[1] D. Cenitta
[2] R. Vijaya Arjunan
[3] N Arul
[4] Bhargav J Bhatkalkar
[5] Shwetha G K
[6] Jayantkumar A Rathod

Optimized Heart Disease Prediction through Fuzzy Rough Set based Missing Data Imputation Techniques

Abstract: - This abstract describes an effort to include fuzzy-rough set-based missing data imputation approaches into heart disease prediction models to improve their accuracy. Missing values are a major difficulty and may jeopardize the accuracy of predictive studies in medical datasets, such as those kept in repositories like the University of California, Irvine (UCI). This paper suggests using fuzzy-rough set-based imputation techniques, which are well-known for their ability to handle the ambiguity and uncertainty included in medical data, to address this problem. Fuzzy-rough sets and their latest expansions are used to introduce the Cardiovascular Disease Multiple Imputation Technique (CVDMIT), a unique technique. By way of extensive testing using a Random Forest classifier, CVDMIT is thoroughly analyzed and compared with well-known methods like as fuzzy roughest, fuzzy C means, and expectation maximization. The results show that using CVDMIT in conjunction with Random Forest classification results in a significant improvement in accuracy, with a precision rate of 94%. With increased accuracy and dependability, this study advances the field by highlighting the potential of fuzzy-rough set-based missing data imputation approaches in optimizing heart disease prediction.

Keywords: UCI, Fuzzy rough Set, Multiple imputation techniques, Machine learning, Random Forest.

I. INTRODUCTION

Heart disease is the leading cause of mortality worldwide, posing significant challenges to individuals and public health systems alike. To lessen the negative effects of cardiac disease and enhance patient outcomes, effective prediction, and early identification are essential. Predictive analytics for heart disease has evolved significantly in recent years thanks to the development of sophisticated computer methods and the accessibility of large-scale medical information. The existence of missing values, which can compromise the accuracy and dependability of predictive models, is a recurring problem when using these datasets for precise prediction.

Missing values are prevalent in medical datasets, particularly those stored in repositories such as the University of California Irvine (UCI). These can be caused by several things, including mistakes in data collection, non-compliance from patients, or restrictions in the system. If missing data is not properly addressed, it can cause uncertainty and result in biased or erroneous forecasts. As a result, strong approaches are desperately needed to deal with missing data, especially regarding heart disease prediction.

Conventional methods for imputation of missing data, including mean imputation or case deletions that are not full, frequently fall short in capturing the intricate patterns and correlations seen in medical data. Furthermore, they have the potential to add bias and alter the data's underlying distribution, which would eventually jeopardize the accuracy of prediction models. Given this, sophisticated imputation methods based on fuzzy-rough sets provide a viable solution to the problem of missing data while maintaining the ambiguity and uncertainty that characterize medical datasets.

Fuzzy-rough sets are ideal for imputation jobs in medical data analysis because they offer an adaptable framework for managing imprecise and missing information. Fuzzy-rough set-based imputation algorithms can more accurately impute missing values and reflect the underlying structure of the data by using the concepts of fuzzy

[1, 2] Department of Computer Science and Engineering, Manipal Institute of Technology, Manipal Academy of Higher Education (MAHE), Manipal, Karnataka 576104, India. [1]cenitta.d@manipal.edu, [2]vijay.arjun@manipal.edu

[3] Department of Computer Science and Engineering, AJ Institute of Engineering and Technology, Mangalore – 575006, Karnataka, India. arul_hi2000@yahoo.com

[4] KFSCIS, Florida International University, Miami, Florida, USA-33199. bhargav.j.bhatkalkar@gmail.com

[5] Department of Computer Science and Engineering, NMAM Institute of Technology, Affiliated to NITTE Deemed to be University, Nitte-5744110 , India. gk.shwetha@nitte.edu.in

[6] Department of Computer Science and Engineering, Alva's Institute of Engineering and Technology, Moodubidire-574225, India. jayantkumarrathod@gmail.com

partitions and rough membership functions. These techniques have proven effective in several fields, including healthcare, where trustworthy and high-quality data are crucial [1][13].

1.1 The Need for Enhanced Missing Data Imputation Techniques in Heart Disease Prediction

Heart disease is a broad term that includes a variety of cardiovascular disorders, such as arrhythmias, heart failure, and coronary artery disease. For prompt intervention and preventative actions to lower morbidity and death rates, early identification and precise prognosis of heart disease are essential. Based on their clinical, lifestyle, and demographic characteristics, predictive analytics approaches like machine learning and statistical modeling have demonstrated promise in identifying people who are at high risk of developing heart disease.

Nevertheless, the accuracy and comprehensiveness of the input data determine how useful predictive models are. Missing values are common in real-world medical datasets and can be caused by several things, such as inaccurate measurements, incomplete patient records, or problems with data entry. prediction modeling has substantial obstacles when missing data is present since it might result in skewed estimates, lower prediction accuracy, and weakened model generalizability.

Medical datasets, where the missingness mechanism is frequently non-random and linked with the result of interest, are unsuitable for traditional approaches to handling missing data, such as complete-case analysis or straightforward imputation techniques like mean replacement. Furthermore, these techniques yield less-than-ideal prediction performance since they do not consider the underlying uncertainty and unpredictability in the data [2].

