International Journal of Advances in Applied Sciences (IJAAS)
Vol. 13, No. 4, December 2024, pp. 948~960
ISSN: 2252-8814, DOI: 10.11591/ijaas.v13.i4.pp948-960

Early detection of coronary heart disease based on risk factors using interpretable machine learning

Wiharto, Farah Nada Mufidah
Department of Informatics, Faculty of Information Technology and Data Science, Universitas Sebelas Maret, Surakarta, Indonesia

Article Info

Article history:

Received Nov 6, 2023
Revised Jul 9, 2024
Accepted Aug 25, 2024

Keywords:

C5.0 algorithm
Coronary heart disease
Early detection
Interpretable machine learning
Risk factors

ABSTRACT

Coronary heart disease (CHD) is the leading cause of death in the world. The risk of coronary heart disease can be reduced or even prevented by early detection. Early detection of CHD has been widely developed using machine learning, but the machine learning algorithms used sometimes have low interpretability. Low interpretability makes it difficult for users to understand the cause of the decision. Referring to this, this research aims to propose an early detection model using machine learning interpretability, which is implemented using the C5.0 algorithm and interpreted using Shapley additive explanations (SHAP). This research method is divided into 3 stages, namely preprocessing, interpretable machine learning, and performance evaluation. This study used 215 patient data from Dr. Moewardi Surakarta Hospital. Testing the resulting model using the k-folds cross-validation method. The test results show that the risk factors that make a high contribution to the output of the coronary heart disease detection model are systolic blood pressure, diastolic blood pressure, and employment level, with the resulting accuracy performance of 84.64%. The proposed model can be an alternative for early prediction of coronary heart disease which can explain the influence of each selected risk factor on the model output.

Corresponding Author:

Wiharto
Department of Informatics, Faculty of Information Technology and Data Science
Universitas Sebelas Maret
Ir. Sutami St., No. 36A, Kentingan, Jebres, Surakarta, Indonesia
Email: wiharto@staff.uns.ac.id

1. INTRODUCTION

Coronary heart disease (CHD) is a leading global cause of death, with more than 9 million deaths attributed to CHD in 2020 [1]. The COVID-19 pandemic has worsened this already dire situation. Although the mortality rate due to COVID-19 varies worldwide, ranging from 1-2%, most patients can recover. Nonetheless, a considerable amount of evidence indicates that COVID-19 can cause various long-term health issues, including one that heightens the possibility of heart problems. A comprehensive study of health records in the United States revealed that individuals who contracted COVID-19 have a 55% greater chance of experiencing extended cardiovascular complications. Complications of coronary heart disease consist of heart rhythm disturbances, heart inflammation, blood clots, stroke, heart attack, heart failure, and possibly death. The sample registration system (SRS) survey from 2014 determined that coronary heart disease has the second highest mortality rate in Indonesia, following stroke, with a total mortality rate of 12.9% of all recorded deaths. Coronary artery disease is the primary and most prevalent cause of mortality in Indonesia, accounting for 26.4% of all deaths. The number of deaths attributed to heart disease outweighs the number of

heart and blood vessel specialists by a significant margin in Indonesia, with only 600 currently practicing in the country.

Coronary artery disease (CAD) is caused by the buildup of plaque in the coronary arteries, leading to decreased blood flow to the heart and increasing the risk of heart attacks and even mortality. Early awareness of a person's likelihood of experiencing CAD can decrease risk. Regular monitoring of CAD risk factors, such as cholesterol levels, blood sugar, blood pressure, and weight, is helpful for early detection. Population-based CAD prediction models have been developed based on risk factors. Many population-based models have been developed for the early prediction of CHD, including the Framingham risk score (FRS), prospective cardiovascular Münster (PROCAM), and systematic coronary risk evaluation (SCORE). These models provide valuable insight into CHD risk and can assist healthcare providers in determining appropriate preventive measures. However, it is important to recognize the limitations of such models and the need for ongoing research in this field. However, certain population-based models are unsuitable for detecting diseases in populations different from the one that the model represents. For instance, FRS is not appropriate for detecting heart disease in Japanese and Koreans, while Koreans cannot use PROCAM and SCORE models. Additionally, models based on urban populations prove ineffective when applied to rural populations [2], [3]. The assigned risk score and QRISK3 score models are comparable to PROCAM, which has only been validated for the German population. Therefore, it may not be suitable for other populations. The CUORE risk score model is also only appropriate for the population in Italy [4].

The rapid development of artificial intelligence, particularly in the domain of machine learning, has resulted in the creation of numerous prediction models that rely on machine learning datasets pertaining to risk factors of coronary heart disease. CHD prediction research utilizing a duo output artificial neural network ensemble (DOANNE) has achieved an accuracy level of 86% [5]. This accuracy rate is based on the use of 12 risk factors. Similar research has employed a combination of principal component analysis (PCA) and support vector machine (SVM) [6] to predict early CHD. The study identified 9 out of the 12 tested risk factors as influential. The model achieved a sensitivity performance of 84.2%. Subsequent developments, referencing FRS, were modeled using a Mamdani fuzzy inference system. The model employed a G medical center dataset from Korea, providing an accuracy performance of 69% [3]. A study similar to the aforementioned research was conducted utilizing the FRS. A predictive model was established using the classification and regression tree (CART) algorithm to create rules, which were subsequently modeled using fuzzy rule-based methods. The resulting model was tested through The Korea National Health and Nutrition Examination Survey-VI (KNHANES-VI) dataset, also known as the Korean National Health and Nutrition Examination Survey VI. Korean National Health and Nutrition Examination Survey VI achieved an accuracy performance of 69.51% and a sensitivity of 93.1%. Unfortunately, C5.0 currently outperforms the CART algorithm [7].

A model for the early detection of coronary heart disease has been developed using 9 risk factors and South African heart disease data-knowledge extraction based on evolutionary learning (KEEL) [8]. The model evaluated three machine learning algorithms: J48, Naïve Bayesian, and SVM. The Naïve Bayesian model showed higher sensitivity with a 63% value compared to J48 and SVM which recorded less than 50% respectively. There was no significant difference in specificity value among the three algorithms. A model has been developed utilizing the CART algorithm and fuzzy logic to detect coronary heart disease in the Korean population. The model employs nine risk factors and was tested with results indicating an accuracy value of 69.51%, a sensitivity of 93.1%, a specificity of 25.64%, a positive prediction value (PPV) of 69.95%, a negative prediction value (NPV) of 66.67%, and an area under the curve (AUC) of 0.594 [9]. Sivaprasad *et al*. [10] conducted a study on the development of deep learning, specifically transfer learning, and compared the results with SVM, neural networks, and random forest algorithms. In addition, Sarma *et al.* [11] and Ghasemieh *et al.* [12] developed a CHD detection model using ensemble stacking and deep learning techniques.

