Science and Technology Indonesia
e-ISSN:2580-4391 *p*-ISSN:2580-4405
Vol. 6, No. 4, October 2021

Research Paper

Handling Missing Data Using Combination of Deletion Technique, Mean, Mode and Artificial Neural Network Imputation for Heart Disease Dataset

Anita Desiani[1]*, Novi Rustiana Dewi[1], Annisa Nur Fauza[1], Naufal Rachmatullah[2], Muhammad Arhami[3], Muhammad Nawawi[4]

[1]*Mathematics Department, Mathematics and Natural Science Faculty, Sriwijaya University, Palembang, 30862, Indonesia*
[2]*Informatics Technique Department, Informatics Faculty, Sriwijaya University, Palembang, 30862, Indonesia*
[3]*Informatics Technique Department, Lhokseumawe State Polytechnic, Aceh, 24301, Indonesia*
[4]*Mechanical Engineering Departement, Graduate School of Science, Engineering and Technology, Istanbul Technical University Maslak Sarıyer, 34467, Turkey*
***Corresponding author**: anita_desiani@unsri.ac.id

Abstract

The University of California Irvine Heart disease dataset had missing data on several attributes. The missing data can loss the important information of the attributes, but it cannot be deleted immediately on dataset. To handle missing data, there are several ways including deletion, imputation by mean, mode, or with prediction methods. In this study, the missing data were handled by deletion technique if the attribute had more than 70% missing data. Otherwise, it were handled by mean and mode method to impute missing data that had missing data less or equal 1%. The artificial neural network was used to handle the attribute that had missing data more than 1%. The results of the techniques and methods used to handle missing data were measured based on the performance results of the classification method on data that has been handled the problem of missing data. In this study the classification method used is Artificial Neural Network, Naïve Bayes, Support Vector Machine, and K-Nearest Neighbor. The performance results of classification methods without handling missing data were compared with the performance results of classification methods after imputation missing data on dataset for accuracy, sensitivity, specificity and ROC. In addition, the comparison of the Mean Squared Error results was also used to see how close the predicted label in the classification was to the original label. The lowest Mean Squared Error was obtained by Artificial Neural Network, which means that the Artificial Neural Network worked very well on dataset that has been handled missing data compared to other methods. The result of accuracy, specificity, sensitivity in each classification method showed that imputation missing data could increase the performance of classification, especially for the Artificial Neural Network method.

Keywords

Missing Data, Artificial Neural Network, Imputation, Mean-mode, Deletion, Heart Disease

Received: 10 June 2021, Accepted: 18 September 2021
https://doi.org/10.26554/sti.2021.6.4.303-312

1. INTRODUCTION

The heart disease dataset is often used in classification or prediction. That is used to determine the pattern of factors that affect a heart disease. Heart disease is one of the biggest causes of death in the world (Rahakbauw et al., 2016). Based on research that has been conducted by Stewart et al. (2017), nearly one billion people worldwide suffer strokes caused by hypertension and heart attacks. Some data on the results of examining heart disease diagnoses are published as a dataset to help various researches. One data warehouse that provides a heart attack dataset is the University of California Irvine (UCI) Machine Learning Repository. The heart attack dataset provided by the UCI is heart disease diagnostic data collected based on four sources, namely the Cleveland Clinic Foundation (Cleveland data), Hungarian Institute of Cardiology, Budapest (Hungarian data), VA Medical Center, Long Beach, CA (long-beach-va data), and University Hospital, Zurich, Switzerland (Switzerland data). Data from these various sources was published by UCI into a dataset of diagnoses of heart disease patients which can be used for prediction of heart disease patients (Jasoni and Steinbrunn, 2013). The heart disease patient dataset contains 76 attributes, but only 14 attributes that affect heart disease (Zriqat et al., 2016). The heart attack dataset has incomplete data about 491 missing data from several attributes (Misir and Samanta., 2017). Although it has weaknesses in the completeness of the data and attributes used, the heart attack dataset has been widely used in various studies to diagnose patients with heart disease.

Missing data problems can occur in various datasets not only in UCI heart disease dataset, such as gene and microarray data (Moorthy et al., 2014), medical data (Karim et al., 2017;

Purwar and Singh, 2015), credit data (Crone and Finlay, 2012; Lan et al., 2020), software quality dataset (Huang et al., 2017; Jing et al., 2016), etc. The Missing data problems are a common problem in real-world data classification. Therefore, a strong classification method is needed when classifying data that has missing data problems in its dataset domain (Somasundaram and Nedunchezhian, 2012). Missing data is caused by several things including errors in manual data entry procedures, equipment errors or wrong measurements (Purwar and Singh, 2015). Incomplete dataset will affect the accuracy of the data mining model, it can give biased results, and reduce the efficiency of the computation process because there is missing information in the dataset (Choudhury and Pal, 2019). The missing data can significantly reduce the accuracy and usefulness of the assessment model especially in missing cases with lots of variations and can also cause errors and confusion in interpreting the data. The Missing data compromises the quality of the data, and in turn affects the accuracy of the model derived from the data (Karim et al., 2017; Silva-Ramírez et al., 2011). Unfortunately, missing data in the dataset have a negative impact on estimation accuracy and hence, may lead to inconsistent results. Many estimation models cannot directly handle missing data values; therefore, the preprocessing stage becomes indispensable for modern estimation processes in software engineering (Huang et al., 2017).