More advanced missing data imputation methods have drawn attention recently as a way to overcome the problems of more conventional approaches and make use of the data's natural structure. Based on the ideas of rough sets and fuzzy logic, fuzzy-rough set theory provides a potential paradigm for managing missing data in medical datasets. Fuzzy-rough set-based imputation approaches effectively impute missing values while maintaining the underlying structure and connections in the data by collecting rough approximations of the data space and conveying uncertainty using fuzzy membership functions.

1.2 Fuzzy-Rough Set-Based Missing Data Imputation Techniques

Fuzzy logic and rough set theory are combined in fuzzy-rough set-based missing data imputation strategies to address incompleteness and uncertainty in medical datasets. Rough membership functions, which describe the extent of an element's membership in a particular set, are the fundamental idea of fuzzy-rough set theory. Fuzzy-rough set-based imputation techniques can reflect the inherent variability and uncertainty in medical data by creating fuzzy partitions across the attribute space .

Finding rough equivalency classes or granules within the dataset based on the information at hand is a popular method for fuzzy-rough set-based missing data imputation. These granules serve as bases for imputing missing values depending on the properties of nearby instances and represent clusters of related data points. Fuzzy-rough set-based imputation approaches minimize the impact of missing data on predictive modeling by utilising the links between characteristics and taking into account the local context of the missing values. This allows for the generation of reliable estimations.

The literature has presented several varieties of fuzzy-rough set-based imputation approaches, each with its benefits and trade-offs in terms of computing cost, accuracy, and resilience. As an illustration, certain techniques could give priority to neighbourhood ties and local information, while others would use global optimisation criteria to direct the imputation process. Furthermore, hybrid techniques that combine fuzzy logic with neural networks or evolutionary algorithms, or extensions of fuzzy-rough set theory, have demonstrated promise in further enhancing imputation performance and scalability [3].

1.3 Experimental Evaluation and Comparative Analysis

Thorough experimental assessments are required to determine the effectiveness of fuzzy-rough set-based missing data imputation algorithms in the context of heart disease prediction. Standardized evaluation measures and benchmark datasets are often used in these assessments to compare the effectiveness of various imputation techniques. Among the crucial metrics are predictive of the receiver operating characteristic curve's area under the curve, sensitivity, specificity, and accuracy.

This research, uses a well-known benchmark dataset for heart disease prediction, such as the Cleveland Heart Disease dataset from the UCI repository, to assess the effectiveness of fuzzy-rough set-based missing data imputation algorithms. To evaluate the relative merits and drawbacks of the suggested imputation methods state-

of-the-art approaches, such as machine learning-based methods and conventional imputation methods, will be compared.

The UCI repository is one popular benchmark dataset for heart disease prediction that we want to use in this study to assess the effectiveness of fuzzy-rough set-based missing data imputation algorithms. To estimate the relative benefits and weaknesses of the suggested imputation methods, we will compare them to cutting-edge strategies, such as machine learning-based methods and conventional imputation methods [4].

The incorporation of fuzzy-rough set-based methods for missing data imputation has significant potential to improve the precision and dependability of heart disease prediction models. These techniques may successfully address incompleteness and uncertainty in medical datasets by utilizing the concepts of rough set theory and fuzzy logic. This improves prediction performance and facilitates more informed clinical decision-making. Our goal is to show how fuzzy-rough set-based imputation approaches may be used to improve heart disease prediction and advance the area of predictive analytics in healthcare through thorough experimental evaluations and comparative studies.

II. Literature Review

Nikfalazar et al.[5] contribute to knowledge and information systems by addressing the challenges of handling missing data in predictive modeling. With the proliferation of data-driven approaches across various domains, the issue of missing data has become increasingly prevalent and problematic. Existing imputation techniques often struggle to effectively handle the complexities and uncertainties inherent in real-world datasets. To overcome this issue, proposing a novel approach that leverages fuzzy-rough set-based imputation techniques to improve the accuracy of predictive models. Their work builds upon prior research in data mining and machine learning, aiming to advance the state-of-the-art in predictive analytics by offering a robust solution for handling missing data challenges.

Razavi-Far et al. contribute to the domain of knowledge-based systems by focusing on the utilization of artificial intelligence techniques for improving decision-making processes. The paper addresses the growing demand for advanced computational methods capable of handling complex and dynamic decision environments. With the rapid advancement of artificial intelligence and machine learning, there is increasing interest in leveraging these technologies to develop intelligent decision support systems. Their research aligns with prior studies in the field, which have emphasized the importance of harnessing the power of artificial intelligence to extract actionable insights from large datasets and facilitate informed decision-making [6]. Pati et al. explore the application of machine learning techniques for addressing classification challenges in complex datasets. The study delves into the realm of knowledge and information systems, where the efficient classification of data plays a pivotal role in decision-making processes across various domains. Their work builds upon prior research in machine learning and data mining, aiming to enhance the accuracy and efficiency of classification algorithms for real-world applications [7].

Idri et al. [8] investigate the application of technology in healthcare systems, focusing on developing robust medical information systems. They address the reliance on technology to streamline healthcare processes and improve patient outcomes, building upon prior research in health informatics. Through a literature review, authors have provided insights into the current state-of-the-art in medical information systems, informing future research directions and facilitating the adoption of technology-driven approaches to improve healthcare quality and efficiency.

