Science and Technology Indonesia
e-ISSN:2580-4391 p-ISSN:2580-4405
Vol. 10, No. 4, October 2025

Research Paper

MICE and ADASYN for Missing Data Imputation and Imbalanced Data Handling on Heart Disease Classification

Anita Desiani[1*], Deshinta Arrova Dewi[2], Ali Amran[1], Ananda Pratiwi[1], Yuli andriani[1], Endro Setyo Cahyono[1]

[1]*Department of Mathematics, Faculty of Mathematics and Natural Science, Universitas Sriwijaya, Ogan Ilir, South Sumatera, 30622, Indonesia*
[2]*Department of Data Science, Faculty of Data Science, Inti International University, Nilai, Negeri Sembilan, 71800, Malaysia*
***Corresponding author**: anita_desiani@unsri.ac.id

Abstract
The quality of data is determined by several things, namely the completeness and balance data. The heart disease dataset from the University of California, Irvine (UCI) has missing and imbalanced data, which if it is not handled, can lead to a lack of accuracy in the prediction model and errors in interpreting the data. To overcome missing data, several methods can be used, one of which is data imputation. Attributes with missing data of 5% or less are handled using imputation methods such as Mean, Mode, and MICE. Attributes with numeric types are handled by Mean. Attributes with categorical types are imputed by Mode. Attributes with more than 5% missing data are imputed using the MICE method. Imbalanced data can be handled by applying an oversampling method using the Adaptive Synthetic Sampling Approach (ADASYN). The effect of imputing missing data and addressing class imbalance on heart disease classification performance was tested using Random Forest, Support Vector Machine (SVM), and Multi-Layer Perceptron (MLP) algorithms. After handling missing values and data imbalance, improvements were observed in the classification results. The accuracy, precision, recall, and F1-score showed excellent performance, above 90% on several classification methods. The results indicate that handling missing and imbalanced data through Mean, Mode, MICE, and ADASYN positively impacts the performance of classifiers on the UCI heart disease dataset.

Keywords
Imbalanced Data, Heart Disease, Health Risk, Missing Data, Public Health

Received: 3 February 2025, Accepted: 3 July 2025
https://doi.org/10.26554/sti.2025.10.4.1020-1030

1. INTRODUCTION

Automatic classification using machine learning has been widely developed today, and one of the benefits is for the classification of heart disease (Bayuaji et al., 2024). To get robust classification results, the data should have good quality (Tan, 2021). However, the numerous low-quality data can affect the classification performance (Osisanwo et al., 2017). Some common problems that affect data quality are missing data and imbalanced data. These issues are also present in the UCI heart disease dataset, which contains both missing values and class imbalance.

Missing data is a problem that refers to the loss of a value content in an attribute or several attributes, causing incomplete data (Liu et al., 2016). Missing data is usually denoted by NaNs, blank spaces, undefined, question marks, or zeros (Hasan et al., 2021). The causes of missing data are errors in writing data, errors in data collection, and inadequate tools used (Pedersen et al., 2017). In contrast, missing data issues can occur in more than one attribute. The imbalanced data problem typically involves one attribute, namely the class label used to categorize the data into multiple classes (Ali et al., 2019). In the UCI heart disease dataset, the data is labeled into two classes, namely the healthy class and the unhealthy class. The missing data and imbalanced data can reduce classification performance, so it need to be handled before applying the classification method (Chen et al., 2017; Desiani et al., 2021b).

Missing data can be handled using two different imputation methods: single imputation and multiple imputation (Pedersen et al., 2017). Single imputation is used to find a single value to replace the contents of all data on one attribute with that one value (Pauzi et al., 2021). Some techniques in single imputation are the Mean and Mode (Lee et al., 2022). Mean imputation works by calculating the average of an attribute with missing values and replacing the missing entries with that mean value (Desiani et al., 2021a). Mean imputation is used for numeric attributes, while for nominal attributes it uses Mode imputation (Wu et al., 2019). Mode imputation is a technique that replaces missing values with the most frequently occurring value for an

attribute (Misir and Samanta, 2017). Desiani et al. (2021a) used missing data imputation by combining Mean, Mode imputation, and ANN to classify heart disease using Multi-Layer Perceptron (MLP). The study indicates that the application of Mean, mode, and ANN can increase accuracy by 13%; however, the resulting accuracy is still below 95%. Seliem (2022) applied Mean imputation to the classification of heart disease using Naïve Bayes. The application of Mean imputation can increase accuracy by 3% however, the accuracy obtained is only 85%.

According to Pauzi et al. (2021), Gabr et al. (2023), and Lee et al. (2022), single imputation should be used on missing data less than 5%. If a single imputation is used on the attribute that has more than 5% missing data, it can cause bias in the values used, which can lead to incorrect interpretation of the data. For attributes with more than 5% missing data, the use of multiple imputation is recommended (Desiani et al., 2021a; Lee et al., 2022). Multiple imputation overcomes missing data by generating several plausible values for each missing entry, forming multiple complete datasets for analysis (Pedersen et al., 2017). The use of multiple imputed values helps reduce the risk of bias in statistical estimates (Austin et al., 2021). Multiple Imputation by Chained Equations (MICE) is one of the techniques commonly used in multiple imputation (Jäger et al., 2021). MICE is used to impute missing data with various values using a regression model approach (Khan and Hoque, 2020). Mera-Gaona et al. (2021) applied MICE to a heart disease dataset by increasing the accuracy by 8%, however, the accuracy obtained was only 85.8%. Wu et al. (2019) applied MICE to breast cancer datasets and increased accuracy by 4% in Random Forest classification. From these studies, it can be seen that MICE imputation can improve classification accuracy on a dataset. However, these studies have not dealt with imbalanced data.

Classification quality is also affected by imbalanced data. Imbalanced data can reduce classification performance, especially in learning patterns in minority classes (Ebenuwa et al., 2019). Undersampling and oversampling are techniques that can be used for imbalanced data handling (Thabtah et al., 2020). Undersampling reduces the majority class by removing some of its data, aiming to balance it with the minority class (Wongvorachan et al., 2023). Oversampling is the opposite of undersampling, where additional data is randomly generated for the minority class to make its size comparable to that of the majority class (Ali et al., 2019).