The development of machine learning models for detecting coronary heart disease has not been able to explain the decision-making process that it produces, commonly referred to as black-box. This is supported by a study conducted by Franklin and Muthukumar [13], which showed that many of the developed detection models still rely heavily on black-box methods, including the use of deep learning and neural networks. Mondal *et al*. [14] and Kim *et al*. [15] proposed a model for detecting coronary heart disease using ensemble learning and a statistical deep belief network for cardiovascular risk prediction. The proposed models do not use interpretable machine learning. This makes the model unable to explain the relationship between model input and model output. The result of this condition will make the resulting model not get full trust by the user [16], [17]. Trust is the main way to increase user confidence in using machine learning [18], as well as their comfort when using and managing it [19]. Trust is related to the ethics and intensity of regulatory activities.

To be able to provide a high level of confidence in the model, an interpretable machine learning (IML) model was developed. The implementation of IML is by using a number of machine learning

algorithms combined with interpretable methods, such as Shapley additive explanations (SHAP), local interpretable model-agnostic explanations (LIME), and class activation mapping (CAM) [20], [21]. SHAP, LIME, and CAM are able to explain the machine learning model so that the relationship between the input and output of the model can be understood. The three methods have their respective advantages, the CAM method is widely applied to explain image problems, LIME for some classification problems, while SHAP can be used for various problems, not only limited to classification [22]. In addition, SHAP has a strong theoretical explanation ability, and a fair distribution in its predictions [23]–[25]. The use of SHAP is also widely applied in various fields, one of which is health [26].

Referring to a number of studies that have been carried out, this study developed a model for early detection of coronary heart disease using interpretable machine learning. The interpretable machine learning model is implemented using a combination of C5.0 and interpreted using SHAP. The early detection model was developed using risk factor medical record data. The risk factors used were taken at Dr. Moewardi Hospital Surakarta, Indonesia.

2. RESEARCH METHOD

2.1. Dataset

This study utilizes medical records of patients from the Heart and Vascular Disease Polyclinic at Dr. Moewardi Hospital in Surakarta, Indonesia. The dataset consisted of 215 patients, with 133 diagnosed with coronary heart disease and 82 without. Non-modifiable and modifiable risk factors were investigated, resulting in 12 risk factors which are categorized in Table 1. The data for these risk factors were separated into two types, categorical and numerical.

Table 1. Representation of clinical data for risk factors

No.	Risk Factors	Category	Number (%)	*Mean* ± SD	*NA (%)*
1.	Age			59.11 ± 14.07	
2.	Gender	Male	121 (56.28)		
		Female	94 (43.72)		
3.	Employment level	Light	96 (44.65)		4 (1.86)
		Medium	85(39.53)		
		Heavy	30 (13.95)		
4.	Total cholesterol			169.17 ± 45.47	
5.	Low-density lipoprotein (LDL)			110.18 ± 35.88	
6.	High-density lipoprotein (HDL)			35.87 ± 12.55	
7.	Triglyceride			128.62 ± 87.18	
8.	Systolic blood pressure			138.61 ± 24.67	
9.	Diastolic blood pressure			86.96 ± 14.65	
10.	Obesity			20.70 ± 7.65	
11.	Smoking	Yes	72 (24.18)		
		No	143 (34.06)		
12.	Diabetes history	Yes	30 (32.97)		
		No	54 (59.34)		

2.2. Method

The study was carried out in three phases: pre-processing, interpretable machine learning, and performance evaluation. Figure 1 displays a complete overview of these phases. The pre-processing phase contained a missing value imputation process and data normalization. The missing value imputation process applied the single center imputation from the multiple chained equations (SICE) algorithm [27]. The SICE algorithm is an extension of the multiple imputation by chained equations (MICE) algorithm [27], [28]. Algorithm 1 shows the MICE algorithm, and Algorithm 2 presents the SICE algorithm.

In the SICE algorithm presented in Algorithm 2, the imputation process is performed by considering the data type of the risk factor. If the risk factor is categorical, the mode is used for calculation, while the average is used for numerical data. Following the imputation process, the data normalization procedure is carried out to rescale the data between 0-1 [5]. The normalization process follows the criteria presented in Table 2. If the data meets the criteria column, the normalization outcomes are then displayed in the value column.

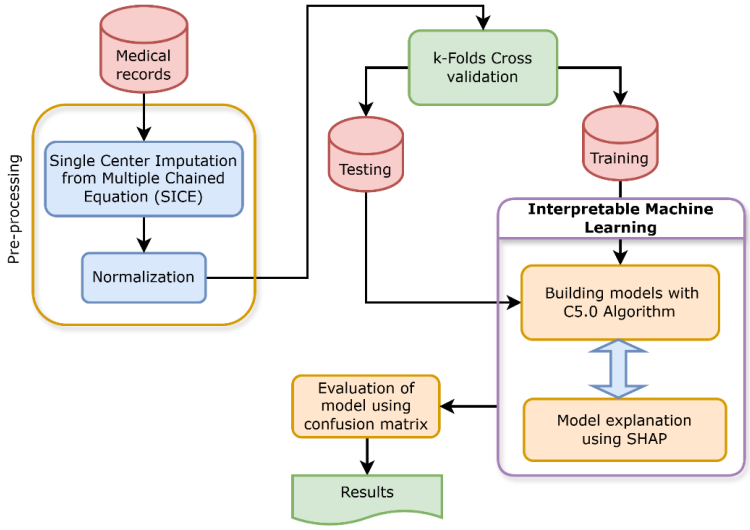

Figure 1. Interpretable model of machine learning

Algorithm 1. MICE

1. Simple mean imputation is performed for every missing value which is referred to as placeholders.
2. The placeholder mean value for one variable is set back to missing
3. Appropriate regression is done between observed values of the missing variable against other variables.
4. The missing value is then predicted using a regression model
5. IF other missing value, THEN to step 2 ELSE to step 6
6. IF iteration end THEN to step 7 ELSE to step 2
7. Results

Algorithm 2. SICE

INPUT:

x: instances with missing values categorical and numeric in medical records
y: instances with no missing value data in the same medical record.
m: number of imputations defined by the user.

