Engineered Science

DOI: https://dx.doi.org/10.30919/es8d681

Ischemic Heart Disease Multiple Imputation Technique Using Machine Learning Algorithm

D. Cenitta,[1,#] R Vijaya Arjunan[1,*] and Prema K V[2,#]

Abstract

Medical datasets in profound data repository like the University of California Irvine (UCI) have missing values. These essential data are used for multiple analyses by researchers in a variety of predictions even though the results could be biased at times. It necessitates an important practice to use missing data imputation methods to fill up missing values for arriving at validated experimental results ensuring unbiased outcomes and predictions especially when the heart disease data set is handled. These methods are a type of treatment for data sets that include uncertainty and vagueness. Methods based on fuzzy-rough sets, on the other hand, offer excellent tools for dealing with ambiguity, with desirable properties such as robustness and noise tolerance. Fuzzy sets can also find minimal data representations and do not need potentially erroneous user inputs which confirms using fuzzy-rough sets for imputation be viable. In this paper we propose a novel Ischemic Heart Disease Multiple Imputation Technique (IHDMIT) missing value imputation methods based on fuzzy-rough sets and their recent extensions. The proposed IHDMIT with Random Forest classifier is compared with fuzzy roughest, fuzzy C means, and expectation maximization. The result shows that the proposed IHDMIT random forest classifier gives better accuracy of 93%.

Keywords: Fuzzy rough Set; Machine learning; Multiple imputation techniques; Random forest; UCI.
Received: 18 November 2021; Revised: 05 March 2022; Accepted: 12 March 2022.
Article type: Research article.

1. Introduction

The computerization of society has significantly improved in collecting and generating data from various sources. A massive amount of data has flooded into nearly all aspects of our lives. This massive increase in transient data has generated an urgent need for automated tools and innovative techniques that can cleverly help us convert vast quantities of information into knowledge and usable information. This leads to data mining development and its applications, a thriving and exciting frontier in computer science. "Data mining also referred to as knowledge discovery from data (KDD), is the automated extraction of patterns considering knowledge captured in large databases, the web, warehouses, other huge information repositories or data streams".[1]

In the medical realm exploring hidden patterns plays a vital role. But, the current raw form of medical data is extensively distributed, heterogeneous and voluminous. These datasets need to be collected in an organized format. Data mining algorithm provides an organized approach to the hidden patterns used for clinical diagnoses, predicting the severity of diseases, characterization of patient clusters, *etc.* "Cardiovascular diseases (CVD) are disorders of the heart and blood vessels, including rheumatic heart disease, cerebrovascular disease, coronary heart disease, and other conditions". CVD is the world's leading cause of death. Every year, many people die as a result of CVD. According to a world health organization (WHO) report,[2] 17.9 million people died from CVD in 2016, accounting for 31% of all deaths globally. Stroke or heart attack accounts for 85 percent of these deaths. CVD fatalities account for three-quarters of all deaths, primarily in developing countries, and low- and middle-income countries are affected. Of the 17 million deaths that occur too soon caused by non-communicable diseases in 2015, 82% were in countries with low and moderate income, and CVDs affected 37%. There is a need to detect CVDs, and data mining algorithms can be used to achieve this.[3]

A factor in data analytics that degrades productivity is missing data. An incorrect imputation may lead to an incorrect prediction of missing values.[4] When a great amount of data is

[1] *Department of Computer Science and Engineering, Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal, Karnataka 576104, India.*

[2] *Department of Computer Science and Engineering, Manipal Institute of Technology, Manipal Academy of Higher Education, Bengaluru, Karnataka 560064, India.*

[#]*These authors contributed to this work equally.*

**E-mail:* vijay.arjun@manipal.edu (V. Arjun)

generated every second in this age of big data, and this data is a major concern for stakeholders, it becomes even more critical to effectively manage missing values.[5] By maintaining good data quality, data cleaning and pre-processing play a critical role in data mining. Tasks for data cleaning include imputing missing values, finding outliers, and identifying and correcting noisy data.[6] For incomplete datasets, missing value imputation (MVI) is a significant challenge in data mining (DM) and large data analysis. The result of research may be affected if the imputed data from incomplete datasets isn't very good.[7] Missing data (MD) is an inevitable and common problem facing decision systems based on e-health data mining (DM) since many missing values are found in many historical medical datasets. Therefore, to deal with MD, a pre-processing stage is usually required before developing any DM-based decision system.[8] There are three kinds of MD:

1. Missing Completely at Random (MCAR): regardless of both observed and unobserved data, data is missing.
2. Missing at Random (MAR): data is lost regardless of ignored data, given the data observed.
3. Missing Not at Random (MNAR): the missing measurements are attributed to non-observed data values themselves.

Single imputation (SI) techniques create a specific value in a dataset for an actual missing value. For this technique, fewer computational expenses are required. There are numerous kinds of SI methods suggested by the investigators. By evaluating other responses, the general practice is to choose the highest possible answer. The value can be derived from the median, mean, and mode of the variable's existing values. The other methods can also be used for single imputation. Such techniques focus on machine learning (ML).

Multiple imputation (MI) methods, using different simulation models, for the imputation of a single missing item, generate numerous values. To find a variety of possible answers, these methods introduce the uncertainty of imputed results. MI methods are dynamic, but biased ideals such as single imputation do not suffer from them. The algorithm for MICE, suggested by Buuren and Groothuis, for multiple imputations, is commonly used.[9] The different types of multiple imputation technique are multivariate imputation by chained equation (MICE).[9], single center imputation from multiple chained equation (SICE).[5], iterative fuzzy clustering (IFC).[10], fuzzy-rough nearest neighbor imputation.[11], *etc*.

Nikfalazar *et al*.[12] have proposed DIFC ("decision trees and fuzzy clustering with iterative learning") missing data imputation method. Compared to DIFC, there are five imputation techniques, six datasets with numeric and category data that are missing, and several quantities and many forms of missing values used to perform experiments. The experimental findings indicate that in terms of imputation precision, the proposed DIFC technique outperforms other techniques.

