Imputation Techniques for Heart Disease Classification using Machine Learning

Deeksha U, Aishwarya Kulal,*Nirmal Kumar Nigam

Department of Data Science and Computer Applications

Manipal Institute of Technology,

Manipal Academy of Higher Education,

Manipal, India-576104

Abstract:

The ever-growing importance of accurate and reliable healthcare data has prompted a comprehensive investigation into various imputation techniques to handle missing data for heart disease classification. This research rigorously explores diverse imputation techniques for heart disease classification, employing the Cleveland, Framingham, and Heart Attack Prediction datasets. The datasets encompass essential health indicators, including age, gender, heart rate, blood pressure, blood sugar, and Test-Troponin. Our study systematically evaluates and compares the effectiveness of imputation methods based on Non-ML based Imputation techniques such as Forward fill, backward fill, hot deck, Deletion imputation techniques and ML based imputation techniques such as KNN imputation and SVM imputation. Furthermore, through the integration of multiple datasets and the application of ensemble techniques, we were able to enhance the completeness of the datasets and improve the performance of classification algorithms. The research contributes valuable insights to enhance the reliability of predictive models in cardiovascular health studies, demonstrating that the accuracy achieved with merged datasets surpasses that of using individual datasets alone.

Keywords: **Missing data, Imputation Methods, Machine Learning.**

1. Introduction

Cardiovascular diseases, particularly heart attacks, stand as pervasive global health concerns, necessitate accurate predictive modeling for effective healthcare interventions [1]. However, the accuracy of these prediction tools depends on how good and completes the data they use. Missing data poses a significant obstacle, prompting the exploration of adept imputation techniques [2].This Research focuses on the various imputation methods tailored for CVD detection utilizing the datasets. Medical dataset such as Cleveland and Heart Attack datasets paves path for scientists to develop a research model that resolves all underlying complexities in health domain. And, Cleveland dataset is a case study in clinical research [3]. These dataset is well-structured leading to the coverage of major attributes for the medical condition such as aged, gender, pulse, BP, blood sugar, CK-MB and Test-Troponin. This system of together, those parameters contribute complete information that is constituted by lots of the most sophisticated provisions and by the way it helps in getting many-sided description of the cardiovascular health of an individual. This wisdom imparted by the Heart Attack Prediction data is not only rounded-up with the detection of the factors that result in the outbreak or the cardiac attack but it is beyond that [4]. A difficult task is to detect and extract the patterns from the data, which turns to be a hard problem for any researcher as we come up with everyday situation of data suffering from incomplete data [5]. Consequently, a wide range of methods handling on the various details of the other essential techniques are of significant relevance, for the accuracy and completeness of the given data. All of the consequences of not taking care of wasteful opinion though are wide ranging from the analysis, a forecast and the decision to arrive at wrong place almost too false information. On the one hand, data poverty may affect the investigation in such a way that the conclusions of that research cannot be sufficiently analyzed and observed in depth and in details as well as it cannot provide precise. Determining the cause of missing values is crucial for success of the result In medical care these gaps are from various root causes including randomness, human error'; missing of items of equipment, or the patients who don't respond or refuse [6]. The random appearance of missing values arises as an outcome of the error done by humans or the omission during data entry or data collection. Such errors are very difficult to forecast and can show up on essential services unpredictably or irregularly. The much of incomplete data make also come from equipment malfunction during data collection, particularly in healthcare datasets [7], where these technical failures may result in either non-values or inaccurate images. Likewise, patients see it fit to dodge or return evasive responses when given survey-type or during clinical interviews which lead for missing values in the data collection. This research paper is targeted at a detailed work of imputation methods that is to be used for classification of heart diseases. This is under the vision of the current increments and technological advancements and to get accurate results.

2. Methodology

Imputation techniques in ML refer to the methods that are used to fill in missing values within a datasets [8]. Missing data is a common issue in the datasets, and handling it correctly is crucial for building accurate and strong result.