OUTPUT:

x': update x with imputed missing data

1. FOR each missing value in x DO
2. use Algoritma 1 to find the missing value;
3. END
4. Repeat for m times;
5. miceResult[i]=imputed data for ith missing value;
6. FOR each row in miceResult DO
7. IF type data of miceResult[i] is numeric THEN
8. siceResult=mean(miceResult[i,1:m])
9. ELSE
10. siceResult=modus(miceResult[i,1:m])
11. END
12. x'=x update with siceResult
13. END

After normalization, the next step involves dividing the data into training and testing sets. The data division process applies k-fold cross-validation, where k is set to 10. Subsequently, the divided data is used to construct a model that incorporates interpretable machine learning (IML). By using IML, it becomes simpler for users to comprehend the decision-making process. The algorithm used in the IML model is a combination of C5.0 with SHAP. The C5.0 algorithm creates a tree diagram model that clarifies the input-output connection and simplifies decision-making [21]. C5.0 modifies the iterative dichotomiser 3 (ID3) and C4.5 algorithms. When building the decision tree, the root for the next node is selected based on the maximum

information gain [29], [30]. This method begins by considering all data as the root of the decision tree, with the chosen attribute serving as the dividing factor for the sample [30].

The final stage involves evaluating the performance of the resultant IML model by referencing the confusion matrix [30], which is displayed in Table 3. Using Table 3, one can calculate the performance parameters including sensitivity, specificity, accuracy, positive prediction value (PPV), and negative prediction value (NPV). These parameters are computed by referencing the confusion matrix generated during the testing phase and applying (1)-(5) for parameter calculation.

$$\text{Accuracy} = \text{ACC} = \frac{(\text{TP}+\text{TN})}{(\text{TP}+\text{TN}+\text{FP}+\text{FN})} \text{ X } 100\% \quad (1)$$

$$\text{Specificity} = \text{SEN} = \frac{\text{TN}}{(\text{TN}+\text{FP})} \times 100\% \quad (2)$$

$$\text{Sensitivity} = \text{SPE} = \frac{\text{TP}}{(\text{TP}+\text{FN})} \times 100\% \quad (3)$$

$$\text{PPV} = \frac{\text{TP}}{(\text{TP}+\text{FP})} \text{x} 100\% \quad (4)$$

$$\text{NPV} = \frac{TN}{(\text{TN}+\text{FN})} \times 100\% \quad (5)$$

Table 2. Data normalization requirements

No	Risk Factors	Criteria	Value	No	Risk Factors	Criteria	Value
1	Age	<41	0	7	Triglyceride	<100	0
		41–50	0.2			100–149	0.25
		51–60	0.4			150–199	0.5
		61–70	0.6			200–499	0.75
		71–80	0.8			≥500	1
		≥81	1				
2	Gender	Male	1	8	Systolic blood pressure	<120	0
		Female	0.5			120–129	0.2
						130–139	0.4
						140–159	0.6
						160–179	0.8
						≥180	1
3	Employment level	Light	0	9	Diastolic blood pressure	<80	0
		Medium	0.5			80–84	0.2
		Heavy	1			85–89	0.4
						90–99	0.6
						100–109	0.8
						≥110	1
4	Total cholesterol	<200	0	10	Obesity	<18.5	0
		200–239	0.5			18.5–22.9	0.25
		≥240	1			23–24.9	0.5
						25–29.9	0.75
						≥30	1
5	LDL	<100	0	11	Smoking	Yes	1
		100–129	0.25			No	0
		130–159	0.5				
		160–189	0.75				
		≥190	1				
6	HDL	<40	0	12	Diabetes history	Yes	1
		40–59	0.5			No	0
		≥60	1				

Table 3. Confusion matrix

Actual Class	Prediction Class	
	Positive	Negative
Positive	True positive (TP)	False negative FN
Negative	False positive (FP)	True negative TN

3. RESULTS AND DISCUSSION

3.1. Results

Early detection model for coronary heart disease utilizing IML, run on hardware with CPU specifications of 2.00 GHz, 1.99 GHz, and 4 GB RAM. The IML model was implemented using the library in version 4.2.2 of the R programming language with R Studio version 2022.07.2 Build 576 and Microsoft Excel version 2210 Build 15726.20202. This research first performed imputation using the SICE algorithm, which is an evolution of MICE. The MICE procedure was executed utilizing the mice library in R. Algorithm 1 displays the outline of the MICE algorithm. In our study, MICE was run with m=35 and iterated four times to give 35 approximated values for each missing value. This equates to 35 MICE processes carried out on the dataset and each process was repeated four times. The frequency distribution of the missing values for each risk factor can be found in Figure 2. The imputation process is categorized into two types of risk factors: categorical and numerical.

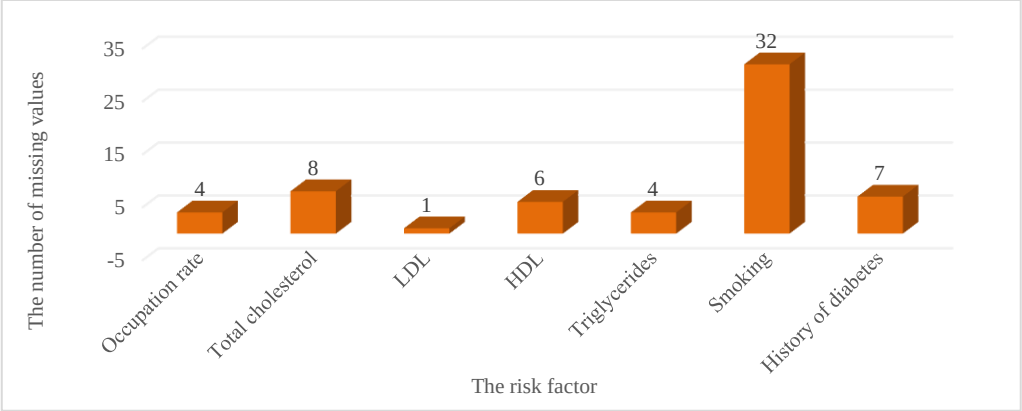

Figure 2. Number of missing data values

For categorical data, such as employment level, smoking, and history of diabetes, missing values are imputed using the mode of the data. Afterward, one of the values is again selected to be missing. Then, all data from patients with missing values undergo regression analysis using the polytomous logistic regression method [27] for employment level data, and the logistic regression method for smoking and diabetes history data. This results in an estimated value for the missing data point. The procedure was systematically executed until all absent values were populated with the computed values from the regression. Subsequently, in the ensuing round, all undone values were populated with the mode from the preceding round's outcomes. This complete process was iterated until the termination of the execution.

Missing values in numerical data, including total cholesterol, LDL, HDL, and triglycerides, are replaced with their respective mean. Subsequently, one of the missing values is chosen to be emptied again and all data from the patient with the missing value is regressed on other data using the predictive mean matching method. This process is repeated until all missing values are replaced with predicted values obtained through predictive mean matching [27]. Then, in each subsequent iteration, missing values are replaced with the mean of the previous iteration results. This process is repeated until the end of the iteration. The entire process is performed 35 times simultaneously, denoted by the value of m, resulting in 35 complete datasets without any missing values.

