Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Hong-Jui Shen Date

Unraveling the Impact of Fuzzy Similarity Algorithms on Missing Data Imputation of Heart Bypass Surgery Cohort

By

Hong-Jui Shen

Master of Science in Public Health

Department of Biostatistics and Bioinformatics

Dr. Rameshbabu Manyam

Thesis Advisor

Dr. McPherson Tarrant

Reader

Unraveling the Impact of Fuzzy Similarity Algorithms on Missing Data Imputation of Heart Bypass Surgery Cohort

By

Hong-Jui Shen

B.A.

Denison University

2022

Advisor: Rameshbabu Manyam, Ph.D.

An abstract of

A thesis submitted to the Faculty of the

Rollins School of Public Health of Emory University

in partial fulfillment of the requirements for the degree of

Master of Science in Public Health

in Department of Biostatistics and Bioinformatics

2024

Abstract

Unraveling the Impact of Fuzzy Similarity Algorithms on Missing Data Imputation of Heart Bypass Surgery Cohort

By

Hong-Jui Shen

Objective: This thesis introduces the Fuzzy C-Means based Random Forest (FCRF) method, developed to address the limitations of existing data imputation techniques in public health datasets. Aimed at enhancing imputation accuracy, FCRF integrates fuzzy logic and similarity learning to navigate complex missing data mechanisms: Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR).

Method: The performance of FCRF is evaluated against traditional imputation methods—Mean, K-Nearest Neighbors (KNN), Multiple Imputation by Chained Equations (MICE), and Iterative Imputation—using metrics like Average RMSE, Normalized RMSE, Mean Absolute Error (MAE), Weighted F1-Score, and Normalized Accuracy. This comparative analysis spans various missing data scenarios to assess each method's effectiveness comprehensively.

Results: Results show that FCRF exhibits competitive performance across all scenarios, particularly excelling in complex MNAR situations where conventional methods falter. Its methodological design, which combines clustering and predictive modeling, offers nuanced capabilities beneficial for public health research.

Conclusion: FCRF marks a significant advancement in data imputation, promising more accurate and reliable analyses for public health research. Future work will explore FCRF's impact on standard error and variance estimates to ensure the method's robustness, aiming to prevent potential biases in statistical inferences. This research contributes to enhancing data integrity, supporting informed decision-making in public health.

Keywords: Data Imputation, Public Health, Fuzzy C-Means, Random Forest, Missing Data, Machine Learning, Similarity Learning.

Unraveling the Impact of Fuzzy Similarity Algorithms on Missing Data Imputation of Heart Bypass Surgery Cohort

By

Hong-Jui Shen

B.A.

Denison University

2022

Advisor: Rameshbabu Manyam, Ph.D.

A thesis submitted to the Faculty of the

Rollins School of Public Health of Emory University

in partial fulfillment of the requirements for the degree of

Master of Science in Public Health

in Department of Biostatistics and Bioinformatics

2024

Chapter 1. Introduction and literature review

1.1 Relevance of Data Imputation In Public Health

In the realm of public health, the analysis of extensive datasets plays a pivotal role in shaping health policies, guiding disease control strategies, and refining healthcare services. The integrity and accuracy of such data are fundamental for monitoring disease outbreaks, evaluating health outcomes, and assessing the impact of public health interventions. Nonetheless, the prevalence of missing data in public health datasets—stemming from non-responses, attrition, or errors during data collection and recording—significantly undermines the reliability of health assessments and the consequent decisions derived from them.

The challenge of missing data extends beyond its mere presence to include the diverse mechanisms through which data can become missing. These mechanisms, broadly classified into Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR), each introduce specific complications for data analysis and necessitate tailored imputation approaches. Traditional imputation techniques, ranging from simple strategies like mean substitution to sophisticated models like Multiple Imputation by Chained Equations (MICE) and k-nearest neighbors (k-NN), have made strides in addressing missing data issues. However, these methods often encounter limitations in handling the complexity of public health data or

substantial proportions of missing information, sometimes failing to adequately represent the uncertainty inherent in the imputation process.

The evolution of machine learning and data mining technologies presents new frontiers for improving data imputation methodologies. In particular, similarity learning—which identifies and utilizes patterns of similarity among data points—emerges as a promising approach for increasing imputation accuracy in datasets characterized by intricate relationships. Complementing similarity learning, fuzzy logic offers a framework for managing ambiguity and uncertainty, enabling a more nuanced classification and imputation of missing data. The synergistic application of these advanced techniques holds the potential for developing refined imputation methods that can more adeptly navigate the complexities of public health datasets.

By integrating fuzzy logic with similarity learning, the Fuzzy C-Means based Random Forest (FCRF) imputation method proposes a novel solution to the pervasive issue of missing data in public health research. This approach seeks to surmount the constraints of existing imputation methods, paving the way for more accurate, reliable, and comprehensive analyses of public health data. As such, the FCRF method represents a significant advancement in the field of data imputation, promising to enhance the quality of public health research and contribute to the development of more effective health policies and interventions.