Razavi-Far *et al*.[13] have proposed two MD imputation techniques, based on the pre-imputation, k-nearest neighbors, and expectation maximization imputation (kEMI) and kEMI+(Extension of kEMI). KNN (K Nearest Neighbours) algorithm and the posterior-imputation expectation-maximization algorithm. This approach is intended to find excellent value for k automatically. By learning global similarities, the latter employs the best KNN to evaluate MD. A novel knowledge fusion mechanism is used by kEMI+. It can accommodate numerical as well as characteristics that can be classified. The proposed data outcome imputation methods are evaluated by applying them to twenty-one publicly available datasets with various missing data and ratios, and then comparing them to other state-of-the-art missing data imputation strategies using standard measurement metrics like the normalized RMS (Root Mean Square) discrepancy and absolute error. It is concluded that the efficacy of the proposed novel missing data imputation methods was successful.

Khan *et al*.[5] have suggested a hybrid approach to multiple and single imputation methods for the MD imputation process, MICE two-variation algorithm for imputing categorical and numerical data. To impute ordinal, numeric, and binary missing values, the experiment tested 12 existing algorithms. Data collected from 65,000 accurate health records protected data privacy from various hospitals and diagnostic centers in Bangladesh. Three public data sets from the Kaggle, UCI ML Repository, Eidgenössische Technische Hochschule Zurich, and Kaggle were also obtained from the data. The findings have been compared to current algorithms. The results of the implementation show that the proposed algorithm achieves a 20 percent higher F-measure for imputing binary data and an 11 percent lower error for numerical data imputation than its competitors with comparable execution time.

To replace the missing values of datasets, Nikfalazar *et al*.[10] built and developed a new algorithm for IFC. To change the substituted data through iterations, fuzzy clustering is used. Experiments observed the IFC algorithm's performance on 3 widely used datasets and a city mobility dataset case study. The results demonstrate that the IFC algorithm is effective for datasets with fewer MDs and provides a large percentage of missing data for datasets with effective imputation performance.

Pati *et al*.[14] have proposed a novel "missing value estimation technique through cluster analysis" (MVETCA). In addition to overcoming the constraints of current approaches, the dataset microarray for imputing missing values also offers substantially better statistical measures such as less "normalized root mean square error" (NRMSE), high "conserved pair proportion" (CPP), and high "biomarker list concordance index" (BLCI). The Euclidian distance metric is used to estimate comparability, even though it is ineffective in high-dimensional datasets. The paper[6] suggests a "fuzzy expectation maximization and fuzzy clustering-based data pre-processing missing value imputation method" (FEMI). This works considerably better than support vector regression, Expectation Maximization Imputation, FKMI, GkNN, and

"Iterative Bi-Cluster based Local Least Square Imputation".

To evaluate the impact of MD techniques, Idri *et al*.[15] compared four classification algorithms' accuracy. The prediction techniques are "SVM (support vector machines), RF (random forest), NB (naïve bayes), and C4.5 DT (decision tree)". Therefore, 216 tests were conducted using 3 missing mechanisms not missing at random, MCAR, and MAR. There are 2 MD methods as KNN imputation and deletion. In a dataset obtained from the "autonomic nervous system" (ANS) unit of the "university hospital Avicenna in morocco", there are nine MD percentages of 10 to 90 percent. The author concluded that using the KNN substitution technique rather than the removal method progresses the higher accuracy rates of the above 4 classifiers.

A hybrid linear regression model (HLRM) was proposed by Srinivas *et al*.[16]. Firstly, the MD was imputed by the methodology of KNN and simple mean imputation. The main component analysis (PCA) is used to extract the maximum contributory disease-causing attributes. Secondly, the linear regression used to record the probability values of dependent variables for access to the relationship between dependent and independent variables is the stochastic gradient descent. The classification precision of the HLR model is observed as 89.13 percent. As a guide for medical professionals and also as a research forum, the author concluded that this study would support the academy.

Salleh *et al*.[17] suggested the technique of fuzzy-based imputation. A fuzzy swarm was used by clustering the complete candidates using particle swarm optimization to fill in and optimize the missing data (PSO). The decision tree classification algorithm then trained the entire data collection. The experimentation was trained with the HD (Heart Disease) dataset and used precision, accuracy, and ROC values to evaluate the results. The results show that, after applying a fuzzy swarm for imputation, the decision tree's efficiency is increased.[18] A substitution method based on "fuzzy C-means and particle swarm optimization" (FCMPSO) was filling in the missing values during the preprocessing step. Results show that with the application for imputation by FCMSPO, the output of the decision tree is improved.

Tran *et al*.[19], introduced a new strategy for recovering accuracy by MD prediction using a combination of replacement feature selection and clustering. The data set was collected using the UCI ML repository. The differential evolution (DE) method was used to solve a problem by iteratively improving a clarification of a particular quality. When compared to GA and "particle swarm optimization" (PSO), DE came out on top. DE produced outcomes that were comparable to PSO and superior to the Genetic Algorithm (GA). While predicting new cases, it enhanced classification accuracy and lowered calculation time for MD. They discovered that the non-imputed approach yields the following result: poorer than the proposed method, which has FS (feature selection) as a restriction. Sudha[20] developed application-specific intelligent computing (ASIC) as a decision support system (DSS) for cardiac problems, fertility diagnosis, and breast cancer. This approach was put to the test. The dataset was retrieved from a repository at UCI. There were two stages to ASIC. Pre-processing is the first step, followed by training and testing. In the first phase, the noisy data and MD in the data set were recognized. For attribute selection, GA was applied. The backpropagation neural network (NN) architecture was used for testing and training. The tests were carried out using the WEKA (Waikato Environment for Knowledge Analysis) tool. According to the findings, the Artificial Neural Network (ANN)-based ASIC outperformed current classifiers such as RF, bayesian network, DT, and others.[21].

An approach to finding missing values was proposed by Shahzad *et al*.[22] Missing values in datasets were found using GA. The performance of a separate result was calculated using information gain (IG). The dataset was found in the UCI repository. The final results showed that GA was an effective technique for the substitution of the MD. When there was a lot of incomplete data and a wide range of values, the proposed method performed better. This study emphasizes the significance of dealing with MD. Nikon *et al*.[23] suggested a technique for determining the danger of coronary artery atherosclerosis. To identify missing values in the atherosclerosis data set, a "ridge expectation-maximization imputation" (REMI) based on an "extreme learning machine" (ELM) technique was developed. The REMI/ELM classifier was estimated using the UCI and STULONG datasets. When compared to REMI/SVM, this classifier had a higher risk identification accuracy.