Some studies suggest using oversampling rather than undersampling for imbalanced data handling. Oversampling has a greater chance of losing important information due to the deletion of some data. This problem can reduce classification performance (Aditsania and Saonard, 2017; Ali et al., 2019; Douzas et al., 2018; Ramadhan, 2021). The Adaptive Synthetic Sampling Approach (ADASYN) is an oversampling method used to address class imbalance by generating synthetic samples for the minority class. The ADASYN works by using the K-nearest neighbors of minority instances to create new samples, allowing the minority class size to approach that of the majority class (Guan et al., 2023). Ramadhan (2021) applied ADASYN to diabetes mellitus disease datasets with Support Vector Machine (SVM) classification, increasing accuracy by 4%. Kurniawati et al. (2018) applied ADASYN to the cervical cancer dataset using the K-Nearest Neighbor (KNN) classification, showing that there was an increase in accuracy of 15.64%. However, these studies have not involved missing data imputation.

This study combines missing data imputation and imbalanced data handling on the UCI heart disease dataset. This study performs single imputation using the Mean and Mode for attributes where missing data is less than or equal to 5%. For numeric attributes, the study used Mean imputation, while for nominal attributes, the study used Mode imputation. For data imputation on attributes with more than 5% missing data, both numeric and nominal, this study uses the MICE method. For imbalanced data Handling, the study applies the ADASYN method. The ADASYN is used to increase the number of samples in the minority class of the label attribute until it approximates the majority class, resulting in a more balanced class distribution. The dataset generated from missing data imputation and imbalanced data handling is used in heart disease qualification using several classification methods. The classification methods applied are Random Forest, Support Vector Machine (SVM), and Multi-Layer Perceptron (MLP). The results of this study are evaluated using accuracy, precision, recall, and F1-score obtained from each classification method. These performance results are to see how far missing data imputation and imbalanced data handling can affect the performance results of heart disease classification.

2. EXPERIMENTAL SECTION

2.1 Materials

This study uses secondary data consisting of records from heart disease patients, which can be accessed through the page (Kaggle dataset). This dataset contains 10 diagnostic attributes and one target attribute (*num*) that determines whether a patient has heart disease. The attributes used to diagnose heart disease are patient age (*age*), patient gender (*sex*), chest pain type (*cp*), blood pressure (*trestbps*), cholesterol (*chol*), blood sugar (*fbs*), electrocardiography results (*restecg*), maximum heart rate (*thalach*), chest pain due to exercise (*exang*), depression caused by exercise relative to rest (*oldpeak*), and num is the attribute that has a diagnosis of heart disease, a class containing healthy and sick categories. The explanation of each attribute is shown in Table 1.

2.2 Method

The workflow of the proposed method in this study is shown in Figure 1. Based on Figure 1, this study has several stages, namely missing data imputation using Mode for attributes that have a missing data rate of 5% or less, and MICE imputation for attributes that have a missing data more than 5%. The new dataset resulting from imputation will be treated for imbalanced data using the ADASYN method. The new dataset will be

Table 1. The Attributes and Information of The Dataset

Attribute Name	Description	Amount of Missing Data	Missing Data (%)	Data Type
Age	Year	-	-	Numerical
Sex	0 = female, 1 = male	-	-	Nominal
Chest pain (Cp)	1 = typical angina, 2 = atypical angina, 3 = non-anginal pain, 4 = asymptomatic	-	-	Nominal
Blood pressure at rest (*trestbps*)	mmHg	33	7.22	Numerical
a	mg/dl	26	5	Numerical
Blood sugar (fbs)	Blood sugar >120mg/dl, 1 = correct, 0 = wrong	54	11.82	Nominal
Electrocardiograph result (*restecg*)	0 = normal, 1 = has ST – T wave, 2 = indicates left ventricular hypertrophy	2	0.44	Nominal
Heart rate (*thalach*)		33	7.22	Numerical
Exercise angina (*exang*)	0 = no, 1 = yes	33	7.22	Nominal
Exercise-induced (*oldpeak*)		36	7.87	Numerical
Diagnosis of heart disease (*num*)	0 = healthy, 1 = sick	-	-	Nominal

examined for its effect on classification performance. The classification methods applied in this study are Random Forest, Support Vector Machine (SVM), and Multi-Layer Perceptron (MLP).

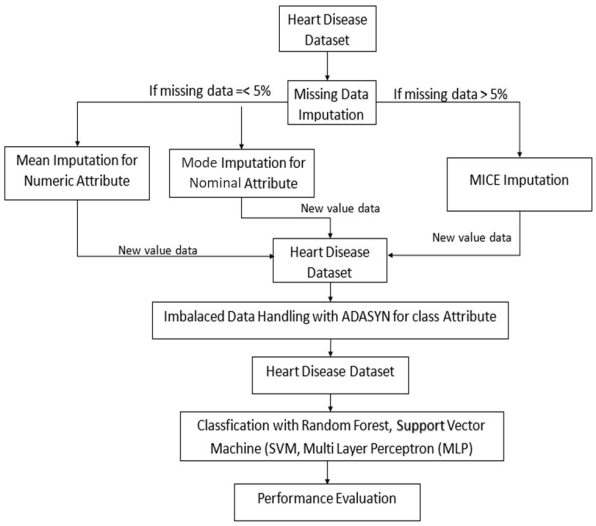

Figure 1. The Stages in the Study for MICE and ADASYN on the Heart Disease Dataset

2.3 Missing Data Imputation

Mean imputation is used on data that has less than or equal to 5% missing data by using the average of these attributes for numeric attributes. Mean imputation uses Equation 1.

$$\bar{A} = \frac{1}{n}\sum_{i=1}^{n} A_{ij} \tag{1}$$

where $\bar{A}$ represents the mean (average) value of the data. Nominal attributes with less than 5% missing data are imputed using the Mode. For the attributes that have more than 5% missing data, both numeric and nominal attributes use MICE Imputation. The steps of the MICE method are (Mera-Gaona et al., 2021):

