Signal and Image Processing Letters
Vol. 1., No. 2, July 2019, pp. 22-32
ISSN 2714-6677
https://simple.ascee.org/

K-Fold Cross Validation for Selection of Cardiovascular Disease Diagnosis Features by Applying Rule-Based Datamining

Dwi Normawati[a,1,*], Dewi Pramudi Ismi [b,2]

[a] Department of Informatics, Universitas Ahmad Dahlan, Yogyakarta, Indonesia
[1] dwi.normawati@tif.uad.ac.id*; [2] dewi.ismi@tif.uad.ac.id
* corresponding author

ARTICLE INFO

Article history
Received April 26, 2019
Revised June 8, 2019
Accepted July 19, 2019

Keywords
Coronary Heart Disease
Feature Selection
K-Fold
VPRS
Cross-validation

ABSTRACT

Coronary heart disease occurs when atherosclerosis inhibits blood flow to the heart muscle in the coronary arteries. This disease is often the cause of human death. The method for diagnosing coronary heart disease that is often a doctor's referral is coronary angiography, but it is invasive, expensive, and high-risk. This study aims to analyze the effect of k-Fold Cross-Validation (CV) on the dataset to create features based on the rules used to diagnose coronary heart disease. This study uses the Cleveland heart disease dataset, where feature selection is performed using a medical expert-based method (MFS) and a computer-based method, Variable Precision Rough Set (VPRS). Evaluation of the classification performance using the k-fold 10-fold, 5-fold and 3-fold methods. The results showed the number of different attribute selection results in each fold, both for the VPRS and MFS methods, with the highest accuracy score in the VPRS method 76.34% with k = 5, while the MTF accuracy was 71.281% with k = 3.

1. Introduction

The number one cause of death in the world is Coronary Heart Disease or Cardiovascular disease (CVD). Every year the number of patients with coronary heart disease increases. Data from the World Health Organization (WHO) recorded that 17.5 million people died in 2012, representing 31% of all global deaths and an estimated 7.4 million caused by Coronary Artery Disease (CAD) [1] or *Ischemic Heart Disease* [2] which is a heart disease that attacks many humans, where the mortality rate of this disease is very high. In 2008 an estimated 7.3 million deaths worldwide were caused by CAD [3].

Medical diagnosis, especially CAD diagnosis is considered a complex task involving many factors and doctors decisions. This is usually done by analysing the result of previous patients tests, while the method used is highly dependent on the knowledge, intuition and experience of the doctor, so that it can affect the service to patients with coronary heart disease, because the decision taken by doctors, have not been done in automated way, so the accuracy of diagnosis hasn't been significant. This practice causes unwanted biases, errors and excessive medical costs [4]. Early diagnosis of coronary heart disease usually uses a patient's medical history, physical examination and then further tests, such as

electrocardiography (ECG), echocardiogram, stress tests, nuclear imaging and coronary angriography [5]. The series of follow-up tests requires carefulness and accuracy of the cardiologist, has a risk to the patient and also requires expensive costs. The difficult process of diagnosing CAD disease is exacerbated by the small number of cardiologists in Indonesia. Along with the development of information technology, the diagnosis of CAD has been developed using computer aided methods.

Research using the VPRS method for the case of coronary heart disease diagnosis has been carried out to find data patterns in the form of rules-based classification [6], which produces fewer rules than the rough set method. whereas the rules produced by VPRS are easier to understand and if the rules are reduced the accuracy value decreases [7]. Research on the diagnosis of coronary heart disease has resulted in an accuracy value of 75.22% with the VPRS classification method [8][9], the trial was carried out by randomizing the data 30 times, but the accuracy performance of each rule is unknown. While the selection of features based on the rules with VPRS for the diagnosis of coronary heart disease has been carried out, in order to select the best features used in the process of diagnosing coronary heart disease [10], this study compared the performance of medical expert-based feature selection (MFS) with computer-based feature selection using the VPRS method, the result of which was an increase in accuracy by selecting features with VPRS compared to diagnoses without feature selection that had been done in previous studies [8][9][11]. The method of selection feature combination of VPRS and MFS, produces fewer Rules compared to MFS, whereas for the accuracy value for VPRS with a combination of VPRS and MFS has the same accuracy value that is 84.84% [10]. However, in this research [10] the treatment of the dataset in the testing process is only by splitting data that is 2/3 data as training data and 1/3 other data as testing data, this is a very large occurrence of data noise because not all data can be tested on the performance of the classification method so that the testing data is limited to just that data.