On datasets from the UCI ML Repository, including the Statlog database and the Cleveland HD dataset, Radhimeenakshi[24] evaluated the performance of SVM and ANN. The replace by the mean method was used to impute the missing data. To minimize the error function in ANN, the gradient descent (GD) technique was applied. The SVM classifier employed the 2D kernel function. SVM can handle both nonlinear and linear functions, as well as kernel functions. Only nonlinear data can be handled by ANN. The accuracy of the SVM model was higher.

Dinesh *et al*.[25] developed a methodology for predicting HD in patients and providing risk level awareness. The UCI machine learning dataset for heart disease was used. When data were missing, the default value was filled in. RF, SVM, NB, Logistic Regression (LR), and gradient-boosting, outcomes were compared using the Confusion matrix. It was discovered that LR performed the best. Dewan *et al*.[26] developed an ensemble model for cardiac disease categorization based on backpropagation and GA techniques. The data was taken from the UCI Machine Learning repository. The filter procedure took care of the missing value. To create hybrid approaches, NB, backpropagation, and J48 were used. ANN is the best classification method for non-linear data. Backpropagation's local minima problem is a drawback of ANN.

Amiri *et al.*[11] introduced three new fuzzy-rough set imputation procedures: implicator/t-norm-based fuzzy-rough sets, OWA(ordered weighted average)-based fuzzy-rough sets, and loosely quantified rough sets. The outcomes indicate that the FRNNI (fuzzy-rough nearest neighbor imputation) executes better than the other 2 techniques and exceeds the maximum of the other techniques of substitution method measured in this research. The OWANNI and the FRNNI are comparable, but the FRNNI is significantly superior. Finally, while VQNNI (vaguely quantified nearest neighbor imputation) performs worse than the other methods, it still outperforms several missing value imputation strategies. Although the methods presented work well, they have a high time complexity that could be decreased by optimization.

Zeinulla *et al.*[27] proposed a data imputation method for building a model from a medical sensor dataset. The major goal is to overcome the problem of putting together a framework for diagnosing heart diseases (HD) using limited medical data. It was decided to use the fuzzy RF approach. Jabbar *e.t al.*[28] suggested a cardiac disease prediction model based on RF and chi-square. The proposed methodology was put to the test using data sets on heart disease. According to the findings of the experiments, the strategy surpasses other methods in terms of classification accuracy, and the given model will aid healthcare practitioners in predicting heart disease.

Data mining classification approaches such as NB, SVM, k-NN, Decision Tree (DT), neural network (NN), LR, and gradient boosting have been proposed to predict the likelihood of coronary HD.[29]. A computerized system that predicts the risk of HD might be considered a major accomplishment. The dataset from the UCI ML repository is used to calculate this work on predicting cardiac disease. Traditional machine learning algorithms perform better with the feature selection strategy. Among the classification approaches, For HD categorization, the RF algorithm with PCA has the best accuracy of 92.85%.

Nilashi *et al.*[30] designed a cardiac disease diagnosis prediction system using machine learning techniques. The proposed solution was built using both unsupervised and supervised learning approaches. This study employed “Principal Component Analysis” (PCA), “Self-Organizing Map” (SOM), fuzzy SVM, and two imputation methods for MD imputation. In addition, to enhance disease prediction computation time, incremental PCA and Fuzzy SVM should be used for incremental data learning. According to the investigation of 2 real-world datasets, Cleveland and Statlog, incremental Fuzzy SVM could considerably improve the accuracy of HD categorization. The testing results showed that the method is successful in lowering the calculation time of disease detection when compared to the non-incremental learning technique. The study's weakness is that it only used two medical datasets to evaluate the technique. The datasets used in this study are too little to accurately demonstrate their utility in processing massive datasets. A huge amount of clinical data will also be used to show the flaws and benefits of the proposed technique in terms of computing time and prediction accuracy.

Jordanov *et al.*[31] developed and suggested two additional metrics in addition to the “overall classification accuracy” (OCA) measure: “inner class accuracy” (IA) and “outer class accuracy” (OA). They concluded that they can be utilized in conjunction with the OCA when determining the top classifier for a given situation. Several supervised ML algorithms were used to classify cardiac disease and their performance and accuracy were compared. Using a Kaggle heart disease dataset, this study discovered that the RF technique outperforms the KNN, DT, and RF methods for three-classification based on KNN, DT, and RF algorithms.[32]

2. Materials and methods

2.1 Dataset

The Data is a collection of connected data, with a report for each instance based on the Data it represents, and an attribute for each attribute in the dataset. This study uses data from “Cleveland, Switzerland, Long Beach, and Hungary”, as well as data from the University of California Irvine (UCI) ML repository.[33] and Kaggle.[34] for data analysis. There are 76 features in the data set, 14 of which are quite beneficial in identifying heart disease. In most cases, the predictive class attribute comes last. The attributes' data set specifications are shown in Table 1.

2.2 Significance of the attribute

Every Single attribute in UCI HD dataset is considered to be valuable for HD prediction, the significance of each of them designated it as a hyperparameter. The significance of the attribute, description, domain value, and hyperparameter is shown in Table 2.

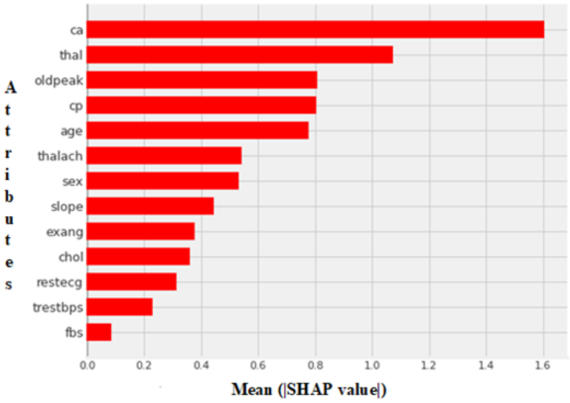


Fig. 1 The average absolute value of the Shapley Additive exPlanations values for each feature's relative importance.

Shapley Additive exPlanations (SHAP) values

Figure 1 depicts the UCI heart disease classification model using the global mean (|Tree SHAP|) approach. When a feature is "hidden" from the model, the x-axis represents the average magnitude change in model output. The term "hidden" refers to removing a variable from the model. Shapley values are employed to maintain consistency and accuracy because the impact of hiding a feature varies depending on what other features are also hidden.

