Journal of Engineering Science and Technology
Vol. 16, No. 2 (2021) 1370 - 1385

CORONARY HEART DISEASE USING SUPPORT VECTOR MACHINE

OKFALISA[1,*], LESTARI HANDAYANI[1,2], DINDA JUWITA P.[1], MUHAMMAD AFFANDES[1], S. S .M. FAUZI[3], SAKTIOTO[4]

[1]Faculty Science and Technology, Universitas Islam Negeri Sultan Syarif Kasim Riau, No. 155, Jalan HR. Soebrantas Panam Km.15, Kabupaten Kampar, 28293, Riau, Indonesia
[2]Lab Prisme, Insa Centre Val de Loire, Bourges, France
[3]Software Engineering Research Group, Universiti Teknologi MARA, Malaysia
[4]Faculty of Mathematics and Natural Sciences, Universitas Riau, Riau, Indonesia
*Corresponding Author: okfalisa@gmail.com

Abstract

The preference of SVM kernel function with optimal features that flexibly applied for dynamic dataset is a new challenge. The restriction of technology and infrastructure support for diagnosing the bioinformatics at rural area is a major concern for developing countries towards excellent health services. Therefore, this study aimed at evaluating the utilization of Support Vector Machine (SVM) in classifying patients of coronary heart disease with Unstable Angina Pectoris (UAP), Non-Segment (ST) Elevation Myocardial Infarction (NSTEMI) and ST-Elevation Myocardial Infarction (STEMI) classes. So far, 280 samples were experimented with 17 attributes by considering four types of dataset, which include the original, reduced, pre-processing and K-Nearest Neighbours (k-NN). To evaluate the optimal parameter pairs in terms of accuracy and processing time for the above dataset types, 10-folds cross-validation and percentage split were carried out on Polynomial and Radial Basis Function (RBF) kernels. Waikato Environment for Knowledge Analysis (WEKA) tool for 10-folds reveals the optimum accuracy of 100% for polynomial kernel and 98.9% for RBF. Also, the percentage split of 70:30 affirms 100% accuracy with 0.06 seconds of processing time as the ideal values of Polynomial kernel test. Meanwhile, RBF exhibits 80:20 split for 100% accuracy with 0.08 seconds in dataset pre-processing. In a nutshell, SVM enhances the data precision and recall as well as minimizes the error possibility for the greatest classification of coronary heart disease patients in Polynomial and RBF kernel than another classifier such as Neural Network (NN). Therefore, the application of SVM improves the accuracy of coronary heart disease diagnostics.

Keywords: Coronary heart diseases, Data mining, K-nearest neighbours, Neural network, Support vector machine.

1. Introduction

Data mining provides various manipulation services to achieve the prediction, classification, clustering, mapping, and anomalous detection of data. The utilization of this technique in various disciplines has evolved and shown a significant contribution to the field of knowledge, including medicine, finance, industry, technology, and even molecular biology as well as bioinformatics. With an emphasis on classification, the advent of methods in disaggregation data improves its usefulness and manoeuvrability in interpreting information, for examples Nijssen and Fromont [1] studied the optimal constraint of Decision tree method induction in pattern mining; Network and Tree-based methods were applied for data mining modeling in the corrosion of concrete sewer [2]; k-NN for scholarship recipient cases [3]; Multilayer Perceptron (MPL) for data mining in healthcare operations [4]; Naive Bayes approach in classifying the analysis of students' performance [5], Artificial Neural Network as a validation tool of Loud Haul Dump (LHD) machine performance characteristics [6], Neural Network in designating the water cycle problems [7] and the utilization of SVM in data mining [8].

Recently, the enforcement of the above methods in analysing the complex bioinformatics data was put into practice. Big data opportunities bring unprecedented potential and challenges in data mining and biological analysis systems in a cost-efficient manner [9]. Also, big data technology ensures that the biologist generates large number of facts and measurement of genomic sequences, images of physiological structures, measuring the messenger Ribonucleic Acid (mRNA) and protein expression, transcription factor binding, and metabolite concentration with limitation of programming skills [10]. In addition, Majhi et al [11] utilized bioinformatics techniques to identify the early stages of diseases such as metabolic and urea cycle disorders, inborn errors and path-aligners through genetics analysing processes and proteomics reports, which are therefore compared with health care data. Furthermore, Dashtban and Balafar [12] found the significance of data mining as artificial intelligent tools in classifying the microarray cancer data.

The adoption of machine learning algorithms in bioinformatics accomplished the reduction of complex data and allocated the feature selection of biomarkers in raw data. Serra et al. [13] verified the successful employment of machine learning techniques as well as clustering, classification, embedding techniques and network-based approaches in addressing bioinformatics problems which include gene expression clustering, patient classification, brain network analysis, and identification of biomarkers. In addition, this technology's ability to capture biomedical data has reformed machine learning into a sophisticated way to solve the complexities of big data. The number of heterogeneity modalities in biological and neurobiological phenomena insists on the multi-view of intelligent data integration from several resources. In addition, multi-view learning and data integration offers greater statistical power analysis [14]. In the process of improvising classification parameters, especially in predicting bioinformatics data, a high level of precision is required to produce the best and most effective classifier tool. The classification techniques that involve data mining, as well as machine learning, reduce computational time and improve categorization precision in determining the optimum values as clarifying in the case of unknown protein sequence classification [15].

