

DEVELOPING AN INTEGRATED MODEL OF CLINICAL DECISION SUPPORT SYSTEM FOR THE EARLY DETECTION OF CORONARY HEART DISEASES

Jenifer Ghai

ABSTRACT

Objectives: This study aims to propose a model of coronary heart disease assessment system based on risk factors. Methods/Statistical Analysis: To achieve these objectives, the model proposed system comprises several processes. First, the dimension reduction using principle component analysis (PCA). Second, classification using support vectormachine. Third, validate using 10-fold cross validation in the process of training and testing. Training and testing using patient data from the Hospital Dr. Moewardi Solo Indonesia, which amounted to 120 with 12 attributes. Fourth, the system performance is measured using several parameters, namely sensitivity, specificity, area under the curve, positive prediction value, negative prediction value and accuracy. Findings: Tests on the proposed system, to process the dimensional reduction by PCA, which is followed by the method of orthogonal rotation with verimax, resulting nine attributes of the 12 attributes of risk factors. Attribute are obtained by using the variance of data from the PCA process of 71.1908%. Attribute risk factors rotation outcome is age, gender, occupation rate, total cholesterol, low-density lipoprotein, triglycerides, systolic and diastolic blood pressure and smoking. System performance prediction is generated for parameter sensitivity 84.20%, specificity 69.09%, accuracy 78.61%, positive prediction value 82.53%, negative prediction value 71.70% and the area under the curve 76.64%. Applications/Improvements: System model of clinical decision support system for the assessment of coronary heart disease based on risk factors can be used by clinicians, as support in making clinical decisions. The proposed system provides the performance of the medium category.

1. INTRODUCTION

In 2014 in Southeast Asia, particularly in Indonesia, the death rate recorded 35% or about 1.8 million cases of deaths due to heart disease. Active smokers will impact very badly on the health of the heart and stroke. World Health Organization (WHO) states that in the world every 6.5 seconds a person dies due to active smoking. It is estimated that a person with an active lifestyle smoked for less than 20 years or more, will die more quickly from heart disease compared to those who do not smoke. This is reinforced in study 1, that a good lifestyle, will have an impact on quality of life, especially health. Coronary heart disease is so dangerous, so precaution is very important 2. Prediction of coronary heart disease events is an important step that must be done to reduce the high number of deaths from the disease. Predictions can be done by assessing the risk factors for coronary heart disease. Risk factors for coronary heart disease can be divided into two groups, the first non-modifiable and modifiable 3. Non-modifiable factors such as age, gender, and heredity. While the modifiable include hypertension, stress, diabetes mellitus, high cholesterol, smoking, and others. Prediction of heart disease can be performed using models such as the Framingham Risk Score. In addition, the development of information technology also brought about changes in the development of coronary heart disease prediction system. It is also reinforced