1. Determine the base imputation for each missing value in the dataset by using the Mean value obtained from equation 1, and after that, the missing values for one feature are rearranged.
2. Build a prediction model using a regression equation where the missing value is used as the dependent variable, while other attributes are used as independent variables. The form of the multiple linear regression equation is shown in Equation 2.

$$\hat{T} = \theta_0 + \theta_1 A_1 + \theta_2 A_2 + \cdots + \theta_n A_n \tag{2}$$

To get θ, form the normal equation as in Equation 3.

$$\theta = \left(A^T A\right)^{-1} A^T T \tag{3}$$

The value of A is obtained from Equation 4.

$$A = \begin{bmatrix} 1 & A_{11} & A_{12} & \cdots & A_{1n} \\ 1 & A_{21} & A_{21} & \cdots & A_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & A_{m1} & A_{m2} & \cdots & A_{mn} \end{bmatrix} \quad (4)$$

The vector T contains the values of the dependent variable, as in Equation 5.

$$T = \begin{bmatrix} T_1 \\ T_2 \\ \vdots \\ T_m \end{bmatrix} \quad (5)$$

The coefficients of the vector θ are obtained from Equation 6.

$$\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \vdots \\ \theta_n \end{bmatrix} \quad (6)$$

The steps to form a Regression Equation are:
a. Calculate $A^T A$ as in Equation 7.

$$A^T A = \begin{bmatrix} m & \sum_{i=1}^{m} A_{i1} & \sum_{i=1}^{m} A_{i2} & \cdots & \sum_{i=1}^{m} A_{in} \\ \sum_{i=1}^{m} A_{i1} & \sum_{i=1}^{m} A_{i1}^2 & \sum_{i=1}^{m} A_{i1}A_{i2} & \cdots & \sum_{i=1}^{m} A_{i1}A_{in} \\ \sum_{i=1}^{m} A_{i2} & \sum_{i=1}^{m} A_{i2}A_{i1} & \sum_{i=1}^{m} A_{i2}^2 & \cdots & \sum_{i=1}^{m} A_{i2}A_{in} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^{m} A_{in} & \sum_{i=1}^{m} A_{in}A_{i1} & \sum_{i=1}^{m} A_{in}A_{i2} & \cdots & \sum_{i=1}^{m} A_{in}^2 \end{bmatrix} \quad (7)$$

b. Calculate $A^T T$ as in Equation 8.

$$A^T T = \begin{bmatrix} \sum_{i=1}^{m} T_i \\ \sum_{i=1}^{m} A_{i1}T_i \\ \sum_{i=1}^{m} A_{i2}T_i \\ \vdots \\ \sum_{i=1}^{m} A_{in}T_i \end{bmatrix} \quad (8)$$

where $\hat{T}$ represents the predicted value of the dependent variable, A denotes the matrix of the independent variable, T denotes the dependent variable, θ denotes the coefficient vector that needs to be estimated, θ_0 intercept, and θ_1 denotes the regression coefficient for A.

3. Replace the missing values using the predictions calculated by the model. Repeat each step until all features that have missing values are filled or, in other words, a complete dataset is obtained.
4. Evaluate the results of MICE imputation by calculating the Mean Absolute Error (MAE) value. MAE is used to see the quality of imputation results by calculating between the imputed value and the true value (Mera-Gaona et al., 2021). The MAE value can be calculated with the Equation 9.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|\hat{T}_i - T_i\right| \quad (9)$$

where $\hat{T}_i$ denotes the i-th predicted values, T_i represents the i-th actual data, and n is the number of data.

Table 2. MAE Results Based on Iteration on MICE Method for Missing Data Imputation

Iteration	*trestbps*	*thalach*	MAE
1	135.57	143.89	6.93
2	135.4178	140.852	7.8964
3	137.4862	141.6579	1.9091
4	138.2092	142.08	0.4107
5	138.485	142.2585	0.1394
6	138.4958	142.2664	0.0095
7	138.4958	142.2664	0.0093
8	138.4877	142.2679	0.0019
9	138.498	142.2682	0.0003
10	138.498	142.2682	0.00006

2.4 Imbalanced Data Handling

Imbalanced data handling with the ADASYN method works with the following steps (Ramadhan, 2021):

1. Calculate the amount of synthesized data to be created using Equation 10.

$$G = (m_l - m_s) \times \theta \quad (10)$$

where G is the number of data syntheses to be made, and β is an equilibrium factor randomly selected from the range [0, 1].
2. Calculate the i-th ratio with Equation 11

$$r_i = \frac{\Delta_i}{K} \quad (11)$$

where r_i is the i-th ratio, Δ_i represents the number of majority class samples among the K nearest neighbors, and K denotes the number of nearest neighbors.

3. Normalize the ratio to get the density distribution using Equation 12.

$$\hat{r}_i = \frac{r_i}{\sum_{i=1}^{m_s} r_i} \quad (12)$$

where r_i is the density distribution, and the value i, which starts from 1 to as many as the amount of data in the minority class.

4. Determine the amount of synthesized data to generate for each minority data using Equation 13.

$$g_i = \hat{r}_i \times G \quad (13)$$

where g_i denotes the amount of synthesized data from the minority class, r_i senotes the density distribution, and G denotes the amount of synthesized data to be created.

5. Create new synthesis data using Equation 14.

$$s_i = a_i + (a_u - a_i) \times \lambda \quad (14)$$

where s_i denotes new synthesized data, x_i denotes the minority class data included in the loop, a_u is the data selected based on the nearest neighbor, λ indicates the ratio between the amount of synthesized data to be created and the actual number of minority class data filled with a random number from 0 to 1.