SVM is a classification method that produces a fairly high degree of accuracy and is commonly used compared with the conventional decision tree, ANN [16, 17] and other classifiers [18]. Furthermore, Sivagami [19] compared SVM, Multilayer Perception (MLP), One R, and Decision Tree J48 methods in the classification of breast cancer. The results showed that SVM with kernel type RBF provided the highest accuracy rate of 95%, 91% in polynomial type, and 90% in linear type. One R exhibited 83%, 80% in J48 and 74.1% in MLP, which is the lowest performance. The comparison of SVM and Left Anterior Descending (LDA) for the classification of Coronary heart showed accuracy at 96.86% and 78.18%, respectively [20]. Furthermore, Mo and Xu [21] attempted to improve the performance of SVM based on the hybrid kernel function using the optimization of the Particle Swarm Optimization (PSO) algorithm in heart disease diagnosis. Meanwhile, the accuracy of SVM in the early diagnosis of a heart condition by modifying the kernel width using trial and error approach significantly increase by 18.2% [22]. This showed that the kernel function on SVM provides the opportunities in enhancing the accuracy. Unfortunately, some difficulties in choosing the SVM kernel function were encountered [23], as well as flexibility in dataset changing [24], selecting optimal features, and time-consumption [25].

Exploring the various advantages and problems in the SVM method, this study aimed at identifying Coronary Heart Disease (CHD) problems in Indonesia cases. This is interesting since most hospitals are equipped with standard technological equipment in terms of features, functions, and images in interpreting a disease, especially in rural areas. Simultaneously, medical doctors are required to provide a proper diagnostic and ensure appropriate treatment for patients. Furthermore, taking advantage of SVM in Bioinformatics, 280 CHD patients at Central Hospital in Pekanbaru, Riau were classified into three class outputs namely UAP, NSTEMI, and STEMI. By emphasizing on the kernel function impact [22, 23], the classification of Polynomial and RBF kernels using one-against-one multiclass SVM was analysed and investigated to explore the optimality of SVM accuracy. Therefore, this classification aids the diagnosis of CHD patients at an early stage flawlessly, to begin further treatment and also in clinical decision making [26].

The organization of this study begins with an introduction that explains the background, previous reviews on the SVM method, the objectives, the research work, and implications. Furthermore, detailed data, instruments, and step processes are elucidated in the research method. The output of Knowledge Discovery and Data mining (KDD) and SVM analysis as well as and SVM evaluation are deliberated in the research result and discussion. Finally, the conclusion is given as a resume and suggestion is made for future studies.

2. Research Method

Related to data resources, CHD patients in Central Hospital taking from 2011 to 2014 with the specification on 134, 64, and 82 cases of UAP, NSTEMI, and STEMI respectively. In selecting data that limits the patient's age beyond 25 years, 17 attributes were exploited, and they were defined based on the reviews of previous research [27-33] as presented in Table 1.

Table 1. Numbers of attributes

Code	Attributes
1	Age
2	gender
3	family history
4	heart history
5	history of diabetes mellitus
6	history of hypertension
7	history of cholesterol
8	obesity
9	systolic blood pressure
10	diastolic blood pressure
11	LDL levels
12	HDL levels
13	total cholesterol levels
14	triglyceride levels
15	blood levels glucose
16	elevation
17	cardiac enzymes

Following the employment of SVM in KDD [34].

Step 1: Pre-processing

This step is to reduce data, therefore there is no missing value. The activity begins with data selection from CHD and then performed as an effort to feature subset selection by ignoring the irrelevant attributes CHD risk factors and missing values. In view of this, k-NN with the Euclidian distance calculation is performed in Eq. (1)

$$dist = \sqrt{\sum_{k=1}^{n}(pk - qk)^2} \quad (1)$$

where *n* is number of attributes, *pk* and *qk* values are the -*k* attribute.

Step 2: Transformation

This step is to produce seventeen attributes and using k-NN and it is driven by discretizing the attributes with an equal width approach. The Equal width is one of the unsupervised discretisation of continuous features to obtain a better precision rate in dealing with data manipulation with high cardinality attributes [35] and its outputs become an input to the classification.

Step 3: Classification using SVM

Subsequently, the core process of data mining, which is the one-against-one SVM multiclass method is defined with a value of *d*, sigma *σ*, and *C* as explained in Table 2.

The variable d is specified as the degree of the polynomial, the value of C is a constant that allows to trade off the influence of the higher and lower-order terms and this is a consideration for varying C values between 0.01 and 1. The selection values of d, and σ impact the performance accuracy, while C is selected based on the C function as a constraint, therefore, a greater value of C implies

more penalty for classification errors. Meanwhile, the values of σ provide a good fit or an overfit to the data, when σ is large compared to the distance between the classes, it results in an overly flat discriminant surface. However, a smaller σ value compared to the distance between classes result in an over-fit [36]. A good choice for σ will be comparable to the distance between the closest members of the two classes. Furthermore, the highest accuracy of parameter pairs during the training session was found at C and σ for kernel RBF as well as C and d for the polynomial kernel. To process the data, WEKA 3.7.10, which is a powerful tool in data mining [37] and machine learning [38] was adopted.

Table 2. The define of SVM value.

d	sigma (σ)	C
1	1	0.01
2	2	0.02
3	3	0.03
4	4	0.04
5	5	0.05
		0.06
		0.07
		0.08
		0.09
		0.1
		0.2
		0.3
		0.4
		0.5
		0.6
		0.7
		0.6
		0.9
		1