The study utilized the k-fold cross-validation method with k=10 to conduct testing. Table 4 displays the results of these tests, including the performance of each fold out of 10 total. The 9th fold had the best performance, but overall, the IML model incorporating the C5.0 algorithm yielded an 84.64% sensitivity rate, indicating its ability to accurately detect positive cases of coronary heart disease [31], [32]. As the sensitivity of the IML model increases, more positive CHD test results are obtained from patients already suffering from CHD, or fewer negative CHD cases are missed in this group. The 9th fold reveals the model's notable performance in accurately detecting CHD patients, achieving a rate of 93.33%. Although this rate is higher when examining the 2nd folds, other performance parameters remain low.

Testing with 10 folds also generates a decision tree model for each fold. The decision tree created with the C5.0 algorithm serves as an embedded feature selection method, allowing the algorithm to perform the feature selection process and use the result as a risk factor in compiling the decision tree. The resulting decision tree in each fold contains the same number of attributes, but the selected risk factors vary. The

frequency of risk factor selection in each fold is displayed in Figure 3. Employment level, systolic blood pressure, and diastolic blood pressure are attributes present in all folds. If a sample is taken from the 9th fold, it yields five significant attributes out of the available twelve. The five attributes that pose a risk factor are smoking, age, employment level, diastolic blood pressure, and systolic blood pressure. Referring to the results of the information gain calculation in the C5.0 algorithm, the usefulness level of the diastolic blood pressure attribute is 100%, the usefulness level of the systolic blood pressure attribute is 48.97%, the usefulness level of the work level attribute is 37.11%, the usefulness level of the smoking attribute is 10.31%, and the usefulness level of the age attribute is 9.97%.

Table 4. Performance of the CHD early detection model

No. Fold	SEN	ACC	SPE	PPV	NPV
1	94.44%	81.82%	25.00%	85.00%	50.00%
2	100.00%	71.43%	0.00%	71.43%	0.00%
3	76.92%	81.82%	88.89%	90.91%	72.73%
4	80.00%	71.43%	63.64%	66.67%	77.78%
5	84.62%	59.09%	22.22%	61.11%	50.00%
6	75.00%	61.90%	44.44%	64.29%	57.14%
7	92.86%	80.95%	57.14%	81.25%	80.00%
8	69.23%	68.18%	66.67%	75.00%	60.00%
9	93.33%	85.71%	66.67%	87.5%	80.00%
10	80.00%	63.64%	50.00%	57.14%	75.00%
Mean	84.64%	72.60%	48.47%	74.03%	60.27%

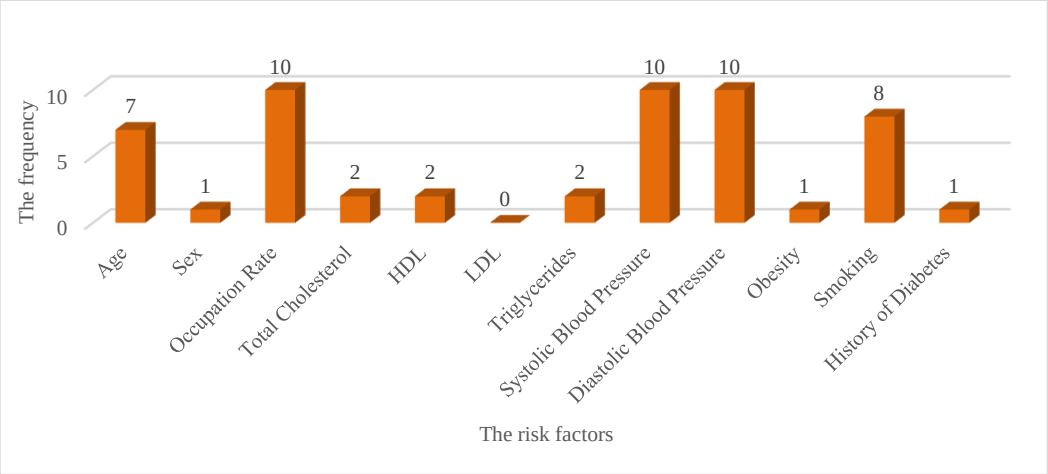

Figure 3. Risk factors influencing CHD detection

The decision tree depicted in Figure 4 indicates that the diastolic blood pressure risk factor attribute serves as the root with attribute values of 0.2, 0.4, 0.8, and 1, or in standard measurement units. This leads to positive CHD prediction for blood pressure measurements of 80-89 or ≥100, resulting in the classification of CHD=1. These findings are supported by 76 positive cases out of a total of 99 cases, with a confidence level of 76.77%. Patients with diastolic blood pressure risk factors of 0 and 0.6, or with values <80 or 90-99 as specified in Table 2, should have their systolic blood pressure checked. Thus, diastolic blood pressure values of 0 or 0.6, or results measuring <80 or 90-99, would predict that patients with systolic blood pressure risk factor values of 0.4 or 1, or results measuring 130-139 or ≥180, would have a negative prediction of coronary heart disease or be placed in the CHD=0 class. This result indicates a confidence level of 86.96%, based on 20 cases of CHD negativity among 23 patients with identical blood pressure levels. It is necessary to assess the occupational risk factor for patients with systolic blood pressure values of 0, 0.2, 0.6, and 0.8, or with measurements of ≤129 or 140-179. When a patient's systolic blood pressure is found to be ≤129 or 140-179 while having a heavy or work level attribute worth 1, they will likely test positive for CHD or be classified under CHD=1. The prediction results show an 81.82% confidence level, based on 9 positive cases of coronary heart disease out of 11 total cases. For patients with light or moderate workloads, or systolic blood pressure values of 0 or 0.5, predictions will be made in the next step based on the risk factors for systolic blood pressure.

If a patient's work level is light or moderate, or if their attribute is worth 0 or 0.5 and they have a systolic blood pressure attribute worth 0.2, or if their measurement results are between 120-129, then they will be classified as CHD=0 or predicted negative. This prediction has a 62.5% confidence level, supported by 10 negative cases out of a total of 16 patient cases with similar attribute values. If the patient's work level is light or moderate, or if the attribute is 0 or 0.5 with a systolic blood pressure attribute of 0.8, or if the measurement result falls between 160-179, then the patient is predicted to be positive for CHD=1 or to have coronary heart disease. This prediction carries a confidence level of 83.33% which is backed by 5 positive cases out of a total of 6 patient cases with the same attribute value. Patients with a systolic blood pressure attribute value of 0 or measurement results <120 will be predicted based on their smoking attribute value. This is due to the smoking risk factor attribute having the largest information gain ratio value in this subset of patient data. Patients with a systolic blood pressure attribute value of 0.6 or a measurement result between 140-159 will be predicted based on their age attribute value. This is due to the age attribute having the highest information gain ratio in this subset of patient data.