1.2 Literature Review

This literature review synthesizes existing research on data imputation, with a special

focus on public health contexts. It underscores the significance of accurate data in health policy and service delivery, examines the impact of missing data, critiques current imputation methods, and explores the potential of similarity learning and fuzzy logic. By reviewing historical to recent publications, vital statistics, and expert communications, this chapter lays the groundwork for introducing the Fuzzy C-Means based Random Forest (FCRF) imputation method.

The integrity of healthcare databases is crucial for informed decision-making and policy formulation in public health. The challenge of missing data in these databases cannot be understated, as highlighted by authors like Schmitt et al. (2015), who discuss the statistical and practical ramifications of missing data in health research. These challenges include compromised data integrity, biased statistical analyses, and ultimately, the potential for misguided public health policies.

Traditional imputation methods, such as mean substitution, mode imputation, and even more sophisticated techniques like Multiple Imputation by Chained Equations (MICE), have been extensively explored in literature. For instance, Rahman and Davis (2013) critique these methods for their inability to adequately handle complex data relationships and their tendency to underestimate imputation uncertainty. Despite their utility, these conventional techniques often fall short in addressing the nuanced challenges presented by healthcare datasets, particularly those with high-dimensional data or significant proportions of missingness.

Recent advancements in machine learning, notably similarity learning and fuzzy

logic, offer promising solutions to these challenges. The work of Amiri and Jensen (2016) introduces fuzzy logic into the domain of data imputation, illustrating its potential to manage the inherent uncertainty and ambiguity of missing healthcare data more effectively than traditional statistical approaches. Similarly, the application of similarity learning, as explored in specific machine learning frameworks, has demonstrated its ability to discern complex patterns within data, suggesting a pathway toward more accurate imputation methods.

Among the novel methodologies emerging in the field, the Fuzzy C-Means based Random Forest (FCRF) imputation method represents a potentially significant advancement. This approach, integrating fuzzy logic with the robust classification and regression capabilities of random forests, aims to address the limitations of existing imputation techniques directly. By leveraging the strengths of both fuzzy logic and similarity learning, the FCRF method proposes a sophisticated solution capable of handling the complexities and uncertainties characteristic of public health and healthcare datasets.

In summary, the literature on data imputation in public health research emphasizes the critical need for more sophisticated methods that can navigate the challenges of missing data with greater accuracy and reliability. The exploration of advanced machine learning techniques, including fuzzy logic and similarity learning, alongside innovative approaches like the FCRF method, highlights a promising direction for future research in this domain. The integration of these methodologies could enhance the quality of data imputation, ultimately leading to more informed public health decisions and

policies.

1.3 Ethical Considerations

This thesis on retrospective patient data from CABG surgeries at Emory Healthcare hospitals (2014-2019) strictly adheres to ethical principles, emphasizing the safeguarding of patient privacy and confidentiality through anonymization of personal information. (Manyam RB et al., 2024) It respects the original informed consent under which data was collected, ensuring transparency and integrity in the presentation of research findings. The study is committed to beneficence, aiming to maximize public health benefits while minimizing potential harm to individuals. Accountability in adhering to ethical standards and legal compliance with health data privacy laws, including HIPAA, is paramount. This approach ensures the research contributes responsibly to public health knowledge, respecting the dignity and rights of individuals involved.

Chapter 2. Methodology

2.1 Software and Package Utilization

The study's computational analyses were performed using Python version 3.9.13, chosen for its extensive support and robust libraries suited to statistical computation and machine learning. The following packages were instrumental in the data processing, imputation, and evaluation phases:

- Pandas: Employed for initial data encoding and preparation, offering powerful data structures like DataFrames, which facilitated the handling of complex, tabular data.

- NumPy: This fundamental package for scientific computing provided support for the high-level mathematical functions necessary for operations on multi-dimensional arrays and matrices.
- SciPy: Utilized for its extensive statistical functions, aiding in the rigorous statistical testing needed for preliminary data analysis.
- Scikit-learn: A versatile machine learning library that provided tools for dimensionality reduction (PCA), Random Forest algorithms for the imputation models, and a suite of metrics for evaluating imputation accuracy such as f1-score and accuracy.
- Fuzzy C-Means: Applied for clustering the dataset into fuzzy clusters before the imputation process, aligning with the clustering approaches highlighted in the literature for handling ambiguities within data groupings (Khan and Hoque, 2020).
- Matplotlib and FancyImpute: These packages were used for visualization and advanced imputation algorithms, respectively.
- Statsmodels and Pytesmo: These provided additional statistical models, tests, and a toolbox for validation metrics such as RMSD and NRMSD.

2.2 Data Collection and Processing

The dataset underpinning this study consists of retrospective patient data collected from Coronary Artery Bypass Grafting (CABG) surgeries performed at Emory Healthcare over a five-year span (2014-2019). This data acquisition, authorized by the Society of Thoracic Surgeons, incorporates an extensive compilation of 12,328 records, each representing a unique surgical case. The initial phase of data preparation involved

a meticulous encoding process. This process included the transformation of categorical variables into numerical codes, enabling efficient manipulation and analysis. In addition, essential continuous variables such as age and Body Mass Index (BMI) underwent standardization to ensure a homogenized data structure suitable for the application of statistical and machine learning techniques.

