputation methods is done	Only NRMSE is used as evaluation criteria. Missing mechanism is not mentioned.
[24]	Performance comparison of six data imputation	Handled only MCAR missingness,

	methods is done	only 4 evaluation criteria used for evaluating imputation methods
[25]	Performance comparison of 8 methods on real-world cardiovascular cohort study	Evaluation criteria MAE, RMSE and AUC, MAR missing mechanism
[26]	Performance comparison of 7 imputation methods on 3 health care datasets	Four MR-10,15,20and 25%, evaluation criteria only MAE and RMSE, MCAR missing mechanism
[27]	Performance comparison of 4 imputation methods	MCAR Missing mechanism, 2 Evaluation criteria model prediction accuracy and imputation error.
[28]	Performance comparison of 7 imputation methods on 5 datasets	MCAR mechanism, MR handled till 20% only. Evaluation criteria used only NRMSE

Most studies compare multiple imputation methods, highlighting their advantages but often focusing on specific criteria or missing data mechanisms, while emphasizing the need to address MAR and MNAR mechanisms. However, there is a need for broader evaluation metrics and more forward-looking research directions.

3. Methodology

Diabetes Dataset [29] from UCI repository was utilized in this research, consisting of 768 records containing discrete and continuous type of numerical data, the categorical outcome variable was complete and excluded from imputation. However, 376 of these records had missing values and were subsequently excluded, resulting in the utilization of 392 complete records. The number of missing values in Glucose, BP, SkinThickness, Insulin and BMI were 5, 35, 227, 374, and 11 respectively. Synthetic missingness in multivariate (missingness in more than one variable) configuration was then introduced into this dataset via the MAR mechanism using the proposed MAR algorithm, generating six incomplete datasets with varying degrees of missingness (5%-30% MR). To address this issue, this work employed three ML-based imputation methods, namely KNNI, MICE, and MissForest. By applying these methods to the six incomplete datasets, authors were able to generate a total of eighteen imputed datasets, comprising six datasets for each of the imputation methods. The experimental design process is shown in Fig 1. This study measures the effectiveness of these three imputation methods in four key areas related to diabetes prediction modeling. The first category examined was the prediction model performance, where authors compared the performance of logistic regression classifiers with eighteen imputed datasets of the above-mentioned imputation methods, and complete dataset using accuracy, precision, recall, F1 score, and Mcoff score. This work analyzed the quality of imputation using MAE, RMSE, R-squared, and SD metrics for KNNI, MICE, and MissForest methods across missing rates from 5% to 30%. A correlation evaluation was carried out to determine the best imputation method for capturing complex relationships and producing accurate results. The best model was chosen based on AIC and BIC scores from the full and step model constructed from complete and imputed datasets. The experimental design process is shown in figure 1. The full model used all variables, while the step model utilized stepwise regression to enhance performance by selecting a subset of variables. To comprehensively evaluate the effect of various imputation techniques, twelve

performance metrics were used: accuracy, precision, recall, F1 score, MCoff score, MAE, RMSE, R-squared, standard deviation, Pearson correlation, AIC, and BIC scores. They helped in discovering the following: 1) What impact does the selection of imputation method have on the accuracy, precision, recall, and F1 score of models built on imputed chronic disease data? 2)Which ML imputation method best preserves the statistical relationships between key biomarkers (e.g., blood glucose levels) in MAR-affected diabetes datasets?3) Does model complexity, as measured by AIC and BIC, play a role in the performance of ML-based imputation methods in medical datasets with MAR missingness? 4) Which machine learning imputation techniques minimize error and maintain predictive performance in chronic disease datasets with MAR missing data.