research conducted by 4 showed that the use of clinical decision support system is able to provide improvements in clinical practice a doctor. Clinical decision support system for prediction of coronary heart disease has been developed using either predictive models such as the Framingham Risk Score, PROCAM and SCORE, also using data mining. Predictions using models, sometimes not appropriate for a particular country, because the models developed for specific populations, such as the Framingham risk score was developed for the population of the United States, Australia and New Zealand⁵. The authors² in his research explained that the Framingham Risk Score and PROCAM not suitable for Korean population, so that in her research to develop coronary heart disease prediction models using data mining². The risk factors used in the study were age, total cholesterol, LDL, HDL, systolic and diastolic blood pressure, sex, smoking and diabetes. Data mining algorithms used are Fuzzy Inference System (FIS). The performance of the system in comparison with the algorithm of Artificial Neural Network (ANN), Support Vector Machine (SVM), logistic regression and C5.0. The comparison results showed FIS better, but the resulting performance difference is not very significant, if compared with SVM and ANN. A similar study has also been developed by⁶, only in the study in the fuzzy rule base using the algorithm C4.5 generates during the training process⁶. The next prediction system proposed by⁷ which uses 12 risk factors for coronary heart disease. The risk factors together with research by^{2,6}, coupled with the use of alcohol attribute, physical activity, diet, obesity, stress and the presence or absence of coronary heart disease descendants. This system was developed using a combination of genetic algorithms and ANN. Genetic algorithms are used to determine the initial weight ANN, resulting ANN training process much faster to achieve convergent⁷. Subsequent research conducted by⁸. In the study to investigate the incidence of Myocardial Infarction (MI), Percutaneous Coronary Intervention (PCI), Coronary Artery Bypass Graft surgery (CABG). Risk factors were used for the investigation was divided into two, namely before and after the incident. Before the occurrence of risk factors are age, gender, family history of coronary heart disease, history of smoking, history of hypertension and diabetes history. Factor after the events are smoking, total cholesterol, High Density Lipoprotein (HDL), Low Density Lipoprotein (LDL), triglycerides, blood glucose, systolic and diastolic blood pressure. Based on these risk factors, an analysis using the algorithm C4.5. The analysis resulted in that MI is influenced by age, smoking, and a history of hypertension. PCI was influenced by family history of CHD, history of hypertension, history of diabetes. While CABG influenced by age, history of hypertension and smoking⁸. The next study the prediction of coronary heart disease with reference to the use of the dataset of the University of California Irvine (UCI). UCI datasets using 13 attributes consisting of risk factors (age, sex, blood pressure, cholesterol and fasting blood sugar), symptoms, Examination Electrocardiography (ECG), scintigraphy and Fluoroscopy⁹. The authors¹⁰ proposed a system of prediction of coronary heart disease by using a combination of genetic algorithm and ANN, the combination of different combinations performed⁷. This study uses a GA to determine the weights while the ANN to determine the fitness of the chromosomes. The process is carried out continuously until a certain generation which resulted the best chromosome. The best chromosomes are used as weights on ANN, ANN is subsequently used to predict coronary heart disease¹⁰. The authors¹¹ also proposed a system of prediction, but there are stages attribute preprocessing to reduce coronary heart disease. The method used is the Principle Component Analysis (PCA). PCA is able to reduce 13 attributes of coronary heart disease to 7 attribute, for the next algorithms using Adaptive Neuro Fuzzy Inference System (ANFIS) for

predicting coronary heart disease [11]. The authors [12] propose a diagnosis system that combines the k-mean clustering, Weighted Association Classifier (WAC) and C5.0 decision tree. K-mean algorithm used to perform clustering in the learning process in the WAC, the next in the classification by using the algorithm C5.0. The use of these three methods capable diagnosis system gives improved performance, compared to just using the algorithm C5.0. Attribute dimension reduction of coronary heart disease is also conducted in the research [13]. The study proposes the use of artificial bee colony to perform dimension reduction. Artificial bee colony can reduce attribute from 13 to 5 and 7 attributes. Attribute reduction results in the analysis using SVM algorithm, the results of the analysis, to attribute 7 is able to provide better performance than 5 attributes. Attribute of the reduction results, if viewed from a group of risk factors are age, cholesterol and fasting blood sugar [13]. Similar studies conducted by [14], on the research dimension reduction using PCA and data mining algorithms SVM. Dimension reduction carried attribute 9 of 13 attributes, the resulting performance, best by using a Kernel function in SVM radial basis function (RBF). SVM use for the diagnosis of coronary heart disease is also carried by [15], a study comparing naïve Bayesian algorithm, SVM, IBK, AdaboostM1, J48 (C4.5) and PART. The resulting performance of SVM algorithm is better than the other, with the testing method k-fold cross validation [15]. This is also supported in research [16, 17] that the SVM is able to deliver performance better accuracy compared ANN, Naive Bayesian, Bayesian network for the diagnosis of heart disease. Research by [18], also perform the analysis of system performance diagnosis using SVM. SVM's performance parameter AUC provide better value compared to some algorithms such as MLP, Radial Basis Function (RBF) and the MLP with the dimensional reduction. Based on previous investigations, this study proposes a clinical decision support system for predicting coronary heart disease based on risk factors. Risk factors are analysed draws on data from the Hospital Dr. Moewardi Solo Indonesia. The data used has 12 risk factors. The risk factor will be analysed using PCA dimension reduction algorithm. Furthermore, after dimension reduction, performed with Verimax orthogonal rotation, to get the risk factors after the transformation. The risk factors obtained from the rotation, then classified using SVM algorithm. The final step of testing the method of k-fold cross validation. The parameters used to measure performance parameters which are commonly used in the medical world, namely sensitivity, specificity, positive prediction value, negative prediction value, accuracy and area under the curve.

