(MIJ) 2019, Vol. No. 5, Jan-Dec **e-ISSN: 2454-924X; p-ISSN: 2454-8103**

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 usingpatient data from the Hospital Dr. Moewardi Solo Indonesia, which amounted to 120 with 12 attributes. Fourth, thesystem performance is measured using several parameters, namely sensitivity, specificity, area under the curve, positiveprediction value, negative prediction value and accuracy. Findings: Tests on the proposed system, to process thedimensional 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 processof 71.1908%. Attribute risk factors rotation outcome is age, gender, occupation rate, total cholesterol, low-densitylipoprotein, triglycerides, systolic and diastolic blood pressure and smoking. System performance prediction is generatedfor parameter sensitivity 84.20%, specificity 69.09%, accuracy 78.61%, positive prediction value 82.53%, negativeprediction value 71.70% and the area under the curve 76.64%. Applications/Improvements: System model of clinicaldecision 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, thedeath rate recorded 35% or about 1.8 million cases ofdeaths due to heart disease. Active smokers will impactvery badly on the health of the heart and stroke. WorldHealth Organization (WHO) states that in the worldevery 6.5 seconds a person dies due to active smoking. Itis estimated that a person with an active lifestyle smokedfor less than 20 years or more, will die more quickly fromheart disease compared to those who do not smoke. Thisis reinforced in study1, that a good lifestyle, will have an impact on quality of life, especially health. Coronary heartdisease is so dangerous, so precaution is very important2.Prediction of coronary heart disease events is animportant step that must be done to reduce the high numberof deaths from the disease. Predictions can be done byassessing the risk factors for coronary heart disease. Riskfactors for coronary heart disease can be divided into twogroups, the first non-modifiable and modifiable3. Nonmodifiablefactors such as age, gender, and heredity. Whilethe modifiable include hypertension, stress, diabetes mellitus,high cholesterol, smoking, and others. Prediction ofheart disease can be performed using modelssuch as theFramingham Risk Score. In addition, the development ofinformation technology also brought about changes inthe development of coronary heart disease prediction system.It is also reinforced research conducted by 4 showedthat the use of clinical decision

support system is able toprovide improvements in clinical practice a doctor.Clinical decision support system for predictionof coronary heart disease has been developed usingeither predictive models such as the Framingham RiskScore, PROCAM and SCORE, also using data mining.Predictions using models, sometimes not appropriatefor a particular country, because the models developedfor specific populations, such as the Framingham riskscore was developed for the population of the UnitedStates, Australia and New Zealand5. The authors2 in hisresearch explained that the Framingham Risk Score andPROCAM not suitable for Korean population, so that inher research to develop coronary heart disease predictionmodels using data mining2. The risk factors used in thestudy were age, total cholesterol, LDL, HDL, systolic anddiastolic blood pressure, sex, smoking and diabetes. Datamining algorithms used are Fuzzy Inference System (FIS).The performance of the system in comparison with thealgorithm of Artificial Neural Network (ANN), SupportVector Machine (SVM), logistic regression and C5.0. Thecomparison results showed FIS better, but the resultingperformance difference is not very significant, if comparedwith SVM and ANN. A similar study has also beendeveloped by6, only in the study in the fuzzy rule baseusing the algorithm C4.5 generates during the trainingprocess6.The next prediction system proposed by 7 which uses12 risk factors for coronary heart disease. The risk factorstogether with research by 2,6, coupled with the use ofalcohol attribute, physical activity, diet, obesity, stress andthe presence or absence of coronary heart disease descendants.This system was developed using a combinationof genetic algorithms and ANN. Genetic algorithmsare used to determine the initial weight ANN, resultingANN training process much faster to achieve convergent7.Subsequent research conducted by 8. In the study toinvestigate the incidence of Myocardial Infraction (MI),Percutaneous Coronary Intervention (PCI), CoronaryArtery Bypass Graft surgery (CABG). Risk factors wereused for the investigation was divided into two, namelybefore and after the incident. Before the occurrence ofrisk factors are age, gender, family history of coronaryheart disease, history of smoking, history of hypertensionand diabetes history. Factor after the events are smoking,total cholesterol, High Density Lipoprotein (HDL),Low Density Lipoprotein (LDL), triglycerides, bloodglucose,systolic and diastolic blood pressure. Based onthese 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 influencedby family history of CHD, history of hypertension,history of diabetes. While CABG influenced by age, historyof hypertension and smoking8.The next study the prediction of coronary heartdisease with reference to the use of the dataset of theUniversity of California Irvine (UCI). UCI datasetsusing 13 attributes consisting of risk factors (age, sex,blood pressure, cholesterol and fasting blood sugar),symptoms, Examination Electrocardiography (ECG),scintigraphy and Flouroscopy9. The authors10 proposed asystem of prediction of coronary heart disease by using acombination of genetic algorithm and ANN, the combinationof different combinations performed 7. This studyuses a GA to determine the weights while the ANN todetermine the fitness of the chromosomes. The processis carried out continuously until a certain generationwhich resulted the best chromosome. The best chromosomesare used as weights on ANN, ANN is subsequentlyused to predict coronary heart disease10. The authors11 also proposed a system of prediction, but there are stagesattribute preproccessing to reduce coronary heart disease.The method used is the Principle Component Analysis(PCA). PCA is able to reduct 13 attributes of coronaryheart disease to 7 attribute, for the next algorithms usingAdaptive Neuro Fuzzy Inference System (ANFIS) for predictingcoronary heart disease11. The authors12 proposesa diagnosis system that combines the k-mean clustering, Weighted Association Classifiern (WAC) and