2.3 Variable Selection

The critical task of variable selection was informed by both clinical expertise and the analytical framework established by the extant literature. 59 variables, directly correlated with the outcomes of CABG surgeries, were meticulously chosen to be included in the analysis. The attributes of these variables can be viewed in Table1.

2.4 Data Preprocessing

In the preprocessing stage, the initial handling of missing values was addressed. Consistent with the guidelines set forth by Jadhav, Pramod, and Ramanathan (2019), variables with a missing data proportion exceeding a 30% threshold were earmarked for exclusion from the dataset to preserve analytical integrity. Using NumPy, a foundational package for scientific computing in Python, and Pandas, this stage also involved cleaning the data by detecting and correcting (or removing) corrupt or inaccurate records from the dataset, ensuring a high-quality dataset conducive for rigorous statistical analysis.

Inspired by the work of Fouad et al. (2021), the preprocessing stage in this study also incorporates a preliminary exploration of similarity-based imputation techniques as a

precursor to the main analysis. This involves the generation of MCAR (Missing Completely At Random), MAR (Missing At Random), and MNAR (Missing Not At Random) datasets to systematically evaluate the efficacy of various imputation strategies, including those based on similarity learning, within the context of different missing data mechanisms.

2.41 Missing Completely At Random (MCAR)

The MCAR datasets were generated by randomly introducing missing values across the dataset without regard to the values of other variables. This method will randomly select cells within the dataset without any bias towards specific rows, columns, or values. Replacing the selected cells' values with 'NaN' to simulate missingness, ensuring that the probability of any data point being missing is uniform across the dataset.

2.42 Missing At Random (MAR)

For the MAR datasets, missingness was introduced in a way that depended on observed data but not on the missing data itself. This method will identify causative variables whose observed values would dictate where missingness would occur in other variables. Then, systematically introducing 'NaN' values in specific variables based on the values of the causative variables, adhering to predetermined conditions (e.g., when a causative variable exceeds a certain threshold).

2.43 Missing Not At Random (MNAR)

The generation of MNAR datasets was the most intricate, given that the missingness

of data depends on unobserved (missing) information. The approach taken in this research is to establish some criteria for missingness that directly relate to the values that would be missing (e.g., values below or above certain thresholds). Applying these criteria across the dataset to introduce 'NaN' values, thereby simulating scenarios where the likelihood of missingness is inherent to the missing data itself.

The framework of creating MCAR, MAR, and MNAR data can be viewed in Figure1.

2.5 Theoretical Foundation and Algorithmic Framework

The Fuzzy C-Means based Random Forest (FCRF) imputation method represents a sophisticated convergence of clustering and predictive modeling, specifically tailored for the complexities of healthcare data. This methodology embraces the principles outlined by Bezdek (1981) in his seminal work on fuzzy clustering, which is particularly resonant with the ambiguous and overlapping nature of clinical data. The Random Forest algorithm, as a robust estimator within this framework, is favored for its capacity to handle high-dimensional spaces and complex data structures without succumbing to overfitting, a quality extensively discussed in Breiman's (2001) comprehensive analysis of ensemble methods. The proposed Fuzzy C-means based random forest imputation method (FCRF) framework can be viewed in Figure2. The main steps for the FCRF algorithm are explained in the following:

Step 1-Sperate the data into the observed data and missing data.

Step 2-Preliminary Imputation for incomplete datasets: For categorical variables, fill missing values with the mode; for continuous variables, use the mean. This step ensures

there are no missing values that might affect the dimensionality reduction and clustering processes.

Step 3- Dimensionality Reduction:

- Apply Standard Scaler: Normalize the dataset to ensure features contribute equally to the analysis.
- Perform PCA: Apply Principal Component Analysis (PCA) to the normalized dataset to reduce dimensions while retaining most of the variance.

Step 4-Perform Fuzzy C-Means (FCM): Cluster the PCA-transformed data into predefined groups (clusters) using FCM, facilitating targeted imputation within similar observation groups.

Step 5- Encode Categorical Variables: Convert categorical variables into numerical format using Label Encoding, preparing them for the Random Forest algorithm.

Step 6-Imputation Process

1. Cluster-Based Separation: Divide the dataset into clusters based on FCM results.
2. Iterate Over Each Cluster: For observations within each cluster, perform the imputation separately.
3. Categorical and Continuous Features: Use Random Forest Classifier for categorical features and Random Forest Regression for continuous features.
4. Direct Imputation: Apply the Random Forest model to impute missing values based on the feature type.

5. Binary Feature Post-processing: Ensure binary features are post-processed to be either 0 or 1.

Step 7- Merge Imputed Data: Combine imputed observations back with the complete observations to form a fully imputed dataset.

Step 10-Post-processing Imputed Data:

1. Re-add excluded unique identifier columns back to the imputed dataset.
2. Ensure binary and categorical variables are correctly formatted.

2.6 Imputation Process

Before the commencement of the actual imputation, the data must be rendered into a form where the complex, multi-dimensional nature of healthcare data is distilled into its most informative components. Initially, the dataset undergoes a PCA transformation, reducing its dimensionality while preserving as much of the variance as possible. The process of PCA, as expounded by Jolliffe (2002), ensures that the data's intrinsic structure is retained, which is pivotal for the effective clustering that follows. Subsequently, Fuzzy C-Means clustering categorizes the data into clusters with varying degrees of membership, a concept that deviates from the rigidity of hard clustering methods by allowing for more flexible data partitioning. This method's capacity for handling imprecise boundaries between clusters makes it highly applicable to the kind of heterogeneous and incomplete datasets prevalent in medical research.