Srinivas et al. contribute to the scientific discourse by exploring advancements in molecular and clinical medicine. The study likely investigates topics relevant to molecular biology, clinical diagnostics, and therapeutic interventions, aiming to contribute to the understanding and treatment of various medical conditions. By synthesizing existing knowledge in the field, their work likely provides a comprehensive overview of current trends and future directions in molecular and clinical medicine, informing both research and clinical practice [9].

Mohd Salleh et al. focus on applying swarm intelligence techniques, specifically utilizing a fuzzy swarm approach for data imputation. Their study likely integrates particle swarm optimization (PSO) within a fuzzy clustering framework to address missing data, then trains a decision tree classification algorithm on the augmented dataset. Authors [10] have concluded that the potential of swarm intelligence techniques, particularly fuzzy-based approaches, in enhancing data imputation and classification tasks. Salleh and Samat [11] introduce FCMPSO, a novel imputation method tailored for handling missing data features in heart disease classification. The FCMPSO technique combines fuzzy clustering (FCM) with particle swarm optimization (PSO) to effectively address missing

data while optimizing dataset integrity. Following the imputation process, a heart disease classification task is undertaken, likely utilizing machine learning algorithms to train and evaluate the dataset's performance.

Sudha [12] likely delves into the application of technology within medical systems to improve healthcare delivery and patient outcomes. The study may focus on the development or evaluation of medical informatics solutions, electronic health records, or telemedicine applications, among other potential topics. To explore the integration of technological advancements into healthcare settings, aiming to enhance the efficiency, accuracy, and accessibility of healthcare services.

A method for locating absent values was put out by Shahzad et al.[14] Using GA, missing data in datasets were located. To determine the performance of an alternative result, information gain (IG) was employed. The UCI repository included the dataset. The end findings demonstrated that GA was a successful method for replacing the MD. The suggested approach worked better when there was a large range of values and a lot of incomplete data. This study highlights how important it is to manage MD. Nikon et al. [15] introduced a method for identifying the risk of atherosclerosis in the coronary arteries. A "ridge expectation-maximization imputation" (REMI) method based on an "extreme learning machine" (ELM) technique was developed to locate missing values in the atherosclerosis data set. To estimate the REMI/ELM classifier, the STULONG and UCI datasets were used. This classifier has a greater accuracy in risk detection than REMI/SVM.

Radhimeenakshi[16] assessed the effectiveness of SVM and ANN using a collection of UCI ML Repository datasets, such as the Cleveland HD dataset and the Statlog database. The missing data was imputed using the replace-by-mean approach. ANN error function minimization was achieved by using the gradient descent method. They employed the SVM classifier's 2D kernel function(KF). In addition to linear and nonlinear functions, SVM can handle kF. The nonlinear data can be handled by ANNs. There is more accuracy in the SVM model.

An approach for classifying HD in patients and supplying risk-level awareness was created by Dinesh et al. [17]. The heart disease machine learning dataset from UCI was utilized. The default value was used in cases when data were absent. With the Confusion matrix, the results of NB, SVM, RF, gradient-boosting, and Logistic Regression (LR) were compared. It was found that LR had the best performance. An ensemble model for categorizing heart diseases was created by Dewan et al. [18] using GA and backpropagation methods. The data came from the UCI ML repository. The filtering procedure took care of the missing data. Hybrid methods were developed with NB, J48, and backpropagation. For non-linear data, ANN is the most effective classification technique. One of the shortcomings of ANNs is the local minima problem in backpropagation.

A model for predicting heart disease based on RF and chi-square was proposed by Jabbar et al. [19]. Heart disease data sets were used to evaluate the suggested technique. The concept outperforms other approaches in terms of predicting accuracy, according to the trial results, and the provided model will help medical professionals forecast cardiac problems. The risk of coronary heart disease (HD) has been predicted using data mining algorithms such as gradient boosting, k-NN, NB, Decision Tree (DT), SVM, LR, and neural network (NN)[20]. A significant achievement might be the creation of a computerized HD risk prediction system. The UCI ML repository provided the dataset used in this work on the prediction of heart disease. Conventional machine learning algorithms perform better when using the feature selection strategy. With an accuracy of 92.85%, the RF algorithm in combination with PCA is the most accurate classification method for HD categorization.

Using machine learning approaches, Nilashi et al. [21] created a prediction system for the detection of heart illness. Both supervised and unsupervised ML techniques were used in the development of the suggested strategy. In this study, two imputation methodologies for MD imputation, fuzzy SVM, Principal Component Analysis (PCA), and Self-Organizing Maps (SOM) are used. It is recommended to employ incremental PCA and fuzzy SVM for incremental data learning to further minimize the computation time for illness prediction. The clinical data evaluation was restricted to two medical datasets, which is one of the study's limitations. This makes it difficult to weigh the advantages and disadvantages of the suggested approach in terms of forecast accuracy and computation time. Additionally, a clinical data set will be utilized to highlight the advantages and disadvantages of the suggested calculation technique.

Two other metrics, "inner class accuracy" (IA) and "outer class accuracy" (OA), were developed and suggested by Jordanov et al.[22] in addition to the "overall classification accuracy" (OCA) measure. They concluded that they might be utilized to determine which classifier is better in a given case in addition to the OCA. Numerous supervised machine-learning algorithms were tested for their accuracy and efficacy in classifying cardiac illnesses. This study

found that for three classifications based on KNN, DT, and RF algorithms, the RF strategy performs better than the KNN, RF, and DT strategies using a Kaggle heart disease dataset[23].