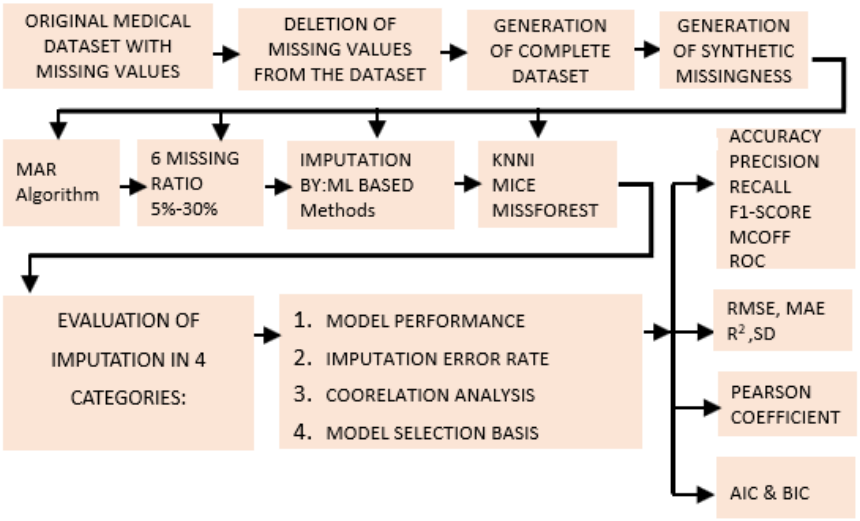

Fig 1: Experimental design process

4. Experimental Setup & Results

This experiment aimed to compare the effectiveness of three imputation methods for six missing rates varying from 5% to 30% for MAR multivariate missing patterns. The study evaluated the performance of the three ML based imputation techniques using twelve performance metrics.

4.1 Generation of MAR missingness in multivariate pattern in the dataset:

MAR missingness is artificially generated in multivariate configuration for 5%-30% MR in the complete dataset, and six incomplete datasets are generated.MAR missingness is generated in multivariate pattern as per the MAR algorithm given below, by calculating the correlations between observed variables. For each missing rate, authors create a copy of the original data and introduce missingness to some of its columns based on their correlation with other columns. Specifically, for each column that does not already have missing values, the code calculates the correlation between that column and all other columns, selects the columns with the highest correlations, and introduces missingness to them at the specified missing rate.

Algorithm I: MAR Missingness generation

Input: Original data **Output:** missing data output files for each MR
1. Load original data from input file. 2. Define missing rates to simulate. 3. Calculate pairwise correlations between observed variables. 4. For each missing rate: 4.1. Generate missingness for each missing rate and create a copy of the original data. a. Make a copy of the original data structure to preserve the original values. b. Repeat over each column and introduce missingness based on the missing rate. c. Randomly select rows to introduce missing values based on the missing rate. 4.2. Determine which variables to introduce missingness based on their correlation with other variables. For each column: • Check if the current column has any missing values. If the column has no missing values: • Compute the correlation between the current column and all other columns. • Choose the columns with the highest correlations. • Introduce missingness in the selected columns. 4.3. Save the missing data to the output file. Repeat steps 4.1 to 4.3 for each MR

4.2 Evaluation of Imputation Methods by Model Performance:

To evaluate the imputation methods by model performance, a full model and step model of a logistic regression classifier are used. The full model includes all variables of the dataset, while the step model is created using stepwise regression. The performance of the model is evaluated using accuracy, precision, recall, F1 score, and Mcoff score. For building the step model of the logistic regression classifier, the model is initialized, and stepwise regression is performed using a while loop. In each iteration, the adjusted R-squared is calculated for each remaining feature when added to the previously selected features. The feature with the highest adjusted R-squared is chosen and added to the list of selected features. Once the list of selected features is complete, the final model is fitted using logistic regression with only those features. This final model is then used to make predictions. The performance of imputation techniques is evaluated by comparing the absolute difference of five model performance metric values between the imputed and original datasets. To accomplish this, the values of accuracy (and other performance metrics like precision, recall, etc.) for both the imputed and original datasets are computed, and the absolute difference is calculated between them. If the absolute difference is zero, it indicates that the accuracy (or other performance metrics like precision, recall, etc.) of a model trained on the imputed dataset is equal to the original dataset, suggesting that the imputation method preserved the original data's quality. A smaller absolute difference suggests that the imputation preserves the model performance, while a larger absolute difference indicates that the imputation significantly alters model performance, potentially introducing more bias or distorting the data. The total difference value of accuracy, precision, recall, F1 score, and Mcoff score across all MR% of

MissForest is smaller than other imputation methods, and the sum of the total difference for all performance metrics of MissForest is less than the other two imputation methods, meaning it preserves model performance most closely across all metrics. As shown in Table 2, MissForest performs better than the other two methods in 6 out of 6 MR cases, i.e., 100% of cases across the five model performance metrics.