Within each fuzzy cluster, Random Forest imputation is executed separately for continuous and categorical variables. The Random Forest algorithm, known for its

ensemble learning approach using multiple decision trees, is particularly adept at capturing the non-linear interactions between variables. By employing it within the homogenous groups defined by fuzzy clustering, the algorithm can make more accurate predictions, as the clusters are likely to contain similar response patterns.

The Random Forest imputation is inherently iterative; it handles missing data by using the observed values to predict the missing ones iteratively. It begins with a simple imputation (e.g., mean imputation) and then fits a Random Forest model on the observed values to predict the missing values. The process is repeated until the predictions stabilize, ensuring that the imputed values are plausible given the observed data.

Continuous variables with missing values initially undergo mean imputation. This step provides a simple yet effective way to maintain dataset integrity before applying more sophisticated imputation techniques. For each cluster identified by the Fuzzy C-Means algorithm, a Random Forest Regression model is trained. Random Forest Regression is tailored for regression tasks, using multiple decision trees to estimate the missing values by averaging the predictions from all trees, providing a single continuous value as the output. By aggregating the predictions from many decision trees, Random Forest Regression reduces the variance of the predictions, leading to more accurate and stable imputed values for numerical variables compared to using a single decision tree. This model leverages the patterns and relationships observed within the cluster's complete cases to predict missing values for continuous variables. Since the imputation process for continuous variables benefits from an iterative approach, after

the initial imputation, the Random Forest Regression model refines the imputation by iteratively predicting missing values based on the observed data, enhancing the accuracy of the imputed values with each iteration.

Categorical variables first receive mode imputation, where missing values are replaced with the most common category within each variable. This step ensures that all categorical variables are complete before proceeding to more advanced imputation methods. To prepare for machine learning-based imputation, categorical variables are transformed into numerical format through Label Encoding. This encoding process assigns a unique integer to each category level, making the data compatible with Random Forest algorithms. Within each cluster, a Random Forest Classifier model is specifically tailored to impute missing values in categorical variables. As it can predict the category for each missing value by looking at the 'votes' from multiple decision trees and selecting the category with the majority vote. Moreover, it can capture these non-linear interactions effectively due to its ensemble nature, combining the outcomes of numerous decision trees trained on various subsets of the data. By training on complete cases within the cluster, the model utilizes the homogeneity of data within clusters to accurately predict missing categorical values.

Once the Random Forest models have imputed the missing values—whether for continuous or categorical variables—within their respective clusters, the imputed values are seamlessly integrated back into the main dataset. This integration process ensures that the imputed dataset retains the structural and statistical properties of the original data as closely as possible.

2.7 Evaluation Metrics

To rigorously assess and compare the performance of various imputation methods—including FCRF, Simple Imputation, KNN, MICE, and Iterative Imputation—this thesis employs a comprehensive suite of evaluation metrics. These metrics are instrumental in quantifying the accuracy and reliability of the imputation methods applied to both continuous and categorical data within the dataset. Specifically, we utilize the following five key metrics:

2.71 Continuous Variables Evaluation

- Root Mean Square Error (RMSE): RMSE measures the standard deviation of the prediction errors, providing insights into the average magnitude of the prediction error. It is particularly sensitive to outliers and thus is a robust indicator of the imputation accuracy for continuous variables.
- Normalized Root Mean Square Error (NRMSE): Normalizing RMSE by the range of the data provides a scale-independent measure of error magnitude, enabling comparisons across different datasets or variables with varying scales. NRMSE is critical for understanding the relative error size in the context of the data's variability.
- Mean Absolute Error (MAE): MAE assesses the average absolute difference between the imputed values and the actual values, offering a straightforward interpretation of the average error magnitude. Unlike RMSE, MAE is not as heavily influenced by outliers, providing a complementary perspective on the

imputation accuracy for continuous variables.

2.72 Categorical and Binary Variables Evaluation

- Weighted F1-Score: The weighted F1-Score combines precision and recall into a single metric, taking into account the label imbalance by weighting the scores of each class according to their presence in the dataset. This metric is particularly valuable for evaluating imputation performance on categorical and binary variables, ensuring that the model's accuracy is not overly influenced by the most frequent class.
- Normalized Accuracy: This metric represents the accuracy of correctly imputed values, adjusted for chance. Normalized accuracy is crucial for assessing the performance of imputation methods on categorical and binary variables, especially in datasets where some classes are significantly more prevalent than others. It provides a more nuanced understanding of the model's predictive power beyond simple accuracy.

Chapter 3. Results

The evaluation of imputation performance was systematically conducted by comparing the Fuzzy C-Means based Random Forest (FCRF) imputation method against four traditional imputation techniques: Mean, K-Nearest Neighbors (KNN), Multiple Imputation by Chained Equations (MICE), and Iterative Imputation. The metrics utilized for this comparison were the Average Root Mean Square Error (RMSE), the normalized root mean square error (NRMSE), the mean absolute error (MAE), the

weighted f1-score, and the normalized accuracy. Each metric sheds light on different aspects of imputation accuracy and reliability.