K-Fold Cross-validation (CV) is a statistical method, where data is divided into two subsets, namely training data for the learning process and data testing for validation or evaluation, which is used to evaluate the performance of models or methods or algorithms. CVs can be selected based on the size of the dataset. Usually K-Fold is used to reduce computing time and also to maintain the accuracy of the estimate [12].

2. Dataset

The dataset used in this study is the Cleveland Heart Disease dataset from the UCI machine learning repository. The amount of data used is 303 data that has 7 missing value data, the missing value data is deleted, so that it does not affect the classification results. The dataset used has 14 attributes and 2 classes, sick and not sick. Table 1 describes the attributes in the Cleveland heart disease dataset [13].

Table 1. Description of Cleveland Heart Disease Dataset

Attribute	Description and Value
Age	Age (Numeric)
Sex	Sex (0:Female; 1:Male)
Cp	Chest pain type **(**1 : typical angina; 2 : atypical angina; 3 : non-anginal pain; 4 : asymptomatic**)**
Trestbps	Resting blood pressure **(**Numeric**)**
Chol	Serum cholesterol **(Numeric)**
Fbs	Fasting blood sugar >120mg/dl **(**0 : false; 1 : true**)**
Restecg	Resting electrocardiographic result **(**0 : normal; 1 : having ST-T wave abnormality; 2 : showing probable or definite left ventricular hypertrophy by Estes' criteria**)**

Thalac	Maximum heart rate achieved **(**Numeric**)**
Exang	Exercise induced angina **(**0 : No; 1 : Yes**)**
Oldpeak	Segment ST depression induced by exercise relative to test **(Numeric)**
Slope	The slope of the peak exercise ST segmen **(**1 : usloping; 2 : flat; 3 : downsloping**)**
Ca	Number of major vessels colored by fluoroscopy **(**0, 1, 2 and 3**)**
Thal	Thal (3 : normal; 6 : fixed defect; 7 : reversible defect)

3. Method

The research method used is there are four main processes consisting of pre-processing, discretization, feature selection, randomization with k-fold cross validation, generating rules and evaluating performance. As shown in the research flow diagram in Figure 1.

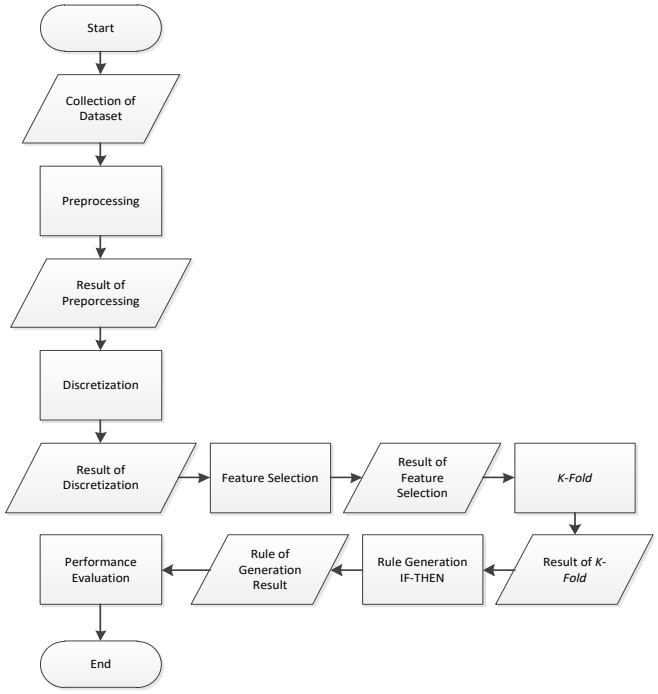

Fig. 1. Research flowchart

3.1. Preprocessing

Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, sc, dc, and rms do not have to be defined. Do not use abbreviations in the title or heads unless they are unavoidable.

3.2. Discretization

- Use either SI (MKS) or CGS as primary units. (SI units are encouraged.) English units may be used as secondary units (in parentheses). An exception would be the use of English units as identifiers in trade, such as “3.5-inch disk drive.”
- Avoid combining SI and CGS units, such as current in amperes and magnetic field in oersteds. This often leads to confusion because equations do not balance dimensionally. If you must use mixed units, clearly state the units for each quantity that you use in an equation.
- Do not mix complete spellings and abbreviations of units: “Wb/m2” or “webers per square meter,” not “webers/m2.” Spell units when they appear in text: “...a few henries,” not “...a few H.”