Table 1. University of California Irvine Heart disease dataset attribute information.

Attribute	Description	Domain of value
Age	Age in year	29 to 77
Sex	Sex	Male (1) Female (0)
Cp	Chest pain type	Typical angina (1) Atypical angina (2) Non-anginal (3) Asymptomatic (4)
Trestbps	Resting blood sugar	94 to 200 mm Hg
Chol	Serum cholesterol	126 to 564 mg/dl
Fbs	Fasting blood sugar	>120 mg/dl True (1) False (0)
Restecg	Resting ECG result	Normal (0) ST-T wave Abnormality (1) LV hypertrophy (2)
Thalach	Maximum heart rate achieved	71 to 202
Exang	Exercise-induced angina	Yes (1) No (0)
Oldpeak	ST depression induced by exercise relative to rest	0 to 6.2
Slope	The slope of peak exercise ST segment	Upsloping (1) Flat (2) Downsloping (3)
Ca	Number of major vessels colored by fluoroscopy	0 – 3
Thal	Defect type	Normal (3) Fixed defect (6) Reversible defect (7)
Num	Heart disease	0 – 4

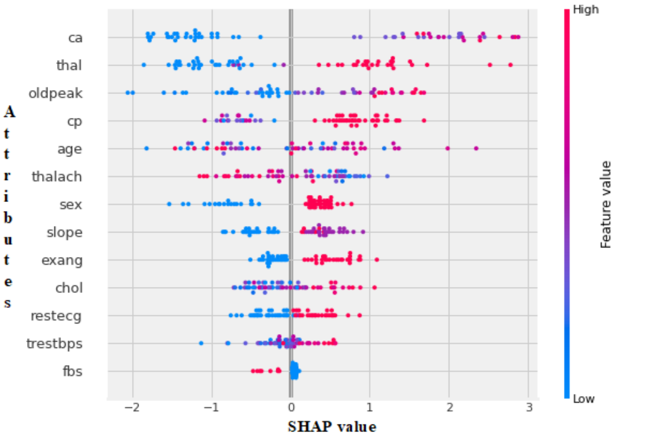


Fig. 2 Explanations generated by Shapley Additive exPlanations summary plot.

Plot the SHAP values of each feature for each sample in Fig. 2 to gain insight into which features are most relevant for a model. The plot shows the distribution of the impacts each feature has on the model output by sorting features by the sum of SHAP value magnitudes across all samples. For the red high and blue low, the color symbolizes the feature value.

2.3 The proposed ischemic heart disease multiple imputation techniques

The proposed methodology to achieve the research objective by appropriate classification of ischemic heart disease is briefly explained through the functional block diagram shown in Fig. 3.

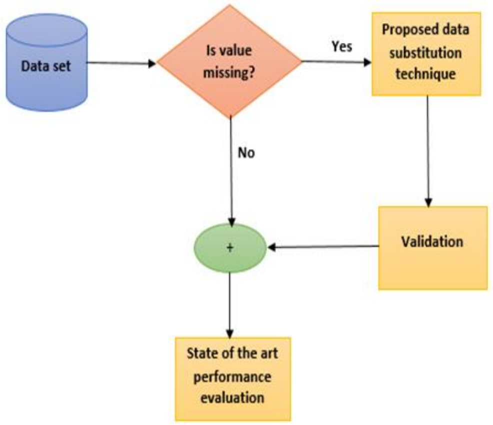

Data set
Is value missing?
Yes
Proposed data substitution technique
No
Validation
+
State of the art performance evaluation

Fig. 3 Flowchart for proposed ischemic heart disease multiple imputation techniques.

In the process, the following functionalities are considered: 1) Data substitution; 2) validation.

The proposed research work will resource the patient data set about various ischemic heart diseases from the data set. The IHDMIT data substitution technique first identifies and substitutes missing values in the data sets. Following that, a validation method based on maximum values with a standard dataset is used to authenticate the determined missing value.

The data set used in this process is taken from the University of California Irvine ML repository with missing values. The IHDMIT data imputation technique is shown in Fig. 4.

The input X is a dataset with missing attributes. The procedure determines the NN and stores them in set N for each instance of the dataset, y, that contains at least one missing value for feature a. The algorithm then uses y's NN to approximate the missing value. The final membership function M is calculated by averaging the lower and upper estimations of y concerning z. The algorithm produces a value, based on these computations after repeating the process for all the instances belonging to N. The value z is determined by all of its neighbors' values. The variables J1 and J2 are initialized for obtaining a particular attribute that is missing in that instance. It is possible, though unlikely, that J2 = 0. If M = 0, in this situation, and since M = 0, J1 = 0 as well. As a result, J1/J2 can't be determined. To tackle this problem, the

Table 2. Significance of the University of California Irvine Heart disease dataset attributes.