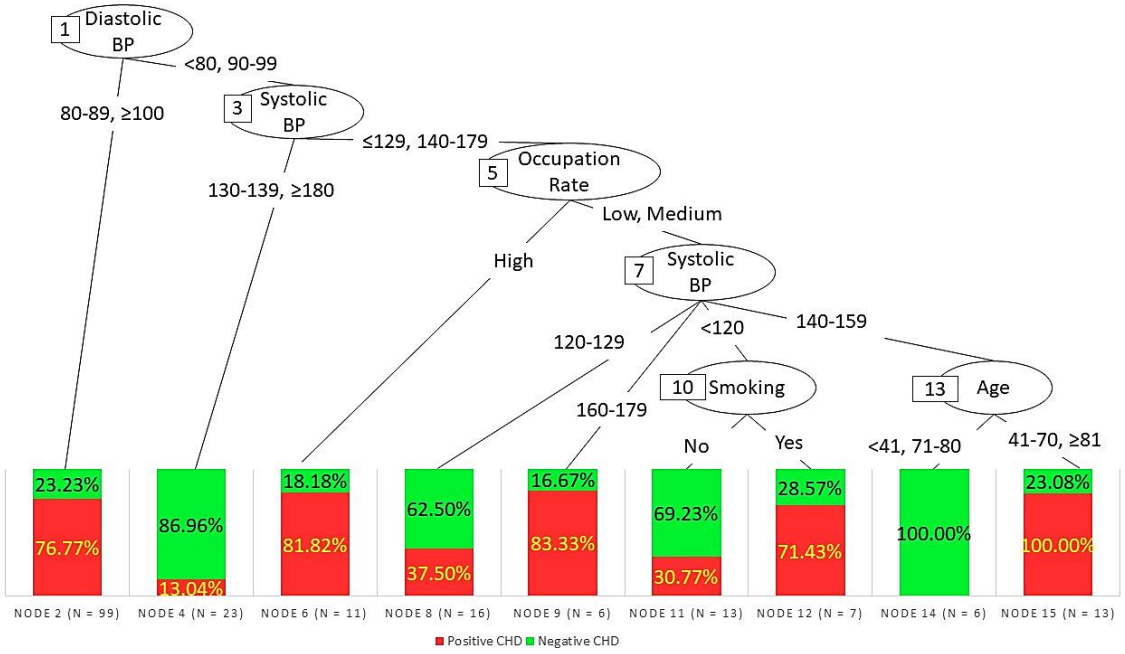

Figure 4. IML model for coronary heart disease detection

When the patient's systolic blood pressure attribute is less than 120 and the patient does not smoke or has a smoking attribute value of 0, they are categorized into the CHD=0 prediction class, indicating a negative prediction for coronary heart disease. This prediction has a confidence level of 69.23%, supported by 9 negative cases out of a total of 13 cases with the same attribute value. Meanwhile, if the patient is a smoker or has a smoking attribute value of 1, they will be predicted positively to suffer from coronary heart disease or enter the CHD prediction class=1. Patients with a systolic blood pressure attribute value of 0.6 (140-149) and an age attribute value of 0 or 0.8 (patients aged <41 years or 71-81 years) will be included in the CHD prediction class=0 or predicted to be "negative" for coronary heart disease. This prediction is based on 6 cases with similar attribute values, all of whom tested negative for coronary heart disease. However, patients aged 41-70 or ≥81 years old (age attribute value 0.2, 0.4, 0.6, or 1) are likely to test positive for coronary heart disease or have an attribute value of 1 for the CHD prediction class. This prediction has a confidence level of 76.92% which is supported by 10 positive cases out of a total of 13 patient cases with similar attribute values.

The IML model built using SHAP can be analyzed for the impact of each feature on the output of the system model. The impact of each feature can be shown in Figure 5. Referring to Figure 6, it can show the contribution of features from each data to the system model, besides that, the average shapley values that show the contribution of features can also be taken, as shown in Figure 6. Figure 6 shows that blood pressure, both diastolic blood pressure (DBP) and systolic blood pressure (SBP) have a high impact on the system model. In addition to these two features, the employment rate feature also has a higher impact than age and smoking.

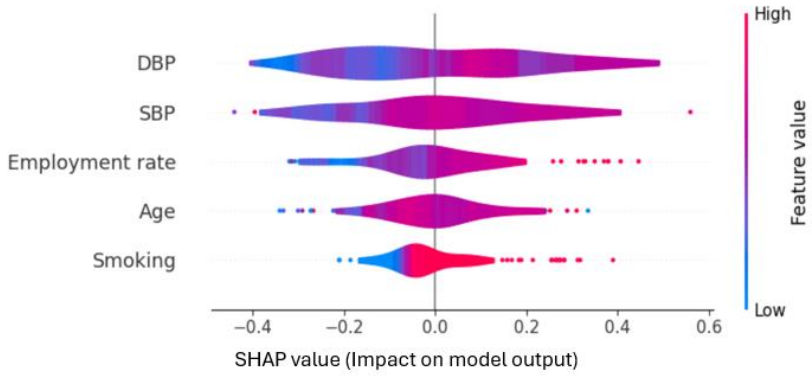

Figure 5. Impact on model output

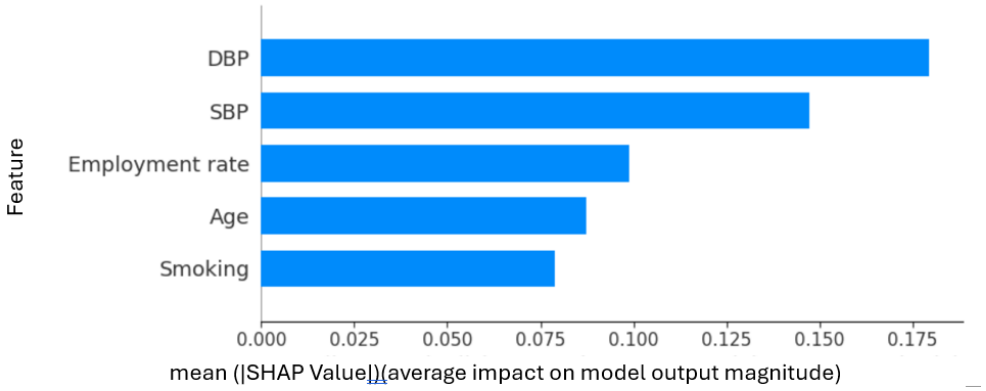

Figure 6. SHAP value (average impact on model output)

3.2. Discussion

There is an anomaly in the prediction pattern of coronary heart disease in the developed IML model. Specifically, diastolic blood pressure exhibits a higher number of positive predictions for CHD than expected, despite being below the usual threshold. This pattern is observed in over 50 cases. The use of blood pressure control medications in coronary heart disease patients may result in a low diastolic blood pressure reading during blood pressure measurement. Anomalies also occurred in systolic blood pressure risk factors, with 20-49 cases total. High systolic blood pressure was associated with negative predictions, particularly in young patients. High systolic blood pressure in patients without coronary heart disease may result from errors during blood pressure measurement when the patient is not adequately rested. Proper diagnosis and lifestyle modifications are crucial for managing high blood pressure. In younger patients who test positive for coronary heart disease, heredity and an unhealthy lifestyle may be causative factors. In patients with high blood pressure, those who are younger have a higher probability of developing coronary heart disease [33].