- Use a zero before decimal points: “0.25,” not “.25.” Use “cm3,” not “cc.” (*bullet list*)

3.3.Feature Selection

Computer based feature selection is done to reduce Cleveland data and also choose features that relevant to decision for the diagnosis of coronary heart disease. This research used two methods to feature selection applied, which are computer based namely VPRS and medical expert based or motivated feature selection (MTF). This research for feature selection with VPRS method uses ROSE2 software.

3.4.Medical expert based feature selection

Medical expert-based feature selection or motivated feature selection (MFS) is based on knowledge possessed by medical experts. In cases of coronary heart disease, medical expert determine eight factors of medical significance that influence on diagnosis process, which are age, chest pain type(angina, abnang, notang, asympt), resting blood pressure, cholesterol, fasting blood sugar, resting heart rate (normal, abnormal, ventricular hypertrophy), maximum heart rate dan exercise induced angina [11][16]. These eight factors are used as result of feature selection based medical expert.

3.5.Variable Precision Rough Set (VPRS)

Computer-based feature selection used VPRS methods. This research used VPRS to feature selection dan classification method.

VPRS is the continuation of classical model of rough set. In this research, it is proposed to analyze and identify the data pattern, which is representing the functional statistic trend [17]. VPRS related to classification of partial precision detection of parameter β. Ziarko defines the value of β as misclassification and ranged in value $0 \leq \beta < 0.5$. Procedures VPRS models have four steps [18], namely :

- step-1: chosen a precision parameter value (β)
- step-2: finding a full set of β-reduce
- step-3: remove duplicate objects
- step-4: rule extraction.

VPRS is an approach to data analysis that relies on two basic concepts, namely β-lower and β-upper approximations which can be expressed in following equation: β-lower approximations of the set in Equation (1), and β-upper approximations of set in Equation (2).

VPRS is an approach to data analysis that relies on two basic concepts, namely β-lower and β-upper approximations which can be expressed in following equation: β-lower approximations of the set in Equation (1), and β-upper approximations of set in Equation (2).

$$\underline{C}_{\beta}(D) = \bigcup_{1-P_r(Z|x_i)\leq\beta} \{x_i \in E(P)\} \quad (1)$$

$$\overline{C}_{\beta}(D) = \bigcup_{1-P_r(Z|x_i)<1-\beta} \{x_i \in E(P)\} \quad (2)$$

Here, $\underline{C}_{\beta}(D)$ and $\overline{C}_{\beta}(D)$ are the lower and upper approximation of D with precision level β, respectively. Where E(P) indicates a set of equivalent classes, and class conditions based on subsets of attributes P, while $Z \subset E(D)$. Mathematically formulated in Eq. 3.

$$P_r(Z \mid x_i) = \frac{Card(Z \cap x_i)}{Card(x_i)} \quad (3)$$

According to [17], the size of classification quality for VPRS models can be defined by the following equation:

$$\gamma(P,D,\beta) = \frac{Card(\bigcup_{1-P_r(Z|x_i)\leq\beta}\{x_i \in E(P)\})}{Card(U)} \quad (4)$$

Where $Z \subset E(D)$ dan $P \subseteq C$, for certain β value. The value of equation (4) measures the proportion of objects on set universe (U) for classification based on decision attribute D, and allowing for certain β value.

The procedure to produce a decision rule of an information system is done by two major steps as follows:

- Step 1 : Selection of the best smallest set of attributes (eg, β-reduct value election)
- Step 2 : Simplification of information systems can be achieved by dropping the specific values of attributes.

Ziarko [17] indicates that every smallest set of attributes is considered as an alternative to group attributes that are used as substitute all attributes available in case based decision making.

3.6.K-Fold Cross Validation

The next process is randomization of the dataset using the k-Fold cross validation method for data testing to evaluate the performance of the VPRS method. The dataset is divided into 'k 'subsets with the same amount of data. This research will use 10-fold, 5-fold and 3-fold. The data is divided into 10 folds that are approximately the same size for each fold, so they have 10 data subsets. For each of the 10 data subsets, the Cross-Validation test will use 9-fold for training and 1-fold for testing as illustrated in Figure 2. The method is also done for 5-fold and 3-fold.