Attribute	Description	Domain of value	Significance of the attribute to designate it as one of the hyperparameters
Age	Age in year	29 to 77	Higher the age, the higher the risk of developing coronary artery disease. This happens irrespective of gender, although women tend to be about a decade older when they develop cardiovascular disease compared to men.
Sex	Sex	Male (1) Female (0)	Male sex is an independent risk factor for developing ischemic heart disease. However, women tend to have poorer outcomes following acute coronary syndromes. Women also have atypical symptoms and delayed presentations compared to men.
Cp	Chest pain type	Typical angina (1) Atypical angina (2) Non-anginal (3) Asymptomatic (4)	The presence of typical angina makes the diagnosis of ischemic heart disease much more likely compared to atypical angina. Non-anginal pain makes it less likely. Although uncommon, ischemic heart disease can present as silent ischemia (no symptoms related to ischemia) in elderly patients, diabetics, especially with neuropathy, *etc*. However, a complete lack of symptoms in a younger, non-diabetic patient usually goes against significant coronary artery disease.
Trestbps	Resting blood sugar	94 to 200 mm Hg	Elevated blood sugar levels esp fasting blood sugar levels indicates either poor control of sugars in a known diabetic or the presence of diabetes in previous non-diabetic patients. Those with diabetes and ischemic heart disease do poorly if their blood sugars are not well controlled with medications.
Chol	Serum cholesterol	126 to 564 mg/dl	Elevated serum cholesterol levels esp. low-density lipoproteins (LDL) are independent risk factors for ischemic heart disease. Control of LDL levels to predefined targets based on the patient's risk profile is one of the main goals of therapy for ischemic heart disease.
Fbs	Fasting blood sugar	>120 mg/dl True (1) False (0)	Elevated blood sugar levels esp fasting blood sugar levels indicates either poor control of sugars in a known diabetic or the presence of diabetes in previous non-diabetic patients. Those with diabetes and ischemic heart disease do poorly if their blood sugars are not well controlled with medications.
Restecg	Resting ECG result	Normal (0) ST-T wave Abnormality (1) LV hypertrophy (2)	Normal resting ECG does not rule out the presence of ischemic heart disease. Stress testing like treadmill exercise testing may be required. The presence of LVH can interfere with the diagnosis of ischemia from both the resting ECG and during treadmill exercise testing. The presence of ST-T wave abnormality in resting ECG (esp. in the absence of LVH) is an important diagnostic clue for ischemic heart disease.
Thalach	Maximum heart rate achieved	71 t0 202	The maximum heart rate achieved during treadmill exercise testing indicates the completeness of the test (in general, we need the person undergoing TMT to achieve a heart rate of more than 85% of the maximum age-predicted heart rate). If the patient achieves a heart rate lesser than this, TMT is regarded as inconclusive. If the target heart rate is achieved, then we can go on to interpret the TMT further, and based on the presence, type, and degree of ST-T changes, the probability of underlying ischemic heart disease is estimated.
Exang	Exercise-induced angina	Yes (1) No (0)	Exercise-induced angina is an important indicator of significant coronary artery disease. However, it can also be seen in aortic stenosis.
Oldpeak	ST depression induced by exercise relative to rest	0 to 6.2	The larger the ST depression esp. if seen in multiple contiguous ECG leads, the higher the likelihood of underlying ischemic heart disease.
Slope	The slope of peak exercise ST segment	Upsloping (1) Flat (2) Downsloping (3)	Order of importance from most to least important: Down sloping, flat followed by upsloping. Upsloping ST depression is the least important.
Ca	Number of major vessels colored by	0 – 3	A coronary angiogram is a diagnostic test used to confirm the presence of coronary artery disease. The more the number of vessels affected, the worse

	fluoroscopy		the clinical outcome for the patient.
Thal	Defect type	Normal (3) Fixed defect (6) Reversible defect (7)	A reversible defect is specific for significant obstruction in one or more coronary arteries. A fixed defect may be seen in those who have infarcted areas indicating previous MI. Normal perfusion study indicates a low chance of having coronary artery disease.

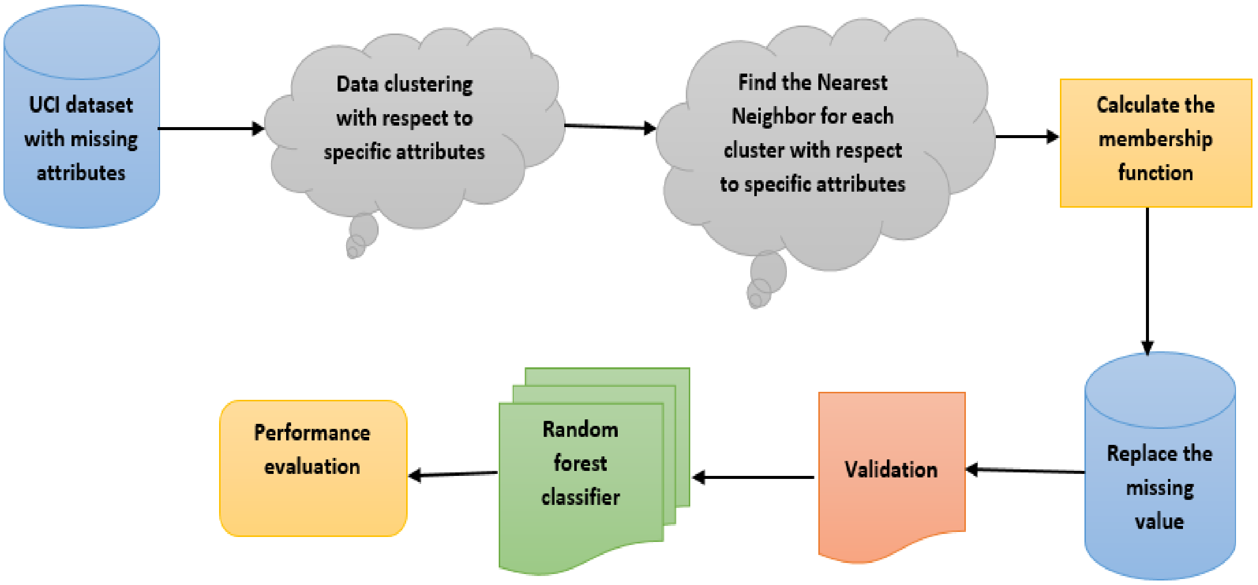

UCI dataset with missing attributes
Data clustering with respect to specific attributes
Find the Nearest Neighbor for each cluster with respect to specific attributes
Calculate the membership function
Replace the missing value
Validation
Random forest classifier
Performance evaluation

Fig. 4 ischemic heart disease multiple imputation technique overall system model.

the average value of the feature for its neighbors is used.

2.4 The IHDMIT algorithm

The key functionality of the proposed IHDMIT is as follows:

Params: *R, a* Fuzzy tolerance relation

The algorithm of the Ischemic Heart Disease Multiple Imputation Technique (Algorithm S1) is shown. The function contains the Missing*(y)* method in this algorithm returns true when y has at least one missing value. The function is missing *(a(y))* method returns the true value if the value is missing in the y test cases. When evaluating the upper and lower approximations for y concerning z, it will be ignored if $t = z$.

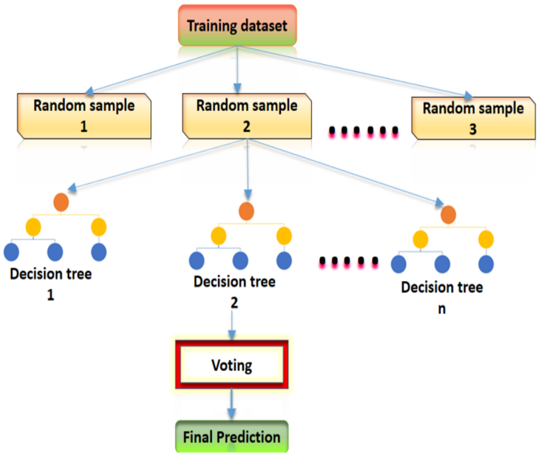


Fig. 5 Machine learning random forest algorithm.