From the calculated results, the coronary heart disease early detection model demonstrated a sensitivity of 84.64%, accuracy of 72.597%, specificity of 48.467%, positive predictive value of 74.03%, and negative predictive value of 60.265%. The model's low specificity and negative predictive values stem from an imbalance between CHD-positive and negative patient data used in the experiment, with a ratio of 5:3. As a result, the model classifies many negative patients as positive, leading to a decrease in true negative value and an increase in false positive value.

Table 4 shows that the sensitivity model developed in this study outperforms previous studies that used data from the same location and only considered 5 risk factors. However, the accuracy of our model is lower than that of previous studies due to the uneven distribution of patient data between positive CHD and negative CHD cases. This phenomenon has been observed in various studies, including research conducted by Chawla *et al*. [34], Thabtah *et al*. [35], and Luque *et al*. [36]. The studies emphasize that machine learning algorithms often prioritize labeling the majority class in predicted data, leading to neglect of the minority class. As a result, these algorithms mostly generate accurate predictions for the majority class.

Table 4. Comparison of model performance with previous research in Indonesia

References	Method	Number of Data	#Risk factor	SEN	ACC
[6]	PCA+SVM	120	9	84.20%	78.61%
[5]	DOANNE	120	12	-	86.87%
Proposed	IML (C5.0+SHAP)	215	5	84.64%	72.597%

Previous research using the PCA+SVM method resulted in 9 risk factors, whereas C5.0 produced 5 risk factors. From both techniques, the C5.0 algorithm eliminated 4 risk factors, which include total cholesterol, LDL, triglyceride levels, and gender. The risk factors for total cholesterol, LDL, and triglyceride levels are interconnected, as demonstrated in the Friedewald formula [37] presented in (6) and the Martin-Hopkins formula [38] depicted in (7). When utilizing the PCA+SVM method to eliminate HDL risk factors, reference to both (6) and (7) reveals that all three risk factors are interconnected and can be eliminated using the C5.0 algorithm. Despite the removal of all cholesterol-related risk factors, the C5.0 algorithm demonstrates higher sensitivity than the use of cholesterol risk factors alone. These results highlight the IML model's ability to detect CHD patients, with a detection rate of 84.64% for C5.0 compared to 84.2% for PCA+SVM.

$$LDL = 0.659 \times cholesterol\ total + 0.182 \times HDL - 0.117 \times trygliserida\ levels + 18.03 \quad (6)$$

$$LDL = cholesterol\ total - HDL - trygliserida\ levels \quad (7)$$

Early detection using IML, which is implemented using the C5.0 algorithm and interpreted using SHAP, is able to provide an explanation that is easy for users to understand the decision-making process. This ability is better than models that use SVM and DOANNE algorithms in previous studies. The ability of the model to explain the decision-making process will make IML with C5.0, trustworthiness, causality, transferability, informativeness, confidence, and fairness. IML model capability can be analyzed for each input data, for example as shown in Figure 7. In Figure 7, each line plotted on the decision plot shows how strongly each feature contributes to a single model prediction, thus explaining what feature values drive the prediction. For the red line showing data with positive CHD, the red line shows that DBP has a high influence in determining the model output, as well as SBP. The blue line shows when the data is negative CHD, it is also highly determined by the DBP and SBP features.

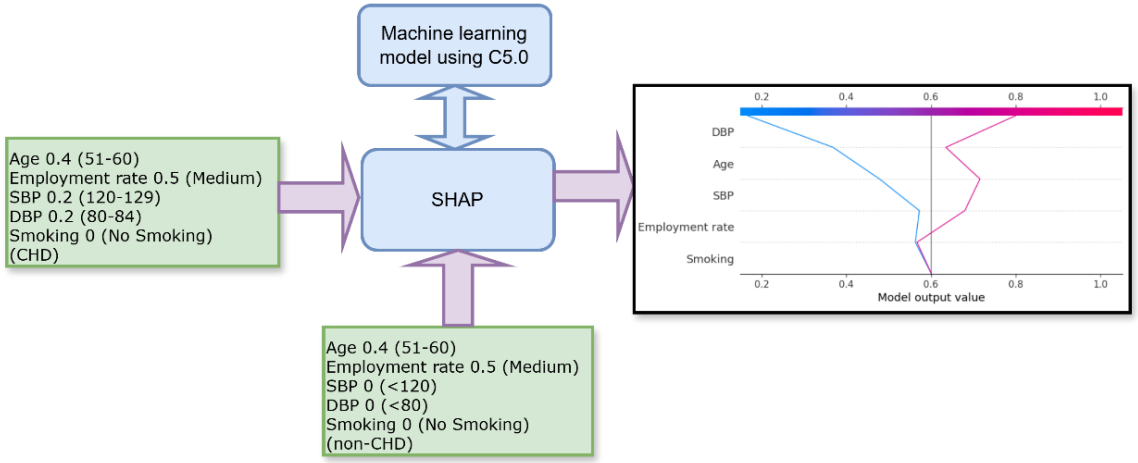

Figure 7. Testing the effect of features from one patient's data on model prediction

The influence of features on prediction results can also be represented in a face plot graph, as shown in Figures 8 and 9. Figure 8 shows how much influence the SBP, DBP, and employment rate features have on prediction results, as shown in red. For age and smoking features, the influence is low, as shown in blue. Figure 9 shows the data tested for CHD negative data, where the results show that DBP and SBP have a low

influence on prediction, which is shown in blue. The low influence is because SBP <120 and DBP <80, while the employment rate has a high influence, but is at the bottom, which is close to the boundary value of positive and negative CHD, which is at -0.04. Referring to Figures 8 and 9, it can be concluded that the SBP, DBP, and employment rate features are features that have a high influence on determining the prediction results of the model.