2. METHODS

This research was conducted by several steps, namely data collection, pre-processing, classification and analysis of the testing results. These steps can also be described in a flowchart, as shown in Figure 1. Data collection was done by taking the collection of data from the outpatient clinic Dr. Moewardi Hospital at Solo Indonesia. The data used were 120 patients with 12 variable risk factors for heart disease. The 12 variables are shown in Table 1. The next steps are pre-processing. In this step was make the selection of the 12 variables to determine which variable has a high influence on coronary heart disease? It is done by using the Principle Components of the Analysis (PCA). Principal Component Analysis is one method of dimension reduction in feature extraction group (transformation). This method performs the reduction by transforming data into a

new dimension. Following the method algorithm 19 Calculate the covariance matrix of the data by using Equation 1 below

$$\text{cov}(xy) = \frac{\sum xy}{n} - (\bar{x})(\bar{y}) \tag{1}$$

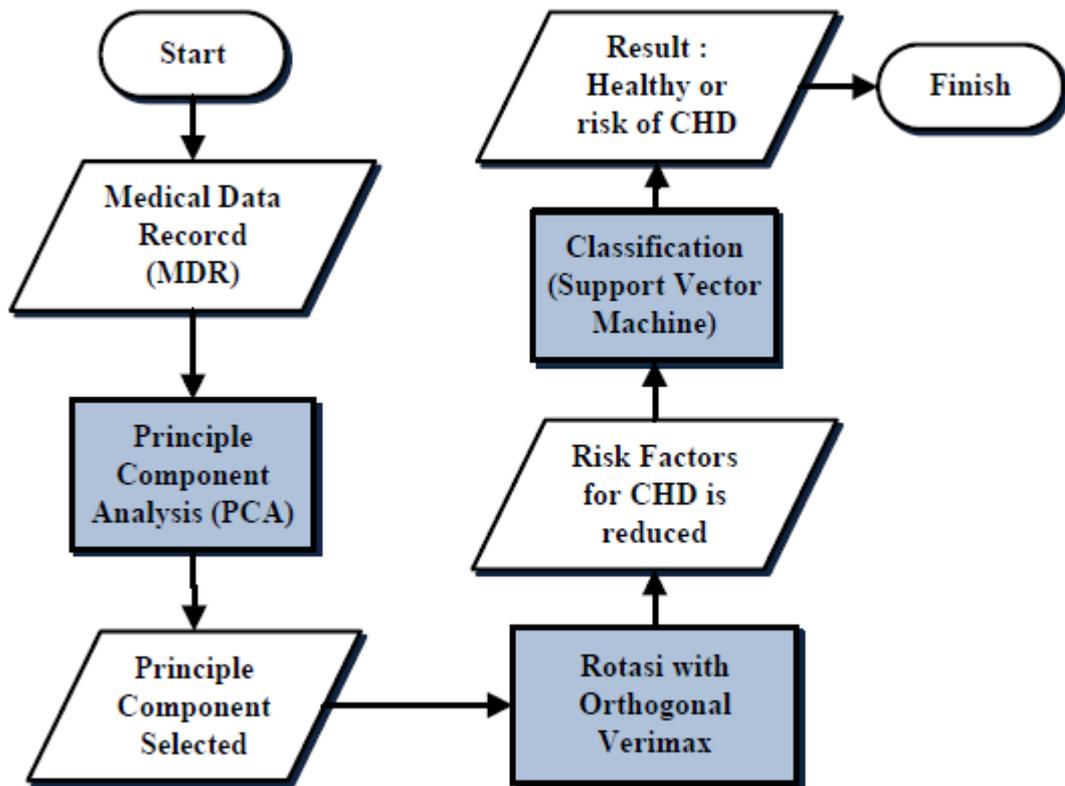


Figure 1. Flowchart system.