The preprocessing step is the process needed to clean and filter target data because data collection is rarely complete and perfect (Salleh and Samat, 2017). So, the preprocessing method is an important role in the data mining task. Preprocessing is an important step for filtering and cleaning the dataset before it can be trained at the data mining stage so that the data used is of best quality (Crone and Finlay, 2012). The ability to handle missing data has become a fundamental requirement for pattern classification, because improper treatment of missing data can lead to misclassification results (García et al., 2015). Nowadays, most of the algorithms in data mining have not been able to directly handle the problem of missing data. According to Eekhout et al. (2014); Poolsawad et al. (2012); Vazifehdan et al. (2019), there are several techniques for handling missing data, namely; the first is Deletion, namely deleting an instant (record) or an incomplete attribute and the classification only uses the complete part of the data. The second is the imputation or estimation of missing data will be used in classification. The third is ignore, Third is ignore, which is using data directly without handling missing data in a data set The simplest way to deal with this problem is deletion, i.e. deleting data that has missing data directly. However, this is only suitable for very small loss rates of 1-5% (Vazifehdan et al., 2019). Otherwise, if there is too much missing data on an attribute that there is little bit of information about the data, then the attribute can be removed from the data set because the information that attribute is incomplete (Shah et al., 2017). Removing missing data in a dataset sometimes has a negative impact on the accuracy of estimates and hence results in inconsistent results (Lan et al., 2020). Deletion has been widely adopted to handle missing data during data preprocessing (Huang et al., 2017; Malarvizhi and Thanamani, 2012). The method of imputation missing data was to replace missing variables with value estimates that can maintain data completeness (Choudhury and Pal, 2019). The imputation method is a solution that can handle the problem of missing data where the missing data attribute is estimated or replaced by using various methods including statistical methods, such as mean or mode, machine learning, and others (García et al., 2015; Luengo et al., 2011; Tsai et al., 2018).

There are some method and technique can use for imputation missing data. The simplest statistical method is mean method for numeric attribute and modes that focuses on the value of an attribute which appears frequently for imputation of category attribute (Eekhout et al., 2014; Mehrotra et al., 2017; Nishanth and Ravi, 2016; Silva-Ramírez et al., 2011). According to Eekhout et al. (2014) the mean imputation can lead to biased estimates for each data scenario when the incidence of missing data in a domain is more than 10%. The mean imputation can also produce biased results if the observational needs a relational value between variables, because it does not consider the existing relational value variables (Pedersen et al., 2017). Another shortcoming is the mean method cannot be used to represent data for the values in the attribute are extreme. Conversely, the mode is the easiest way to impute categorical data, but the results given will be biased if the mode value is more than one or even the mode value is not found in the attribute that has missing data. Another disadvantage of the imputation mode is that it ignores the variance of the population or sample that exists (Nishanth and Ravi, 2016). The mean and mode methods are very suitable for imputation of missing data at a single value where the percentage of missing data is not too large. Currently, more complex imputation approaches using machine learning approaches, such as Random Forest (Stekhoven and Bühlmann, 2012), Neural Networks (Nishanth and Ravi, 2016; Rahman and Davis, 2012), K-Nearest Neighbor (KNN) (Manimekalai and Kavitha, 2018), K-Means (Poolsawad et al., 2012), Decision Tree (Chauhan et al., 2013), Deep Learning (Ting et al., 2020). Approaches with mechanical learning such as Neural Networks is an alternative to best imputation results but take more time than the statistical approach and it is not effective for small amounts of missing data (Tsai et al., 2018). Apart from deleting and estimating missing data by imputation, attribute selection also influences the classification or prediction results. Irrelevant attributes do not affect the description of target class. The redundant attributes do not contribute to anything but they create bias in the classification results (Shilaskar and Ghatol, 2013).

According to Pedro, at least 45% of the data set provided by the UCI had a problem with missing data, including the heart data dataset. Several studies to predict or classify heart disease disordered in the heart disease dataset both deal with missing data by signing it, deletion data, imputation data and elimination by using the selection attribute (attribute). Al Khaldy and

Kambhampati (2016) predicted the pattern of heart disease by applying machine learning to predict heart disease problems regardless of the existence of missing data in the dataset. Salleh and Samat (2017) applied Fuzzy C-Means and Particle Swarm Optimization to import missing data on heart disease dataset regardless of the percentage of missing data on each attribute. Choudhury and Pal (2019) showed that the Neural Network method had stable performance for attribute that have 1-10% missing data, but their performance greatly improved when working on attribute that have 50% missing data. Unfortunately it did not explain how the data handler handled more than 50% of missing data. Tsai et al. (2018) implemented imputation of missing data using class center method, namely by finding the center of each class and measuring the distance between classes to estimate the missing data threshold in the dataset. The accuracy in this study was only 78% without explaining the classification or prediction method used to detect cardiac disorders. Silva-Ramírez et al. (2011) applied a Neural Network to impute the missing data which amounts to no more than 5% but it did not explain the difference in results if the missing data was only 1% or greater than 5%. Subbalakshmi et al. (2011) combined several imputation methods, namely LOCF, Mean-Mode and IV but it did not explain the imputation method used by each attribute. Hernández-Pereira et al. (2015) compared Mean or Mode Imputation, Multiple Linear Regression, Hot-deck, K-NN, and Neural Network (NN) to handle missing data. The result of in this study showed that the performance of NN provides the best performance for handling missing data. Several studies conducted deletion of data on attributes that deemed to have less significant influence on heart disease dataset (El-Bialy et al., 2015; Long et al., 2015).