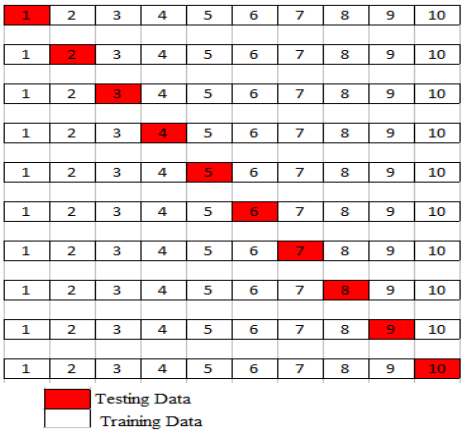

Fig. 2. Scheme 10-fold cross validation

3.7.Generate Rules IF-THEN

This research uses VPRS method to generate rules which by using ROSE2 software.

3.8.Performance Evaluation

Medical Performance evaluation is done by doing classification. Evaluation is done by analyzing confusion matrix [19], which consist of accuracy, sensitivity and specificity. Confusion matrix is shown in table 2.

Table 2. Confusion Matriks

	Predicted Positive	**Predicted Negative**
Actual Positive	True Positive (TP)	False Negative (FN)

Actual Negative	False Positive (FP)	True Negative (TN)

The accuracy is the success rate of classification or classification accuracy was measured by counting the number of correct classifications divided by the total classification. Sensitivity is the probability the patient said to suffer from coronary heart disease was diagnosed positive illness (Sick), while specificity was diagnosed negative illness (Health).

4. Results and Discussion

In this study 296 data were used taken from the Cleveland Heart Disease dataset. Data discretization is carried out by the ROSETTA software. Feature selection is done to choose features that are relevant to the results of the diagnosis of coronary heart disease. The feature selection process and the rule making process for the VPRS method are carried out using the ROSE2 software. Classification process is calculated manually by using Microsoft Excel.

4.1.Preprocessing

The first step in the preprocessing data is cleaning data process. Cleaning data is removed missing value data in dataset Cleveland, and then converted multiclass dataset into binary class dataset with assumed that positive class is healthy (0) and negative class is sick (1).

The second step is discretization data. Discretization changes the data type of attributes from numeric into discrete. Some attributes with type of numeric which have Cleveland dataset are age, trestbps, chol, thalach, oldpeak and ca, they are transformed into discrete type using Entropy/MDL algorithm. Table 3 shows the result of discretization data.

The last step in preprocessing data is splitting data. Splitting data process split dataset into two parts. Dataset is split into two part. The three-quarter of data is used to train data and the other is used to test data. Splitting data is done to split dataset into training and testing datasets. Training datasets is used to find rules and knowledge on dataset for diagnosing coronary heart disease. While, testing dataset is used to test data with matching class prediction result of rules knowledge with class dataset.

Table 3. The Result Of Data Discretisation

	Numeric type attributes					
	Age	***Trestbps***	***Chol***	***Thalach***	***Oldpeak***	***Ca***
DISCRETE VALUES	[*, 71) [71, 77) [77, *)	[*, 186) [186, *)	[*, 276) [276, 277) [277, 280) [280, 295) [295, 299) [299, 301) [301, 319) [319, 320) [320, 322) [322, 324) [324, 326) [326, 338) [338, 341) [341, 342) [342, 348) [348, 354) [354, 401) [401, 413) [413, *)	[*, 148) [148, 151) [151, 162) [162, 170) [170, 172) [172, 175) [175, 176) [176, 178) [178, 183) [183, 195) [195, 199) [199, *)	[*, 2.5) [2.5, 2.7) [2.7, 3.1) [3.1, 3.5) [3.5, 3.6) [3.6, 4.3) [4.3, *)	[*,3) [3, *)

4.2.Feature Selection

The first step in feature selection process uses training dataset to reduce data and tested according to the features that have been selected. Table IV shows the result of feature selection with MTF and VPRS methods [13]. This process is the same as done in previous studies.

Table 4. Rule with vprs classifier

Feature selection methods	Feature selection result
MTF	age, cp, trestbps, chol, fbs, restecg, thalach, exang
VPRS	cp, chol, restecg, thalach, exang, oldpeak, slope, thal
MTF+VPRS	age, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, thal

4.3.K-Fold Cross-Validation

The next step is randomizing the dataset with the k-Fold validation method for testing data to evaluate the performance of the VPRS method. The dataset is divided into ‘k ’subsets with the same amount of data. This research will use 10-fold, 5-fold and 3-fold subsets.