C5.0 decisiontree. K-mean algorithm used to perform clusteringin the learning process in the WAC, the next in the classificationby using the algorithm C5.0. The use of thesethree methods capable diagnosis system gives improvedperformance, compared to just using the algorithm C5.0.Attribute dimension reduction of coronary heartdisease is also conducted in the research13. The study proposes the use of artificial bee colony to perform dimensionreduction. Artificial bee colony can reduce etribute from13 to 5 and 7 attributes. Attribute reduction results in theanalysis using SVM algorithm, the results of the analysis,to attribute 7 is able to provide better performance than5 attributes. Attribute of the reduction results, if viewedfrom a group of risk factors are age, cholesterol and fastingblood sugar13. Similar studies conducted by 14, onthe research dimension reduction using PCA and datamining algorithms SVM. Dimension reduction carriedattribute 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 heartdisease is also carried by 15, a study comparing naïve Bayesian algorithm, SVM, IBK, AdaboostM1, J48 (C4.5)and PART. The resulting performance of SVM algorithmis better than the other, with the testing method k-foldcross validation15. This is also supported in research16,17that the SVM is able to deliver performance better accuracycompared ANN, Naive Bayesian, Bayesian networkfor the diagnosis of heart disease.Research by 18 , also performthe analysis of system performance diagnosis usingSVM. SVM's performance parameter AUC provide bettervalue compared to some algorithms such as MLP, RadialBasis Function (RBF) and the MLP with the dimensionalreduction.Based on previous investigations, this study proposesa clinical decision support system for predicting coronaryheart disease based on risk factors. Risk factors areanalysed draws on data from the Hospital Dr. MoewardiSolo Indonesia. The data used has 12 risk factors. The riskfactor will be analysed using PCA dimension reductionalgorithm. Furthermore, after dimension reduction, performed with Verimax orthogonal rotation, to get the riskfactors after the transformation. The risk factors obtainedfrom the rotation, then classified using SVM algorithm.The final step of testing the method of k-fold cross validation.The parameters used to measure performanceparameters which are commonly used in the medicalworld, namely sensitivity, specificity, positive predictionvalue, negative prediction value, accuracy and area underthe curve.