Table I: University of California Irvine Information on the attributes of the heart disease dataset.

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 t0 202
Exang	Exercise induced angina	Yes (1)
		No (0)
Oldpeak	ST depression induced by exercise relative to rest	0 to 6.2
Slope	Slope of peak exercise ST segment	Upsloping (1)
		Flat (2)
		Downsloping (3)
Ca	Number of major vessels coloured by fluoroscopy	0 – 3
Thal	Defect type	Normal (3)
		Fixed defect (6)
		Reversible defect (7)
Num	Heart disease	0 – 4

III. MATERIALS AND METHODS

3.1 Dataset

The Dataset is a group of linked data that, depending on the Data it represents, has a report for each instance and a feature for each dataset attribute. This study draws upon dataset from the UCI ML repository as well as from "Cleveland, Switzerland, Long Beach, and Hungary".Use Kaggle [24] and [25] to analyse your data. Of the 76

variables in the data collection, 14 are very helpful in the diagnosis of heart disease. Usually, the attribute for the predictive class is listed last. The dataset descriptions for the characteristics are shown in Table 1.

3.2 Significance of the attribute

Each feature in the UCI HD dataset has been given a hyperparameter based on how essential it is thought to be for HD prediction. Table 2 displays the characteristic's relevance, description, domain value, and hyperparameter.

Table II: Relevance of the characteristics in the University of California, Irvine heart disease dataset.

Attribute	**Description**	**Domain of value**	**Significance of the attribute to designate it as one of the hyper parameter**
Age	Age in year	29 to 77	Higher the age, 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 CVD. 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)	Presence of typical angina makes the diagnosis of 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, 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) is an independent risk factor 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. Presence of LVH can interfere with the diagnosis of ischemia from both the resting ECG and during treadmill exercise testing. Presence of ST-T wave abnormality in resting ECG (esp. in the absence of LVH) is an important diagnostic clue for ischemic heart disease.

Attribute	Description	Domain of value	Significance of the attribute to designate it as one of the hyper parameter
Thalach	Maximum heart rate achieved	71 t0 202	Maximum heart rate achieved during treadmill exercise testing indicates completeness of the test (in general, we need the person undergoing TMT to achieve a heart rate of more than 85% of 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, 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	Larger the ST depression esp. if seen in multiple contiguous ECG leads, higher the likelihood of underlying ischemic heart disease.
Slope	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 coloured by fluoroscopy	0 – 3	Coronary angiogram is a diagnostic test used to confirm the presence of coronary artery disease. More the number of vessels affected, worse the clinical outcome for the patient.
Thal	Defect type	Normal (3)	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 chances for having coronary artery disease.

3.3 The proposed cardiovascular disease multiple imputation technique

The functional block diagram displayed in Figure 1 specifies a brief description of the suggested technique to accomplish the study aim by appropriately classifying CVD. The features are regarded as data validation and substitution in the process. The patient data set concerning different ischemic heart conditions would be accessed by the proposed research project. The significance, domain value, explanation, and hyperparameter of the characteristic using the CVDMIT data replacement technique, the missing values in the data sets are first located and replaced. Subsequently, a standard dataset is employed in a validation procedure based on maximum values to validate the identified missing value. The missing value data set used in this process, which is shown in Table 2, was contributed by the machine learning repository at the University of California, Irvine.

To find the nearest neighbor for each cluster concerning specific attributes, to calculate the proximity between cluster centroids and individual data points or other clusters. This involves computing the distances and identifying the closest data points or clusters based on the selected attributes, ensuring the nearest neighbor within the context of those specific attributes is determined for each cluster. The membership function calculates the degree to which a given data point belongs to a specific cluster, typically ranging from 0 to 1, where higher values indicate stronger membership or belongingness to the cluster. Predicting heart disease (HD) can be enhanced by combining multiple techniques: Fuzzy Rough Set and Fuzzy C-Means provide robust methods for handling uncertainty and clustering, respectively, while Expectation Maximization aids in probabilistic clustering, and Random Forests offer a powerful ensemble method for classification. By integrating these methods, the prediction model can effectively leverage diverse strengths, improving accuracy and reliability in diagnosing heart disease.