2.5 Analysis and Evaluation

The testing process is conducted using the entire dataset. Testing in this study consists of statistical testing and classification performance testing. Statistical testing includes the Shapiro-Wilk Test, Komogorov-Smirnov Test, and Levene's Test. For classification performance testing, the study uses classification methods, namely Random Forest, Support Vector Machine (SVM), and Multi-Layer Perceptron (MLP). To evaluate the classification performance, the dataset is split into training and testing data. In this study, the dataset is divided into 70% training data and 30% testing data. It means the number of training data becomes 320 data while the testing data becomes 137 data. To evaluate classification performance after handling missing and imbalanced data, this study uses evaluation metrics such as accuracy, precision, recall, and F1-score. Accuracy represents the percentage of data correctly classified during the testing phase across the entire dataset (De Diego et al., 2022). Precision is the ratio of correctly predicted positive observations to the total predicted positives, while recall is the ratio of correctly predicted positives to all actual positives (Desiani et al., 2024). F1-score is the harmonic mean of precision and recall which is used to assess the performance of a classification model in a balanced manner (Gabr et al., 2023). The metrics of accuracy, precision, recall, and F1-score were computed based on Equations 15, 16, 17, and 18 (Salamah et al., 2024).

$$\text{Accuracy} = \frac{P_{\text{True}} + N_{\text{True}}}{P_{\text{True}} + N_{\text{False}} + P_{\text{False}} + N_{\text{True}}} \times 100\% \quad (15)$$

$$\text{Precision} = \frac{P_{\text{True}}}{P_{\text{False}} + P_{\text{True}}} \times 100\% \quad (16)$$

$$\text{Recall} = \frac{P_{\text{True}}}{P_{\text{True}} + N_{\text{False}}} \times 100\% \quad (17)$$

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \times 100\% \quad (18)$$

where, P_{True} denotes the correctly predicted positive cases (True Positive). P_{False} denotes incorrectly predicted positive cases (False Positive). N_{True} denotes the correctly predicted negative cases (True Negative). N_{False} denotes incorrectly predicted negative cases (False Negative).

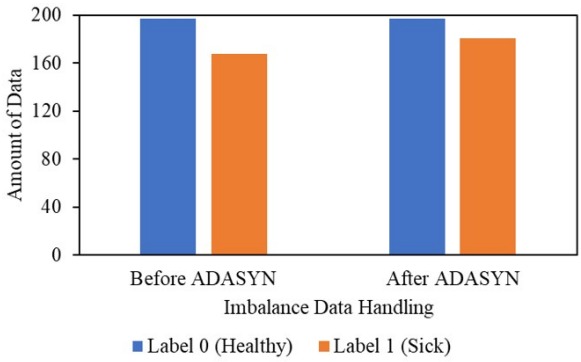

Figure 2. The Comparison of the Amount of Data Before and After the ADASYN Process on Imbalanced Data Handling

3. RESULTS AND DISCUSSION

3.1 Missing Data Imputation

3.1.1 Missing Data Imputation on Features with Missing Data Less Than or Equal to 5%

Based on Table 1, the attributes restecg and chol are identified as having missing data of 5% or less. The restecg is a nominal attribute, and it is sufficient to use Mode imputation. There are three classes in the restecg attribute: class 0 denotes normal results, class 1 indicates wave abnormalities, and class 2 reflects left ventricular hypertrophy. Most entries in the restecg attribute belong to class 0, so missing values are replaced with 0. The chol attribute is a numeric variable, so it is sufficient to use Mean imputation. The average (mean) result of the data in the chol attribute is 200.05, so the missing data in each row in the chol attribute is replaced with the value.

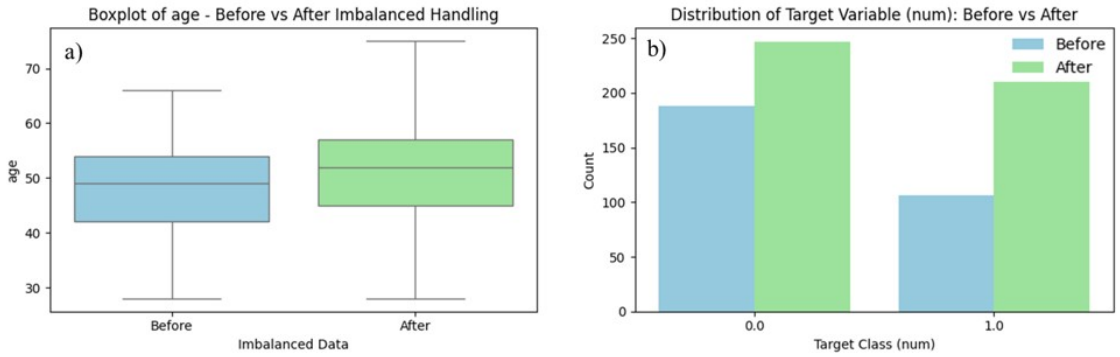

Figure 3. Comparison of Data Distributions Before and After Handling Imbalanced Data (a), Age Variable Num Variable (b)

Table 3. Statistical Test Results before and after Imputation Missing and Handling Data Imbalance

Variable	Shapiro-Wilk Test		Kolmogorov-Smirnov Test	Levene's Test
	p-Value (Before Imputation)	*p*-Value (After Imputation)	*p*-Value	*p*-Value
Age	0.01	0.01	1.00	1.00
Sex	0.00	0.00	1.00	1.00
Cp	0.00	0.00	1.00	1.00
trestbps	0.00	0.00	0.95	1.00
Chol	0.00	0.00	1.00	1.00
Fbs	0.00	0.00	0.15	1.00
restecg	0.00	0.00	1.00	1.00
thalach	0.00	0.01	1.00	1.00
exang	0.00	0.00	0.88	1.00
oldpeak	0.00	0.00	0.76	1.00
Num	0.00	0.00	1.00	1.00

3.1.2 Missing Data Imputation on Features with Missing Data More Than 5%

Six attributes have more than 5% missing data, namely *trestbps*, *fbs*, *thalach*, *exang*, and *oldpeak*. The *trestbps* attribute has 33 missing data. The *fbs* attribute has the most missing data, 54 data with a percentage of 11.82%. The *thalach* and *exang* attributes have 33 missing data. For missing data on the *oldpeak* attribute, as much as 36 data.