The heart disease dataset has several attributes that have different amounts of missing data. The study focused how to handle the missing data on heart disease dataset and tried to use multiple ways to overcome the problem. This study was not only use one technique or method for some attributes but it used multiple ways namely Deletion technique, Mean, Mode and Artificial Neural Network methods for imputation missing data. This study tried to get the advantages of each technique and methods to solve the problem of missing data. The Deletion technique was used in the study for attributes where the amount of missing data was more than 70% because the information available for the attribute was considered insufficient. The Mean and Mode methods used for attribute that had missing data lower or equal with 1%, because mean or mode method are suitable for missing data with a single value and the amount of missing data is not too much (Nishanth and Ravi, 2016). For attribute that had amount of missing data more than 1%, the study used artificial neural network because some researches has been showed than Artificial Neural Network (ANN) method was greatly to impute missing data that had amount more than 1% (Choudhury and Pal, 2019; Tsai et al., 2018). The utilization for each technique and methods in the study was adjusted to the percentage amount of missing data from each attribute in the dataset at the pre-processing step to get best data quality and provide best results in classification. Performance measurements of a classification method that were usually used include accuracy, specificity, sensitivity and ROC (Desiani et al., 2021; Resti et al., 2021). This study used accuracy, specificity, sensitivity and ROC to measure the performance classification in the proposed method.

2. EXPERIMENTAL SECTION

2.1 Materials

In this research used secondary data, namely data on patients with heart disease obtained from the University of California Irvine (UCI) Machine Learning Repository which can be downloaded at their official website (Statlog, 2004). Data were obtained from 294 patients suspected of having heart disease. The data contains 14 attributes that are used as influential attributes in diagnosing heart disease, among others age, gender, type of chest pain, blood pressure, cholesterol, sugar levels, electrocardiography, maximal heart rate, induced angina, old peak, slope, fluoroscopy, heart rate and attributes as labels containing the categories healthy and sick. All attributes and values and types of each attribute can be seen in Table 1.

2.2 Methods

All of stages of the proposed Method in the study can be seen in Figure 2. The stages are :

2.2.1 Handling Missing Data

Deletion, the attributes that have a percentage of missing datas of more than 70% are deleted and dropped from the data set because they are considered not representative and have insufficient information. Attributes that have missing data of less than 70% will be imputed missing data. Mean and Mode Imputation, the mean method that has number of missing data less than 1% will be predicted using calculate mean of the attribute or attribute with categorical type that has number of missing datas less than 1% of the total data will be predicted using the mode imputation. For attribute with numeric (integer or real) that has number of missing data less than 1% will be predicted using calculate mean of the attribute. Artificial Neural Network Imputation, for missing data with an amount of more than 1%, it will be imputed using a Artificial Neural Network (ANN). In this step, data will be split into 2 group, first was data training and the second as data testing. The ANN for this step was used 3 layers, namely input layer, one hidden layer and output layer (Figure 1).

The first step that must be done in the training stage is normalizing the input data only for the continue or numeric attribute on the dataset, because the data range for the continue attribute is different so the data must be normalized by

$$x'_i = \frac{x_i - a}{b - a} \tag{1}$$

Next initialize the weights for each input and bias in Figure 1 associated with the hidden layer. Then Calculate the input

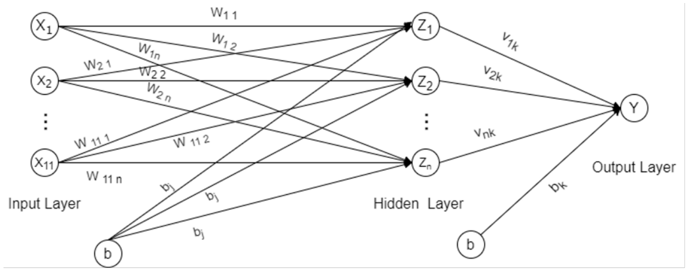

Figure 1. Artificial Neural Network with 3 Layers

value for each unit using Equation (2) based on Figure 1, is the jth neuron in the hidden layer.

$$z_{inj} = \sum_{i=1}^{n} w_{ji} x'_i + b_j \tag{2}$$

Where w_{ji} is the weight for the i-th neuron in the input layer and the j-th neuron in the hidden layer and b_j is the bias to calculate z_j . After that, calculate all the activation values as the output of each hidden layer (ϕ_h) to the output layer (y) with the sigmoid function as the activation function in equation (3). Next step is calculate the input value (y_{inj}) for the output layer using Equation (4).

$$\phi_h = \frac{1}{1 + e^{-z_{inj}}} \tag{3}$$

$$y_{inj} = \sum_{j=1}^{p} v_{kj} z_j + b_k \tag{4}$$

Where v_{kj} is weight for the j-th on hidden layer and the k-th neuron on output layer. b_k is bias is a bias to calculate y_k . After that, calculate activation function (ϕ_o) as input for output layer (y_{inj}) on every the j-th of input using Equation (5).

$$\phi_o = \frac{1}{1 + e^{-y_{inj}}} \tag{5}$$

Perform the backpropagation of error stage by calculating the unit error factor (δ) based on the error on input (y_{ink}) on output layer for each output)y_k with Equation (6).

$$\delta_k = (t_k - \phi_o) f'(y_{ink}) = (t_k - \phi_o) y_k (1 - \phi_o) \tag{6}$$

δ_k is the error unit that will be used in changing the layer weight with (t_k) being the k-th output target. Next step is calculate the value of the weight change ($\Delta \mathrm{v}_{kj}$) in Equation (7) which is used to update the weight value of v_{kj} on the hidden layer z_j based on the activation value (ϕ_o) that has been calculated previously, with the learning rate acceleration (α=0.1)

$$\Delta v_{kj} = \alpha \delta_k z_j \tag{7}$$

Calculate the value of the change in bias ($\Delta \mathrm{b}_k$) which is used to update the bias value b_k at the output layer y based on the value of learning rate and unit error (δ_k) and Calculate the unit error (δ_{inj}) that comes from the output layer to the hidden layer using Equation (9).