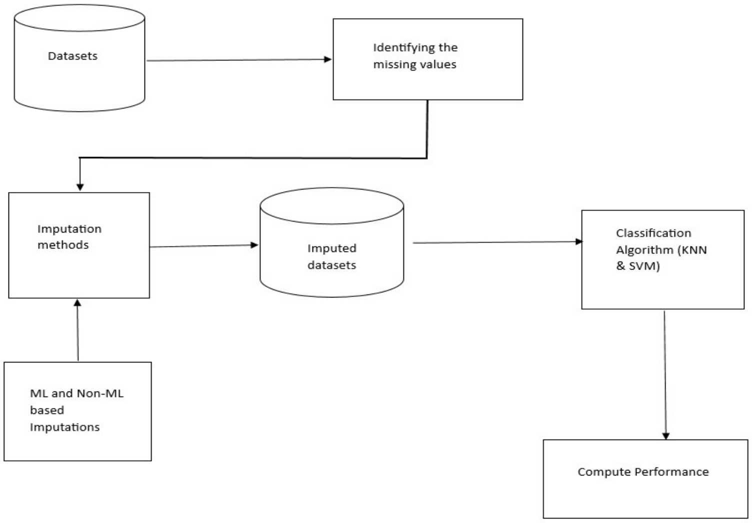
Datasets
Identifying the missing values
Imputation methods
Imputed datasets
Classification Algorithm (KNN & SVM)
ML and Non-ML based Imputations
Compute Performance

Figure 1. Block Diagram of Imputation Methods before Merging Data

In Figure 1, individual datasets are depicted, each undergoing a meticulous process of identifying missing values and subsequently applying various imputation techniques. These techniques encompass both machine learning (ML) and non-ML methods, ensuring a comprehensive approach to handling missing data. Following imputation, the imputed datasets serve as inputs for further analysis. The imputed datasets are subjected to classification algorithms such as K-Nearest Neighbors (KNN) and Support Vector Machine (SVM), elucidating the effectiveness of the imputation methods in enhancing the predictive capabilities of subsequent models.

2.1 Non-Ml based Imputation Techniques

2.1.1 Forward Fill: Forward fill imputation is applied to solve the missing value from the datasets [9]. This imputation technique entails adding additional value through having values that was closer to that which was taken as ideal for calculating non null value which was observed with strong assumption that there is a chance of continuity in those values[10]. Implementation: Implementing forward fill, we have utilized its method from respective library. This approach efficiently propagates the recent observed values in the dataset, same way as addressing the missing values too.

2.1.2 Backward Fill: Backward fill Imputation Technique is used to handle missing values from the dataset. This technique includes filling missing values with their respective observed value with the specified axis [11]. That is particularly effective in the case where missing values exhibits a temporal data patterns**.** Implementation: Implementing backward fill, the method from panda's library is used to instruct for missing values with the next observed values along the respected axis, using the last observed value backward to replace missing values.

2.1.3 Hot Deck Imputation Techniques: This Imputation technique selects a value from similar non-missing data points from the dataset, a simple yet useful method for imputing missing values [12]. It is useful when there's an expectation that similar observations should have similar data values. Implementation: For Hot Deck Imputation, a custom logic can be used to perform the imputation. The process includes choosing a suitable metric to measure similarity and choosing suitable features for matching.

2.1.4 Deletion Imputation: Deletion Imputation includes the removal of observations or features with missing values from the dataset [13]. While simple to implement, this method can lead to information loss and biased results, especially if missing values are not completely at random (MCAR). Implementation: Deletion Imputation can be carried out by removing either entire rows (list wise deletion) or specific columns (feature-wise deletion) containing missing values. Careful consideration should be given to the extent of missing ness and its potential impact on the analysis.

2.2 Ml based Imputation Techniques

2.2.1 K-Nearest Neighbors (KNN) Imputation: The KNN imputation is centered not only on those K values of each of the data points but it's also by replacing the missing values with average or mean values. The diversity of the methods may vary from numerical data and also using mode for categorical data values [14]. Implementation:

Scikit-learn's KNN Imputer is utilized for KNN-based imputation. Implementation includes selecting an suitable value for k, which determines the number of neighbors considered during imputation.