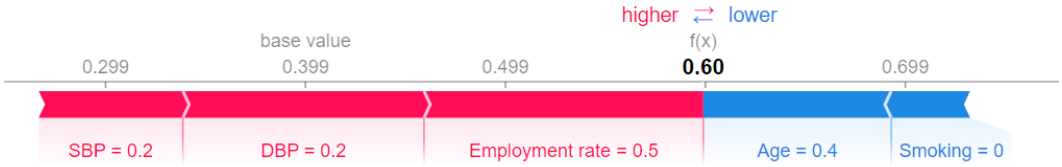

Figure 8. Effect of features on model output, for positive CHD patient data

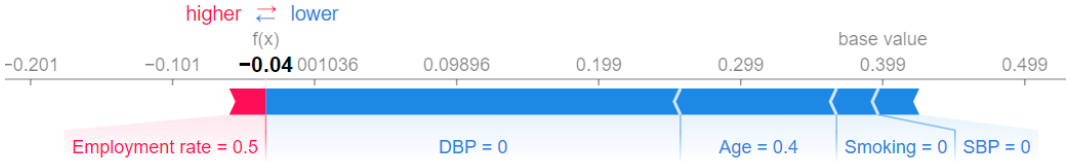

Figure 9. Effect of features on model output, for negative CHD patient data

4. CONCLUSION

In this study, it can be concluded that the early detection model of coronary heart disease using IML can explain the decision-making process. The performance of the IML model with the C5.0 algorithm can provide good performance, namely sensitivity of 84.64% and accuracy of 72.597%, with testing using the 10-fold cross-validation method. High-performance parameters for sensitivity, indicating that the IML model with C5.0 can be used for early detection because the model has a high ability for patients who are positive for CHD, by the model also detected positive CHD. In addition, the results of interpretation using SHAP explain that the risk factors of diastolic blood pressure, systolic blood pressure, and employment level are the most influential attributes in detecting coronary heart disease from the 12 risk factor attributes used.

ACKNOWLEDGEMENTS

We would like to express our gratitude to Sebelas Maret University for the research funding provided through the Group Research Grant scheme, as outlined in contract No. 228/UN27.22/PT.01.03/2023. In addition, we are grateful to the Faculty of Information Technology and Data Science for the provision of facilities and infrastructure that enabled the successful completion of this research project.

REFERENCES

[1] M. Naghavi *et al.*, "Global burden of 288 causes of death and life expectancy decomposition in 204 countries and territories and 811 subnational locations, 1990–2021: a systematic analysis for the global burden of disease study 2021," *Lancet*, vol. 403, no. 10440, pp. 2100–2132, May 2024, doi: 10.1016/S0140-6736(24)00367-2.

[2] K. Nishimura *et al.*, "Predicting coronary heart disease using risk factor categories for a Japanese urban population, and comparison with the Framingham risk score: the Suita study," *J. Atheroscler. Thromb.*, vol. 21, no. 8, pp. 784–98, 2014, doi: 10.5551/jat.19356.

[3] J.-K. Kim, J.-S. Lee, D.-K. Park, Y.-S. Lim, Y.-H. Lee, and E.-Y. Jung, "Adaptive mining prediction model for content recommendation to coronary heart disease patients," *Cluster Comput.*, vol. 17, no. 3, pp. 881–891, 2014, doi: 10.1007/s10586-013-0308-1.

[4] A. Sofogianni, N. Stalikas, C. Antza, and K. Tziomalos, "Cardiovascular risk prediction models and scores in the era of personalized medicine," *J. Pers. Med.*, vol. 12, no. 7, 2022, doi: 10.3390/jpm12071180.

[5] W. Wiharto, E. Suryani, and V. Cahyawati, "The methods of duo output neural network ensemble for prediction of coronary heart disease," *Indones. J. Electr. Eng. Informatics*, vol. 7, no. 1, pp. 51–58, 2019, doi: 10.52549/ijeei.v7i1.458.

[6] W. Wiharto, H. Kusnanto, and H. Herianto, "Clinical decision support system for assessment coronary heart disease based on risk factor," *Indian J. Sci. Technol.*, vol. 10, no. 22, pp. 1–12, 2017, doi: 10.17485/ijst/2017/v10i22/84940.