$$\Delta b_k = \alpha \delta_k z_j \tag{8}$$

$$\delta_{inj} = \sum_{k=1}^{m} \delta_k v_{kj} \tag{9}$$

Next, it should calculate the hidden unit error (δ_j) in the hidden layer using Equation (10) and Calculate the change weight value of w which is then used to update the weight value of w_{ij}.

$$\delta_j = \delta_{inj} f'(z_{inj}) \tag{10}$$

$$\Delta w_{ij} = \alpha \delta_j x_i \tag{11}$$

For the bias, calculate the value of the change in bias ($\Delta \mathrm{b}_j$) based on the unit error (δ_j) on hidden layer. After that, Update each bias and weight on the hidden layer with the Equation (13) and Update each bias and weight on the hidden layer with the Equation (14).

$$\Delta b_j = \alpha \delta_j \tag{12}$$

$$w_{ji}(\text{new}) = w_{ji}(\text{old}) + \Delta w_{ji} \tag{13}$$

$$v_{kj}(\text{new}) = v_{kj}(\text{old}) + \Delta v_{kj} \tag{14}$$

Update the weight on the bias by using Equation (15) to obtain the new bias weight value (b_j(new)) in the hidden layer and Equation (16) for the new bias weight value (b_k(new)) in the output layer. The calculation steps for the training phase are carried out on all existing input data until the weights no longer experience significant differences or depend on the epoch specified for each attribute.

$$b_j(\text{new}) = b_j(\text{old}) + \Delta b_j \tag{15}$$

$$b_k(\text{new}) = b_k(\text{old}) + \Delta b_k \quad (16)$$

At this testing phase , the weights generated by the ANN were applied to the testing data to test the performance results of the ANN. The steps needed were to take the last weight at the training stage and classify it with Equations (2) and (4) then compared the labels of the classification results with the original labels of the data. The comparison is used to measure the results in the accuracy, specificity (SP), and sensitivity (SN) performance for ANN. Another way to The measure the ANN performance for every attribute, the label classification in each method was compared with the original label to calculate the difference or error that occurs. The error is calculated based on MSE. The smaller value of MSE, it can be considered that the network architecture is better (Saputra et al., 2017). The value of MSE could be calculated using the following Equation (17).

$$MSE = \sum_{i=1}^{n} \frac{(y_i - \widehat{y})}{n} \quad (17)$$

Where y_i was the output label on the data set which had n quantities and $(\widehat{y})$ was the predictive value of the model.

2.2.2 Filling Imputation Result

The results of imputation missing data were put back to each attribute that has a missing data, so that the dataset had no more missing data or in another word the dataset was completed.

2.2.3 Analysis

At this stage, testing of the complete dataset is carried out by calculating the resulting performance of classification methods namely accuracy, specificity (SP), sensitivity (SN) and ROC. The methods that used for classification on complete dataset were Artificial Neural Network (ANN), Naïve Bayes, Support Vector Machine (SVM) and K-Nearest Neighbor (KNN). The completed data was divided using a percentage split, which was 80% as training data and 20% as testing data. In Naïve Bayes, it used Gaussian Naïve Bayes. Gaussian Naïve Bayes was used because the dataset consists of categories and continuous types data, while in SVM, it use the One Agains One method because the label on the data was binary (2 labels). In the KNN method, it used k = 3. This has been proven during trials by giving different k values (1-8). At the value of k = 3 the performance value of KNN gave the highest results and decreased for k>3. The ANN at this stage used an architecture similar to the ANN at the imputation stage of missing data using Figure 2.

3. RESULTS AND DISCUSSION

3.1 Attributes Deletion

The 14 attributes in the dataset were calculated for the presentation of each of the total number of existing data. The attributes that had missing data and their percentages could be seen in Table 1. From Table 1 it could be seen that the backup attribute had a very large missing data, namely Ca attribute and Thal attribute. Ca attribute had 98.97% missing data and Thal had 90.47% missing. It mean that there were a lot of lack of information that we got from these two attributes so that the attributes were dropped and deleted from the dataset. Meanwhile, 7 other attributes had missing data which was still quite low below 70%.