2.2.2 Support Vector Machine (SVM) Imputation: SVM Imputation is the predictive power of SVM algorithms to estimate missing values based on the relationships between features [15]. SVM imputation is effective when there are complex non-linear relationships in the data. Implementation: For SVM Imputation, custom logic can be utilized. The process involves training an SVM model on the complete data, using non-missing values as training data, and then predicting missing values based on the trained patterns.

2.3 Classification for Handling Missing Values

2.3.1 K-Nearest Neighbors (KNN): KNN is a versatile algorithm used for imputing missing values in datasets. It operates by identifying the nearest neighbors to the missing values based on a chosen distance measure, typically the Euclidean distance [16]. The Euclidean distance formula given by distance *xy* calculates the distance between instances with complete and incomplete data attributes [17]. It is given by:

$$d = \sqrt{(x_2 - x_1)^2 + (y_1 - y_2)^2} \tag{1}$$

Here, x_{ik} is the value of attribute j containing missing data values, and X_{jk} is the value of the j_{th} attribute with complete data from the datasets.

Once the K nearest neighbors is identified, the Weighted Mean Estimation is calculated to impute the missing values. The mean estimation X_k is determined using the formula:

$$X_k = \frac{\in_{j=1}^{J} w_j . v_j}{\text{C}_{j=1}^{J} w_j} \tag{2}$$

Here, J is the number of parameters, v_j the complete values on attributes containing missing data points, and w_j is weight assigned to the nearest neighbours observed.

The weighted value is given by the equation:

$$W_j = \frac{1}{d_j} \tag{3}$$

KNN imputation technique is versatile for discrete and continuous values, it compromise precision and introduce false associations [18], additionally, increased computational time due to its exhaustive search through the data points from the datasets.

2.3.2 Support Vector Machine (SVM): SVM is a powerful ML algorithm utilized for missing data handling. SVM seeks to find an optimal separating hyper plane that maximizes the distance from the hyper plane to nearest data points.

Hyper planes are defined by the equation:

$$w . x + b = 0 \tag{4}$$

Here, w is weight vector x is input vector, and b is bias.

SVM regression-based methods have been employed for missing data imputation, where decision attributes are set as condition attributes, and SVM regression predicts the condition attribute values. The precision of SVM regression has been demonstrated in various experiments [19], with some studies showing superior performance, particularly on specific datasets.

In conclusion, both KNN and SVM classification offer valuable tools for handling missing values in datasets. While KNN excels in identifying nearest neighbors and imputing missing values based on their attributes, SVM focuses on finding optimal hyper planes for regression or classification tasks. The choice between these methods depends on the dataset's characteristics and the specific requirements of the analysis [20].

3. Results

Implementing a variety of imputation techniques, both non-ML and ML-based, provided valuable insights into handling missing data within the dataset. Non ML methods like forward fill backward fill effectively maintained dataset by utilizing temporal patterns. And hot deck imputation has its efficacy towards missing value imputation based on the similar non missing values. But the simplicity of Deletion imputation has a cost of important information loss and not accurate results. And the ML based imputation techniques such as K-Nearest Neighbor Imputation capitalized on basic structure of the data set values. Moreover, SVM Imputation shows its effectiveness in handling the missing values, particularly in case of complex nonlinear dependencies. The Resulted dataset will perform enhanced completeness, with missing values effectively filled in through the training data of SVM model on complete data.

Table 1. Accuracy Assessment of Non Ml based Imputation Techniques

Datasets	Forward fill		Backward fill		Deletion		Hotdeck	
	KNN	SVM	KNN	SVM	KNN	SVM	KNN	SVM
HEART	0.65	0.83	0.64	0.82	0.63	0.86	0.65	0.83
HEART_UCI	0.54	0.54	0.55	0.57	0.56	0.68	0.53	0.53
FRAMINGHAM	0.83	0.85	0.82	0.84	0.83	0.83	0.82	0.85
HEART_ATTACK	0.64	0.79	0.64	0.80	0.58	0.76	0.63	0.80

4. Comparative Analysis

On Comparing various non-ml based approaches like Forward fill, back fill, hot deck and deletion methods with the ml based approaches which use techniques like KNN and SVM methods, we obtained long tails among them, the Non-Ml and Ml allows a large scale computational advantage, that can be caught up among them is ml methods. And a temporal continuity is maintained by using forward and backward fill equivalent that they are intensifying the errors. On the other way the hot deck imputation is biased at the same time they do provide an ease to same case of data with many similarities among them. And the actual decision shall be dependent on the various conditions such as if the data is changing based on seasonal conditions or the found data is perfect in all cases.