For K = 10, the data is divided into 10-folds that are approximately the same size for each fold, so that they have 10 data subsets. For each of the 10 data subsets, Cross Validation testing will use 9-fold for training and 1-fold for testing, as shown in Figure 3.

10-Fold	Dataset (1-296 record)									
	1	2	3	4	5	6	7	8	9	10
1	1 - 30	31 - 60	61 - 90	91 - 120	121 -150	151 - 180	181 - 210	211 - 240	241 - 270	271 - 296
2	1 - 30	31 - 60	61 - 90	91 - 120	121 -150	151 - 180	181 - 210	211 - 240	241 - 270	271 - 296
3	1 - 30	31 - 60	61 - 90	91 - 120	121 -150	151 - 180	181 - 210	211 - 240	241 - 270	271 - 296
4	1 - 30	31 - 60	61 - 90	91 - 120	121 -150	151 - 180	181 - 210	211 - 240	241 - 270	271 - 296
5	1 - 30	31 - 60	61 - 90	91 - 120	121 -150	151 - 180	181 - 210	211 - 240	241 - 270	271 - 296
6	1 - 30	31 - 60	61 - 90	91 - 120	121 -150	151 - 180	181 - 210	211 - 240	241 - 270	271 - 296
7	1 - 30	31 - 60	61 - 90	91 - 120	121 -150	151 - 180	181 - 210	211 - 240	241 - 270	271 - 296
8	1 - 30	31 - 60	61 - 90	91 - 120	121 -150	151 - 180	181 - 210	211 - 240	241 - 270	271 - 296
9	1 - 30	31 - 60	61 - 90	91 - 120	121 -150	151 - 180	181 - 210	211 - 240	241 - 270	271 - 296
10	1 - 30	31 - 60	61 - 90	91 - 120	121 -150	151 - 180	181 - 210	211 - 240	241 - 270	271 - 296

Keterangan :
: Data Testing
: Data Training

Fig. 3. Randomize 10-fold cross validation

For K = 5, the data is divided into 5-folds that are approximately the same size for each fold, so that they have 5 subsets of data. For each of the 5 data subsets, Cross Validation testing will use 4-fold for training and 1-fold for testing, as shown in Figure 4.

5-Fold	Dataset (1-296 record)				
	1	2	3	4	5
1	1 - 59	60 - 118	119 -177	178 - 236	237 - 296
2	1 - 59	60 - 118	119 -177	178 - 236	237 - 296
3	1 - 59	60 - 118	119 -177	178 - 236	237 - 296
4	1 - 59	60 - 118	119 -177	178 - 236	237 - 296
5	1 - 59	60 - 118	119 -177	178 - 236	237 - 296

Keterangan :
: Data Testing
: Data Training

Fig. 4. Randomize 5-fold cross validation

For K = 3, the data is divided into 3 folds that are approximately the same size for each fold, so that they have 3 subsets of data. For each of the 3 data subsets, Cross Validation testing will use 2-fold for training and 1 fold for testing, as shown in Figure 5.

3-Fold	Dataset (1-296 record)		
	1	2	3
1	1 - 99	100 - 198	199 - 296
2	1 - 99	100 - 198	199 - 296
3	1 - 99	100 - 198	199 - 296

Keterangan :
: Data Testing
: Data Training

Fig. 5. Randomize 3-fold cross validation

4.4.Generate Rule IF-THEN

The next step is generating IF-THEN rules by using result feature selection datasets with MTF and VPTS, which have randomized with k-fold cross validation.

4.5.Variable Precision Rough Set (VPRS)

In order to get IF-THEN rules or decision rules for VPRS method, the value β = 0.15 is used by using ROSE2 software [19]. In the research work, 3 datasets are used which resulted from the feature selection process. Each dataset produces different rules and numbers. Table VIII, IX, X, XI, and XII shows the number of rules result datasets.

4.6.Classification

Classification based on VPRS is done. The resulting rules are tested on a test dataset that has been randomized by the k-fold cross validation method. The test is applied into the test dataset from the result feature selection dataset, so a confusion matrix is obtained for each dataset.