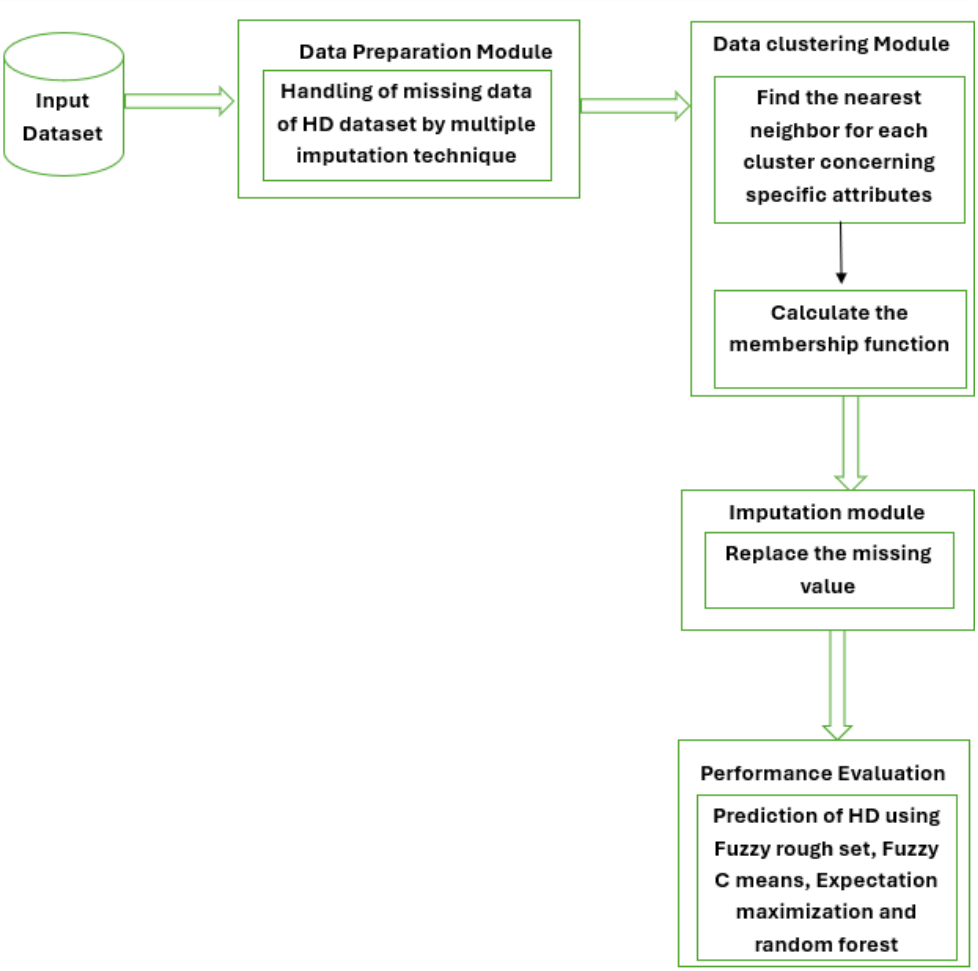

Figure 1. Overall system model for multiple imputation approach for cardiovascular disease

3.4 The CVDMIT algorithm

The algorithm for the Ischemic Heart Disease Multiple Imputation Technique starts by taking a dataset D with missing values and a fuzzy tolerance relation R as inputs. For each instance y in the dataset, it checks for missing values using the `Missing(y)` method. For each attribute a in y, it checks if the specific attribute value is missing using the `missing(a(y))` method. If a missing value is found, an imputation set Ia is initialized to store potential imputed values. The algorithm then evaluates each other instance z in the dataset, ensuring evaluations are only performed when $z \neq t$. It calculates the upper and lower approximations for the missing value using fuzzy rough sets and the fuzzy tolerance relation R, adding these values to the imputation set Ia. The missing attribute value is imputed by computing the mean or another appropriate central tendency measure of the values in Ia, and the computed imputed value replaces the missing attribute value in the instance y. Finally, the algorithm1 returns the dataset with the imputed values.

Algorithm1: Cardiovascular Disease Multiple Imputation Technique

Input: Dataset D with missing values, Fuzzy tolerance relation R.
Output: Dataset D′ with imputed values.
Procedure:
For each instance y in D:
If Missing(y) **is true**:
For each attribute a in y:
If missing(a(y)) **is true**:
Initialize imputation set Ia to an empty set.
For each instance z in D:
If $z \neq t$:
Calculate the upper approximation Ua(y) and lower approximation La(y) for a(y) using fuzzy rough sets and R.

Add Ua(y) and La(y) values to the imputation set Ia.
Compute the imputed value for a(y) as the mean of the values in Ia.
Replace the missing value a(y) with the computed imputed value.
End If
End For
Return the imputed dataset D′.

3.5 Results and Discussion

In the HD dataset, missing values are shown as "blank space." 'KNNimputer,' 'Mean,' 'fill,' and 'bill' are the most often utilized Single Imputation techniques. Since DataFrame interprets these missing values as "NaN" values, they are removed from the dataset using the imputation approach, as shown below. The limitation of the variance between features in both approaches is a downside, and the presence of outliers in the dataset will have an impact on the mean value. results show that the KNNimputer has an accuracy rate of 0.80 and a recall rate of 0.79.
The accuracy and recall rates of several single imputation methods when utilising the Decision Tree (DT) algorithm are displayed in Table 3. KNNimputer is the most effective of these techniques, with an accuracy rate of 0.81 and a recall rate of 0.80. A comparative analysis of the accuracy and recall rates, both with and without the imputation approach, indicates a significant enhancement in performance when imputation is employed. Specifically, the use of imputation techniques consistently improves the accuracy and recall rates, highlighting their importance in handling missing data within the DT algorithm framework. These findings underscore the efficacy of imputation methods in boosting the predictive performance of machine learning models.

Table III: Comparison of different approaches' accuracy and recall rates with and without single imputation.