Step 4: Evaluation using SVM

The evaluation process was carried out to ensure the performance of the classification methods in the SVM with two kernel trick types on polynomial and RBF. The value of accuracy and time in the building model is thoroughly investigated to achieve the superlative one. Also, the 10-folds validation and confusion matrix with percentage splits on the portion of training data compare to test data in 40:60, 50:50, 60:40, 70:30, and 80:20 is applied to support the assessment process. However, there are no specific rules in the distribution of training-data and test-data, therefore, a large number of the former will represent the diversity of the data [39]. Furthermore, to calibrate the testing procedure and the overcoming of various issues related to percentage splits in defining the best C and parameter values, 10-folds validation was exploited. Also, the test simulation took place in four stages, viz the original dataset (with missing values), the reduced (no missing values), the k-NN (with Euclidian distance calculation), and the Pre-processing (with KDD formation). Therefore, the success rate of classification, the determination of accuracy, error rate, precision,

and recall values are performed based on the confusion matrix as depicted in Eqs. (2-5) [40] given by,

$$Accuracy = \frac{TP+TN}{P+N} \times 100\% \quad (2)$$

$$Error\text{-}rate = \frac{FP+FN}{P+N} \times 100\% \quad (3)$$

$$Precision = \frac{TP}{TP+FP} \quad (4)$$

$$Recall = \frac{TP}{TP+FN} \quad (5)$$

TP *(True Positive)* = The amount of correctly classified data (*Actual class (yes), Predicted class (yes)).*

TN *(True Negative)* = The amount of correctly classified data (*Actual class (no), Predicted class (no)).*

FN *(False Negative)* = The amount of incorrectly classified data (*Actual class (yes), Predicted class (no)).*

FP *(False Positive)* = The amount of incorrectly classified data (*Actual class (no), Predicted class (yes)).*

P = Total of TP and FN

N = Total of FP and TN

To scrutinize the performance of SVM with other classifiers, the calculation of confusion matrix in NN Multilayer perceptron is measured by considering the values of accuracy, error rate, precision, and recall. This was adopted because research has steadily built on the accuracy and efficiency of data mining using NN and SVM for medical prediction and classification tasks [41, 42]. NN methods were extensively adopted in classifying problems and as one of the most active research and application areas. Furthermore, SVM and NN have been used with high accuracy in classification with relatively small sample data [43, 44].

3. The Results Result and Discussion

3.1. The Result of KDD analysis

3.1.1. Pre-processing data analysis

The data were manually selected from the medical record of 280 CHD patients at Central Hospital by paying special attention to the feature related to attributes and missing value treatments. The diversity of data based on the feature is shown in Table 1 and missing value consideration in Fig. 1. Table 3 explains that the increasing numbers of training data from 40%, 50%, 60%, 70%, and 80% are directly proportional to the diversity of data in accordance with seventeen attributes and three classes (UAP=1, NSTEMI=2, and STEMI=3). Consequently, the new pattern tested data is recognized easily.

Also, Fig. 1, describes the transformation of pre-processing activity before and after manipulating the missing values by referring to k-NN distance calculation in Eq. (1). The missing values in the dataset at number 28 column 11, 12, and 14 is

replaced by 93, 57, and 84 respectively as well as the missing values at dataset number 69, and 71.

Table 3. Data diversity according to the feature.

Feature	Training Data Composition								
	40%				…	80%			
	Area	Classes				Area	Classes		
		1	2	3	…		1	2	3
Age (1)	37-44	2	3	2	…	25-31	0	1	0
	45-51	14	5	4	…	32-37	2	1	3
	52-58	16	10	9	…	38-43	1	3	3
	59-65	13	6	5	…	44-49	19	11	10
	66-72	5	5	6	…	50-55	32	22	13
	73-79	3	2	0	…	56-61	20	10	8
	80-86	2	0	0	…	62-67	15	8	8
					…	68-73	6	7	6
					…	74-79	6	5	0
					…	80-86	4	0	0
Gender (2)	M	36	23	21	…	M	67	54	36
	F	19	8	5	…	F	38	14	15
…	…	…	…	…	…	…	…	…	…
Cardiac Enzymes (17)	Norm	55	0	0	…	Norm	105	0	0
	High	0	31	26	…	High	0	68	51

No	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	Case
27	68	F	Yes	No	Yes	Yes	No	No	144	93	160.5	31.1	207	77	381	Yes	High	STEMI
28	63	M	No	Yes	Yes	Yes	No	No	140	90	?	?	212	?	262	Yes	High	STEMI
31	56	M	No	No	No	Yes	No	No	150	90	138.6	38.4	198	105	83	Yes	High	STEMI
69	68	M	No	No	No	No	No	No	100	70	?	?	164	241	76	Yes	High	STEMI
71	54	M	No	No	No	Yes	No	No	120	90	139.1	70	226	?	88	Yes	High	STEMI

No	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	Case
27	68	F	Yes	No	Yes	Yes	No	No	144	93	160.5	31.1	207	77	381	Yes	High	STEMI
28	63	M	No	Yes	Yes	Yes	No	No	140	90	93	57	212	84	262	Yes	High	STEMI
31	56	M	No	No	No	Yes	No	No	150	90	138.6	38.4	198	105	83	Yes	High	STEMI
69	68	M	No	No	No	No	No	No	100	70	90.3	20.1	164	241	76	Yes	High	STEMI
71	54	M	No	No	No	Yes	No	No	120	90	139.1	70	226	140	88	Yes	High	STEMI

Fig. 1. Pre-processing with missing value.

3.1.2. Transformation data analysis

The medical records of CHD patients were collected in a variety of formats. Consequently, the discretization with the equal width approach was applied in expressing the standard range values from 0 to 1 as in Eq. (6).

$$\text{Series of range} = \frac{\text{the highest area - the lowest area}}{\text{The number of categories}} \quad (6)$$

The discretization of attributes is depicted in Tables 4 and 5. Table 4 defines the values of attribute 1 for age discretization, attribute 9 for systolic blood pressure (BP), attribute 10 for diastolic blood pressure, attribute 11 for LDL, attribute 12 for HDL, attribute 13 for Total cholesterol, attribute 14 for Triglyceride, and attribute 15 for a glucose level. The rest of the attributes (2, 3, 4, 5, 6, 7, 8, 16, and 17) were categorized into two series and discretized into 0 value for "No" and 1 for "Yes" as shown in Table 5. This discretization value will be the format for SVM input. The sample of format SVM input is described in Fig. 2.