2. METHODS

This research was conducted by several steps, namely datacollection, pre-processing, classification and analysis ofthe testing results. These steps can also be described in aflowchart, as shown in Figure 1. Data collection was doneby taking the collection of data from the outpatient clinic Dr. Moewardi Hospital at Solo Indonesia. The data usedwere 120 patients with 12 variable risk factors for heartdisease. The 12 variables are shown in Table 1.The next steps arepre-processing. In this step was makethe selection of the 12 variables to determine which variablehas a high influence on coronary heart disease? It isdone by using the Principle Components of the Analysis(PCA).Principal Component Analysis is one method ofdimension reduction in feature extraction group (transformation).This method performs the reduction bytransforming data into a new dimension. Following themethod algorithm19Calculate the covariance matrix of the data by usingEquation 1 below

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

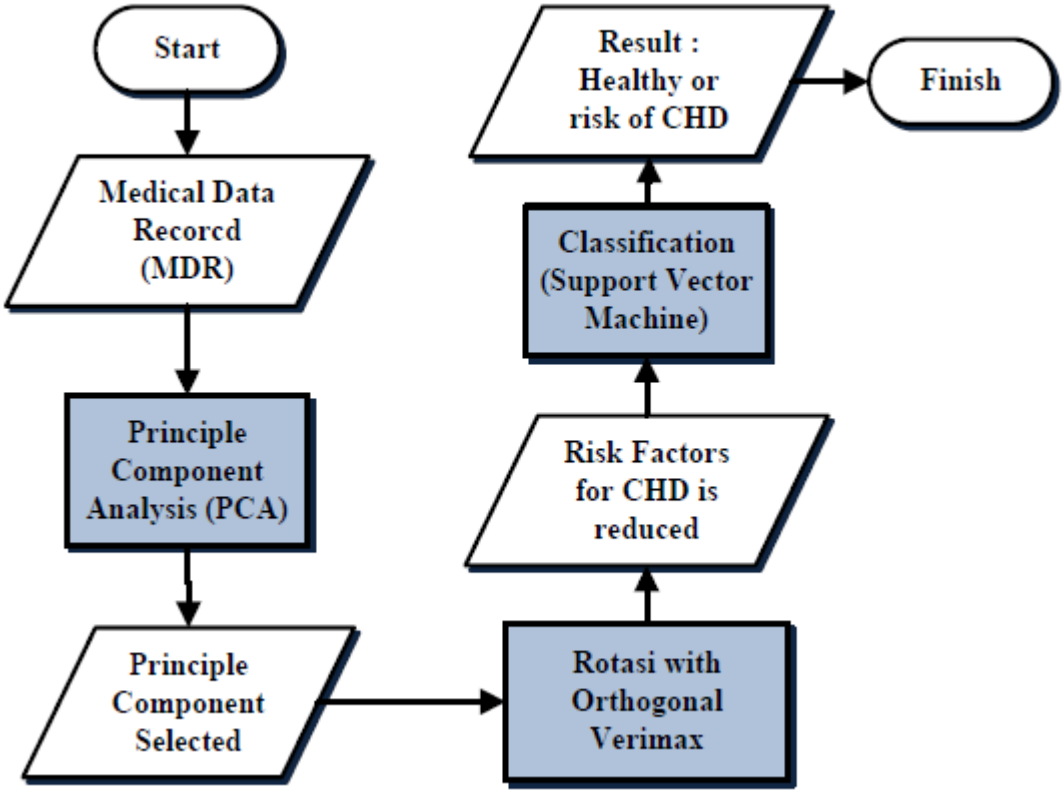

Figure 1. Flowchart system.

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.

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

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

Calculate new variables (principal component) bymultiplying the original variables with the matrix ofeigenvectors value.Principal component selected by looking eigenvaluewhich is worth more than 1.The results of the PCA is having a new variablewhich have different dimension to the original variable.To determine the variables which included in the newvariable (principle component) done by rotate the factorsusing varimax vector. The results of the rotation willindicate variables that influence the occurrence of coronaryheart disease. After the selection the nextis scaling selected variables into a scale of 0-1 (normalization).The next step is classification, the process consistsof training and testing. Classification system that willbe used is a binary classification model which consist oftwo outputs, a healthy and risk of coronary heart disease.Classification method used is by using Support VectorMachine (SVM)20,21. As for training and testing usingthe data of 120 outpatients Dr.Moewardi Hospital at SoloIndonesia. Training and testing method is by using10-fold cross validation. The method works by dividingthe data into 10 subsets randomly, then 9 subset for trainingand one subset for testing, the process is performed 10times at random. Each subset alternately so that all subsetsare used for testing. The end result is obtained fromthe experiment 10 times, then analyzed using Metrics confusion, as shown in Table 2.