3.1 MCAR Data Scenario

In the MCAR scenario, the FCRF method demonstrated commendable performance with an Average RMSE of 1.474 and an Average AE of 0.316, closely aligning with the results of Simple Imputation and surpassing those of KNN. Notably, the F1-score and Accuracy metrics indicated robust capabilities in accurately imputing categorical and binary variables, with FCRF achieving a F1-score of 0.928 and an Accuracy of 0.882. This showcases the FCRF method's balanced performance across both continuous and categorical data under the MCAR mechanism. The results are in Table 2.

3.2 MAR Data Scenario

The MAR data scenario, characterized by missingness dependent on observed variables, further highlighted the strengths of the FCRF method. Achieving an Average RMSE of 1.068 and an Average AE of 0.281, FCRF performed on par with Simple Imputation and demonstrated improved efficiency over KNN. The FCRF method also showed competitive classification performance, with an F1-score of 0.926 and an Accuracy of 0.883, underscoring its effectiveness in contexts where missing data is related to other observed data. The results are in Table 3.

3.3 MNAR Data Scenario

The MNAR scenario, often considered the most challenging due to missingness being related to the missing data itself, saw the FCRF method achieving an Average

RMSE of 0.571 and an Average AE of 0.136. These results not only exhibit the method's superior performance in accurately imputing missing values but also emphasize its precision in handling complex missing data patterns. The F1-score of 0.918 and an Accuracy of 0.875 further affirm FCRF's robustness across varied data types within the MNAR context. The results are in Table 4.

3.4 Discussion

The comparative analysis reveals the FCRF method's consistently strong performance across all three missing data scenarios. While the Iterative and MICE methods showed lower average RMSE and AE in some cases, the FCRF method's advantages become evident when considering its higher F1-scores and Accuracy in the challenging MNAR scenario. This suggests that while FCRF may not always have the lowest error metrics, it maintains a high level of predictive accuracy, particularly for categorical and binary data—a crucial aspect in many real-world applications.

The FCRF method's balanced approach to imputation, leveraging both clustering to identify inherent data structures and Random Forest models to predict missing values, offers a nuanced capability to handle missing data. This method proves especially valuable in scenarios where the relationship between data points and missingness is complex, showcasing its potential as a versatile and effective tool for data imputation.

Consider the inherent challenge of data imbalance, which significantly impacts imputation accuracy, especially for categorical variables. In datasets where certain classes are underrepresented, traditional imputation methods may inadvertently skew

towards the majority class, leading to biased imputations. The application of oversampling techniques, such as Synthetic Minority Over-sampling Technique (SMOTE), before the imputation process can alleviate this bias by creating synthetic instances of the minority class, thus providing a more balanced dataset for the Random Forest models within the FCRF framework. While beneficial, oversampling must be applied judiciously to prevent the introduction of bias or overfitting, which could adversely affect imputation accuracy.

The performance of the FCRF method heavily relies on the quality of clustering achieved by Fuzzy C-Means. Effective clustering that accurately reflects the underlying structure of the data can significantly improve the precision of the subsequent Random Forest imputation. Clusters that capture genuine groupings within the data allow for more targeted and contextually appropriate imputations. Assessing and optimizing the quality of clustering is therefore crucial. This might involve selecting the optimal number of clusters, adjusting clustering parameters, or incorporating additional preprocessing steps to enhance the distinctiveness and relevance of the clusters formed.

Chapter 4. Conclusion and Future work

This thesis presented a comprehensive analysis of the Fuzzy C-Means based Random Forest (FCRF) imputation method's performance compared to conventional imputation methods in handling missing data within a CABG surgery dataset. While the FCRF method did not outperform all traditional methods according to the chosen evaluation metrics, it demonstrated a robust potential in managing the complex nature of clinical

datasets.

The results highlighted the nuanced requirements of imputation in medical data, where preserving underlying relationships can be as critical as achieving low error rates. The comparative analysis provided valuable insights into the strengths and weaknesses of various imputation techniques, emphasizing that the choice of method must be tailored to the specific characteristics of the dataset and the research objectives.

Future research should investigate alternative approaches to aligning sample sizes post-imputation, focusing on methods that retain the original dataset's characteristics to avoid potential bias in performance evaluation. Also required to explore the impact of different initial imputation techniques on clustering and dimensionality reduction. There is a need for strategies that can maintain the inherent structure and variability of the data without resorting to mean imputation. Exploring advanced clustering techniques that can accommodate missing data may offer improvements in the pre-imputation phase, leading to more accurate clusters and, consequently, more precise imputation. Incorporating cross-validation methods that are robust to variations in sample size could provide a more accurate measure of an imputation method's performance, ensuring that results are not skewed by the way the data is partitioned. Research into more effective ways of handling oversampling is needed to prevent the overestimation of imputation model performance, particularly in datasets where similar records are numerous. The ultimate test of any imputation method lies in its application to real-world data and its validation through practical deployment in clinical settings. Subsequent research could apply the FCRF method to other medical datasets, validate

its effectiveness in various clinical scenarios, and compare its performance with the outcomes of actual patient records. Another promising avenue is integrating the imputation process with predictive modelling to directly measure the impact of different imputation strategies on the predictive accuracy of clinical outcome models.