Table 4. Attribute discretization

Age discretization (1)		**Systolic TD discretization (Sis) (9)**	
Age (years)	**Discretization**	**Systolic BP (mmHg)**	**Discretization**
25 ≤ U <35	0	Sis<120	Optimal (0)
35 ≤ U <45	0.2	120< Sis <130	Normal (0.2)
45 ≤ U <55	0.4	130< Sis <140	Normal Height (0.4)
55 ≤ U <65	0.6	140< Sis <150	Low hypertension (0.6)
65 ≤ U <75	0.8	150< Sis <160	Moderate hypertension (0.8)
U ≥ 85	1	Sis >160	Severe hypertension (1)
Diastolic TD (Dias) discretization (10)		**Discretization of LDL (LDL) levels (11)**	
Diastolic BP (mmHg)	**Discretization**	**LDL levels (mg / dL)**	**Discretization**
Dias<80	Optimal (0)	LDL<100	Optimal (0)
80≤Dias<85	Normal (0.2)	100< LDL <130	Approaching optimal (0.25)
85≤Dias<90	Normal Height (0.4)	130< LDL <160	Borderline high (0.5)
90≤Dias<100	Low hypertension (0.6)	160< LDL <190	High (0.75)
100≤Dias<110	Moderate hypertension (0.8)	LDL >190	Very high (1)
Dias≥110	Severe hypertension (1)		
Discretization of HDL(HDL) (12)		**Discretization of total cholesterol (Chol)(13)**	
HDL levels (mg / dL)	**Discretization**	**Chol levels (mg / dL)**	**Discretization**
HDL<40	Low (0)	Chol <200	Desirable (expected to be safe) (0)
40≤ HDL <60	Normal (0.5)	200≤ Chol <240	Borderline (must be aware- begin to control) (0.5)
HDL ≥60	High (1)	Chol ≥240	High (1)
Triglyceride discretization (14)		**Glucose Level discretization (Glu) (15)**	
Triglyceride levels (mg / dL)	**Discretization**	**Glucose Levels (mg/dL)**	**Discretization**
trig <150	Normal (0)	Glu<40	Optimal (0)
150≤ trig <200	Borderline high (0.33)	40≤ Glu <60	Normal (0.2)
200≤ trig <500	High (0.66)	60≤ Glu <125	Normal Height (0.4)
trig ≥500	Very High (1)	125≤ Glu <145	Low hypertension (0.6)
		145≤ Glu <200	Moderate hypertension (0.8)

	Glu ≥200	Severe hypertension (1)

Table 5. Attributes with two series discretization

Attributes	Discretization	
Gender (2)	Male	1
	Female	0
Family History (3)	None	0
	Yes	1
Heart History (4)	None	0
	Yes	1
DM History (5)	None	0
	Yes	1
Hypertension History (6)	None	0
	Yes	1
Cholesterol History (7)	None	0
	Yes	1
Obesity (8)	None	0
	Yes	1
Elevation (16)	None	0
	Yes	1
Cardiac Enzymes (17)	None	0
	Yes	1

	Attributes discretization																	
No	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	Cases
1	0,4	1	0	0	0	0	0	0	0,4	0,2	0	0	0	0	0,4	0	0	UAP
2	0,4	1	0	0	0	0	0	0	0,8	0,8	0	0	0	0	0,4	0	0	UAP
3	0,4	0	0	1	0	0	0	0	0,2	0,2	0	0,5	0	0	0,4	0	0	UAP
4	0,6	1	0	1	0	0	0	0	0,2	0,2	0,25	0,5	0	0	0,8	0	0	UAP
5	0,6	1	0	0	1	0	0	0	1	0,2	0,25	0,5	0	0	0,4	0	0	UAP
6	0,6	1	0	1	1	1	1	0	0,2	0,2	0,5	1	0,5	0	0,4	0	1	NSTEMI
7	0,6	1	0	0	0	0	0	0	0,6	0,2	0	0	0	0,33	0,8	0	0	UAP
8	0,4	0	0	1	0	0	0	0	0,6	0,8	0,25	1	0,5	0	0,4	0	0	UAP
9	0,6	1	0	0	1	1	0	0	0,6	0,2	0,75	0	0,5	0	0,4	0	0	UAP
10	0,4	0	0	1	1	1	0	0	0,4	0,2	0,25	0,5	0,5	0,66	0,4	0	0	UAP
11	0,2	1	0	1	1	1	0	0	0,2	0	0,25	0	0	0	0,4	0	0	UAP
12	0,6	1	0	1	1	1	0	0	1	0,8	0,25	0	0	0,33	0,4	0	0	UAP
13	0,4	0	0	0	0	0	0	0	0	0	1	0	1	0	0,4	0	0	UAP
14	0,4	1	0	1	1	1	0	0	0,6	0,6	0	0	0	0	0,4	0	0	UAP
15	0,4	1	0	1	1	1	0	0	0,6	0,8	0,5	0	0,5	0,66	0,4	0	0	UAP
16	0,4	0	0	0	1	1	0	0	0,6	0,8	0,25	0	0	0	0,8	0	1	NSTEMI
17	0,8	0	0	0	0	0	0	0	1	0,8	0,25	0	0,5	0	0,8	1	1	STEMI
18	0,4	0	1	0	1	1	0	0	0,6	0,6	0	0	0	0	0,4	1	1	STEMI
19	0,6	1	0	0	0	0	0	0	0	0	1	0	1	0	0,4	1	1	STEMI
20	0,4	1	0	1	1	1	0	0	0,4	0,8	0,5	0	0	0	0,4	1	1	STEMI

Fig. 2. The sample of SVM input.