Table 2. Accuracy Assessment of Ml based Imputation Techniques

Datasets	KNN Imputation		SVM Imputation	
	KNN	SVM	KNN	SVM
HEART	0.66	0.83	0.68	0.82
HEART_UCI	0.42	0.54	0.56	0.62

FRAMINGHAM	0.82	0.85	0.82	0.86
HEART_ATTACK	0.63	0.79	0.68	0.78

In Summary, the non-ml and ml based imputation techniques comparison highlights the diverse options for handling the missing data, each with its own set of limitation and usefulness, Non-ml method has its own simplicity and it leads to lack of capturing the data patterns , whereas ml based approaches has much more advanced techniques in relationship to parameter tuning. On this selecting the best imputation method to fill in the missing values should depend on the dataset characteristics [21].

5. Integrating the Datasets

In this study, we showed the data insufficiency by merging the 4 separate datasets into one dataset. To achieve the dataset completeness and increase the performance of ML algorithms in the classification tasks [22]. To achieve this, we applied both ML and non-ML based imputation techniques to handle missing values, followed by classification using K-Nearest Neighbors (KNN) and Support Vector Machine (SVM) algorithms. Our objective was to demonstrate that the integrated dataset, along with appropriate imputation and classification techniques, can yield higher accuracy compared to individual datasets.

5.1 Simple Ensemble through Random Sampling and Merging of Datasets

Simple ensemble methods serve as foundational strategies in the pursuit of enhancing predictive model performance through the combination of multiple models. While more complex ensemble techniques like Bagging, Boosting, and Stacking exist, simple ensemble remains a valuable and accessible approach, particularly for scenarios where computational resources or model interpretability are crucial considerations [23].

Random Sampling: The initial step involves random sampling of each original dataset (Framingham, heart-attack, heart, heart-uci) using the sample method. The fraction of data to be sampled is controlled by the parameter sample-size, ensuring flexibility in adjusting the sample size based on memory constraints or other considerations.

5.2 Results from Merged Datasets

The Classifications which was done resulted into sharper accuracy which was above the results of dataset that were done without the imputation techniques. And we managed to capture a more complete result of the data dimensionality by merging the dataset that gave us an excellent classification model. Additionally, the ensemble of ml and non-ml fill in techniques prevented the inaccuracy that was created by missing data from the classification algorithm. Altogether employing such an integrated data, together with successfully implementing and classification techniques the higher performance of both techniques has gained, namely KNN and SVM.

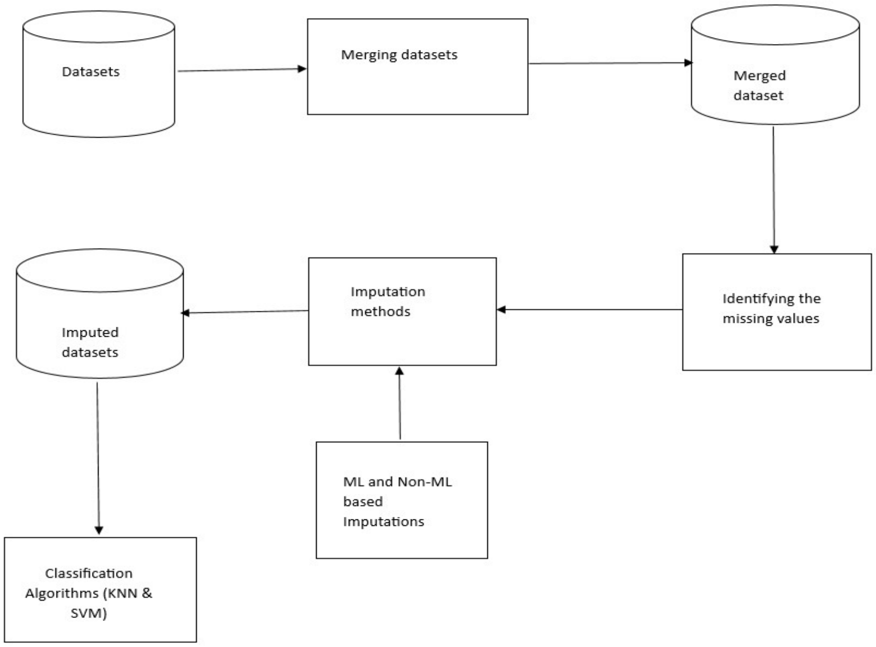
Datasets
Merging datasets
Merged dataset
Identifying the missing values
Imputation methods
Imputed datasets
ML and Non-ML based Imputations
Classification Algorithms (KNN & SVM)