Accuracy = (TP+TN)/(TP+FP+TN+FN)
Sensitivity = TP/(TP+FN)
Specificity = TN/(TN+FP)
Positive Prediction Value (PPV) = TP/(TP+FP)
Negative Prediction Value (NPV) = TN/(TN+FN)
Area Under the Curve (AUC) = (Sensitivity + Specificity)/2

Table 2. Confusion matric

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. Theyare dimention reduction and testing result. The first sectiondimension reduction using a feature extractionmethod Principle Component Analysis (PCA). The data which will be reducted in the reduction of dimensions is data from Dr.Moewardi Hospital with 12 variables. Stepsof dimension reduction is done by using equation 1-3,from the step produced a variant of a new variable, whichis called the principle component 1-12, as shown in Figure2. The new variable number of 12 were selected based onthe 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 componentselected 1-5.

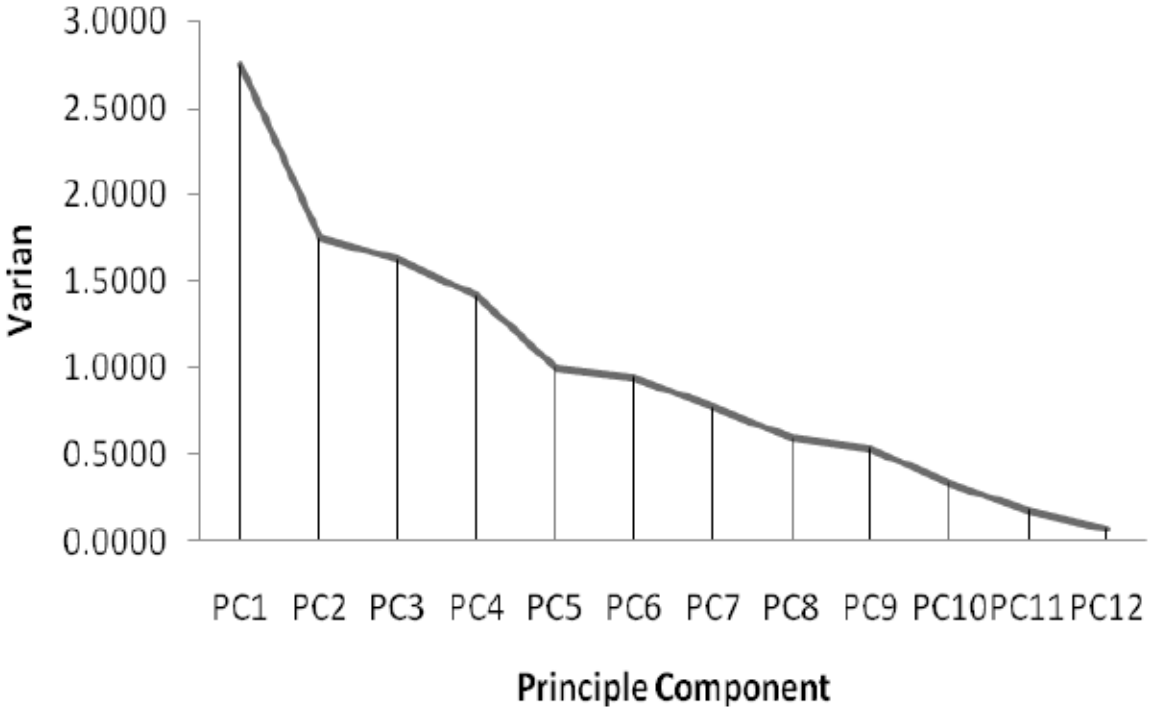

Figure 2. Varian principle component.