Sl.no	Imputation method	Without using imputation method	Accuracy rate with DT	Recall rate with DT
1	ffill	76.2	0.79	0.78
2	bfill		0.79	0.78
3	mean		0.80	0.79
4	KNNimputer		0.81	0.80

Figures 2 and 3 show the sample imputed data and the anticipated class outcomes. The binary classification approach used in this investigation has class values of 0 and 1. Within this context, the absence of heart disease is indicated by a class value of 0, and its presence is shown by a class value of 1 in the dataset. The classification process considers the nearest neighbor approach, where the class value assigned to a data point is influenced by the class value of its nearest neighbor. This method ensures that similar instances, particularly those in close proximity, are classified consistently, thereby enhancing the accuracy of the predictions.

```
     0    1    2    3    4    5  ...  8    9   10   11   12   13
1   63.0 1.0  3.0  145.00000 233.0 1.0  ... 0.0  2.300000 0.0  0.0  1.0  1.0
2   37.0 1.0  2.0  134.50495 250.0 0.0  ... 0.0  3.500000 0.0  0.0  2.0  1.0
3   41.0 0.0  1.0  130.00000 204.0 0.0  ... 0.0  1.400000 2.0  0.0  2.0  1.0
4   56.0 1.0  1.0  120.00000 236.0 0.0  ... 0.0  4.784158 2.0  0.0  2.0  1.0
5   57.0 0.0  0.0  120.00000 354.0 0.0  ... 1.0  0.600000 2.0  0.0  2.0  1.0
..  ...  ...  ...  ...  ...  ...  ... ...  ...  ...  ...  ...  ...  ...
299 57.0 0.0  0.0  140.00000 241.0 0.0  ... 1.0  0.200000 1.0  0.0  3.0  0.0
300 45.0 1.0  3.0  110.00000 264.0 0.0  ... 0.0  1.200000 1.0  0.0  3.0  0.0
301 68.0 1.0  0.0  144.00000 193.0 1.0  ... 0.0  3.400000 1.0  2.0  3.0  0.0
302 57.0 1.0  0.0  130.00000 131.0 0.0  ... 1.0  1.200000 1.0  1.0  3.0  0.0
303 57.0 0.0  1.0  130.00000 236.0 0.0  ... 0.0  0.000000 1.0  1.0  2.0  0.0
[303 rows x 14 columns]
```

Figure 2. Imputed Data

Figure 4 displays the dataset with missing values, highlighting the gaps in the data that need to be addressed for accurate analysis. Following the imputation process, the resulting complete dataset is presented in Figure 5. This imputed dataset fills in the previously missing values, enabling a more robust and reliable analysis by ensuring that

all data points are considered. The transformation from Figure 4 to Figure 5 demonstrates the effectiveness of the imputation techniques in restoring the integrity of the dataset.

```
[1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1.
 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1.
 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1.
 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1.
 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1.
 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1.
 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 0. 0. 0.
 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.
 0. 0. 1. 0. 1. 0. 0. 0. 1. 0. 0. 0. 0. 0. 0. 0. 0. 0. 1. 0. 0. 0. 0. 0.
 0. 0. 0. 0. 0. 0. 1. 0. 0. 0. 0. 0. 1. 0. 1. 0. 0. 0. 0. 0. 0. 0. 0. 0.
 0. 1. 0. 0. 0. 0. 0. 0. 1. 0. 0. 0. 0. 0. 1. 0. 0. 0. 0. 1. 0. 1. 0. 0.
 0. 0. 0. 1. 0. 0. 0. 1. 1. 0. 0. 0. 0. 1. 1. 0. 0. 0. 1. 1. 0. 0. 1. 1.
0. 0. 0. 0. 0. 1. 0. 0. 0. 0. 0. 1. 0. 0. 1.]
```

Figure 3. Predicted Class

1	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
2	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
3	37	1	2		250	0	1	187	0	3.5	0	0	2	1
4	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
5	56	1	1	120	236	0	1	178	0		2	0	2	1
6	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
7	57	1	0	140	192	0	1	148	0	0.4	1	0	1	1
8	56	0	1	140	294	0	0	153	0	1.3	1	0	2	1
9	44	1	1	120	263	0	1	173	0	0	2	0	3	1
10	52	1	2	172	199	1	1	162	0	0.5	2	0	3	1
11	57	1	2	150	168	0	1	174	0	1.6	2	0	2	1
12	54	1	0	140	239	0	1	160	0	1.2	2	0	2	1
13	48	0	2	130	275	0	1	139	0	0.2	2	0	2	1
14	49	1	1	130	266	0	1	171	0	0.6	2	0	2	1
15	64	1	3	110	211	0	0	144	1	1.8	1	0	2	1
16	58	0	3	150	283	1	0	162	0	1	2	0	2	1
17	50	0	2	120	219	0	1	158	0	1.6	1	0	2	1
18	58	0	2	120	340	0	1	172	0	0	2	0	2	1
19	66	0	3	150	226	0	1	114	0	2.6	0	0	2	1
20	43	1	0	150	247	0	1	171	0	1.5		0	2	1
21	69	0	3	140	239	0	1	151	0	1.8	2	2	2	1
22	59	1	0		234	0	1	161	0	0.5	1	0	3	1
23	44	1	2	130	233	0	1	179	1	0.4	2	0	2	1

Figure 4. Dataset with missing value

ImputedData - DataFrame

Index	0	1	2	3	4	5	6	7	8	9
1	63	1	3	145	233	1	0	150	0	2.3
2	37	1	2	134.505	250	0	1	187	0	3.5
3	41	0	1	130	204	0	0	172	0	1.4
4	56	1	1	120	236	0	1	178	0	4.78416
5	57	0	0	120	354	0	1	163	1	0.6
6	57	1	0	140	192	0	1	148	0	0.4
7	56	0	1	140	294	0	0	153	0	1.3
8	44	1	1	120	263	0	1	173	0	0
9	52	1	2	172	199	1	1	162	0	0.5
10	57	1	2	150	168	0	1	174	0	1.6

Format Resize Background color Column min/max Save and Close Close

Figure 5. Imputed Dataset

Table IV: The suggested multiple imputation methodology for ischemic heart disease is compared to alternative imputation techniques.