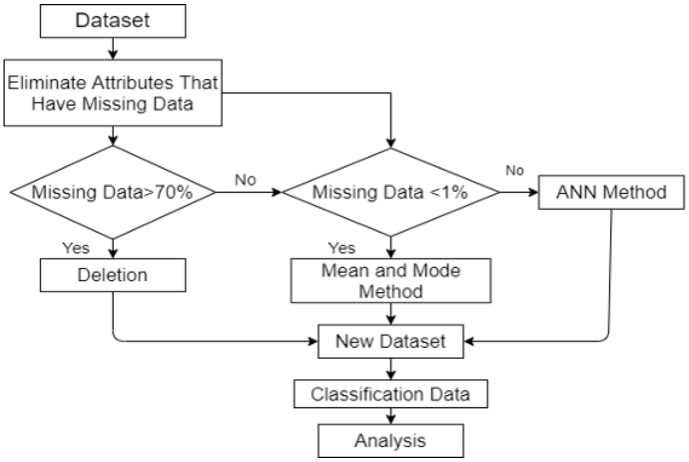

Figure 2. The Stages in Proposed Method

3.2 Imputation of Missing data

3.2.1 Attributes with less than 1% Missing Data

From Table 1, it was known that the very few attributes that had missing data (under 1%) were the Trestbps, Restecg, Thalac and Exang attributes so that to predict missing data in the data set, it was enough to use a simple method, namely the mean (for continuous or numeric data types) and the mode for data of type category. Trestbps and Thalac imputed missing data using the mean method because they each has only had 1 missing data of the total 249 available data. By mean XXX , the missing data for Trestbps was obtained $\widehat{x}$=132.58 and missing data for Thalac was $\widehat{x}$=139.13 . For attributes of type category such as restecg and exang, the mode value was be used to import the missing data. The Restecg attribute had 3 types of data, namely 0 for normal. 1 for ST-T wave abnormalities and 2 for left ventricular hypertrophy. In the Restecg attribute the value of the most data was owned by the normal label (0), then the missing data in the restecg attribute was filled with the label 0.The Exang attribute had two labels 0 for no and 1 for yes. 0 label was as many as 204 data and 1 label was as many as 89 data. The mode in Exang attribute was 0 label then the missing data was imputed with a 0 label.

3.2.2 Attributes with more than 1% Missing Data

For imputation on attributes that use ANN, the attribute of type numeric or continuous must be changed to categorical form because they would be used as classification labels on training stage. There were three attributes with missing data above

Table 1. Attributes, Values, Total and Percentage of Missing data of Heart Disease Dataset

Attributes	Values	Number of Missing Data	Percentage of Missing Data (%)
Age	In years	0	0
Sex	Male, Female	0	0
Chest Pain Type (cp)	Abnang, Angina, Asympt, Notang	0	0
Trestbps	94,0 - 200,0	1	0.34
Cholesterol (chol)	126,0 - 564,0	23	7.82
Fasting Blood Sugar (Fbs)	True, False	8	2.72
Resting ECG (restecg)	Norm, Hyp, Abn	1	0.34
Max Heart Rate (thalach)	99,0 - 103,0	1	0.34
Exercise Induced Angina (exang)	True, False	1	0.34
Oldpeak	0,0 - 6,2	0	0
Slope	Down, Flat, Up	190	64.4
Number of Vessels Colored (ca)	0,0 - 3,0	291	98.97
Thal	Normal, Rever, Fixed	266	90.47
Diagnosis of heart disease (Num)	Healthy, Sick	0	0

1%, namely attributes of fasting blood sugar (Fbs), cholesterol (Chol) and Slope so the ANN imputation were applied for the attributes. The Fbs attribute had missing data as much as 8 data and has two categories, namely patients who had blood sugar > 120 mg/dl with 2 categories, there were 1 to state the patient had blood sugar > 120 mg/dl and 0 to state the patient's blood sugar <120 mg/dl. Cholesterol (Chol) attribute was cholesterol of a patient that had in mg/dl and the attribute had 23 missing data. Chol had a continuous data type, which could be predicted using Artificial Neural Network (ANN). Data on Chol attribute should be converted into categories as label data. Chol attribute could be categorized into 3 labels very high(0) : > 200 mg / dl, High(1): 160-200 mg / dl, Normal (2): <160 mg / dl (Rahakbauw et al., 2016). For Slope attribute, it was used to represent the Slope of the ST segment at (peak). There were three labels, namely 1: up, 2: flat, 3: down. Missing data on the Slope attribute was 114 data. The imputation using ANN were applied for each attribute (Fbs, Chol and Slope) to guess missing data on the attribute.

The results of imputation using ANN were measured based on the values of accuracy, SP, SN and MSE generated by each attribute. The results of these measurements could be seen in Figure 3. In Figure 3 it could be seen that ANN worked very well in imputing missing data on the Fbs attribute with an accuracy of 95.46, SP of 97.69, SN of 62.5 and MSE of 0.06. While on the Chol attribute the measurement results were quite good with an accuracy of 75.75% and an MSE of 0.0765, but the specificity and sensitivity were still low below 60%. The results of the ANN imputation on the Slope attribute were quite low where the accuracy and sensitivity values were still below 60% but the resulting specificity was quite good at 74.62%. The result of attributes deletion and imputation method would applied into heart disease dataset. The data that has been obtained from the handling missing data was returned into initial dataset. Thus, the new dataset did not have missing data and total number of data used for classification was 294 patient data with 12 attributes where 11 attributes as input and 1 attribute as labels (Num attribute). The label Num contained 2 categories, 0 for healthy and 1 for sick. After the process of filling in the imputed data, the results are entered into the data set. The next process was to apply classification methods into heart disease dataset to see the effect of handling missing data with deletion and imputation methods on classification process.

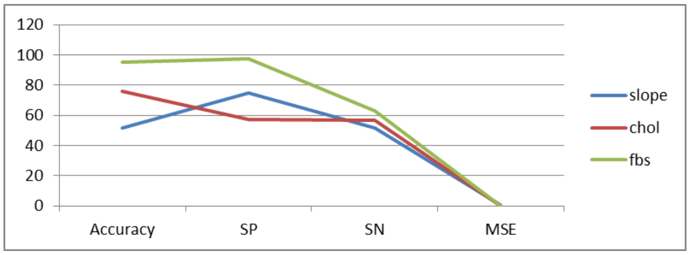

Figure 3. Comparison of Accuracy, Specificity (SP), Sensitvity (SN) and MSE for Imputation

3.3 Filling in Imputation Missing data to Dataset

The data that has been obtained from the results of technique and method of handling missing data on returned into the

Table 2. Comparison of Research Results on The Proposed Method with Previous Research