Appendix

Table 1 Descriptive statistics of CABG data

Variables	Missing data (N)	Missing data (%)	Values
Race	0	0	0-4 race group
Age	0	0	In Years
Gender	0	0	Male, Female
BMI	0	0	kg/m2
RF-Diabetes	2	0.02	Binary
Dialysis	0	0	Binary
Hypertn	0	0	Binary
TobaccoU	0	0	None, Current, Former
ChrLungD	3	0	Binary
Hct	0	0	Continues
PostopReinM_IVes_Simp	0	0	Binary
LOS	17	0.14	In Days
A1cLvl	0	0	Mmol/mol
CreatLst	0	0	Mg/dL
HDEFD	9	0.07	Binary
PrCVInt	0	0	Binary
NumDisV	0	0	0-3 numbers
SurSInf	0	0	Binary
COpReBld	0	0	Binary
COpReVlv	0	0	Binary
COpReOth	0	0	Binary
CSepsis	0	0	Binary
CNStrokP	0	0	Binary
CNParal	0	0	Binary
CPVntLng	0	0	Binary
CPPneum	0	0	Binary
CRenFail	0	0	Binary
COtArrst	0	0	Binary

COtAFib	0	0	Binary
Readmit	0	0	Binary
CardRef	0	0	Binary
DisLoctn	0	0	Binary
FTR	0	0	Binary
RF-Endocarditis	3	0.02	Binary
HmO2	2356	19.11	Binary
SlpApn	2356	19.11	Binary
Pneumonia	2358	19.13	Binary
RF-IVDrugU	2357	19.12	Binary
AlcoholU	2361	19.15	Binary
LiverDis	2356	19.11	Binary
ImmSupp	4	0.033	Binary
MediastRad	2358	19.13	Binary
PAD	3	0.024	Binary
Syncope	2355	19.10	Binary
CVD	4	0.03	Binary
WBC	41	0.33	Million/mm
MELDScr	3011	24.42	Continuous
PrevMI	0	0	Binary
CarShock	0	0	Binary
CarCathPer	2354	19.10	Binary
Status	0	0	Binary
OpOCard	1	0.01	Binary
CPBUtil	0	0	Binary
CircArr	1	0.01	Binary
AortOccl	1614	13.09	Binary
IBldProd	32	0.26	Binary
NumRadDA	694	5.63	0-5 distal anastomoses
Complics	1	0.01	Binary
s_no	0	0	12328

Figure 1. The generation framework for MCAR (Missing Completely At Random), MAR (Missing At Random), and MNAR (Missing Not At Random).