Table 1. Risk factors for coronary heart disease

No.	Parameters	Category	No. (%)	Mean±SD
1	Age			60,37±13,84
2	Gender	Male	70 (58,3)	
		Famale	50 (41,7)	
3	Occupation rate	Low	45 (37,5)	
		Medium	61 (50,8)	
		High	14 (11,7)	
4	Cholesterol Total			171,24±45,28
5	LDL			112,17±35,82

6	HDL			34,61±11,09
7	Triglyceride levels			122,45±80,00
8	Sistolic blood pressure			132,68±25,69
9	Diastolic blood pressure			85,14±14,84
10	Obesity			21,41±2,61
11	Smooking	Yes	25(20,8)	
		No	95(79,2)	
12	History of diabetes	Yes	13 (10,8)	
		No	107 (89,2)	

Calculate the Eigen value by solving Equation 2 and eigenvectors by using the Equation 3.

$$(\mathbf{S} - \lambda\mathbf{I}) = \mathbf{0} \quad (2)$$

$$[\mathbf{S} - \lambda\mathbf{I}][\mathbf{X}] = [\mathbf{0}] \quad (3)$$

Calculate new variables (principal component) by multiplying the original variables with the matrix of eigenvectors value. Principal component selected by looking eigenvalue which is worth more than 1. The results of the PCA is having a new variable which have different dimension to the original variable. To determine the variables which included in the new variable (principle component) done by rotate the factors using varimax vector. The results of the rotation will indicate variables that influence the occurrence of coronary heart disease. After the selection the next is scaling selected variables into a scale of 0-1 (normalization). The next step is classification, the process consists of training and testing. Classification system that will be used is a binary classification model which consist of two outputs, a healthy and risk of coronary heart disease. Classification method used is by using Support Vector Machine (SVM) [20,21]. As for training and testing using the data of 120 outpatients Dr. Moewardi Hospital at Solo Indonesia. Training and testing method is by using 10-fold cross validation. The method works by dividing the data into 10 subsets randomly, then 9 subset for training and one subset for testing, the process is performed 10 times at random. Each subset alternately so that all subsets are used for testing. The end result is obtained from the experiment 10 times, then analyzed using Metrics confusion, as shown in Table 2.

$$\text{Accuracy} = (TP+TN)/(TP+FP+TN+FN)$$

$$\text{Sensitivity} = TP/(TP+FN)$$

$$\text{Specificity} = TN/(TN+FP)$$

$$\text{Positive Prediction Value (PPV)} = TP/(TP+FP)$$

$$\text{Negative Prediction Value (NPV)} = TN/(TN+FN)$$

$$\text{Area Under the Curve (AUC)} = (\text{Sensitivity} + \text{Specificity})/2$$

Table 2. Confusion matrix

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

3. RESULTS

The results of research are explained into two parts. They are dimension reduction and testing result. The first section dimension reduction using a feature extraction method Principle Component Analysis (PCA). The data which will be reduced in the reduction of dimensions is data from Dr. Moewardi Hospital with 12 variables. Steps of dimension reduction is done by using equation 1-3, from the step produced a variant of a new variable, which is called the principle component 1-12, as shown in Figure 2. The new variable number of 12 were selected based on the value of the variant

for each principle component, theselected variables are variables that have a value greaterthan 1. Thus, the new variable is the principle componentsselected 1-5.