2.5 Importance of random forest algorithm

To improve prediction capability, RF employs the notion of bagging to aggregate many DTs. Individual learners are trained independently in bagging. It uses replacement to generate several samples of data at random from the unique dataset, and each DT is trained on distinct models of data. During tree creation, features are also chosen at random. A majority vote is used to integrate predictions provided by various trees.[28] Fig. 5 depicts the random forest's operation.

By optimizing parameters like the total number of estimators, size of the smallest node, and amount of characteristics utilized to separate nodes, the RF can be modified for higher accuracy.

3. Results and discussion

Missing values in the HD dataset are represented as 'blank space.' The pandas DataFrame considers those missing values to be ‘NaN' values, and they are subsequently erased from the dataset using the imputation method, as illustrated below. 'KNNimputer,' 'Mean,' 'fill,' and 'bill' are the most common Single Imputation methods. These methods, on the other hand, have the drawback of limiting the variance between features, and the mean value will be affected if the dataset contains outliers. Table 3 shows the accuracy and recall rate of various single imputation techniques utilizing the DT algorithm. According to the results, KNNimputer has an accuracy rate of 0.80 and a recall rate of 0.79.

Table 3. Accuracy and recall rate of different single imputation techniques.

Sl. no	Imputation method	Accuracy rate with DT	Recall rate with DT
1	ffill	0.78	0.77
2	bfill	0.78	0.77
3	mean	0.79	0.78
4	KNNimputer	0.80	0.79

Table 4 shows a comparison of different accuracy and recall rates using the DT algorithm with and without the imputation approach. The results of the comparison suggest that using the imputation method improves accuracy and recall rate.

Table 4. Accuracy and Recall rate of various methods with and without the use of Single Imputation.

Sl.no	Imputation method	Without using the imputation method	Accuracy rate with DT	Recall rate with DT
1	ffill	76.2	0.78	0.77
2	bfill		0.78	0.77
3	mean		0.79	0.78
4	KNNimputer		0.80	0.79

The sample imputed data and predicted class is shown in S6.1 and S6.2. It is a binary classification method. There are two class values 0 and 1. The class value 0 indicated by there is no heart disease and the class value 1 indicated by there is a heart disease in the dataset. Here the nearest neighbor is considered as the same class value.

The Dataset with a missing value is shown in Fig. S4. The imputed dataset is shown in Fig. S5. The standard threshold values are used in the authentication method for the deduced missing value, as shown in Table 5. The outcome demonstrates that the imputed values achieved using the proposed IHDMIT are within the threshold value.

The comparison of the proposed IHDMIT methodology with different imputation approaches. Table 6 shows the results for different imputation techniques. A confusion matrix regulates the inputs in conjunction with various qualitative parameters like Accuracy, Sensitivity, and Specificity, which will provide a more accurate classification result for the classification of Ischemic Heart Diseases.

Table 7 details a typical confusion matrix. The mathematical background of various parameters associated with the confusion matrix are:

Accuracy: The proportion of instances that are correctly classified.

Table 5. The suggested ischemic heart disease multiple imputation technique validation findings.

Record no	Attribute name	Threshold value	Original value	Imputed values obtained using the proposed IHDMIT
8	fbs	0-1	0	0
262			1	0
5	oldpeak	0-6.2	nil	0.6
186			2.6	0
4	cp	0-3	1	1
303			0	1
17	chol	126-564	219	340
35			273	213
7	slope	1-3	1	1
20			nil	3
99			2	2
3	Trestbps	94-200	nil	130
22			nil	135

Table 6. Comparison of proposed ischemic heart disease multiple imputation techniques with other imputation methods.

	Fuzzy rough set imputation	Fuzzy C means imputation	Expectation maximization imputation	Proposed IHDMIT using RF
Accuracy	90.4	91.2	88.8	93
Sensitivity	96.4912	95	93.2203	96.4912
Specificity	85.2941	87.6923	84.8485	89.7059
Precision	84.6154	87.6923	84.6154	88.7097
Recall	96.4912	95	93.2203	96.4912
F measure	0.9016	0.912	0.8871	0.9244

Table 7. A typical confusion matrix.

	Predicted: Negative	Predicted: positive
Actual: negative	True Negative (TN)	False Positive (FP)
Actual: positive	False Negative (FN)	True Positive (TP)

Accuracy = TP / (TP + FP + TN + FN)

Sensitivity: The proportion of positive instances that are correctly classified as positive.

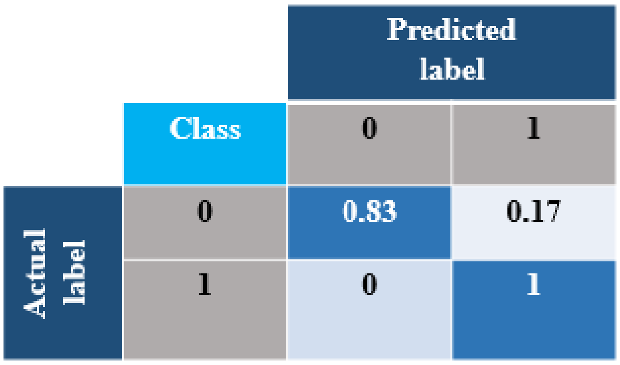
Sensitivity = TP / (TP + FN)

Specificity: The proportion of negative instances that are correctly classified as negative.

Specificity = TN / (TN + FP)

On the UCI heart disease data set, Table 8 illustrates the normalized confusion matrix for a two-class classification issue. The techniques for creating the data collection and the classifier are identical to those described in Table 7. The X-axis of the graphic represents the predicted label, while the Y-axis represents the true label for classes 0 and 1. The true class members of a row are split across columns, and the matrix elements are normalized row by row, that is, the sum of fractions along a row equals 1.

Table 8. Normalized confusion matrix for ischemic heart disease multiple imputation techniques.



		Predicted label	
	Class	0	1
Actual label	0	0.83	0.17
	1	0	1

Table 9 shows the results for the comparison of the proposed ischemic heart disease multiple imputation techniques compared with the fuzzy rough set,[35] fuzzy C means,[36] and expectation-maximization multiple imputation techniques using UCI HD Dataset.[23,37] By using RF classification in the proposed technique, better accuracy of 93% is achieved.