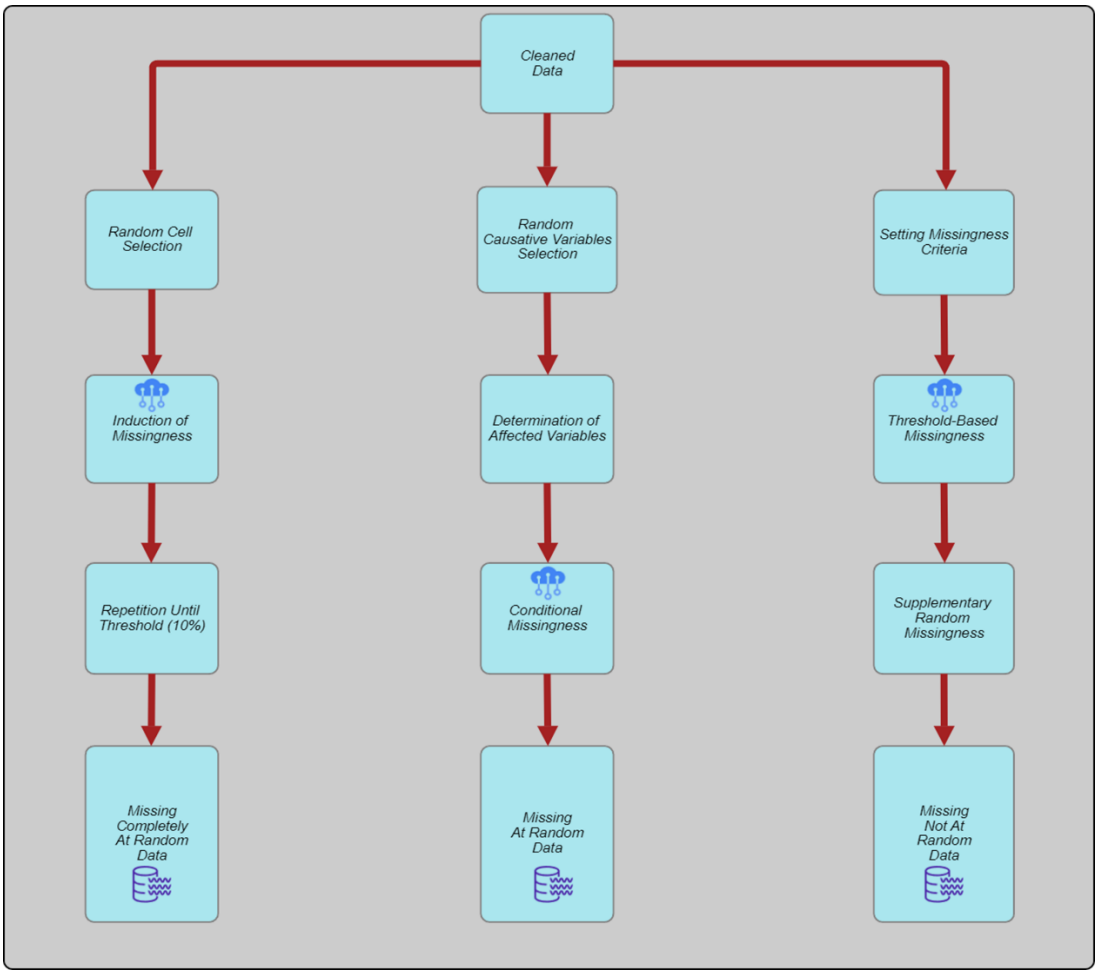

Figure 2. Proposed Fuzzy C-means based random forest imputation method (FCRF) framework.

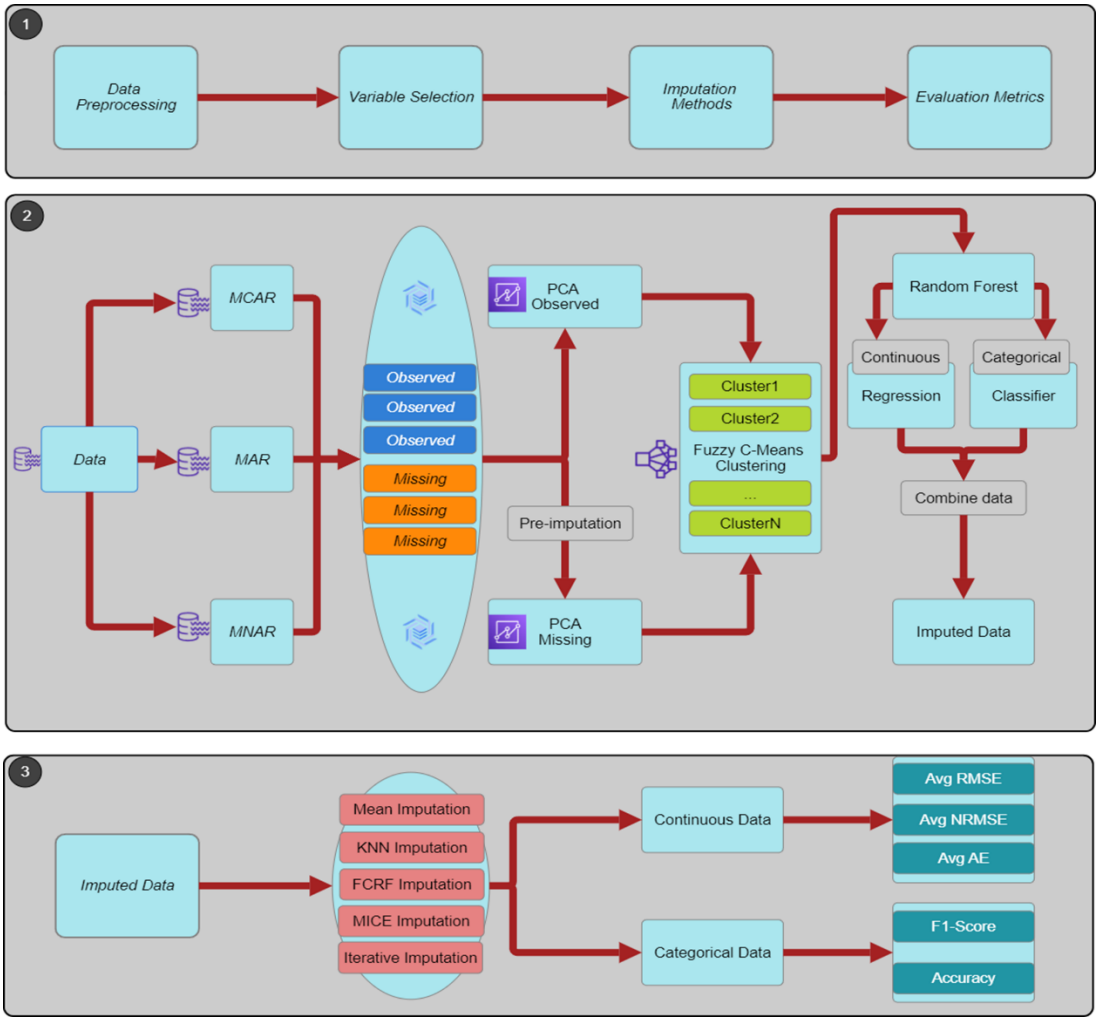

Table 2. Evaluation Metrics over five experiments for MCAR data.