Table 2 Comparing performance metrics values for various Imputation methods

Evaluation Criteria	Imputation	MR 5%	MR 10%	MR 15%	MR 20%	MR 25%	MR 30%	Total Difference
Accuracy	KNNI	0	0	0.26	2.55	0.51	1.79	5.11
	MICE	0.26	0	4.34	1.28	1.17	2.81	9.86
	MISSFOREST	**0**	0.51	**0.26**	**1.02**	**0.26**	**1.02**	**3.07**
Precision	KNNI	0	0.45	0.71	5.4	1.44	1.29	9.29
	MICE	4.63	**0**	0.52	2.18	3.18	4.04	14.55
	MISSFOREST	**0**	0.09	**0.17**	**1.94**	**0.7**	**1.02**	**3.92**
Recall	KNNI	0	0.77	**0**	3.08	0	6.15	10
	MICE	5.97	**0**	42.62	2.31	41.69	6.15	98.74
	MISSFOREST	**0**	2.31	1.54	**1.54**	**0**	**3.08**	**8.47**
F1-Score	KNNI	0	0.31	**0.28**	4.02	0.56	4.22	9.39
	MICE	5.55	**0**	39.38	2.29	44.38	5.37	96.97
	MISSFOREST	**0**	1.5	0.88	**1.72**	**0.28**	**2.29**	**6.67**
MCoff	KNNI	**0**	0	0.01	0.06	0.01	0.05	0.13
	MICE	0.05	0	0.22	0.03	0.25	0.07	0.62
	MISSFOREST	**0**	0.01	**0.01**	**0.02**	**0.01**	**0.03**	**0.06**

4.3 *Evaluation of Imputation Methods by R-Squared, MAE, RMSE and SD Score:*

Performance metrics like MAE, RMSE, R-squared, and SD assess imputation methods by measuring discrepancies between imputed and actual values. MAE and RMSE indicate accuracy, with lower values being better, while R-squared shows the proportion of variance explained, with higher values being better. SD measures deviation from actual values, with smaller values indicating higher accuracy. In a comparison of three ML methods across datasets with 5%-30% missing data, MissForest consistently produced lower MAE, SD and RMSE values in 5 out of 6 MR cases i.e. 83.33% of cases and higher R-squared values in 100% of cases outperforming the KNNI and MICE methods. Results are shown in Fig 2.

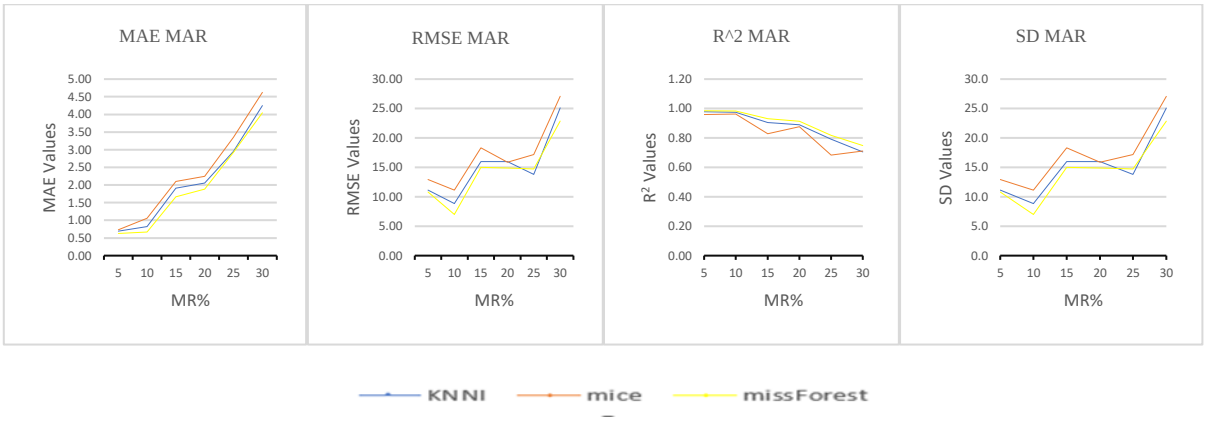

Fig 2: Comparing MAE, RMSE, R-Square & SD values

4.4 Evaluation of Imputation Methods by AIC & BIC Scores:

AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) are model selection methods that estimate the quality of each input model, with AIC derived from frequentist probability and BIC from Bayesian probability. AIC evaluates model efficiency by measuring information preservation, favoring models with lower AIC values, thus penalizing less for complexity. In contrast, BIC imposes a higher penalty for model complexity, leading to higher scores for complex models. The formulas for AIC and BIC are $AIC = -2ln(maximum\ likelihood) + 2m$ and $BIC = -2ln(maximum\ likelihood) + m\ ln(n)$. AIC and BIC calculated using the number of estimated parameters (m) and the total number of observations. (n) [30][31][32][33]. Both metrics are used for logistic regression models. When comparing full models and stepwise regression models (created by iteratively including/excluding variables) on original and imputed datasets (using KNNI, MICE, and MissForest methods with 5%-30% missing data), MissForest imputation demonstrated closer performance to the original dataset than MICE and KNNI in terms of AIC and BIC scores as shown Table 3.If the AIC/BIC score of the imputed dataset is closer to original dataset indicates that the imputation method has preserved the quality of original data. This indicates MissForest's imputation method has preserved the quality of data after imputation.