Figure 2. Block diagram illustrating imputation techniques applied post data integration.

This Study has therefore proved the integration of dataset, imputation and classification which was a key component that leads to improved performance. Overall the data insufficiency and achievement of high accuracy in classification has gained my merging the datasets.

Table 3. Combined Accuracy Assessment of Non-ML and ML Imputation Approaches on Merged Dataset

Parameters	NON ML BASED IMPUTATION								ML BASED IMPUTATION			
	Forward fill		Backward fill		Deletion		Hotdeck		KNN Imputation		SVM Imputation	
	KNN	SVM	KNN	SVM	KNN	SVM	KNN	SVM	KNN	SVM	KNN	SVM
Test size=0.1, Random state=42	0.96	0.96	0.97	0.98	0.99	0.96	0.98	0.92	0.98	0.92	0.98	0.94
Test size=0.2, Random state=52	0.97	0.95	0.97	0.99	0.98	0.95	0.98	0.89	0.98	0.89	0.98	0.93

6. Conclusion

In the present study, Non-Ml and ML imputation methods were used to perform this exhaustive research of heart disease classification using various imputation techniques. Data accurateness is core to the whole process of predictive models in healthcare. Uncovering how missing values influence the quality of predictions assist in designing methods that could possibly detect or eliminate these values. We saw practically the difference in trade-offs between non-ml input and in the text. Non-ml methods provide simple and fast tools well adapted to data with low complexity, while ml-based approaches convey highly advanced techniques for complex data patterns, but need more computational resources and hyper parameters tuning. Moreover, we used ensemble techniques to merge varied datasets and provide additional data completion; following leading to better outcomes for the algorithms used a classification step. The result implies, that the merged dataset accompanied with proper imputation and classification techniques, has the highest precision compared to apply on each source dataset.

Acknowledgments

Sincere gratitude to Mr. Nirmal Kumar Nigam for their invaluable assistance and support throughout the course of this research project. Their expertise and insights have greatly enriched our work and contributed significantly to its success.

REFERENCES

[1] G.A. Roth et al., "Global Burden of Cardiovascular Diseases and Risk Factors, 1990-2019: Update From the GBD 2019 Study," J. Am. Coll. Cardiol., vol. 76, no. 25, ***(2020)*** *December, pp. 2982-3021.*

[2] Q. Song and M.J. Shepperd, "Missing Data Imputation Techniques," Int. J. Bus. Intell. Data Min., vol. 2, no. 3, ***(2007)*** *October, pp. 261-291.*

[3] M. Rahman and D. N. Davis, "Machine Learning Based Missing Value Imputation Method for Clinical Datasets," Lecture Notes in Electrical Engineering, vol. 229, ***(2012)*** *January.*

[4] A. A. Ahmad and H. Polat, "Prediction of Heart Disease Based on Machine Learning Using Jellyfish Optimization Algorithm," Diagnostics (Basel), vol. 13, no. 14, ***(2023)*** *July, pp. 2392.*

[5] S.I. Khan and A.S.M.L. Hoque, "SICE: an improved missing data imputation technique," J. Big Data, vol. 7, ***(2020)*** *June.*

[6] T. Muntinova, "Data Analysis of Heart Attack Risk Factors: Insights from Machine Learning," presented at the VI International Scientific Conference, Toronto, Canada, ***(2024)*** *February.*