The Handle Missing Data Method	Data set	Prediction Method	SN	SP	Accuracy	ROC
A hybrid Bayesian network and tensor factorization approach (Vazifehdan et al., 2019)	Breast Cancer	C45	78.55	92.83	89.29	
Fuzzy C-Means and Particle Swarm Optimization (Salleh and Samat, 2017)	Framingham Heart	Decision Tree	-	0.846	86.3	0.83
Not handling missing data (Apurb et al., 2020)	Heart Disease	Naïve Bayes	82.3	0.845	81.97	-
ANN and T-test attribute selection (Poolsawad et al., 2012)	Heart Failure	ANN	76.8	0.803	-	-
ANN and nonlinear gain analysis attribute selection (Poolsawad et al., 2012)	Heart Failure	ANN	69.5	0.769	-	-
SVM Imputation (Al Khaldy and Kambhampati, 2016)	Cleveland Heart Failure	Random Forest	97.1	48.51	84.97	-
chaos firefly algorithm and rough sets based attribute reduction (Long et al., 2015)	Heart disease dataset	Interval Type-2 Fuzzy Logic System	84.9	93.3	88.3	
K-means Clustering (Purwar and Singh, 2015)	Wisconsin Breast Cancer	ANN	99.91	99.54	99.39	1
Artificial Neural Network (Hernández-Pereira et al., 2015)	Respiratory MIASOFT	Neural Network feed	69.63	84.44	79.03	80
Maximum Likelihood (Misir and Samanta., 2017)	Hungarian data set	Batch backpropagation	36.51	99.86	99.86	-
Fuzzy K-Mean Clustering (Rahman and Davis, 2012)	Cardiovascular	Decision tree	30	0.7	0.64	-
a combination of split data and FKmeans (Vangipuram et al., 2020)	Internet of Thing	Random forest	99	1	99.43	0
Refined Mean Substitution (Somasundaram and Nedunchezhian, 2012)	Breast Cancer	Fuzzy C-means	95.29	85.75	91.73	
Proposed method	Heart disease	Neural Network	94.2	94.2	94.23	0
Proposed Method	Heart disease	Naïve Bayes	87.76	87.6	87.5	0
Proposed Method	Heart disease	SVM	90	90	90	84
Proposed Method	Heart disease	KNN	90.48	0.905	90	90

initial dataset. Thus, the new dataset did not have missing value. After the process of filling in the data imputation results into the data set, the next process was to classify heart disease on the new dataset to see the effect of handling missing data on classification of heart disease. The total number of data used for classification was 294 patient data with 12 attributes whereas 11 attributes as input and 1 attribute as label namely Num attribute. The Num contained 2 categories, 0 for healthy and 1 for sick as a label in dataset.

3.4 Disccussion

In this study the data from the proposed method were applied to several kinds of classification methods. The results of the imputation of missing data using Deletion, Mean, Mode and ANN techniques were analyzed to see if they were able to improve performance on classification using the ANN, Naïve Bayes, SVM, and KNN methods. To analyze the effect of imputation on missing data. The new dataset was tested using several methods, namely the ANN, Naïve Bayes, Support Vector Machine (SVM) and KNN to see the effect of missing data imputation on the performance of classification methods. The results of classification testing showed that missing data imputation increased the accuracy, sensitivity (SN) and specificity (SP) for each method. Figure 4 showed that there was an increase for accuracy, SN and SP in the new dataset. From Figure 4, it could be seen that the highest increase was obtained in classification using the ANN.

The comparison of results classification on the data before and after handling missing data was not sufficient to evaluate the

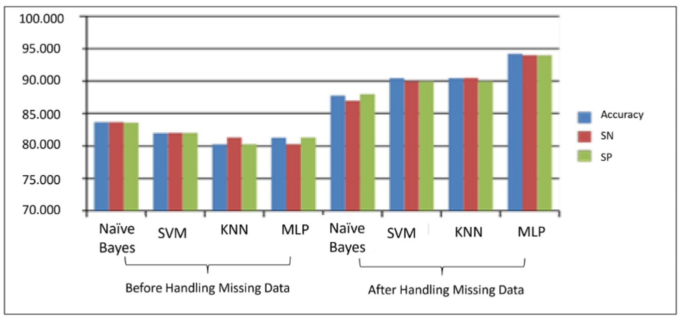

Figure 4. Comparison of Accuracy, Specificity (SP) and Sensitivity (SN) in Prediction of Heart Disease Dataset

success without comparing the results of the proposed method with other methods carried out in other studies. The results of this research were also compared with some other studies. Table 2 showed several studies using various techniques to deal with the problem of missing data by using various data sets, either heart disease by UCI or other datasets. From Table 2, the highest accuracy, SN and SP values were obtained by Purwar and Singh (2015), even the ROC value obtained at the highest value, but the missing data in the study was only 16 out of 569 data. The study by Vangipuram et al. (2020) also had the highest test results compared to other studies, but the total missing data was only 0.072% of 2050 total data from 12 attributes. The test results on Al Khaldy and Kambhampati (2016) also had higher accuracy than the accuracy of the proposed method, but the specificity value was lower than the specificity on the proposed method. The accuracy in the Subbalakshmi et al. (2011) was also very high, but unfortunately the SN value obtained was very small. From Table 2 it could be seen that the results of the accuracy of the proposed method were better than several other studies. The results of the sensitivity and specificity of the proposed method were also good, it was seen that the values obtained were higher and balanced than the other studies. Several previous studies did not show the sensitivity value obtained, some other studies also did not show the accuracy or ROC result that was successfully obtained in the study. From this comparison, it could be concluded that the proposed method was very suitable to be used for imputing missing data and could increase the accuracy, sensitivity, and specificity values which were very good above 85% by different classification methods.