4.7.Performance Evaluation

In the medical contexts, there are only two classes “sick” or “healthy”, which “sick” is more important than “healthy”. The medical diagnosis purpose is to focus on the improvement of the accuracy of “sick” class or sensitivity and maintain the accuracy of “healthy” class or specificity. The accuracy, sensitivity and specificity values can be calculated from confusion matrix for each method by using Equation (5) to Equation (7).

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} x100\% \quad (5)$$

$$Sensitivity = \frac{TP}{TP + FN} x100\% \quad (6)$$

$$Specificity = \frac{TN}{TN + FP} x100\% \quad (7)$$

Table 5 to Table 10 shows the result performance evaluation from datasets feature selection MFS and VPRS. From tables show that Classification performance by implementing k-fold for feature selection of the VPRS method is better than the MFS method. The feature selection method with VPRS that implements K-Fold Cross Validation on the evaluation of classification performance produces the highest accuracy of 76.34% at k = 5 subset dataset, because this research performs the k-fold randomization phase where all records in the dataset have a role as data testing and also training data, so all data plays a role in the process of generating rules and also performance evaluation, but randomization is still structured as many folds have been determined. So, The results of diagnosis of coronary heart

disease by implementing k-fold cross validation with feature selection using the VPRS method have decreased the accuracy value compared to diagnosis with feature selection without k-fold implementation [11].

Table 5. MFS feature selection results for Age, Cp, Trestbps, Chol, Fbs, Restecg, Thalach, Exang attributes with 10-fold

K-Fold	Number of Rules 10-fold	Accuracy	Sensitivity	Specificity
1	77	57,14	63,64	46,15
2	85	64,71	72,22	56,25
3	84	62,86	66,67	57,14
4	87	71,43	66,67	76,47
5	83	52,94	55,56	50
6	83	48,48	61,90	25
7	83	55,88	71,43	30,77
8	80	51,52	66,67	25
9	81	66,67	80,95	41,67
10	89	47,06	63,64	16,67
	Average	**57,87**	**66,93**	**42,51**

Table 6. MFS feature selection results for Age, Cp, Trestbps, Chol, Fbs, Restecg, Thalach, Exang attributes with 5-fold

K-Fold	Number of Rules 5-fold	Accuracy	Sensitivity	Specificity
1	67	58,33	81,82	29,63
2	83	76,27	90	62,07
3	83	67,80	80,65	53,57
4	72	54,24	82,86	12,50
5	70	70,31	85,71	51,72
	Average	**65,39**	**84,21**	**41,90**

Table 7. MFS feature selection results for Age, Cp, Trestbps, Chol, Fbs, Restecg, Thalach, Exang attributes with 3-fold

K-Fold	Number of Rules 3-fold	Accuracy	Sensitivity	Specificity
1	62	70,71	79,25	60,87
2	66	72,73	83,33	60
3	59	70,41	90,57	46,67
	Average	**71,28**	**84,38**	**55,85**

Table 8. VPRS feature selection results for Cp, Chol, Restecg, Thalach, Exang, Oldpeak, Slope, Thal attributes with 10-fold

K-Fold	Number of Rules 10-fold	Accuracy	Sensitivity	Specificity
1	70	62,50	60	66,67
2	72	65,71	63,16	68,75
3	73	60	66,67	50
4	76	80	72,22	88,24
5	69	71,43	72,22	70,59
6	71	68,57	76,19	57,14
7	77	70,59	81,82	50
8	64	48,48	60	30,77
9	73	69,70	64,71	75
10	72	61,76	68,18	50
	Average	**65,87**	**68,52**	**60,72**

Table 9. VPRS feature selection results for Cp, Chol, Restecg, Thalach, Exang, Oldpeak, Slope, Thal attributes with 5-fold

K-Fold	Number of Rules 5-fold	Accuracy	Sensitivity	Specificity
1	79	78,33	81,82	74,07
2	80	86,44	90	82,76
3	76	76,27	90,32	60,71
4	65	67,80	82,86	45,83
5	65	72,88	87,10	57,14
	Average	**76,34**	**86,42**	**64,10**

Table 10. VPRS feature selection results for Cp, Chol, Restecg, Thalach, Exang, Oldpeak, Slope, Thal attributes with 3-fold

K-Fold	Number of Rules 3-fold	Accuracy	Sensitivity	Specificity
1	74	79,80	84,91	73,91
2	68	70,71	87,04	51,11
3	54	72,45	84,91	57,78
	Average	**74,32**	**85,62**	**60,93**

5. Conclusion

The feature selection method with VPRS that implements K-Fold Cross Validation on the evaluation of classification performance produces the highest accuracy of 76.34% at k = 5 subset dataset. This happens because this research performs the k-fold randomization phase where all records in the dataset have a role as data testing and also training data, so all data plays a role in the process of generating rules and also performance evaluation, but randomization is still structured as many folds have been determined.