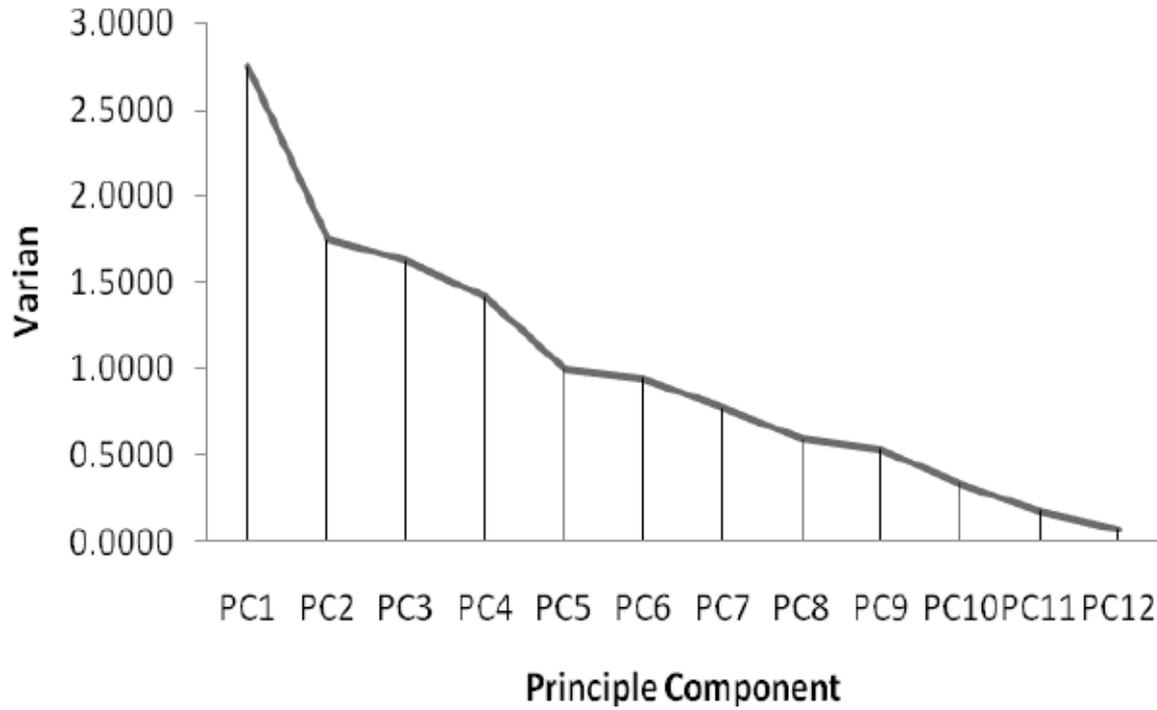


Figure 2. Variance principle component.

Table 3. Rotation with varimax method of principal component variables to risk factors

No	Variabel	Principle Component				
		1	2	3	4	5
1	Age	0,0348	-0,6757	-0,0396	0,0357	-0,0081
2	Gender	0,0279	-0,0263	-0,0870	0,5873	0,0987
3	Occupation rate	0,0161	0,6671	-0,0669	-0,0351	-0,0466
4	Cholesterol Total	0,6434	-0,0211	-0,0081	-0,0072	0,0916
5	LDL	0,6324	0,0240	-0,0272	0,0509	-0,0431
6	HDL	0,3575	-0,0783	0,0674	-0,1889	-0,5243
7	Trygliserida Levels	0,1868	0,0130	0,0094	-0,0056	0,7572
8	Sistolic Blood Pressure	-0,0393	-0,0423	0,6383	-0,1150	-0,0030
9	Diastolic Blood Pressure	-0,0299	-0,0351	0,6623	0,0172	-0,0021
10	Obesity	0,1343	0,2420	0,3563	0,2831	0,0642
11	Smooking	-0,0263	-0,0886	0,0602	0,6809	-0,1346
12	History of Diabetes	0,0007	-0,1446	0,0655	-0,2386	0,3276

Results of principal component analysis is a new dimension that is different from the origin variables, so we do not know the origin of the reduced variable by the PCA process. To determine the variables which are reduced, then the rotation is done using Varimax. Results rotation with reference to the principle of selected components 1-5, which mapped by original variables are shown in Table 3. The value of the new variable mapping with the old variable is called the loading rate. Loading value of greater value than 0.5, then the old variable is a variable that affects the coronary heart disease. Results of mapping a new variable with the original variable, can be obtained variables that influence the occurrence of coronary heart disease as shown in Table 4. There are three variables that do not significantly affect the occurrence of coronary heart disease, which is a history of diabetes, HDL and obesity rates. Use the original variable 9 or 5 new variable (principle component) showed that the principle component 1-5 gives the effect of 71.1908% of the variance of the data. In total effect for each principle component shown in Table 4.

Table 4. List of variables influencing the occurrence of CHD

Principle Component	Varian (%)	Variable
1	22,8917 %	Cholesterol total LDL
2	14,5983 %	Age Occupation rate
3	13,5467 %	Sistolic Blood Pressure Diastolic Blood Pressure
4	11,7925 %	Gender Smooking
5	8,3617 %	Trygliserida Levels
Sum	71,1908 %	

The second part shown by the test resulted using 10-fold cross validation. Performance variables measured include sensitivity, specificity, accuracy, true positive and true negative. The test results are shown in Figure 3. Based on Figure 3 we can calculate the average for each performance parameter. They are sensitivity 84.20%, specificity 69.09%, accuracy 78.61%, positive prediction value of 82.53%, negative prediction value 71.70% and area under the curve 76,65%.