4. CONCLUSIONS

The handling missing data in the study used 3 ways, first deletion technique for attributes that had missing data more than 70%. The second was the mean for numeric or continue data and mode imputation method for category data to handle missing data which the amount missing data was less or not more than 1%, namely the Trestbps and Thalac attributes by mean method, Restecg and Exang attributes by mode method. The third method was Artificial Neural Network (ANN) for attributes that had total missing data more than 1%, namely Fbs, chol, and Slope attributes. The resulting MSE shows that ANN was very good to impute missing data on the FBS and Chol attributes where the resulting MSE results were relatively small. But for imputing missing data on the Slope attribute, ANN was less suitable for use because the MSE result for the attribute was still relatively big. The performance showed the imputation results of the Fbs and Chol attributes by ANN better than other methods. It could be seen from the accuracy obtained above 75%. Unfortunately, the accuracy obtained by the Slope attribute was not very satisfying, it is only 52%. Although the performance results on the Slope attribute from both MSE and confusion matrix measurements were not satisfactory, the results of imputation carried out with the proposed method could improve and increase the accuracy, sensitivity (SN) and specificity (SP) of classification performance on the UCI heart disease data set. The classification performance of ANN, Naïve Bayes, SVM and KNN proofed that their performance has been increased when the methods worked on new dataset that has been handled the missing data problem compared theirs performance on original dataset. This research can be developed further by applying other imputation methods for missing data, especially for the slope attribute which has a low accuracy value and a large MSE.

5. ACKNOWLEDGEMENT

The authors thank to the deep learning discussion group for all the support on our research.

REFERENCES

Al Khaldy, M. and C. Kambhampati (2016). Performance analysis of various missing value imputation methods on heart failure dataset. *Proceedings of SAI Intelligent Systems Conference*; 415–425

Apurb, Rajdhan, S. Milan, A. Avi, and R. Dundigalla (2020). Heart Disease Prediction Using Machine Learning Classifiers. *International Journal of Advanced Science and Technology*, **29**(6); 1700–1707

Chauhan, H., V. Kumar, S. Pundir, and E. S. Pilli (2013). A comparative study of classification techniques for intrusion detection. *Proceedings - 2013 International Symposium on Computational and Business Intelligence*; 40–43

Choudhury, S. J. and N. R. Pal (2019). Imputation of missing data with neural networks for classification. *Knowledge-Based Systems*, **182**; 1–9

Crone, S. F. and S. Finlay (2012). Instance sampling in credit scoring: An empirical study of sample size and balancing. *International Journal of Forecasting*, **28**(1); 224–238

Desiani, A., S. Yahdin, A. Kartikasari, and I. Irmeilyana (2021). Handling the imbalanced data with missing value elimination SMOTE in the classification of the relevance education background with graduates employment. *IAES International Journal of Artificial Intelligence*, **10**(2); 346–354

Eekhout, I., H. C. de Vet, J. W. Twisk, J. P. Brand, M. R.

de Boer, and M. W. Heymans (2014). Missing data in a multi-item instrument were best handled by multiple imputation at the item score level. *Journal of Clinical Epidemiology*, **67**(3); 335–342

El-Bialy, R., M. A. Salamay, O. H. Karam, and M. E. Khalifa (2015). Feature analysis of coronary artery heart disease data sets. *Procedia Computer Science*, **65**; 459–468

García, S., J. Luengo, and F. Herrera (2015). *Data Preprocessing in Data Mining*. Springer

Hernández-Pereira, E. M., D. Álvarez-Estévez, and V. Moret-Bonillo (2015). Automatic classification of respiratory patterns involving missing data imputation techniques. *Biosystems Engineering*, **138**; 65–76

Huang, J., J. W. Keung, F. Sarro, Y. F. Li, Y. T. Yu, W. Chan, and H. Sun (2017). Cross-validation based K nearest neighbor imputation for software quality datasets: an empirical study. *Journal of Systems and Software*, **132**; 226–252

Jasoni, A. and W. Steinbrunn (2013). *Heart Disease Data Set*. UCI Machine Learning Repository

Jing, X. Y., F. Qi, F. Wu, and B. Xu (2016). Missing data imputation based on low-rank recovery and semi-supervised regression for software effort estimation. *Proceedings - International Conference on Software Engineering*; 607–618

Karim, M. N., C. M. Reid, L. Tran, A. Cochrane, and B. Billah (2017). Missing value imputation improves mortality risk prediction following cardiac surgery: an investigation of an Australian patient cohort. *Heart, Lung and Circulation*, **26**(3); 301–308

Lan, Q., X. Xu, H. Ma, and G. Li (2020). Multivariable data imputation for the analysis of incomplete credit data. *Expert Systems with Applications*, **141**; 1–12

Long, N. C., P. Meesad, and H. Unger (2015). A highly accurate firefly based algorithm for heart disease prediction. *Expert Systems with Applications*, **42**(21); 8221–8231

Luengo, J., A. Fernández, S. García, and F. Herrera (2011). Addressing data complexity for imbalanced data sets: analysis of SMOTE-based oversampling and evolutionary undersampling. *Soft Computing*, **15**(10); 1909–1936

Malarvizhi, M. and A. Thanamani (2012). K-NN classifier performs better than K-means clustering in missing value imputation. *IOSR Journal of Computer Engineering*, **6**(5); 12–15

Manimekalai, K. and A. Kavitha (2018). Missing value imputation and normalization techniques in myocardial infarction. *ICTACT Journal on Soft Computing*, **8**(3); 1655–1662