Table 9. Comparison of proposed ischemic heart disease multiple imputation techniques with the University of California Irvine (UCI) Heart Disease dataset.

Imputation method	Dataset	Accuracy %
Fuzzy Rough set.[35]	UCI Heart Disease Dataset	90
Fuzzy C Means.[36]		92
REMI.[23]		89
Proposed IHDMIT		93

Because in the proposed method each neighborhood is compared to all other neighbors only once, and finding the nearest neighbor is based on the class value.[38,39] To further validate the efficacy of the suggested model, the AUC of ROC charts was used to compare its performance to that of other models. Fig. 6 depicts the suggested model, fuzzy roughest, fuzzy C means, REMI model ROC charts, respectively. As indicated in the figures, the proposed model's ROC curve AUC is 0.93, the fuzzy rough set model's AUC is 0.90, the fuzzy C means model's AUC is 0.92, and the REMI model's AUC is 0.89. Based on the assessment criteria employed, the ROC curve's accuracy and AUC, the suggested Ischemic Heart Disease Multiple Imputation Technique outperforms the Fuzzy rough set, Fuzzy C means, REMI

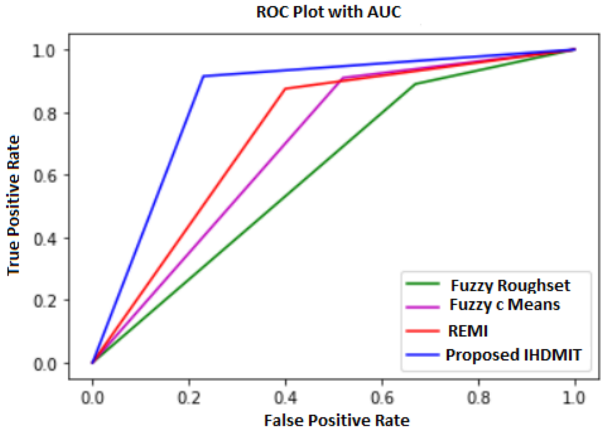


Fig. 6 AUC charts of Fuzzy roughest, Fuzzy C means, REMI, and proposed model.

4. Conclusions

Missing Data Imputation Technique has been developed by using a benchmark secondary UCI Heart Disease dataset. By using RF classification in the proposed technique, an accuracy of 93% is achieved. It can handle numerical values. The proposed IHDMIT is compared with the fuzzy rough set, fuzzy C means, and, expectation-maximization multiple imputation methods. The proposed IHDMIT conforms better in terms of accuracy, sensitivity, specificity, precision, recall, and F_measure when compared to fuzzy rough set, fuzzy C means, and, expectation-maximization imputation techniques. This type of research will be extremely useful in assisting healthcare practitioners in the prediction of HD.

Acknowledgments

We thank the Manipal Academy of Higher Education (MAHE) for the financial support and the resources provided during the research work. Sincere thanks to the Department of Computer Science Engineering, Manipal Institute of Technology, Manipal, India, for the guidance and technical support provided.

Conflict of Interest

The authors declare no conflict of interest.

Supporting information

Applicable.

References

[1] R. Soundharya, D. Cenitta, R. V. Arjunan, *Journal of Advanced Research in Dynamical and Control Systems*, 2018, **10**, 22-26.

[2] World Health Organization web, https://www.who.int/newsroom/fact-sheets/detail/Cardiovascular-diseases-(cvds).

[3] S. Yogaamrutha, Chandana, D. Cenitta, R. V. Arjunan, *Journal of Advanced Research in Dynamical and Control Systems*, 2019, **11**, 25-36.

[4] D. Cenitta, R. V. Arjunan, K. V. Prema, Missing Data Imputation using Machine Learning Algorithm for Supervised Learning, *International Conference on Computer Communication and Informatics (ICCCI)*, 2021, 1-5, doi: 10.1109/ICCCI50826.2021.9402558.

[5] S. Khan, A. Hoque, *Journal of Big Data*, 2020, **7**, 1-21, doi: 10.1186/s40537-020-00313-w.

[6] M. G. Rahman, M. Z. Islam, *Knowledge and Information Systems*, 2016, **46**, 389-422, doi: 10.1007/s10115-015-0822-y.

[7] W-C. Lin, C-F Tsai, *Artificial Intelligence Review*, 2020, **53**, 1487-1509, doi: 10.1007/s10462-019-09709-4.

[8] P. S. Raja, K. Thangavel, *Soft Computing*, 2020, **24**, 4361-4392, doi: 10.1007/s00500-019-04199-6.

[9] S. Van Buuren, K. Groothuis-Oudshoorn, *Journal of Statistical Software*, 2011, **45**, 1-67, doi: 10.18637/jss.v045.i03.

[10] S. Nikfalazar, C. H. Yeh, S. Bedingfield, H. A. Khorshidi, A new iterative fuzzy clustering algorithm for multiple imputation of missing data, *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 2017, 1-6, doi: 10.1109/FUZZ-IEEE.2017.8015560.

[11] M. Amiri, R. Jensen, *Neurocomputing*, 2016, **205**, 152-164, doi: 10.1016/j.neucom.2016.04.015.

[12] S. Nikfalazar, C. H. Yeh, S. Bedingfield, H. A. Khorshidi, *Knowledge and Information Systems*, 2020, **62**, 2419-2437, doi: 10.1007/s10115-019-01427-1.

[13] R. Razavi-Far, B. Cheng, M. Saif, M. Ahmadi, *Knowledge-Based Systems*, 2020, **187**, 104805, doi: 10.1016/j.knosys.2019.06.013.

[14] S. K. Pati, A. K. Das, *Knowledge and Information Systems*, 2017, **52**, 709-750, doi: 10.1007/s10115-017-1025-5.

[15] A. Idri, I. Kadi, I. Abnane, J. L. Fernandez-Aleman, *Medical & Biological Engineering & Computing*, 2020, **58**, 2863-2878, doi: 10.1007/s11517-020-02266-x.

[16] K. Srinivas, B. K. Rani, M. V. P. Rao, R. K. Patra, G. Madhukar, A. Mahendar, *European Journal of Molecular & Clinical Medicine*, 2020, **7**, 1159-1171.