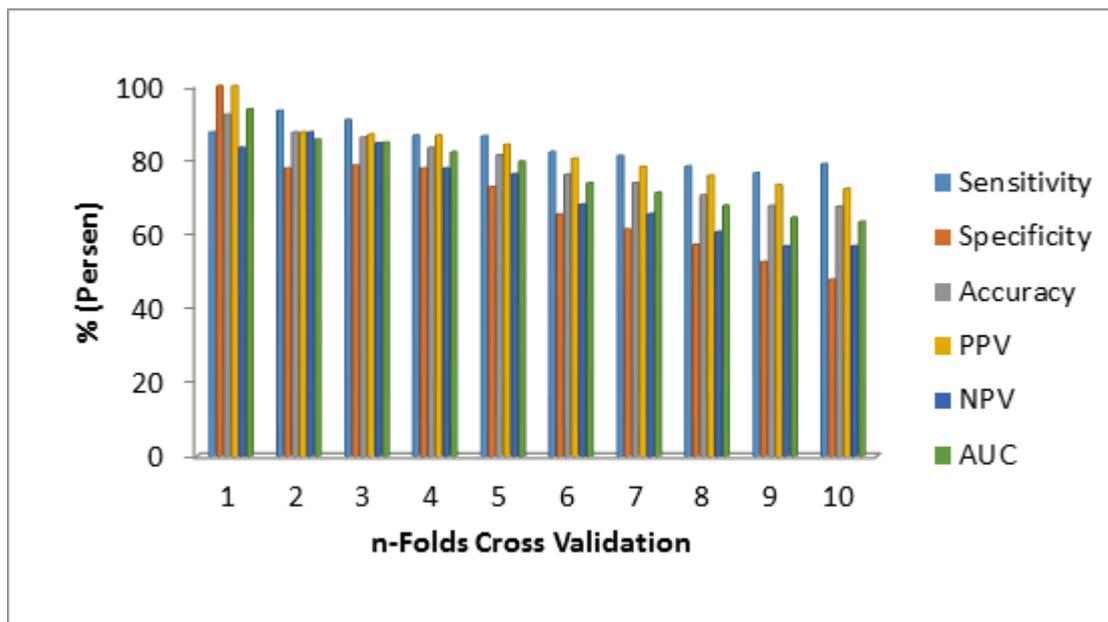


Figure 3. Testing result of 10-folds cross validation.

4. DISCUSSIONS

Based on these results, we can discuss about several things related variables influencing the occurrence of coronary heart disease, and the performance of the system diagnosis coronary heart disease. The results of the PCA process produces 5 principle component. Principle component-1 consists of total and LDL cholesterol. Principle component-1 gives contribution to 22.8917% of the variance of the data CHD. Referring to equation 4 consists of total cholesterol LDL, HDL and tryglyserida²³. HDL is the good cholesterol, which is the higher value is better. LDL and HDL values influence each other, if the LDL cholesterol increases, it will lower HDL cholesterol levels. As for tryglyserida levels, also have contributed to the LDL, if tryglyserida up then going down LDL. Decreasing the value of LDL is only 20% of the value tryglyserida. PCA results showed that tryglyserida contribute to the variance of data coronary heart disease by 8.3617%, which is represented in the principle component-5.

$$\text{LDL} = \text{Cholesterol}_{\text{total}} - \text{HDL} - \frac{1}{5} \text{Triglyceride} \quad (4)$$

Principle component-2 consists of age and occupation rate. Both variables contributed 14.5983% of the data variance of coronary heart disease. Growing the age, a person more susceptible to coronary heart disease, but rarely cause serious disease before 40 years old and 5-fold increase in the age of 40-60 years²⁴. While variable occupation rate will affect the level of stress. Stress has a relationship with lipid metabolism abnormality²⁵. In addition, stress also stimulates the cardiovascular system with the release of catecholamine that increase heart rate and cause vasoconstriction. Besides dealing with stress, occupation rate can also be associated with the work too much sitting, and lack of exercise. A study led by the 26, clarified that the development, for example, high impact exercise will help improvement of coronary illness hazard factors, for example, LDL, HDL and absolute cholesterol. It shows occupation rate would influence hazard factors for coronary illness more. Head segment 3 comprises of systolic and diastolic pulse, giving the impact of 13.5467% of the difference of the information. Pulse is a hazard factor that is legitimately identified with the rate of coronary illness, for each decline in diastolic circulatory strain by 5 mmHg, the danger of coronary illness was decreased 15%²⁵. Next variables smoking and gender, including two factors are included in the principle component of 4 on the incidence of coronary disease. Smoking is a major risk factor for heart disease and have a close relationship with the occurrence of coronary heart disease, so that by quitting smoking will reduce the risk of heart attack²⁷. Even cigarette smoking increases the risk of heart attack between 2-3 times. Approximately 24% of deaths from coronary heart disease in men and 11% in women due to smoking habits²⁵. Men have a greater risk of heart attack and it happened earlier than in women²⁷. Morbidity disease coronary heart disease in men two times greater than in women, and this condition occurs almost 10 years earlier in men than women²⁵. Clinical decision support systems assessment of coronary heart disease has included in the fair performance classification. This study compared with some previous studies. The first study conducted by 7, the study proposes a hybrid system with genetic algorithm neural network. The resulting performance that accuracy for testing 92% and for validation 89%⁷. The accuracy of the