Mehrotra, D. V., F. Liu, and T. Permutt (2017). Missing data in clinical trials: control-based mean imputation and sensitivity analysis. *Pharmaceutical Statistics*, **16**(5); 378–392

Misir, R. and R. K. Samanta. (2017). A Study on Performance of UCI Hungarian Dataset Using Missing Value Management Techniques. *International Journal of Computer Sciences and Engineering*, **5**(3); 40–44

Moorthy, K., M. Saberi Mohamad, and S. Deris (2014). A review on missing value imputation algorithms for microarray gene expression data. *Current Bioinformatics*, **9**(1); 18–22

Nishanth, K. J. and V. Ravi (2016). Probabilistic neural network based categorical data imputation. *Neurocomputing*, **218**; 17–25

Pedersen, A. B., E. M. Mikkelsen, D. Cronin-Fenton, N. R. Kristensen, T. M. Pham, L. Pedersen, and I. Petersen (2017). Missing data and multiple imputation in clinical epidemiological research. *Clinical Epidemiology*, **9**; 157–166

Poolsawad, N., L. Moore, C. Kambhampati, and J. G. Cleland (2012). Handling missing values in data mining-A case study of heart failure dataset. *Proceedings - International Conference on Fuzzy Systems and Knowledge Discovery*; 2934–2938

Purwar, A. and S. K. Singh (2015). Hybrid prediction model with missing value imputation for medical data. *Expert Systems with Applications*, **42**(13); 5621–5631

Rahakbauw, D., F. K. Lembang, and Y. Taihuttu (2016). Analisis dan Prediksi Penyakit Jantung Koroner di Kota Ambon Menggunakan Jaringan Saraf Tiruan. *Barekeng: Jurnal Ilmu Matematika dan Terapan*, **10**(2); 97–105

Rahman, M. M. and D. N. Davis (2012). Fuzzy unordered rules induction algorithm used as missing value imputation methods for k-mean clustering on real cardiovascular data. *Lecture Notes in Engineering and Computer Science*, **2197**(1); 391

Resti, Y., E. S. Kresnawati, N. R. Dewi, N. Eliyati, et al. (2021). Diagnosis of Diabetes Mellitus in Women of Reproductive Age using the Prediction Methods of Naive Bayes, Discriminant Analysis, and Logistic Regression. *Science and Technology Indonesia*, **6**(2); 96–104

Salleh, M. N. M. and N. A. Samat (2017). FCMPSO: An imputation for missing data features in heart disease classification. *IOP Conference Series: Materials Science and Engineering*, **226**(1); 1–8

Saputra, W., M. Zarlis, R. W. Sembiring, D. Hartama, et al. (2017). Analysis resilient algorithm on artificial neural network backpropagation. *Journal of Physics: Conference Series*, **930**(1); 12035

Shah, S. M. S., S. Batool, I. Khan, M. U. Ashraf, S. H. Abbas, and S. A. Hussain (2017). Feature extraction through parallel probabilistic principal component analysis for heart disease diagnosis. *Physica A: Statistical Mechanics and its Applications*, **482**; 796–807

Shilaskar, S. and A. Ghatol (2013). Feature selection for medical diagnosis: Evaluation for cardiovascular diseases. *Expert Systems with Applications*, **40**(10); 4146–4153

Silva-Ramírez, E.-L., R. Pino-Mejías, M. López-Coello, and M.-D. Cubiles-de-la Vega (2011). Missing value imputation on missing completely at random data using multilayer perceptrons. *Neural Networks*, **24**(1); 121–129

Somasundaram, R. and R. Nedunchezhian (2012). Missing value imputation using refined mean substitution. *International Journal of Computer Science Issues*, **9**(4); 306–313

Statlog (2004). *Heart Data Set*. UCI Machine Learning Repository

Stekhoven, D. J. and P. Bühlmann (2012). MissForest-non-parametric missing value imputation for mixed-type data.

Bioinformatics, **28**(1); 112–118

Stewart, J., G. Manmathan, and P. Wilkinson (2017). *Primary Prevention of Cardiovascular Disease: A Review of Contemporary Guidance and Literature*. JRSM Cardiovascular Disease

Subbalakshmi, G., K. Ramesh, and M. C. Rao (2011). Decision support in heart disease prediction system using naive bayes. *Indian Journal of Computer Science and Engineering*, **2**(2); 170–176

Ting, P. Y., T. Wada, Y. L. Chiu, M. T. Sun, K. Sakai, W. S. Ku, A. A. K. Jeng, and J. S. Hwu (2020). Freeway Travel Time Prediction Using Deep Hybrid Model–Taking Sun Yat-Sen Freeway as an Example. *IEEE Transactions on Vehicular Technology*, **69**(8); 8257–8266

Tsai, C. F., M. L. Li, and W. C. Lin (2018). A class center based approach for missing value imputation. *Knowledge-Based Systems*, **151**; 124–135

Vangipuram, R., R. K. Gunupudi, V. K. Puligadda, and J. Vinjamuri (2020). A machine learning approach for imputation and anomaly detection in IoT environment. *Expert Systems*, **37**(5); 1–16

Vazifehdan, M., M. H. Moattar, and M. Jalali (2019). A hybrid Bayesian network and tensor factorization approach for missing value imputation to improve breast cancer recurrence prediction. *Journal of King Saud University-Computer and Information Sciences*, **31**(2); 175–184

Zriqat, I. A., A. M. Altamimi, and M. Azzeh (2016). A comparative study for predicting heart diseases using data mining classification methods. *International Journal of Computer Science and Information Security*, **14**(12); 868–879