	Fuzzy rough set imputation	**Fuzzy C means imputation**	**Expectation maximization imputation**	**Proposed CVDMIT using RF**
Accuracy	90.4176	91.2176	88.8176	94.011
Sensitivity	96.5088	95.0176	93.2379	96.5088
Specificity	85.3117	87.7099	84.8661	89.7235
Precision	84.633	87.7099	84.633	88.7273
Recall	96.5088	95.0176	93.2379	96.5088
F measure	0.9192	0.9296	0.9047	0.942

Table 5 presents a comparative analysis of various IHDMIT using the UCI HD Dataset. The proposed technique, which utilizes Random Forest (RF) classification, demonstrates superior performance, achieving an impressive accuracy of 93%. This is in comparison to other established methods such as the Expectation-maximization multiple imputation approaches [28], fuzzy C-means [27], and fuzzy rough set method [26]. The results underscore the efficiency of the proposed technique in improving the accuracy of ischemic heart disease predictions, highlighting its potential for enhancing clinical decision-making and patient outcomes.

Table V: The University of California Irvin heart disease dataset is used to compare the suggested multiple imputation approach for ischemic heart disease.

Imputation method	**Dataset**	**Accuracy %**
Fuzzy Roughset.	UCI Heart Disease Dataset	90
Fuzzy C Means.		92
REMI.		89
Proposed CVDMIT		94

Due to the fact that the suggested technique only compares each neighbourhood to every other neighbour once and uses the class value to determine which neighbour is closest to you, it ensures computational efficiency and accuracy in identifying relevant patterns. This innovative approach significantly reduces redundancy and enhances the model's precision in handling ischemic heart disease data [29][30]. The ROC charts were utilised as a critical performance parameter to further confirm the effectiveness of the proposed model. The use of ROC AUC values provides a robust assessment of the model's ability to distinguish between different classes, thus highlighting its diagnostic capabilities.

Figure 6 showcases the ROC curves for the proposed model, the REMI model, the fuzzy rough set, and the fuzzy C means. The suggested model has the greatest ROC value (0.94), as shown in the figure, compared to the fuzzy rough set model's 0.90, the fuzzy C means model's 0.92, and the REMI model's 0.89. These findings highlight the suggested IHDMIT's excellent performance in terms of ROC curve accuracy. The higher value of the proposed model indicates its greater effectiveness in correctly classifying ischemic heart disease cases, thus demonstrating its potential for improving clinical decision-making and patient outcomes.

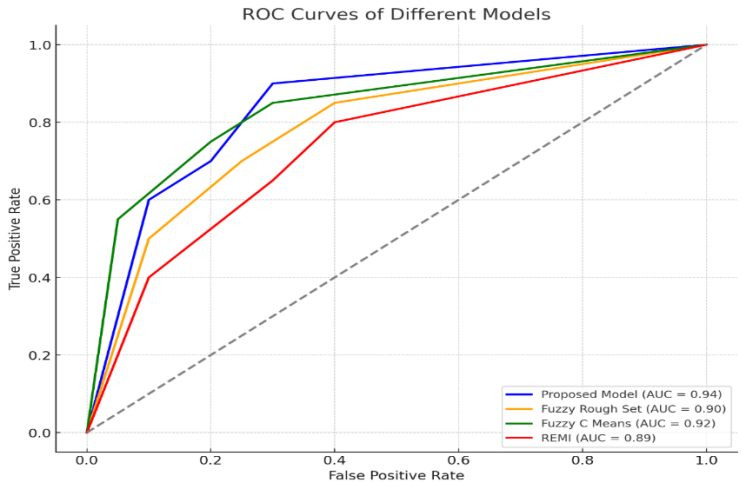

Figure 6. ROC curves of Fuzzy roughest, Fuzzy C means, REMI, and proposed model.

IV. Conclusion

Significant progress has been made in medical data analysis in recent years, especially in the area of predictive analytics for heart disease. One notable addition to this field is the suggested Missing Data Imputation Technique (CVDMIT), which was created using the benchmark UCI Heart Disease dataset. Utilising Random Forest (RF) classification, the CVDMIT attains a remarkable 94% accuracy rate, showcasing its resilience and dependability while managing numerical data. Handling missing data is a key difficulty in medical data analysis, since it can greatly affect the prediction model's effectiveness. This problem has been widely addressed by traditional imputation approaches such expectation-maximization (EM) methods, fuzzy rough set, and fuzzy C means. On the other hand, the suggested CVDMIT has performed better on a number of assessment criteria, including as F_measure, accuracy, sensitivity, specificity, precision, and recall. This better performance highlights how useful the CVDMIT is in producing trustworthy and accurate imputations, which are essential for further predictive modelling.

V. Conflict of Interest

The authors declare no conflict of interest.

References

[1] Cenitta, D., R. Vijaya Arjunan, and K. V. Prema. "Ischemic heart disease multiple imputation technique using machine learning algorithm." *Engineered Science* 19 (2022): 262-272.