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 newdimension that is different from the origin variables, sowe do not know the origin of the reduced variable bythe PCA process. To determine the variables which arereduced, then the rotation is done using Varimax. Resultsrotation with reference to the principle of selected components1-5, which mapped by original variables areshown in Table 3. The value of the new variable mappingwith the old variable is called the loading rate. Loadingvalue of greater value than 0.5, then the old variable is avariable that affects the coronary heart disease.Results of mapping a new variable with the originalvariable, can be obtained variables that influence theoccurrence of coronary heart disease as shown in Table4. There are three variables that do not significantly affectthe occurrence of coronary heart disease, which is a historyof diabetes, HDL and obesity rates. Use the originalvariable 9 or 5 new variable (principle component)showed that the principle component 1-5 gives the effectof 71.1908% of the variance of the data. In total effect foreach 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 using10-fold cross validation. Performance variables measuredinclude sensitivity, specificity, accuracy, true positive andtrue negative. The test results are shown in Figure 3. Basedon Figure 3 we can calculate the average for each performanceparameter. They are sensitivity 84.20%, specificity69.09%, accuracy 78.61%, positive prediction value of 82.53%, negative prediction value 71.70% and area underthe curve 76,65%.

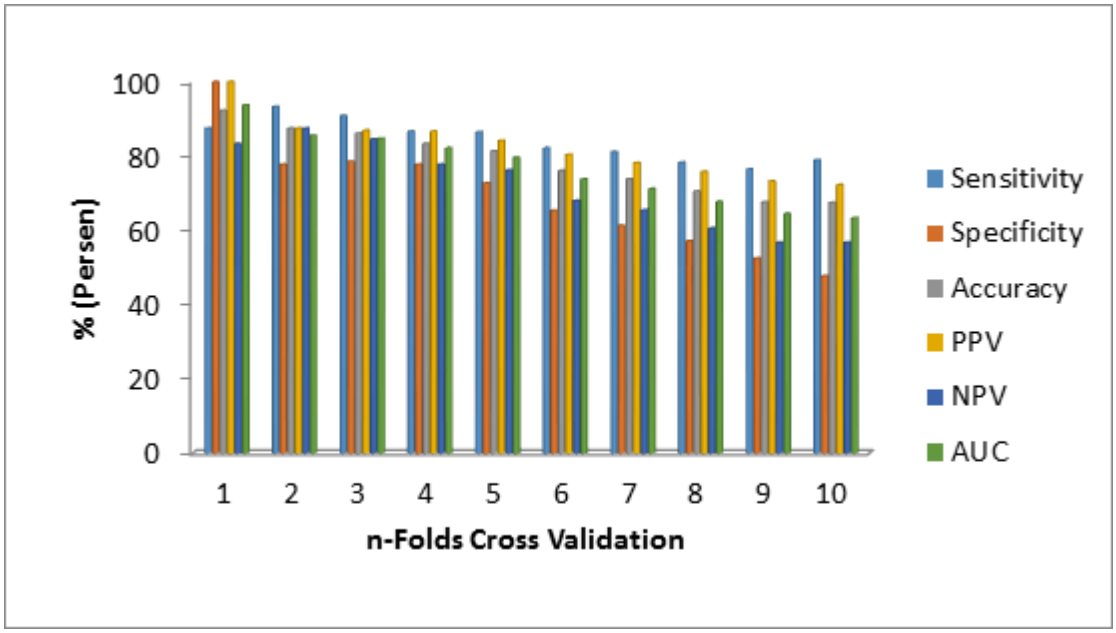

Figure 3. Testing result of 10-folds cross validation.