	FCRF	Simple	KNN	Iterative	MICE
AVG RMSE	1.474	1.474	1.631	1.252	1.252
AVG NRMSE	0.001	0.001	0.001	0.000	0.000
AVG AE	0.316	0.317	0.336	0.266	0.265
F1-score	0.928	0.985	0.979	0.947	0.947
Accuracy	0.882	0.986	0.962	0.900	0.900

Table 3. Evaluation Metrics over five experiments for MAR data.

	FCRF	Simple	KNN	Iterative	MICE
AVG RMSE	1.068	1.068	1.124	0.827	0.827
AVG NRMSE	0.000	0.000	0.000	0.000	0.000
AVG AE	0.281	0.283	0.297	0.217	0.217
F1-score	0.926	0.983	0.978	0.947	0.947
Accuracy	0.883	0.984	0.962	0.903	0.903

Table 4. Evaluation Metrics over five experiments for MNAR data.

	FCRF	Simple	KNN	Iterative	MICE
AVG RMSE	0.571	0.557	0.638	0.461	0.461
AVG NRMSE	0.000	0.000	0.000	0.000	0.000
AVG AE	0.136	0.120	0.161	0.107	0.107
F1-score	0.918	0.988	0.916	0.934	0.934
Accuracy	0.875	0.990	0.895	0.890	0.890

References

1. Hühn, J., & Hüllermeier, E. (2009). FURIA: an algorithm for unordered fuzzy rule induction. Data Mining and Knowledge Discovery, 19, 293-319.
2. Jerez, J. M., Molina, I., García-Laencina, P. J., Alba, E., Ribelles, N., Martín, M., & Franco, L. (2010). Missing data imputation using statistical and machine learning

methods in a real breast cancer problem. Artificial intelligence in medicine, 50(2), 105-115.

3. Rahman, M. M., & Davis, D. N. (2013). Machine learning-based missing value imputation method for clinical datasets. In IAENG Transactions on Engineering Technologies: Special Volume of the World Congress on Engineering 2012 (pp. 245-257). Springer Netherlands.
4. Schmitt, P., Mandel, J., & Guedj, M. (2015). A comparison of six methods for missing data imputation. Journal of biometrics & biostatistics, 6(1), 1.
5. Huang, J., Keung, J. W., Sarro, F., Li, Y. F., Yu, Y. T., Chan, W. K., & Sun, H. (2017). Cross-validation based K nearest neighbor imputation for software quality datasets: an empirical study. Journal of Systems and Software, 132, 226-252.
6. Austin, P. C., White, I. R., Lee, D. S., & van Buuren, S. (2021). Missing data in clinical research: a tutorial on multiple imputation. Canadian Journal of Cardiology, 37(9), 1322-1331.
7. Desiani, A., Dewi, N. R., Fauza, A. N., Rachmatullah, N., Arhami, M., & Nawawi, M. (2021). Handling missing data using combination of deletion technique, mean, mode and artificial neural network imputation for heart disease dataset. Science and Technology Indonesia, 6(4), 303-312.
8. Fouad, K. M., Ismail, M. M., Azar, A. T., & Arafa, M. M. (2021). Advanced methods for missing values imputation based on similarity learning. PeerJ Computer Science, 7, e619.
9. Ge, Y., Li, Z., & Zhang, J. (2023). A simulation study on missing data imputation for

dichotomous variables using statistical and machine learning methods. Scientific Reports, 13(1), 9432.

10. Honghai, F., Guoshun, C., Cheng, Y., Bingru, Y., & Yumei, C. (2005, September). A SVM regression-based approach to filling in missing values. In International Conference on Knowledge-Based and Intelligent Information and Engineering Systems (pp. 581-587). Berlin, Heidelberg: Springer Berlin Heidelberg.
11. Jadhav, A., Pramod, D., & Ramanathan, K. (2019). Comparison of performance of data imputation methods for numeric dataset. Applied Artificial Intelligence, 33(10), 913-933.
12. Amiri, M., & Jensen, R. (2016). Missing data imputation using fuzzy-rough methods. Neurocomputing, 205, 152-164.
13. Pelckmans, K., De Brabanter, J., Suykens, J. A., & De Moor, B. (2005). Handling missing values in support vector machine classifiers. Neural Networks, 18(5-6), 684-692.
14. Khan, S. I., & Hoque, A. S. M. L. (2020). SICE: an improved missing data imputation technique. Journal of big Data, 7(1), 37.
15. McKinney, W., & others. (2010). Data structures for statistical computing in python. In Proceedings of the 9th Python in Science Conference (Vol. 445, pp. 51–56).
16. Bezdek, J. C. (2013). Pattern recognition with fuzzy objective function algorithms. Springer Science & Business Media.
17. Breiman, L. (2001). Random forests. Machine learning, 45, 5-32.
18. Jolliffe, I. T. (2002). Principal component analysis for special types of data (pp. 338-

372). Springer New York.

19. Van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. Statistical methods in medical research, 16(3), 219-242.
20. Oliphant, T. E. (2007). Python for scientific computing. Computing in science & engineering, 9(3), 10-20.
21. Manyam RB, Shen H. Liu Z, Zhang Y, Hu X, Keeling WB., 2024, Machine Learning Algorithms Accurately Predict Risk Factors for Failure to Rescue After Coronary Artery Bypass Grafting, Abstract accepted at: 60th Annual Meeting of the Society of Thoracic Surgeons.