3.1.3. SVM mining analysis

To investigate the implication of pre-processing against SVM, the analysis is conducted by comparing the accuracy within dataset changes in the original data (without missing values), the reduced (with missing values), k-NN (with distance calculation), and pre-processing (KDD formatted). This was executed through the selection of the best parameters for 10-fold cross-validation in Table 6 and percentage split in Table 7 for four scenarios dataset. The graphical views of performances are shown in Figs. 3-5.

The execution of 10-folds cross-validation in Table 6 explained that the pre-processing dataset improved accuracy level up to 100% and 98.9% in kernel polynomial and RBF, respectively with the superior parameters at C = 0.02 and d = 2, C = 0.8 and σ = 1, respectively. Similarly, Table 7 shows that the pre-processing dataset with the percentage split treatment also provided a significant growth of accuracy in polynomial and RBF kernel. Moreover, the execution time in model development considerably impacts the performance of pre-processing both in Polynomial and RBF kernel at the data composition of 70:30 and 80:20, respectively. Figs. 3-5, explained that the pre-processing dataset increases its performance in terms of time (s) and accuracy for Polynomial and RBF kernel.

Table 6. The accuracy of the best parameter - 10-fold cross validation.

Kernel	**Polynomial**			**RBF**		
Parameters	**C**	**d/σ**	**Accuracy**	**C**	**d/σ**	**Accuracy**
Original Dataset	0.03	1	100%	0.01	1	47.9%
Reduced Dataset	0.03	1	100%	0.01	1	48.1%
k-NN Dataset	0.03	1	100%	0.01	1	47.9%
Pre-processing Dataset	0.02	2	100%	0. 8	1	98.9%

Table 7. The accuracy of the best parameter pairs -percentage split.

Kernel	**Polynomial**			**RBF**		
Parameters	**DC**	**T (s)**	**Accuracy**	**DC**	**T(s)**	**Accuracy**
Original Dataset	70:30	27.49	100%	40:60	0.06	49.4%
Reduced Dataset	70:30	21.37	100%	70:30	0.13	53.8%
k-NN Dataset	80:20	24.55	100%	40:60	0.08	49.4%
Pre-processing Dataset	70:30	0.06	100%	80:20	0.08	100%

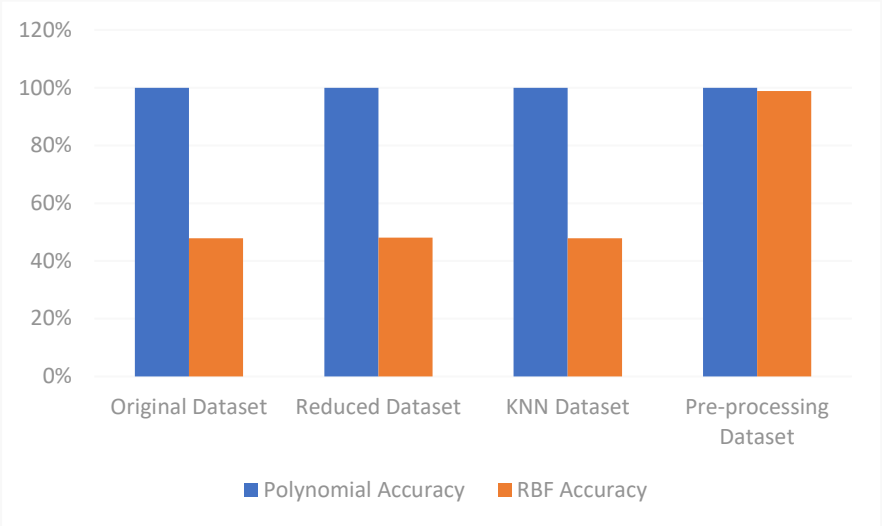

Fig. 3. Dataset performance based on accuracy - 10-fold cross validation.

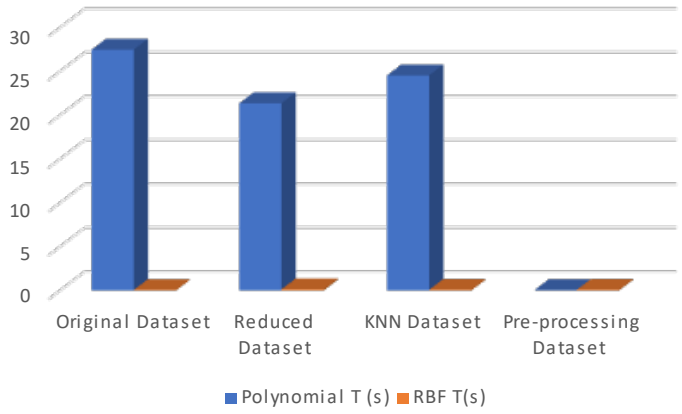

Fig. 4. Dataset performance based on time (s) - percentage split.