proposed system reached 78.61%, but the results of accuracy generated by the proposed system is the average yield of the method 10-fold cross validation. This is different to that carried out by 7, the study is not the result of a number of experiments, but only one-time trial, so it is possible if done the test again with different data might be change in the extreme up or down. The next weakness is related to the use of neural network. Neural network does not generate global optimum solution, but locally, so if trained with repeating with the same data would produce different result. In contrast to the SVM which generate global optimum solution, in addition the SVM is a powerful binary classification method 15. Composition data which will be used in research 7 is 34 data for training which consist of 8 validations and 8 testing.

Table 5. Comparison of the performance of the proposed system with previous research

Author	Method	Sensitivity	Specificity	AUC	Accuracy
7	GA+ANN				89%
28	Optimasi Fuzzy	80%	65,2%	72,6%	73,4%
28	ANFIS	58,2%	55,1%	56,6%	56,8%
28	ANN	80%	66,3%	73,1%	73,9%
29	Framingham Risk Score	100%	2,8%	51,4%	64,65%
29	ANN	85,7%	52,8%	69,2%	73,74%
2	Fuzzy Logic	93,1%	25,64%	59,4%	69,51%
2	SVM	89,66%	26,92%	58,3%	67,71%
Proposed	PCA+SVM	84.20%	69.09%	76,6%	78,61%

The next comparison, the research conducted by 28. The study proposed an automatic diagnosis system based on fuzzy system optimization using 10-fold cross which validated. The system does not perform variable reduction, so there is no elimination of the variables determining the case of coronary heart disease (CHD). Performance was measured by three parameters, namely sensitivity 80.0%, specificity 65.2% and accuracy of 73.4% 28. In addition to the system of fuzzy optimization method, the study also tested for ANFIS method and Artificial Neural Network with 10-fold cross validation. The results obtained sensitivity 58.2%, specificity 55.1% and 56.8% for ANFIS accuracy and sensitivity of 80 %, specificity 66.3% and accuracy of 73.9% for the neural network. The third method gives performance lower when compared with the proposed system with a combination of PCA and SVM. Comparison with previous studies can be summarized in a table, as shown in Table 5. Research conducted by 29 proposed ANN for prediction of coronary heart disease. On the stretch research proved that ANN is better at predicting than using the Framingham risk

score, it is shown by AUC values. The study also confirmed by 2 which explains that the Framingham Risk Score does not correspond to the Korean population. The lack adjust the Framingham Risk Score was due to be used for the prediction of coronary heart disease events in the population of the United States, Australia and New Zealand⁵. Referring to previous studies shown in Table 5, with the parameters AUC performance demonstrates the use of data mining algorithms are relatively better compared to Framingham Risk Score for prediction of coronary heart disease. In addition, by using the performance parameters AUC, showed that the proposed system, capable of providing a relatively better performance. It seen from the comparison of several data mining algorithms used in previous studies.

5. CONCLUSION

Clinical choice emotionally supportive networks for evaluation of coronary illness by utilizing PCA and SVM can convey execution at a reasonable degree of arrangement. The exhibition by estimating the precision of 78.61% variable, affectability 84.20%, particularity 69.09% and 76,64% AUC. The utilization of PCA can be utilized to diminish the component of CHD hazard factors. Dimension reduction produces⁹ of coronary heart disease risk factors, namely total cholesterol, LDL, age, occupation rate, blood pressure (systolic and diastolic), gender, smoking and tryglicerida.