The missing values will be replaced using the MICE imputation method, where at each iteration a regression will be run between each attribute that has missing data. Each missing value to be filled will be treated as a dependent variable, while the other values will be treated as independent variables. So that the equation will be $\hat{Y} = \theta_0 + \theta_1 A_1 + \theta_2 A_2 + \ldots + \theta_6 A_6$. $A_1, \ldots, A_6$ will be filled with independent variables, and a replacement value for the missing value will be generated. In iteration 1, the value of *trestbps* (x_1) is filled with 135.57 and 143.89 for *thalach*, but the MAE results are still not convergent, so calculations need to be carried out for the next iteration using a regression approach until the resulting MAE approaches zero. In Table 2, it can be seen that in the 10th iteration, the MICE results for both *trestbps* and *thalach* have converged. The *trestbps* value produced is 138.4958 and the *thalach* value produced is 142.2682 with MAE results approaching zero and converging.

Based on Table 2, it shows that at iteration 10, the MICE imputation value no longer changes. The value difference between the last iteration and the previous iteration is 0.00024 and close to 0. So, it can be concluded that the MICE iteration stops and has reached a convergent value. In this case, if the MAE value has converged, it means that the quality of the imputation results is good and stable. The total amount of data from the missing data imputation results is 457 data. The obtained data is returned to the initial dataset. Thus, there is no more missing data in the new dataset. After filling in the missing data, the next step is handling the unbalanced data.

3.2 Imbalanced Data Handling

In the UCI heart disease dataset, the num attribute has two data classes, namely 0 for the healthy category and 1 for the sick category. Class of 0 has 247 data, while class of 1 has 210 data. Unbalanced data can result in classification performance only focusing on the majority of data, so it needs to be handled using the ADASYN method. In the handling process using ADASYN, the data will be randomly divided into training data. The training dataset shows a reduction in class 0 from 247 to 197 data, and in class 1 from 210 to 168 data. After ADASYN,

the amount of data on the minority class has increased, and it approaches the amount of data on the majority class. To address the imbalance, class 1, as the minority class is increased from 168 to 181 data. The result can be observed in Figure 1.

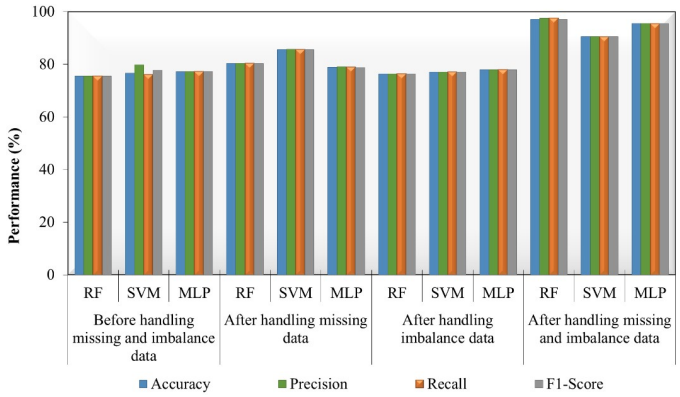

Figure 4. The Comparison Performance of Classification Methods with Handling Missing Data Imputation and Imbalanced Data Handling

Figure 2 shows an increase in class 1 after applying ADASYN. The amount of data in class 1 (minority class) approximates that of data in class 0 (majority class). The dataset has been balanced, and classification on the heart disease dataset can be executed. The new synthetic data generated by ADASYN for handling imbalance is added to the original dataset. The resulting new dataset has no missing or imbalanced data issues. After completing the missing data imputation and handling data imbalance, classification is performed on the updated dataset to evaluate the impact of these processes on heart disease classification. The amount of data used for classification is 378 data, with 10 attributes as input and 1 attribute for classes. The classification process used 378 data samples consisting of 10 input attributes and 1 target class attribute.

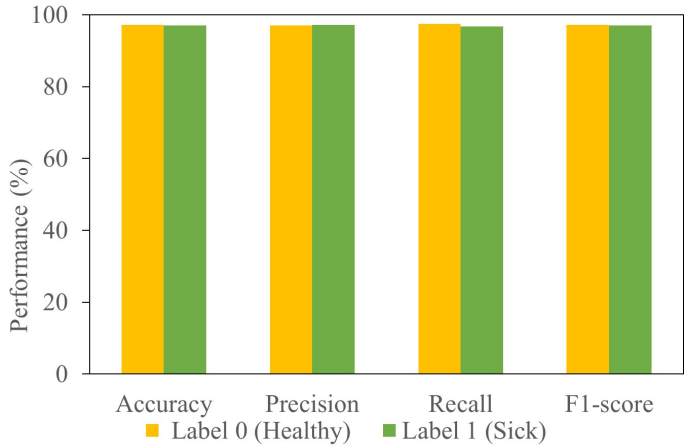

Figure 5. The Comparison Graph of Accuracy, Precision, Recall, and F1-Score on Random Forest (RF) for Each Class

3.3 Statistical Test on New Dataset

The distribution of data before and after missing data imputation and imbalance handling can be visualized using boxplots and distribution plots. If there is no significant difference, it means that the characteristics of each data are well maintained. Figure 3 is an example of a boxplot for the age variable and the distribution that occurs in the class variable.

Figure 3a shows that the age distribution is relatively stable after handling missing data and imbalanced data handling, marked by an increase in the median and maximum range of the data, which means that the age characteristics are maintained even though the data is balanced. Figure 3b demonstrates that the method proposed in this study effectively balances the class proportions in the target variable, so there is no longer a certain class dominance that can affect the modeling results.