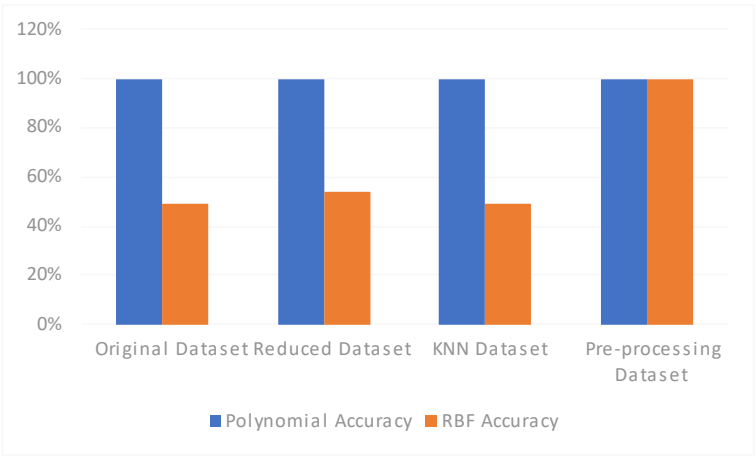

Fig. 5. Dataset performance based on accuracy - percentage split

3.2. Testing

To evaluate the classification of CHD patient's dataset in SVM, the testing procedure was undertaken according to the Test Option Supplied on the Confusion Matrix formula [39]. The pre-processing dataset was put in place on 20% of tested data at C = 0.02 and d = 2 in the polynomial kernel and the values of C and σ are 0.8 and 1 respectively, in the RBF.

In addition, the resemblance of SVM with another classifier, namely Multilayer perceptron Neural Network (NN) is operated to deeply observe the effectiveness of SVM. The confusion matrix for the above dataset of SVM and NN was explained in Table 8. This table showed that the classification in the pre-processing dataset for SVM is more accurate compared to NN, especially for RBF kernel.

By comparing the values for error rate, precision and recall between polynomial kernel and RBF based on the confusion matrix computation as a side of SVM and NN, Fig. 6, is obtained. The figure showed that SVM for Polynomial

kernel has 100% accuracy, "0" for error rate, and "1" for precision, and recall. Meanwhile, RBF kernel discharged from 51.79% into 100% accuracy, 0.48 into 1 for error rate, undefined into 1 for precision, and 0.52 into 1 for recall. Also, NN for polynomial kernel achieved 89% accuracy, "0.11" for error rate, and "0.89" for precision and recall.

Table 8. Confusion Matrix for SVM and NN-Polynomial and RBF.

SVM: Dataset Pre-processing						
	Polynomial			**RBF**		
Class	**Prediction Class**			**Prediction Class**		
	UAP	**NSTEMI**	**STEMI**	**UAP**	**NSTEMI**	**STEMI**
UAP	29	0	0	29	0	0
NSTEMI	0	13	0	0	13	0
STEMI	0	0	14	0	0	14
Accuracy		100%			100%	
Error rate		0			0	
Precision		1			1	
Recall		1			1	

NN-Multilayer Perceptron			
Class	**Prediction Class**		
	UAP	**NSTEMI**	**STEMI**
UAP	**18**	**0**	**1**
NSTEMI	**2**	**18**	**1**
STEMI	**0**	**2**	**14**
Accuracy	**89%**		
Error rate	**0.11**		
Precision	**0.89**		
Recall	**0.89**		

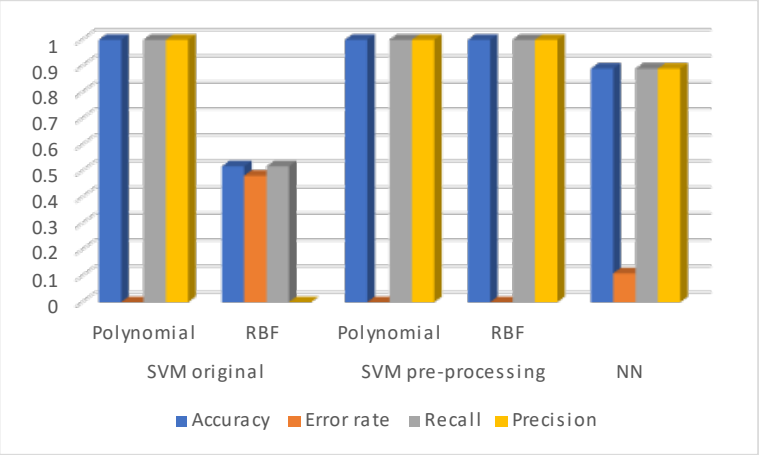

Fig. 6. Performance Polynomial and RBF kernel.

3.3. Discussion

This result reveals that the pre-processing dataset in SVM provides significant values on the accuracy, error rate, precision, and recall, even though it exceeds NN capacity. As studied by Shao and Lunetta [42], the SVM approach gives better predictive capability than other models, including NN. This, of course, has far-

reaching implications in the medical context that require increasing sensitivity, specificity (the ability to predict the absence of the condition when it is not present) as well as discriminatory power of the classifier as key features to consider when comparing classifiers and diagnostic methods [45]. In the reviews on kernel type, the simulation presented that SVM polynomial is more reliable on the dataset changes compare to RBF. Consequently, the pre-processing prescription on SVM-RBF will undoubtedly boost RBF performance. Furthermore, selecting the specific kernel is an important research issue for kernel-based learning in the data mining area and the problem of SVM kernels is found in fitting the appropriate parameter values [46]. This investigation revealed that the SVM polynomial kernel mediates the accuracy and efficiency of the diagnostic results based on the parameters defined in CHD.

4. Conclusions

This study successfully employed the SVM method in classifying the CHD patient's dataset. The simulation of the original, reduced, k-NN, and pre-processing datasets have shown the potential differences between polynomial and RBF kernel in terms of accuracy and processing time. The analysis of 10-folds cross-validation and percentage splits revealed the optimal pairs of parameters and data composition for polynomial and RBF kernel. Furthermore, the confusion matrix presented evidence that the pre-processing dataset delivered greater values in accuracy, precision, error rate, recall, and time model consumption than others. A comparative analysis between SVM and NN has shown the efficiency and accuracy of SVM in accelerating the unsurpassed classification of the CHD dataset with minimal errors. Therefore, this classification practically aids the doctors in suggesting medical assistance and taking a curative action. This result methodically answered the difficulties in choosing the SVM kernel function, which is flexible in changing the data set, optimal functionality, and time-consumption with high performance. Nevertheless, integrating SVM with other methods is a new solution to increase SVM performance for future work.