[17] M. N. Mohd Salleh, N. A. Samat, *International Conference on Swarm Intelligence*, 2017, 285-292, doi: 10.1007/978-3-319-61833-3_30.

[18] M. N. M. Salleh, N. A. Samat, FCMPSO: an imputation for missing data features in heart disease classification, *IOP Conference Series: Materials Science and Engineering*, 2017, **226**, 012102, doi: 10.1088/1757-899x/226/1/012102.

[19] C. T. Tran, M. Zhang, P. Andreae, B. Xue, L. T. Bui, *Applied Soft Computing*, 2018, **73**, 848-861, doi: 10.1016/j.asoc.2018.09.026.

[20] M. Sudha, *Journal of medical systems*, 2017, **41**, 1-10, doi: 10.1007/s10916-017-0823-3.

[21] S. Yogaamrutha, D. Cenitta, R. V. Arjunan, *Journal of Advanced Research in Dynamical and Control Systems*, 2018, **10**, 137-142.

[22] W. Shahzad, Q. Rehman, E. Ahmed, *International Journal of Advanced Computer Science and Applications*, 2017, **8**, 438-445.

[23] O. Gervasi, B. Murgante, S. Misra, G. Borruso, C. Torre, A. M. A. C. Rocha, D. Taniar, B. Apduhan, E. Stankova, A. Cuzzocrea, *International conference on computational science and computational intelligence (CSCI)*, 2020.

[24] S. Radhimeenakshi, Classification and prediction of heart disease risk using data mining techniques of Support Vector Machine and Artificial Neural Network, *2016 3rd International Conference on Computing for Sustainable Global Development*, 2016, 3107-3111.

[25] K. G. Dinesh, K. Arumugaraj, K. D. Santhosh, V Mareeswari, Prediction of Cardiovascular Disease Using Machine Learning Algorithms, *Proceeding of 2018 IEEE International Conference on Current Trends toward Converging Technologies, Coimbatore*, 2018, 1-7, doi: 10.1109/ICCTCT.2018.8550857.

[26] A. Dewan, M. Sharma, Prediction of heart disease using a hybrid technique in data mining classification, *2nd International Conference on Computing for Sustainable Global Development*, 2015, 704-706.

[27] B. Bouchon-Meunier, *IEEE international conference on fuzzy systems (FUZZ-IEEE)*, 2020.

[28] M. A. Jabbar, B. L. Deekshatulu, P. Chandra, Prediction of heart disease using random forest and feature subset selection, *In Innovations in bio-inspired computing and applications, Springer*, 2016, 187-196, doi: 10.1007/978-3-319-28031-8_16.

[29] F. Tasnim, S. Umme Habiba, A comparative study on heart disease prediction using data mining techniques and feature selection, International Conference on Robotics, *Electrical and Signal Processing Techniques*, 2021, 338-341, doi: 10.1109/ICREST51555.2021.9331158.

[30] M. Nilashi, *International Journal of Fuzzy Systems*, 2020, **22**, 1376-1388, doi: 10.1007/s40815-020-00828-7.

[31] I. Jordanov, N. Petrov, A. Petrozziello, *Journal of Artificial Intelligence and Soft Computing Research*, 2018, **8**, 31-48, doi: 10.1515/jaiscr-2018-0002.

[32] M. M. Ali, B. K. Paul, K. Ahmed, F. M. Bui, J. M. Quinn, M. A. Moni, *Computers in Biology and Medicine*, 2021, **136**, 104672, doi: 10.1016/j.compbiomed.2021.104672.

[33] Cleveland, Hungary, Switzerland, The VA Long Beach, https://archive.ics.uci.edu/ml/datasets/heart+disease.

[34] Kaggle, https://www.kaggle.com/ronitf/heart-disease-uci.

[35] D. P. Acharjya, *Journal of Medical Systems*, 2020, **44**, 1-16, doi: 10.1007/s10916-019-1497-9.

[36] R. Chitra, V. Seenivasagam, Heart Attack Prediction System

Using Fuzzy C Means Classifier, *IOSR Journal of Computer Engineering (IOSR-JCE)*, 2013, **14**, 23-31.

[37] H. Khan, X. Wang, H. Liu, *Computers & Electrical Engineering*, 2021, **93**, 107230, doi: 10.1016/j.compeleceng.2021.107230.

[38] D. Li, H. Zhang, T. Li, A. Bouras, X. Yu, T. Wang, Hybrid Missing Value Imputation Algorithms Using Fuzzy C-Means and Vaguely Quantified Rough Set, *IEEE Transactions on Fuzzy Systems*, 2021, 30, 1396-1408, doi: 10.1109/TFUZZ.2021.3058643.

[39] J. Tang, *Journal of Intelligent Transportation Systems*, 2020, **25**, 439-454, doi: 10.1080/15472450.2020.1713772.

Author Information

D. Cenitta*, is presently working as Assistant Professor – Senior Scale in the Department of Computer Science and Engineering, Manipal Institute of Technology, MAHE, Manipal. Her area of research is Machine Learning. She has 14 years of teaching experience and has published 8 papers in reputed Journals and Conferences.*

Dr. R Vijaya Arjunan *is presently working as Associate Professor in the Department of Computer Science and Engineering, Manipal Institute of Technology, MAHE, Manipal. He had worked on deputation with School of Engineering and IT, Manipal, Dubai campus from 2014 until 2017. He obtained his Masters and PhD in Computer Science and Engineering from Sathyabama Institute of Science, Technology, and Sankara University in 2005 and 2013 respectively. He has published 40+ research articles in various International Conferences and Journals. He is a lifetime member in various professional societies like Indian Society for Technical Education, Broadcast Engineering Society, International Association of Computer Science and Information Technology and computer Society of India. He is also an elected co-opted member from academics by Broadcast Engineering Society (I), Chennai chapter for the period 2008 to 2010. His research interest includes Image Processing, Machine Learning, Deep learning, and Data Mining.*

Dr. Prema K V*, is a Professor and Head in the Department of Computer Science & Engineering at Manipal Institute of Technology, Bengaluru, MAHE. Her areas of research interest are Soft computing, Computer Networks, and Security. She has 30 years of teaching experience, 22 years of research experience and has published more than 120 papers in reputed Journals and Conferences.*

Publisher's Note: Engineered Science Publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.