Table 3 Comparing AIC and BIC values for various Imputation methods

	Evaluation Criteria	Original dataset	Imputation	MR 5%	MR 10%	MR 15%	MR 20%	MR 25%	MR 30%
FULL Model	AIC Score	437.8	KNNI	436.1	437.4	438.2	445.4	420.0	443.3
			MICE	421.2	436.3	342.2	449.3	176.2	433.4
			MISSFOREST	**436.6**	**437.7**	**432.4**	**440.6**	**425.1**	**440.9**
	BIC Score	473.5	KNNI	471.8	473.1	**473.9**	481.2	455.8	479.0
			MICE	456.9	472.1	377.9	485.1	212.0	469.2
			MISSFOREST	**472.3**	**473.4**	468.1	**476.4**	**460.9**	**476.6**
STE	AIC	356.9	KNNI	361.1	356.0	363.4	361.7	369.9	358.2

P Model	Score		MICE	361.0	356.6	321.8	362.5	236.7	358.5
			MISSFOREST	**359.9**	354.4	**354.4**	**356.2**	**366.0**	**355.8**
	BIC Score	380.7	KNNI	384.9	379.9	383.2	385.5	389.7	382.0
			MICE	380.9	380.4	341.7	390.3	256.5	378.3
			MISSFOREST	383.8	374.3	**378.2**	**376.1**	**385.9**	**379.7**

4.5 Evaluation of Imputation Methods by Comparing Correlation Among Parameters:

The original dataset displayed some correlation between the parameters, which needs to be preserved in the imputed dataset as well. To evaluate how well the imputed dataset preserves this correlation, it is necessary to determine whether the imputed values were able to maintain the original correlation with other parameters in the dataset. This needs to be done for all three imputation methods, namely KNNI, MICE, and MissForest. To accomplish this, Pearson correlation coefficient values are computed for all the parameters in both the complete dataset and imputed datasets, and then the absolute difference between them is calculated, followed by the total difference for every parameter across all the MR%, as shown in Table 4. For a few parameters, KNNI shows better performance than MissForest; however, the sum of the total absolute difference of the correlation coefficient values of all the dataset parameters for KNNI, MICE, and MissForest methods is 1.31, 2.38, and 1.19, respectively. This suggests that the MissForest method performs better than the other two methods in capturing the intricate relationships between the dataset parameters. These results indicate that the MissForest imputation method could be a valuable approach for accurately imputing missing values in diabetes datasets and other medical datasets with complex inter-relationships between parameters, as it maintains correlations between various parameters like glucose, insulin, and other biomarkers in the diabetes dataset. This method can be applied for handling missing data in departments of big hospitals, especially those monitoring chronic diseases like diabetes.

Table 4 Comparing Pearson correlation coefficient values of Diabetes dataset variables

Diabetes parameters	Imputation	MR 5%	MR 10%	MR 15%	MR 20%	MR 25%	MR 30%	Total Difference
GLUCOSE	KNNI	0	0	0	0	0	0	0
	MICE	0.04	0	0.18	0	0.32	0	0.54
	MISSFOREST	**0**	**0**	**0**	**0**	**0**	**0**	**0**
AGE	KNNI	0	0.01	0.05	0.03	0.02	0.03	0.14
	MICE	0.04	0	0.22	0.06	0.11	0.02	0.45
	MISSFOREST	**0**	**0.01**	**0.01**	**0.01**	**0**	**0.01**	**0.04**
INSULIN	KNNI	0.02	0.01	**0.05**	0.02	**0**	0.03	0.13
	MICE	**0**	0.02	0.08	0.05	0.16	0.03	0.34
	MISSFOREST	0.03	**0**	0.06	**0.01**	0.02	0.04	0.16
BMI	KNNI	0.01	0.01	0.03	0.03	0.13	0.02	0.23
	MICE	0.03	0.03	0.05	0	0.16	0.02	0.29