References

1. Nijssen, S.; and Fromont, E. (2010). Optimal constraint-based decision tree induction from itemset lattices. *Data Mining and Knowledge Discovery*, 21(1), 9-51.
2. Zounemat-Kermani, M.; Stephan, D.; Barjenbruch, M.; and Hinkelmann, R. (2020). Ensemble data mining modeling in corrosion of concrete sewer: A comparative study of network-based (MLPNN & RBFNN) and tree-based (RF, CHAID, & CART) models. *Advanced Engineering Informatics*, 43(101030), 1-12.
3. Okfalisa.; Fitriani, R.; and Vitriani, Y. (2018). The comparison of linear regression method and k-nearest neighbors in scholarship recipient. *Proceedings of IEEE/ACIS 19th International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, SNPD*. Busan, Korea, 194-199.
4. Amin, M.Z.; and Ali, A. (2017). Application of multilayer perceptron (MLP) for data mining in healthcare operations. *Proceeding of the 3rd International Conference on Biotechnology*. University of South Asia, Lahore, Pakistan, 2-11.

5. Makhtar, M.; Nawang, H.; and Wan Shamsuddin, S.N. (2017). Analysis on students performance using naïve bayes classifier. *Journal of Theoretical and Applied Information Technology*, 95(16), 3993-4000.
6. Jakkula, B.; Mandela, G.R.; and Chivukula, S.M. (2020). Application ANN tool for validation of LHD machine performance characteristics. *Journal of The Institution of Engineers (India): Series D*, 101(1), 27-38.
7. Alweshah, M.; Al-Sendah, M.; Dorgham, O.M.; Al-Momani, A.; and Tedmori, S. (2020). Improved water cycle algorithm with probabilistic neural network to solve classification problems. *Cluster Computing*, 23, 2703-2718.
8. Cortez, P. (2012). *Data mining with multilayer perceptron and support vector machines.* Berlin Heidelberg: Springer-Verlag, 24, 9-25.
9. Greene, C.S.; Tan, J.; Ung, M.; Moore, J.H.; and Cheng, C. (2014). Big data bioinformatics. *Journal of Cellular Physiology*, 229(12), 1896-1900.
10. Koboldt, D.C.; Fulton, R.S.; McLellan, M.D.; Schmidt, H.; Kalicki-Veizer, J.; McMichael, J.F.; and Palchik, J.D. (2012). Comprehensive molecular portraits of human breast tumors. *Nature*, 490(7418), 61-70.
11. Majhi, V.; Paul, S.; and Jain, R. (2019). Bioinformatics for healthcare applications. *Proceedings of Amity International Conference on Artificial Intelligence*, Dubai, United Arab Emirates, 204-207.
12. Dashtban M.; and Balafar, M. (2017). Gene selection for microarray cancer classification using a new evolutionary method employing artificial intelligence concepts. *Genomics*, 109(2), 91-107.
13. Serra, A.; Galdi, P.; and Tagliaferri, R. (2018). Machine learning for bioinformatics and neuroimaging. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(5), 1-33.
14. Deepthi, P.N.; Anitha, R.; and Swathi, K. (2019). A review on bioinformatics using data mining techniques. *Proceeding of the International Conference on Computer Vision and Machine Learning*. Andhra Pradesh, India, 1-11.
15. Saha, S.; and Bhattacharya, T. (2019). A novel approach to find the saturation point of n-gram encoding method for protein sequence classification involving data mining. *Proceeding of International Conference on Innovative Computing and Communications*. Barasat, Kolkata, India, 101-108.
16. Zhang, Y.; and Wang, S. (2015). Detection of Alzheimer's disease by displacement field and machine learning. *PeerJ*, 3 (e1251), 1-29.
17. Zhang, Y.; Dong, Z.; Philips, P.; Wang, S.; Ji, G.; Yang, J.; and Yuan, F.T. (2015). Detection of subjects and brain regions related to Alzheimer's disease using 3D MRI scans based on eigenbrain and machine learning. *Frontiers in Computational Neuroscience*, 9(66), 1-15.
18. Collins, M.P.; and Pape, S.E. (2011). The potential of support vector machine as the diagnostic tool for schizophrenia: A systematic literature review of neuroimaging studies. *European Psychiatry*, 26(S2),1363.
19. Sigavimi, P. (2012). Supervised learning approach for breast cancer classification. *International Journal of Emerging Trends & Technology in Computer Science*, 1(4), 125-129.
20. Hongzong, S.; Tao, W.; Xiaojun, Y.; Huanxiang, L.; Zhide, H.; Mancang, L.; and BoTao, F. (2007). Support vector machines classification for

discriminating coronary heart disease patients from non-coronary heart disease. *West Indian Medical Journal*, 56(5), 451-457.