[2] Soundharya, R., D. Cenitta, and R. Vijaya Arjunan. "Information concealment and redemption through data anonymization technique." *Journal of Advanced Research in Dynamical and Control Systems* 10.7 (2018): 22-26.

[3] Tang, Jinjun, et al. "Missing data imputation for traffic flow based on combination of fuzzy neural network and rough set theory." *Journal of Intelligent Transportation Systems* 25.5 (2021): 439-454.

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

[5] Nikfalazar, Sanaz, et al. "Missing data imputation using decision trees and fuzzy clustering with iterative learning." *Knowledge and Information Systems* 62 (2020): 2419-2437.

[6] Razavi-Far, Roozbeh, et al. "Similarity-learning information-fusion schemes for missing data imputation." *Knowledge-Based Systems* 187 (2020): 104805.

[7] Pati, Soumen Kumar, and Asit Kumar Das. "Missing value estimation for microarray data through cluster analysis." *Knowledge and Information Systems* 52 (2017): 709-750.

[8] Idri, Ali, et al. "Missing data techniques in classification for cardiovascular dysautonomias diagnosis." *Medical & Biological Engineering & Computing* 58 (2020): 2863-2878.

[9] Srinivas, K., et al. "Prediction of heart disease using hybrid linear regression." *Eur J Mol Clin Med* 7.05 (2020).

[10] Salleh, Mohd Najib Mohd, and Nurul Ashikin Samat. "An imputation for missing data features based on fuzzy swarm approach in heart disease classification." *Advances in Swarm Intelligence: 8th International Conference, ICSI 2017, Fukuoka, Japan, July 27–August 1, 2017, Proceedings, Part II 8*. Springer International Publishing, 2017.

[11] Salleh, Mohd Najib Mohd, and Nurul Ashikin Samat. "FCMPSO: An imputation for missing data features in heart disease classification." *IOP Conference Series: Materials Science and Engineering*. Vol. 226. No. 1. IOP Publishing, 2017.

[12] Sudha, M. "Evolutionary and neural computing based decision support system for disease diagnosis from clinical data sets in medical practice." *Journal of medical systems* 41.11 (2017): 178.

[13] Cenitta, D., and R. Vijaya Arjunanarjunan. "A technical analysis on data mining classification algorithms for cvd prediction." *Journal of Advanced Research in Dynamical and Control Systems* 10.7 (2018): 137-142.

[14] Shahzad, Waseem, Qamar Rehman, and Ejaz Ahmed. "Missing data imputation using genetic algorithm for supervised learning." *International Journal of Advanced Computer Science and Applications* 8.3 (2017).

[15] 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.

[16] Radhimeenakshi, S. "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 (INDIACom)*. IEEE, 2016.

[17] Dinesh, Kumar G., et al. "Prediction of cardiovascular disease using machine learning algorithms." *2018 International Conference on Current Trends towards Converging Technologies (ICCTCT)*. IEEE, 2018.

[18] Dewan, Ankita, and Meghna Sharma. "Prediction of heart disease using a hybrid technique in data mining classification." *2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom)*. IEEE, 2015.

[19] Jabbar, M. Akhil, Bulusu Lakshmana Deekshatulu, and Priti Chandra. "Prediction of heart disease using random forest and feature subset selection." *Innovations in Bio-Inspired Computing and Applications: Proceedings of the 6th*

International Conference on Innovations in Bio-Inspired Computing and Applications (IBICA 2015) held in Kochi, India during December 16-18, 2015. Springer International Publishing, 2016.

[20] Tasnim, Farzana, and Sultana Umme Habiba. "A comparative study on heart disease prediction using data mining techniques and feature selection." *2021 2nd International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST)*. IEEE, 2021.

[21] Nilashi, Mehrbakhsh, et al. "Coronary heart disease diagnosis through self-organizing map and fuzzy support vector machine with incremental updates." *International Journal of Fuzzy Systems* 22 (2020): 1376-1388.

[22] Jordanov, Ivan, Nedyalko Petrov, and Alessio Petrozziello. "Classifiers accuracy improvement based on missing data imputation." *Journal of Artificial Intelligence and Soft Computing Research* 8.1 (2018): 31-48.

[23] Ali, Md Mamun, et al. "Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison." *Computers in Biology and Medicine* 136 (2021): 104672.

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

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

[26] Acharjya, D. P. "A hybrid scheme for heart disease diagnosis using rough set and cuckoo search technique." *Journal of Medical Systems* 44.1 (2020): 1-16.

[27] Chitra, R., and V. Seenivasagam. "Heart attack prediction system using fuzzy C means classifier." *IOSR Journal of Computer Engineering* 14.2 (2013): 23-31.

[28] Khan, Hufsa, Xizhao Wang, and Han Liu. "Missing value imputation through shorter interval selection driven by Fuzzy C-Means clustering." *Computers & Electrical Engineering* 93 (2021): 107230.

[29] Li, Daiwei, et al. "Hybrid missing value imputation algorithms using fuzzy c-means and vaguely quantified rough set." *IEEE Transactions on Fuzzy Systems* 30.5 (2021): 1396-1408.

[30] Tang, Jinjun, et al. "Missing data imputation for traffic flow based on combination of fuzzy neural network and rough set theory." *Journal of Intelligent Transportation Systems* 25.5 (2021): 439-454.