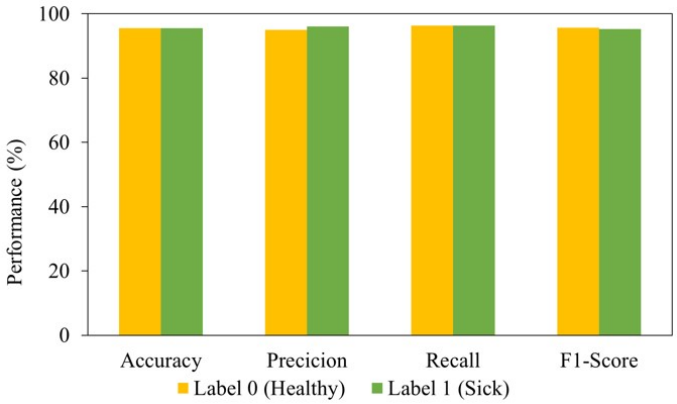

Figure 6. The Comparison Graph of Accuracy, Precision, Recall, and F1-Score on Multilayer Perceptron (MLP) for Each Class

To evaluate the impact of applying missing data imputation and data imbalance handling, the new dataset is compared with the old dataset and tested using statistical tests, namely the Shapiro-Wilk Test, Komogorov-Smirnov Test, and Levene's Test. These tests are used to see the character and distribution of the new dataset, whether it is the same or produces a different distribution and character from the original or old dataset. To evaluate the impact of missing data imputation and imbalance handling, a statistical test was performed by comparing the results before and after these processes. The test results are shown in Table 3.

Based on the results of statistical tests conducted using the Shapiro-Wilk, Kolmogorov-Smirnov, and Levene tests, it was found that the Shapiro-Wilk test indicated that most variables were not normally distributed, both before and after imputation, as indicated by a p-value < 0.05. This indicates that a non-parametric approach is more suitable for further analysis, such as Random Forest, Support Vector Machine (SVM), and Multi-Layer Perceptron (MLP). The Kolmogorov-Smirnov test produced a p-value ≥ 0.05 for all variables, indicating that

Table 4. The Comparison of Performance Results of Several Studies for Missing Data Imputation and Imbalanced Data Handling

Author	Missing Data Imputation	Imbalanced Data handling	Classification methods	Acc (%)	Pre (%)	Rec (%)	F1-score (%)
(Reddy et al., 2021)	SMO Imputation	-	SMO	86.468	86.5	86.5	-
(Misir and Samanta, 2017)	Maximum Likelihood	-	Batch Backpropagation	99.86	99.86	36.51	-
(Poolsawad et al., 2012)	ANN dan T-test attribute selection	-	ANN	-	76.8	80.3	-
(Desiani et al., 2021a)	Mean, Mode, ANN Imputation	-	SVM	90	90	90	-
(Al Khaldy and Kambhampati, 2016)	CMC Imputation	-	Random Forest	85.71	97.56	50.11	-
(Mamilla et al., 2025)	-	SMOTE	KNN	92	91	91	91
		ADASYN	KNN	93	92	92	92
Proposed Method	Mode, MICE Imputation	-	Random Forest	80.3	80.3	80.3	80.3
			SVM	85.6	85.7	85.6	85.6
			MLP	78.9	79	78.9	78.8
	-	ADASYN	Random Forest	76.3	76.3	76.3	76.3
			SVM	77.1	77.1	77.1	77.1
			MLP	77.9	77.9	77.9	77.9
	Mode, MICE Imputation	ADASYN	Random Forest	97	97.5	97.5	97.1
			SVM	90.47	90.5	90.45	90.5
			MLP	95.5	95.5	95.4	95.5

there is no significant difference in distribution between the data before and after imputation. Thus, the imputation process does not change the overall shape of the data distribution. The Levene test indicates that all variables have a p-value = 1.00, indicating that there is no difference in variance between the data before and after imputation. It means that the level of data distribution remains stable. The imputation process does not provide significant changes to the distribution or variance of the data. Therefore, the imputed data can be considered valid and can be used for further analysis without reducing the main statistical characteristics of the original or old data.

3.4 Discussion

This study employs the data generated by the proposed method is applied in several kinds of classification. This study analyzes the effects of imputing missing data using Mode and MICE methods, along with addressing data imbalance through the ADASYN technique, to evaluate improvements in classification performance. The classification methods used are Random Forest, SVM, and MLP. The classification results shows that the missing data imputation and imbalanced data handling can improve accuracy, precision, recall, and F1-score for each method. Figure 3 shows the improvement in accuracy, precision, recall, and F1-score on the new dataset.

Figure 4 shows the heart disease dataset before missing data imputation and imbalanced data handling has accuracy, precision, recall, and F1-score below 80%. After missing data imputation using Mode and MICE, the accuracy, precision, recall, and F1-score have increased. In the Random Forest method, there was an increase of 4.8% for accuracy, precision, recall, and F1-score values. In the SVM method, accuracy increased by 9%, precision by 6%, recall by 9.5%, and F1-score by 7.8%. In the MLP method, the biggest increase in precision was 1.8% and increased by 1.7% for accuracy, recall, and F1-score values.

Handling imbalanced data can also improve the performance of accuracy, precision, recall, and F1-score in the classification method, but these results are still below 80%. The combination of missing data imputation using Mean, Mode, and MICE with ADASYN for addressing data imbalance achieved

a significant improvement in classification performance. The results of the accuracy, precision, recall, and F1-score performance are above 90% after missing data imputation and imbalanced data handling. The increase in accuracy, precision, recall, and F1-score in each class of each classification method is shown in Figures 5, 6, and 7.