[7] M. Balamurugan and S. Kannan, "Performance analysis of cart and c5.0 using sampling techniques," *2016 IEEE Int. Conf. Adv. Comput. Appl.*, pp. 72–75, 2016, doi: 10.1109/ICACA.2016.7887926.
[8] A. H. Gonsalves, F. Thabtah, R. M. A. Mohammad, and G. Singh, "Prediction of coronary heart disease using machine learning: an experimental analysis," in *ICDLT '19: Proceedings of the 2019 3rd International Conference on Deep Learning Technologies*, Association for Computing Machinery, 2019, pp. 51–56. doi: 10.1145/3342999.3343015.
[9] J. Kim, J. Lee, and Y. Lee, "Data-mining-based coronary heart disease risk prediction model using fuzzy logic and decision tree," *Healthc. Inform. Res.*, vol. 21, no. 3, pp. 167–174, 2015, doi: 10.4258/hir.2015.21.3.167.
[10] R. Sivaprasad, M. Hema, B. N. Ganar, D. M. Sunil, V. Mehta, and M. Fahlevi, "Heart disease prediction and classification using machine learning and transfer learning model," in *2022 International Conference on Automation, Computing and Renewable Systems (ICACRS)*, Pudukkottai, India: IEEE, 2022, pp. 595–601. doi: 10.1109/ICACRS55517.2022.10029279.
[11] N. Sharma, L. Malviya, A. Jadhav, and P. Lalwani, "A hybrid deep neural net learning model for predicting coronary heart disease using randomized search cross-validation optimization," *Decis. Anal. J.*, vol. 9, p. 100331, 2023, doi: 10.1016/j.dajour.2023.100331.
[12] A. Ghasemieh, A. Lloyed, P. Bahrami, P. Vajar, and R. Kashef, "A novel machine learning model with stacking ensemble learner for predicting emergency readmission of heart-disease patients," *Decis. Anal. J.*, vol. 7, p. 100242, 2023, doi: 10.1016/j.dajour.2023.100242.
[13] R. G. Franklin and B. Muthukumar, "Survey of heart disease prediction and identification using machine learning approaches," in *2020 3rd International Conference on Intelligent Sustainable Systems (ICISS)*, Thoothukudi, India: IEEE, 2020, pp. 553–557. doi: 10.1109/ICISS49785.2020.9316119.
[14] S. Mondal, R. Maity, Y. R. Singh, S. Ghosh, and A. Nag, "Early prediction of coronary heart disease using boosting-based voting ensemble learning," in *2022 IEEE Bombay Section Signature Conference (IBSSC)*, IEEE, 2022, pp. 1–5. doi: 10.1109/IBSSC56953.2022.10037445.
[15] J. Kim, U. Kang, and Y. Lee, "Statistics and deep belief network-based cardiovascular risk prediction," *Heal. Inf. Res.*, vol. 23, no. 3, pp. 169–175, 2017, doi: 10.4258/hir.2017.23.3.169.
[16] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining explanations: an overview of interpretability of machine learning," in *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, Turin, Italy: IEEE, 2018, pp. 80–89. doi: 10.1109/DSAA.2018.00018.
[17] J. Petch, S. Di, and W. Nelson, "Opening the black box: the promise and limitations of explainable machine learning in cardiology," *Can. J. Cardiol.*, vol. 38, no. 2, pp. 204–213, 2022, doi: 10.1016/j.cjca.2021.09.004.
[18] J. D. Olden and D. A. Jackson, "Illuminating the 'black box': a randomization approach for understanding variable contributions in artificial neural networks," *Ecol. Modell.*, vol. 154, no. 1–2, pp. 135–150, Aug. 2002, doi: 10.1016/S0304-3800(02)00064-9.
[19] S. Athey and G. Imbens, "Recursive partitioning for heterogeneous causal effects," *Proc. Natl. Acad. Sci.*, vol. 113, no. 27, pp. 7353–7360, Jul. 2016, doi: 10.1073/pnas.1510489113.
[20] G. Zheng, Y. Zhang, X. Yue, and K. Li, "Interpretable prediction of thermal sensation for elderly people based on data sampling, machine learning and Shapley additive explanations (SHAP)," *Build. Environ.*, vol. 242, p. 110602, Aug. 2023, doi: 10.1016/j.buildenv.2023.110602.
[21] C. Molnar, *Interpretable machine learning: a guide for making black box models explainable*. Independently published, 2022.
[22] H. T. T. Nguyen, H. Cao, V. T. K. Nguyen, and D. K. N. Pham, "Evaluation of explainable artificial intelligence:shap, lime, and cam," in *FAIC 2021*, 2021. [Online]. Available: https://www.researchgate.net/publication/362165633_Evaluation_of_Explainable_Artificial_Intelligence_SHAP_LIME_and_CAM
[23] A. A. Biswas, "A comprehensive review of explainable ai for disease diagnosis," *Array*, vol. 22, p. 100345, Jul. 2024, doi: 10.1016/j.array.2024.100345.
[24] A. N. Juscafresa, "An introduction to explainable artificial intelligence with lime and SHAP," University of Barcelona, 2022. [Online]. Available: http://hdl.handle.net/2445/192075
[25] M. M. Hasan, "Understanding model predictions: a comparative analysis of SHAP and lime on various ml algorithms," *J. Sci. Technol. Res.*, vol. 5, no. 1, pp. 17–26, 2024, doi: 10.59738/jstr.v5i1.23(17-26).eaqr5800.
[26] V. Vimbi, N. Shaffi, and M. Mahmud, "Interpreting artificial intelligence models: a systematic review on the application of lime and SHAP in Alzheimer's disease detection," *Brain Informatics*, vol. 11, no. 1, p. 10, Dec. 2024, doi: 10.1186/s40708-024-00222-1.
[27] S. I. Khan and A. S. M. L. Hoque, "SICE: an improved missing data imputation technique," *J. Big Data*, vol. 7, no. 1, p. 37, Dec. 2020, doi: 10.1186/s40537-020-00313-w.
[28] A. Z. Alruhaymi and C. J. Kim, "Why can multiple imputations and how (MICE) algorithm work?," *Open J. Stat.*, vol. 11, no. 05, pp. 759–777, 2021, doi: 10.4236/ojs.2021.115045.
[29] M. Salman Saeed *et al.*, "An efficient boosted c5.0 decision-tree-based classification approach for detecting non-technical losses in power utilities," *Energies*, vol. 13, no. 12, p. 3242, Jun. 2020, doi: 10.3390/en13123242.
[30] J. Han, M. Kamber, and J. Pei, *Data mining: concepts and techniques (The Morgan Kaufmann series in data management systems)*, 3rd ed. Morgan Kaufmann, 2011.
[31] R. Parikh, A. Mathai, S. Parikh, G. Chandra Sekhar, and R. Thomas, "Understanding and using sensitivity, specificity and predictive values," *Indian J. Ophthalmol.*, vol. 56, no. 1, p. 45, 2008, doi: 10.4103/0301-4738.37595.
[32] G. Rao, "Remembering the meanings of sensitivity, specificity, and predictive values," *J Fam Pr.*, vol. 53, no. 1, p. 53, 2004.
[33] M. H. Yamada *et al.*, "Associations of systolic blood pressure and diastolic blood pressure with the incidence of coronary artery disease or cerebrovascular disease according to glucose status," *Diabetes Care*, vol. 44, no. 9, pp. 2124–2131, Sep. 2021, doi: 10.2337/dc20-2252.
[34] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002, doi: 10.1613/jair.953.
[35] F. Thabtah, S. Hammoud, F. Kamalov, and A. Gonsalves, "Data imbalance in classification: experimental evaluation," *Inf. Sci. (Ny).*, vol. 513, pp. 429–441, Mar. 2020, doi: 10.1016/j.ins.2019.11.004.
[36] A. Luque, A. Carrasco, A. Martín, and A. de las Heras, "The impact of class imbalance in classification performance metrics based on the binary confusion matrix," *Pattern Recognit.*, vol. 91, pp. 216–231, Jul. 2019, doi: 10.1016/j.patcog.2019.02.023.
[37] H. C. Çubukçu and D. İ. Topcu, "Estimation of low-density lipoprotein cholesterol concentration using machine learning," *Lab. Med.*, vol. 53, no. 2, pp. 161–171, Mar. 2022, doi: 10.1093/labmed/lmab065.
[38] M. K. Palmer, P. J. Barter, P. Lundman, S. J. Nicholls, P. P. Toth, and B. W. Karlson, "Comparing a novel equation for calculating low-density lipoprotein cholesterol with the friedewald equation: a voyager analysis," *Clin. Biochem.*, vol. 64, pp. 24–29, Feb. 2019, doi: 10.1016/j.clinbiochem.2018.10.011.

BIOGRAPHIES OF AUTHORS

Wiharto received obtained a bachelor's degree in electrical engineering (B.E.) from Universitas Telkom, Indonesia, in 1999. He obtained a master's degree in computer science from Universitas Gadjah Mada, Indonesia, in 2004 and a Doctoral degree from the same University, in 2017. Currently, he works as a lecturer in the Department of Informatics, Faculty of Information Technology and Data Science, Universitas Sebelas Maret, Surakarta, Indonesia. His experience and areas of interest focus on artificial intelligence, computational intelligence, machine learning, and expert systems. He can be contacted at email: wiharto@staff.uns.ac.id.

Farah Nada Mufidah received a bachelor's degree in informatics from the Department of Informatics, Faculty of Information Technology and Data Science, Universitas Sebelas Maret, Surakarta, Indonesia, in 2023. His research interests include data mining, artificial intelligence, machine learning, and computational intelligence. She can be contacted at email: farahnadamufidah@student.uns.ac.id.