Classification performance by implementing k-fold for feature selection of the VPRS method is better than the MFS method. Implementation of k-fold for the diagnosis of heart disease by the VPRS method still results in lower accuracy since the distribution of the k-fold subset is only 10-fold, 5-fold, and 3-fold. The comparison result from testing process shows that the results of diagnosis of coronary heart disease by implementing k-fold cross validation with feature selection using the VPRS method have decreased the accuracy value compared to diagnosis with feature selection without k-fold implementation.

Acknowledgment

This research is supported by LPPM Universitas Ahmad Dahlan research

References

[1] O. S. R. N.M. Segerson And D.S. Romaine, *The Encyclopedia of The Heart Diseases*, 2nd ed. 2010.

[2] Arthur Selzer, *Understanding Heart Disease*. 1992.

[3] S. Mendis, P. Puska, and B. Norrving, "Global atlas on cardiovascular disease prevention and control," *World Heal. Organ.*, pp. 2–14, 2011.

[4] M. Kumari and S. Godara, "Comparative Study of Data Mining Classification Methods in Cardiovascular Disease Prediction," *Ijcst*, vol. 4333, no. 2229, pp. 304–308, 2011.

[5] B. Phibbs, *A Basic Guide to Heart Disease*. Philadelphia: Lippincott Williams & Wilkins, 2007.

[6] R. P. Sanjaya, "Deteksi Penyakit Jantung Koroner Menggunakan Model Variable Precision Rough Set dan Logika Fuzzy," University of Gadjah Mada, 2014.

[7] B. . Tripathy, D. . Acharjya, and V. Cynthya, "A Framework for Intelligent Medical Diagnosis Using Rough Set with Formal Concept Analysis," *Int. J. Artif. Intell. Appl.*, vol. 2, no. 2, pp. 45–66, 2011.

[8] D. Normawati, "Diagnosis penyakit jantung koroner menggunakan penambangan data berbasis variable precision rough set (vprs) dan repeated incremental pruning to produce error reduction (ripper)," university of gajah mada, 2015.

[9] H. A. Nugroho, D. Normawati, N. A. Setiawan, and W. K. Z. Oktoeberza, "Rule-Based Data Mining for Diagnosis of Coronary Heart Disease," *JTEC*, vol. 9, no. 3, pp. 93–97, 2017.

[10] D. Normawati *et al.*, "Data Berbasis Variable Precision Rough Set (Vprs) Untuk Diagnosis Penyakit Jantung," vol. 3, no. 2, 2017.

[11] D. Normawati and S. Winarti, "Feature selection with combination classifier use rules-based data mining for diagnosis of coronary heart disease," *Proceeding 2018 12th Int. Conf. Telecommun. Syst. Serv. Appl. TSSA 2018*, pp. 2–7, 2019.

[12] S. Iii, "K -Fold Cross-Validation," 2009.

[13] UCI, "Heart Disease Dataset," 2017. [Online]. Available: https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/. [Accessed: 24-Mar-2017].

[14] Dwi Wahyu Prabowo, "SELEKSI FITUR BERBASIS KOMPUTER UNTUK DIAGNOSIS PENYAKIT JANTUNG KORONER," University of Gadjah Mada, 2014.

[15] Fathul Ihsan and Noor Akhmad Setiawan, "Perbandingan Metode Diskretisasi Untuk Berbagai Macam Algoritma Machine Learning," University of Gadjah Mada, 2013.

[16] T. Herawan, W. Maseri, W. Mohd, and A. Noraziah, "Applying Variable Precision Rough Set for Clustering Diabetics Dataset."

[17] W. Ziarko, "Probabilistic Decision Tables in the Variable Precision Rough Set Model," *Comput. Intell.*, vol. 17, no. 3, pp. 593–603, 2001.

[18] C. T. Su and J. H. Hsu, "Precision parameter in the variable precision rough sets model: An application," *Omega*, vol. 34, no. 2, pp. 149–157, 2006.

[19] Jinwei Han and Michaline Kamber, *Data Mining, south asia edition : Concept and Technology*. Morgan Kaufmann, 2006.