[7] H. Kang, "The prevention and handling of the missing data," Korean J. Anesthesiol., vol. 64, no. 5, ***(2013)*** *May, pp. 402–406.*

[8] C.E. Brodley and M.A. Friedl, "Identifying Mislabeled Training Data," J. Artif. Intell. Res., vol. 11, ***(1999)*** *August, pp. 131–167.*

*[9] D.M.P. Murti, U. Pujianto, A. Wibawa, and M.I. Akbar, presented "K-Nearest Neighbor (K-NN) based Missing Data Imputation" at the 2019 5th International Conference on Science in Information Technology (ICSITech),****(2019)*** *October.*

[10] K.T. Do et al., "Characterization of missing values in untargeted MS-based metabolomics data and evaluation of missing data handling strategies," Metabolomics, vol. 14, no. 10, ***(2018)*** *September, pp. 128.*

[11] M. Ichikawa et al., "Handling missing data in an FFQ: multiple imputation and nutrient intake estimates," Public Health Nutr., vol. 22, no. 8, ***(2019)*** *June, pp. 1351–1360.*

[12] R. R. Andridge and R. J. A. Little, "A Review of Hot Deck Imputation for Survey Non-response," Int Stat Rev., vol. 78, no. 1, ***(2010)*** *April, pp. 40–64.*

[13] U. Yelipe, S. Porika, and M. Golla, "An efficient approach for imputation and classification of medical data values using class-based clustering of medical records," Computers & Electrical Engineering, vol. 66, ***(2018)*** *February, pp. 487-504.*

[14] A.B. Pedersen et al., "Missing data and multiple imputation in clinical epidemiological research," Clin. Epidemiol., vol. 9, ***(2017)*** *March, pp. 157–166.*

[15] S. Alam, M. S. Ayub, S. Arora, and M. A. Khan, "An Investigation of the Imputation Techniques for Missing Values in Ordinal Data Enhancing Clustering and Classification Analysis Validity," Decision Analytics J., vol. 9, ***(2023)*** *December, p. 100341.*

[16] Plaia and A.L. Bondi, "Single imputation method of missing values in environmental pollution data sets," Atmos. Environ., vol. 40, no. 38, ***(2006)*** *December, pp. 7316–7330.*

[17] G. Seni and J.F. Elder, "Ensemble Methods in Data Mining: Improving Accuracy Through Combining Predictions," Synth. Lect. Data Min. Knowl. Discov., vol. 2, no. 1, ***(2010)*** *January, pp. 1-126.*

[18] D. Shah, S. Patel, and S.K. Bharti, "Heart Disease Prediction using Machine Learning Techniques," SN Comput. Sci., vol. 1, ***(2020)*** *October, p. 345.*

[19] K. Sethia, A. Gosain, and J. Singh, presented "Review of Single Imputation and Multiple Imputation Techniques for Handling Missing Values" at the Proceedings of Third Emerging Trends and Technologies on Intelligent Systems (ETTIS 2023), Singapore, ***(2023)*** *September.*

[20] S. Venkatraman, A. Yatsko, A. Stranieri, and H. F. Jelinek, presented "Missing data imputation for individualised CVD diagnostic and treatment" at the 2016 Computing in Cardiology Conference (CinC), Vancouver, BC, Canada, ***(2016)*** *September.*

[21] N. Louridi, S. Douzi, and B. El Ouahidi, "Machine learning-based identification of patients with a cardiovascular defect," J. Big Data, vol. 8, ***(2021)*** *October, p. 133.*

[22] D. Ravì, C. Wong, F. Deligianni, M. Berthelot, J. Andreu-Perez, B. Lo, and G.-Z. Yang, "Deep Learning for Health Informatics," IEEE Journal of Biomedical and Health Informatics, vol. 21, no. 1, ***(2017)*** *January, pp. 4-21.*

[23] A. Jazayeri, O.S. Liang, and C.C. Yang, "Imputation of Missing Data in Electronic Health Records Based on Patients' Similarities," J. Healthc. Inform. Res., vol. 4, ***(2020)*** *September, pp. 295–307.*