4. DISCUSSIONS

Based on these results, we can discuss about severalthings related variables influencing the occurrence of coronaryheart disease, and the performance of the systemdiagnosis coronary heart disease. The results of the PCAprocess produces 5 principle component. Principle component-1 consists of total and LDL cholesterol. Principlecomponent-1 gives contribution to 22.8917% of the varianceof the data CHD. Referring to equation4 consists oftotal cholesterol LDL, HDL and tryligserida23. HDL is thegood cholesterol, which is the higher value is better. LDLand HDL values influence each other, if the LDL cholesterolincreases, it will lower HDL cholesterol levels. Asfor trygliserida levels, also have contributed to the LDL,if trygliserida up then going down LDL. Decreasing thevalue of LDL is only 20% of the value tryligserida. PCAresults showed that tryligserida contribute to the varianceof data coronary heart disease by 8.3617%, which is representedin the principle component-5.

$$\mathbf{LDL = Cholesterol_{total} - HD - \frac{1}{5}Triglyceride} \quad \ldots \quad (4)$$

Principle component-2 consists of age and occupationrate. Both variables contributed 14.5983% of thedata 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 and5-fold increase in the age of 40-60 years24. While variableoccupation rate will affect the level of stress. Stress hasa relationship with lipid metabolism abnormality25. Inaddition, stress also stimulates the cardiovascular systemwith the release of catecholamine that increase heart rateand cause vasoconstriction. Besides dealing with strees,occupation rate can also be associated with the work toomuch 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 factorsare included in the principle component of 4 on the incidenceof coronary disease. Smoking is a major risk factorfor heart disease and have a close relationship with theoccurrence of coronary heart disease, so that by quittingsmoking will reduce the risk of heart attack27. Even cigarettesmoking increases the risk of heart attack between2-3 times. Approximately 24% of deaths from coronaryheart disease in men and 11% in women due to smokinghabits 25. Men have a greater risk of heart attack andit happened earlier than in women27. Morbidity diseasecoronary heart disease in men two times greater than inwomen, and this condition occurs almost 10 years earlierin men than women 25.Clinical decision support systems assessment of coronaryheart disease has included in the fair performanceclassification. This study compared with some previousstudies. The first study conducted by 7, the study proposesa hybrid system with genetic algorithm neural network.The resulting performance that accuracy for testing 92%and for validation89%7. The accuracy of the

proposed systemreached 78.61%, but the results of accuracy generatedby the proposed system is the average yield of the method10-fold cross validation. This is different to that carriedout by 7, the study is not the result of a number of experiments,but only one-time trial, so it is possible if donethe test again with different data might be change in theextreme up or down. The next weakness is related to theuse of neural network. Neural network does not generateglobal optimum solution, but locally, so if trained with repeating with the same data would produce differentresult. In contrast to the SVM which generate global optimumsolution, in addition the SVM is a powerful binaryclassification method 15. Composition data which will beused 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. Thestudy proposed an automatic diagnosis system based onfuzzy system optimization using 10-fold cross which validated.The system does not perform variable reduction,so there is no elimination of the variables determiningthe case of coronary heart disease (CHD). Performancewas measured by three parameters, namely sensitivity80.0%, specificity 65.2% and accuracy of 73.4%28. Inaddition to the system of fuzzy optimization method,the study also tested for ANFIS method and ArtificialNeural Network with 10-fold cross validation. The resultsobtained sensitivity 58.2%, specificity 55.1% and 56.8%for ANFIS accuracy and sensitivity of 80 %, specificity66.3% and accuracy of 73.9% for the neural network. Thethird method gives performance lower when comparedwith the proposed system with a combination of PCA andSVM. Comparison with previous studies can be summarizedin a table, as shown in Table 5.Research conducted by 29 proposed ANN for predictionof coronary heart disease. On the stretcher researchproved that ANN is better at predicting than using theFramingham risk score, it is shown by

AUC values.The study also confirmed by 2 which explains that theFramingham Risk Score does not correspond to theKorean population. The lack adjust the FraminghamRisk Score was due to be used for the prediction of coronaryheart disease events in the population of the UnitedStates, Australia and New Zealand5. Referring to previousstudies shown in Table 5, with the parameters AUC performancedemonstrates the use of data mining algorithmsare relatively better compared to Framingham Risk Scorefor prediction of coronary heart disease. In addition, byusing the performance parameters AUC, showed that theproposed system, capable of providing a relatively betterperformance.It seen from the comparison of several datamining 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, smooking and tryligserida.