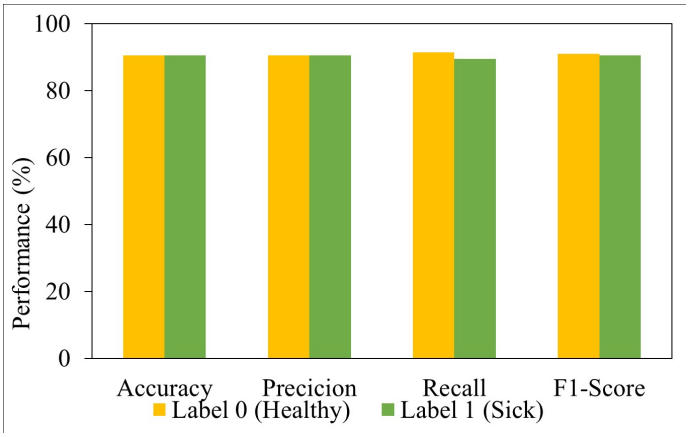

Figure 7. The Comparison Graph of Accuracy, Precision, Recall, and F1-Score on Support Vector Machine (SVM) for Each Class

Figure 5, Figure 6, and Figure 7 show the results of accuracy, precision, recall, and F1-score obtained in each class. All accuracy, precision, recall, and F1-score values obtained in each class are above 90%. From the accuracy value, the largest accuracy is 97%, indicating that the model managed to predict correctly 97% of the data. The largest precision is 97.5%, indicating that the model has a very high ability to identify the class of 0 as the class of 1. The largest recall is 97.5%, indicating that the model can identify 97.5% of the data from all classes of 1. The largest F1-score is 97.1%, demonstrating that the model achieves an excellent balance between precision and recall.

Comparison of classification results before and after the proposed data handling is not enough to assess its success without comparing it with other methods from previous studies. Several studies that used various methods to manage missing and imbalanced data in the UCI heart disease dataset are shown in Table 4. Table 4 compares the results of Accuracy (Acc), Precision (Pre), Recall (Rec), and f1-score from other studies.

Table 4 shows that the highest accuracy and precision are obtained by Misir and Samanta (2017), but the recall obtained is lower than the recall value in the proposed method. Al Khaldy and Kambhampati (2016) achieved precision higher than the proposed method, but the recall is bad. The proposed method demonstrates better accuracy than some studies, as shown in Table 4. Some previous studies did not show the f1-score, precision, and recall results obtained, only comparing the accuracy. However, precision and recall are important in addressing missing and unbalanced data to accurately identify and retrieve information for each class, even when some classes have smaller samples than others. The comparison shows that the proposed method is excellent for heart disease prediction by missing data imputation and imbalanced data handling can improve the accuracy, precision, recall, and f1-score values, which are excellent above 90% with several classification methods. MICE and ADASYN have the risk of overfitting. MICE has the risk of overfitting because MICE uses a regression model that only learns from the available data the noise can be considered as important data that must be imputed. It causes overfitting to noise. ADASYN has the risk of overfitting because ADASYN produces synthetic data around minority data. If there are outliers or noise around the minority, ADASYN will increase the synthetic data around the noise, so it can increase the risk of overfitting. Further study can combine MICE with feature selection for missing data imputation and combine sampling techniques in ADASYN by combining oversampling and undersampling for imbalanced data handling.

4. CONCLUSIONS

This study combines missing data imputation and balanced data handling. Missing data imputation is applied in two ways, namely single value imputation for attributes that have missing data below 5% and multiple value imputation for attributes that have missing data above 50%. In this study, the attributes *chol* and *restecg* were performed with single imputation: Mean and Mode. Chol is approached with mean imputation because the attribute type is numeric, while restecg is approached with Mode imputation because the attribute type is nominal. Multiple value imputation is performed using MICE for the attributes *trestbps, chol, fbs, thalach, exang*, and *oldpeak*. For imbalanced data handling, this study uses the ADASYN method. Based on missing data imputation with MICE and imbalanced data handling using ADASYN, it can be concluded that these techniques improve classification performance across several models, including Random Forest, Support Vector Machine (SVM), and Multi-Layer Perceptron (MLP). The results of classification performance show that the average improvement in accuracy, precision, recall, and f-1 score is above 90%. It demonstrates that applying missing data imputation and balancing techniques strengthens classification performance, especially for heart disease. This method can be developed as one of the methods to help improve the quality of data to be used in other classification methods, such as deep learning.

5. ACKNOWLEDGEMENT

The authors thank the deep learning discussion group and Computation Laboratory, Mathematics and Natural Science Faculty, Universitas Sriwijaya, and Department of Data Science, Faculty of Data Science, Inti International University, Malaysia, for support and facilities in the study.

REFERENCES

Aditsania, A. and A. L. Saonard (2017). Handling Imbalanced Data in Churn Prediction Using ADASYN and Backpropaga-

tion Algorithm. In *2017 3rd International Conference on Science in Information Technology (ICSITech)*. IEEE, pages 533–536

Al Khaldy, M. and C. Kambhampati (2016). Performance Analysis of Various Missing Value Imputation Methods on Heart Failure Dataset. In *Proceedings of SAI Intelligent Systems Conference*. pages 415–425

Ali, H., M. N. M. Salleh, K. Hussain, A. Ahmad, A. Ullah, A. Muhammad, R. Naseem, and M. Khan (2019). A Review on Data Preprocessing Methods for Class Imbalance Problem. *International Journal of Engineering & Technology*, **8**(3); 390–397

Austin, P. C., I. R. White, D. S. Lee, and S. van Buuren (2021). Missing Data in Clinical Research: A Tutorial on Multiple Imputation. *Canadian Journal of Cardiology*, **37**(9); 1322–1331

Bayuaji, L., Kusnadi, M. Y. Amzah, and D. Pebrianti (2024). Optimization of Feature Selection in Support Vector Machines (SVM) Using Recursive Feature Elimination (RFE) and Particle Swarm Optimization (PSO) for Heart Disease Detection. In *2024 9th International Conference on Mechatronics Engineering (ICOM)*. IEEE, pages 304–309

Chen, M., Y. Hao, K. Hwang, L. Wang, and L. Wang (2017). Disease Prediction by Machine Learning over Big Data from Healthcare Communities. *Ieee Access*, **5**; 8869–8879

De Diego, I. M., A. R. Redondo, R. R. Fernández, J. Navarro, and J. M. Moguerza (2022). General Performance Score for Classification Problems. *Applied Intelligence*, **52**(10); 12049–12063

Desiani, A., Y. Andriani, I. Ramayanti, S. Priyanta, B. Suprihatin, C. N. Apriyani, and M. Arhami (2024). RIB-Net as Modification of CNN Architecture for Semantic Segmentation of Optic Disc and Optic Cup. *Biomedical Engineering: Applications, Basis and Communications*, **36**(06); 2450036