	MISSFOREST	0.01	0.01	0.03	0.03	0.04	0.02	**0.14**
PREGNANCIES	KNNI	0	0	0.02	0.02	0.02	0.08	0.14
	MICE	0.01	0	0.07	0.07	0.03	0.11	0.29
	MISSFOREST	0	0	0.01	0.03	0.01	0.02	**0.07**
SKIN THICKNESS	KNNI	0.06	0	0.09	0.02	0.04	0.01	**0.22**
	MICE	0.04	0	0.04	0.04	0.12	0.02	0.26
	MISSFOREST	0.07	**0**	0.1	**0.01**	0.06	0.02	0.26
DIABETES PEDIGREE FUNCTION	KNNI	0.11	0	0.14	0	**0.09**	**0**	0.34
	MICE	0.02	0	0.08	0	0.21	0	0.31
	MISSFOREST	0.12	**0**	0.15	**0**	0.11	0	0.38
BP	KNNI	0	0	0	0	**0**	0.11	0.11
	MICE	0.04	0	0.09	0	0.05	0.01	0.19
	MISSFOREST	**0**	**0**	**0**	**0**	0.03	**0.11**	0.14

5. Conclusion

In this paper, the effectiveness of three ML-based imputation methods for handling MAR missingness in a multivariate configuration, on diabetes dataset, was evaluated. Comparison between the original complete dataset and eighteen imputed datasets, was performed, where missing values were imputed using KNNI, MICE, and MissForest techniques. The experimentation results indicate that the MissForest imputation method achieved lower SD, MAE, and RMSE in 83.33% of cases MR cases and higher R-squared values in all (100%) MR cases. MissForest performed better in 100% of MR cases for all the five performance metrics of model performance. The absolute difference between the computed Pearson correlation coefficients values for all the parameters in both the complete dataset and imputed datasets indicate that the MissForest performs better, than the other 2 methods, in preserving correlation between parameters. The results demonstrate that the MissForest imputation method outperformed KNNI, MICE in maintaining the overall data quality after imputation, making it the preferred choice for handling MAR missingness in chronic disease like diabetes. This study be used for accurate imputation of incomplete medical datasets of chronic diseases for improved data driven clinical decisions. Future research needs to address the problem MNAR missingness across diverse domains, for varying percentages of MR, various sample sizes, and different data types.

Refrences

[1] R. J. A. Little and D. B. Rubin, *Statistical Analysis with Missing Data*, vol. 151, no. 2, 1988.
[2] D. J. Stekhoven and P. Bühlmann, "MissForest non-parametric missing value imputation for mixed-type data," *Bioinformatics*, vol. 28, no. 1, pp. 112-118, 2011.
[3] J. M. Jerez et al., “Missing data imputation using statistical and machine learning methods in a real breast cancer problem,” *Artificial intelligence in medicine*, vol. 50, no. 2, pp. 105-115, 2010, doi: 10.1016/j.artmed.2010.05.002.
[4] D. B. Rubin, "Inference and missing data," *Biometrika*, vol. 63, pp. 581–592, 1976.
[5] X. Miao, Y. Wu, L. Chen, Y. Gao, and J. Yin, "An Experimental Survey of Missing Data Imputation Algorithms," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 7, pp. 6630-6650, 1 July 2023, doi: 10.1109/TKDE.2022.3186498.
[6] M. S. Santos, R. C. Pereira, A. F. Costa, J. P. Soares, J. Santos, and P. H. Abreu, "Generating Synthetic Missing Data: A Review by Missing Mechanism," *IEEE Access*, vol. 7, pp. 11651-11667, 2019, doi: 10.1109/ACCESS.2019.2891360.