Desiani, A., N. R. Dewi, A. N. Fauza, N. Rachmatullah, M. Arhami, and M. Nawawi (2021a). Handling Missing Data Using Combination of Deletion Technique, Mean, Mode, and Artificial Neural Network Imputation for Heart Disease Dataset. *Science and Technology Indonesia*, **6**(4); 303–312

Desiani, A., S. Yahdin, A. Kartikasari, and I. Irmeilyana (2021b). 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

Douzas, G., F. Bacao, and F. Last (2018). Improving Imbalanced Learning Through a Heuristic Oversampling Method Based on K-Means and SMOTE. *Information Sciences*, **465**; 1–20

Ebenuwa, S. H., M. S. Sharif, M. Alazab, and A. Al-Nemrat (2019). Variance Ranking Attributes Selection Techniques for Binary Classification Problem in Imbalance Data. *IEEE Access*, **7**; 24649–24666

Gabr, M. I., Y. M. Helmy, and D. S. Elzanfaly (2023). Effect of Missing Data Types and Imputation Methods on Supervised Classifiers: An Evaluation Study. *Big Data and Cognitive Computing*, **7**(1); 55

Guan, S., H. Yang, and T. Wu (2023). Transformer Fault Diagnosis Method Based on TLR-ADASYN Balanced Dataset. *Scientific Reports*, **13**(1); 23010

Hasan, M. K., M. A. Alam, S. Roy, A. Dutta, M. T. Jawad, and S. Das (2021). Missing Value Imputation Affects the Performance of Machine Learning: A Review and Analysis of the Literature (2010–2021). *Informatics in Medicine Unlocked*, **27**; 100799

Jäger, S., A. Allhorn, and F. Bießmann (2021). A Benchmark for Data Imputation Methods. *Frontiers in Big Data*, **4**; 693674

Khan, S. I. and A. S. M. L. Hoque (2020). SICE: An Improved Missing Data Imputation Technique. *Journal of Big Data*, **7**(1); 37

Kurniawati, Y. E., A. E. Permanasari, and S. Fauziati (2018). Adaptive Synthetic-Nominal (ADASYN-N) and Adaptive Synthetic-KNN (ADASYN-KNN) for Multiclass Imbalance Learning on Laboratory Test Data. In *2018 4th International Conference on Science and Technology (ICST)*. IEEE, pages 1–6

Lee, D.-H., S.-E. Woo, M.-W. Jung, and T.-Y. Heo (2022). Evaluation of Odor Prediction Model Performance and Variable Importance According to Various Missing Imputation Methods. *Applied Sciences*, **12**(6); 2826

Liu, D., D. Liang, and C. Wang (2016). A Novel Three-Way Decision Model Based on Incomplete Information System. *Knowledge-Based Systems*, **91**; 32–45

Mamilla, M. Y., R. Al-Haddad, and S. Chowdhury (2025). Resampling Imbalanced Healthcare Data for Predictive Modelling. *International Journal of Advanced Computer Science and Applications*, **16**(2); 36–44

Mera-Gaona, M., U. Neumann, R. Vargas-Canas, and D. M. López (2021). Evaluating the Impact of Multivariate Imputation by MICE in Feature Selection. *PLoS ONE*, **16**(7); 1–28

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

Osisanwo, F. Y., J. E. T. Akinsola, O. Awodele, J. O. Hinmikaiye, O. Olakanmi, and J. Akinjobi (2017). Supervised Machine Learning Algorithms: Classification and Comparison. *International Journal of Computer Trends and Technology (IJCTT)*, **48**(3); 128–138

Pauzi, N. A. M., Y. B. Wah, S. M. Deni, S. K. N. A. Rahim, and Suhartono (2021). Comparison of Single and MICE Imputation Methods for Missing Values: A Simulation Study. *Pertanika Journal of Science and Technology*, **29**(2); 979–998

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. F. Cleland (2012). Handling Missing Values in Data Mining: A Case Study of Heart Failure Dataset. In *2012 9th International Conference on Fuzzy Systems and Knowledge Discovery*.

IEEE, pages 2934–2938
Ramadhan, N. G. (2021). Comparative Analysis of ADASYN-SVM and SMOTE-SVM Methods on the Detection of Type 2 Diabetes Mellitus. *Scientific Journal of Informatics*, **8**(2); 276–282
Reddy, K. V. V., I. Elamvazuthi, A. A. Aziz, S. Paramasivam, H. N. Chua, and S. Pranavanand (2021). Heart Disease Risk Prediction Using Machine Learning Classifiers with Attribute Evaluators. *Applied Sciences*, **11**(18); 8352
Salamah, U., S. P. Sakti, A. Naba, and H. Soetedjo (2024). Identification of CO_2, SO_2, and a Mixture of Both Gases Using Optical Imaging Combined with Convolutional Neural Network (CNN). *Science and Technology Indonesia*, **9**(2); 371–379
Seliem, M. M. (2022). Handling Outlier Data as Missing Values by Imputation Methods: Application of Machine Learning Algorithms. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, **13**(1); 273–286
Tan, H. (2021). Machine Learning Algorithm for Classification. *Journal of Physics: Conference Series*, **1994**(1); 12016
Thabtah, F., S. Hammoud, F. Kamalov, and A. Gonsalves (2020). Data Imbalance in Classification: Experimental Evaluation. *Information Sciences*, **513**; 429–441
Wongvorachan, T., S. He, and O. Bulut (2023). A Comparison of Undersampling, Oversampling, and SMOTE Methods for Dealing with Imbalanced Classification in Educational Data Mining. *Information*, **14**(1); 54
Wu, X., H. Akbarzadeh Khorshidi, U. Aickelin, Z. Edib, and M. Peate (2019). Imputation Techniques on Missing Values in Breast Cancer Treatment and Fertility Data. *Health Information Science and Systems*, **7**(1); 19