[7] Y. Zhou, S. Aryal, and M. R. Bouadjenek, "Review for Handling Missing Data with special missing mechanism," *arXiv preprint arXiv:2404.04905*, 2024.
[8] Zhang, Zhongheng. "Multiple imputation with multivariate imputation by chained equation (MICE) package." *Annals of translational medicine* 4.2 (2016).
[9] van Buuren, Stef. "Item imputation without specifying scale structure." *Methodology* (2010).
[10] S. Zhang, "Nearest neighbor selection for iteratively kNN imputation," *J. Syst. Softw.*, vol. 85, no. 11, pp. 2541-2552, 2012.
[11] Troyanskaya, Olga, et al. "Missing value estimation methods for DNA microarrays." Bioinformatics 17.6 (2001): 520-525
[12] S. Tiwaskar, M. Rashid, and P. Gokhale, "Impact of machine learning-based imputation techniques on medical datasets-a comparative analysis," *Multimedia Tools and Applications*, pp. 1-21, 2024.
[13] R. J. A. Little and D. B. Rubin, *Statistical analysis with missing data*, Wiley, 2002.
[14] S. van Buuren, *Flexible imputation of missing data*, CRC Press, 2018.
[15] C. K. Enders, *Applied missing data analysis*, Guilford Press, 2010.
[16] I. R. White, P. Royston, and A. M. Wood, "Multiple imputation using chained equations: Issues and guidance for practice," *Statistics in medicine*, vol. 30, no. 4, pp. 377-399, 2011.
[17] S. van Buuren and K. Groothuis-Oudshoorn, "MICE: Multivariate Imputation by Chained Equations in R," *Journal of Statistical Software*, vol. 45, no. 3, pp. 1-67, 2011.
[18] J. W. Graham, "Missing data analysis: Making it work in the real world," *Annual Review of Psychology*, vol. 60, pp. 549-576, 2009.
[19] D. A. Bennett, "How can I deal with missing data in my study?," *Australian and New Zealand journal of public health*, vol. 25, no. 5, pp. 464-469, 2001.
[20] W. C. Lin and C. F. Tsai, "Missing value imputation: a review and analysis of the literature (2006–2017)," *Artif. Intell. Rev.*, pp. 1487–1509, 2020.
[21] M. Kokla, J. Virtanen, M. Kolehmainen et al., "Random forest-based imputation outperforms other methods for imputing LC-MS metabolomics data: a comparative study," *BMC Bioinformatics*, vol. 20, pp. 492, 2019, doi: 10.1186/s12859-019-3110-0.
[22] A. Choudhury and M. R. Kosorok, "Missing Data Imputation for Classification Problems," *arXiv preprint arXiv:2002.10709*, 2020. [Online]. Available: https://arxiv.org/abs/2002.10709.
[23] A. Jadhav, D. Pramod, and K. Ramanathan, "Comparison of Performance of Data Imputation Methods for Numeric Dataset," *Applied Artificial Intelligence*, vol. 33, no. 10, pp. 913–933, 2019, doi: 10.1080/08839514.2019.1637138.
[24] P. Schmitt, J. Mandel, and M. Guedj, "A Comparison of Six Methods for Missing Data Imputation," *Journal of Biometrics & Biostatistics*, vol. 6, pp. 0-0, 2015. [Online]. Available: https://api.semanticscholar.org/CorpusID:63461528.
[25] J. Li et al., "Comparison of the effects of imputation methods for missing data in predictive modelling of cohort study datasets," *BMC Medical Research Methodology*, vol. 24, no. 1, pp. 41, 2024.
[26] L. O. Joel, W. Doorsamy, and B. S. Paul, "On the Performance of Imputation Techniques for Missing Values on Healthcare Datasets," *arXiv preprint arXiv:2403.14687*, 2024.
[27] A. K. Waljee et al., "Comparison of imputation methods for missing laboratory data in medicine," *BMJ open*, vol. 3, no. 8, pp. e002847, 2013.
[28] P. Keerin, "A Comparative Study of Missing Value Imputation Methods for Education Data," *Proceedings of the 29th International Conference on Computers in Education*. Asia-Pacific Society for Computers in Education, 2021.
[29] M. Lichman, "Pima Indians diabetes database," ed. Center for machine learning and intelligent systems: UCI Machine Learning repository.
[30] H. Akaike, "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, vol. 19, no. 6, pp. 716-723, December 1974, doi: 10.1109/TAC.1974.1100705.
[31] "Estimating the Dimension of a Model," *Ann. Statist.*, vol. 6, no. 2, pp. 461-464, March 1978, doi: 10.1214/aos/1176344136.
[32] AIC, BIC and Recent Advances in Model Selection,Editor(s): Prasanta S. Bandyopadhyay, Malcolm R. Forster, In Handbook of the Philosophy of Science, Philosophy of Statistics, North-Holland, Volume 7, 2011, Pages 583-605, ISSN 18789846, ISBN 9780444518620.
[33] Y. Sakamoto, M. Ishiguro, and G. Kitagawa, Akaike information criterion statistics, Dordrecht, The Netherlands: D. Reidel, 1986.