21. Mo, Y.; and Xu, S. (2010). Application of SVM based on hybrid kernel function in heart disease diagnoses, *Proceedings of International Conference on Intelligent Computing and Cognitive Informatics*. Kuala Lumpur, Malaysia, 462-465.
22. Tabesh, P.; Lim, G.; Khator, S.; and Dacso, C. (2010). A support vector machine approach for predicting heart conditions. *Proceedings of IIE Annual Conference and Expo 2010 Proceedings*. Cancun, Mexico, 916-921.
23. Parveen.; and Singh, A. (2015). Detection of brain tumor in MRI images, using combination of fuzzy c-means and SVM. *Proceedings of 2nd International Conference on Signal Processing and Integrated Networks, SPIN*. New Delhi, India, 98-102.
24. Machhale, K.; Nandpuru, H.B.; Kapur, V.; and Kosta, L. (2015). MRI brain cancer classification using hybrid classifier (SVM-KNN). *Proceedings of International Conference on Industrial Instrumentation and Control, ICIC*. Pune, India, 60-65.
25. Ahmmed, R.; Swakshar, A.S.; Hossain, M.F.; and Rafiq, M.A. (2017). Classification of tumors and it stages in brain MRI using support vector machine and artificial neural network. *Proceedings of International Conference on Electrical, Computer and Communication Engineering*. Bazar, Bangladesh, 229-234.
26. Aydin, E.; Aypar, E.; Oktem, A.; Ozyuncu, O.; Yurdakok, M.; Guvener, M.; Demircin, M.; and Beksac, M.S. (2018). Congenital heart defects: the 10-year experience at a single center. *The Journal of Maternal-Fetal & Neonatal Medicine*, 33(3), 368-372.
27. Dirjen Bina Kefarmasian dan AlKes DepKes RI. (2006). *Pharmaceutical Care Untuk Pasien Penyakit jantung Koroner : Fokus Sindrom Koroner Akut*. Jakarta: Departmen Kesehatan RI.
28. Magesh, G.; and Swarnalatha, P. (2020). Optimal feature selection through a cluster-based DT learning (CDTL) in heart disease prediction. *Evolutionary Intelligence*.
29. Arad, Y.; Goodman, K.J.; Roth, M.; Newstein, D.; and Guerci, A.D. (2005). Coronary calcification, coronary disease risk factors, c-reactive protein, and atherosclerotic cardiovascular disease events. *Journal of the American College of Cardiology*, 46(1), 158-165.
30. Hand, D.; Mannila, H.; and Smyth, P. (2001). *Principles of data mining*, Massachusetts London: The MIT Press.
31. Nauta, S.T.; Deckers, J.W.; Boon, R.M.; Akkerhuis, K.M.; and Domburg, R.T. (2014). Risk factors for coronary heart disease and survival after myocardial infarction. *European Journal of Preventive Cardiology*, 21(5), 576-583.
32. Mannsverk, J.; Wilsgaard, M.D.T.; Mathiesen, E.B.; Løchen, M.; Rasmussen, K.; Thelle, D.S.; Njolstad, I.; Hopstock, L.A.: and Bonna, K.H. (2015). Trends in modifiable risk factors are associated with declining incidence of hospitalized and non-hospitalized acute coronary heart disease in a population. *Circulation*, 133(1), 74-81.
33. Sharma, P.; Choudhary, K.; Gupta, K.; Chawla, R.; Gupta, D.; and Sharma, A. (2019). Artificial plant optimization algorithm to detect heart rate and presence

of heart disease using machine learning. *Artificial Intelligence in Medicine*, 102 (101752), 1-14.

34. Ivezic, Z.; Connoly, A.J.; VanderPlas, J.T.; and Gray, A. (2011). *Data Mining and Machine Learning in Astronomy: A Practical Guide*. Princeton: Princeton University Press.
35. Dash, R.; Paramguru, R.L.; and Dash, R. (2011). Comparative analysis of supervised and unsupervised discretization techniques. *International Journal of Advances in Science and Technology*, 2(3), 29-37.
36. Kaytez, F.; Taplamacioglu, M.C.; Cam, E.; and Hardalac, F. (2015). Forecasting electricity consumption: A comparison of regression analysis, neural network and least square support vector machines. *International Journal of Electrical and Power & Energy Systems*, 67(2015), 431-438.
37. Russell, I.; and Markov, Z. (2017). An introduction to the WEKA data mining system. *Proceedings of the 2017 ACM SIGCSE Technical Symposium on Computer Science Education*. Seattle, Washington, USA, 742-742.
38. Rokach, L.; and Maimon, O. (2010). Data mining using decomposition methods. *Data Mining and Knowledge Handbook*. Boston, Springer. 981-998.
39. Polat, K.; Akdemir, B.; and Gunes, S. (2008). Computer aided diagnosis of ECG data on the least square support vector machine. *Digital Signal Processing*, 18(1), 25-32.
40. Sun, L.; Yu, G.; and Sun, Z. (2010). Bi-parameters method for structural vulnerability analysis. *Intelligent Automation and Soft Computing*, 16(5), 747-754.
41. Raczko, E.; and Zagajewski, B. (2017). Comparison of support vector machine, random forest and neural network classifiers for tree species classification on airborne hyperspectral APEX images. *European Journal of Remote Sensing*, 50(1), 144-154.
42. Shao, Y.; and Lunetta, R.S. (2012). Comparison of support vector machine, neural network, and CART algorithms for the land-cover classification using limited training data points. *ISPRS Journal of Photogrammetry and Remote Sensing*, 70(2012), 78-87.
43. Paulo de Magalhaes, O.J.P.; Ricardo, N.; Geraldo, B.; Carlos, B.; Ricardo, S.J.; and Edson, A.J. (2010). Use of SVM methods with surface-based cortical and volumetric subcortical measurements to detect Alzheimer's disease. *Journal of Alzheimer's Disease: JAD*, 19(4), 1263-1272.
44. Jahandideh, S.; Abdolmaleki, P.; and Movahedi, M.M. (2010). Comparing performances of logistic regression and neural networks for predicting melatonin excretion patterns in the rat exposed to ELF magnetic fields. *Bioelectromagnetics*, 31(2),164-171.
45. Maroco, J.; Silva, D.; Rodrigues, A.; Guerreiro, M.; Santana, I.; and de Mendonca, A. (2011). Data mining methods in the prediction of dementia: a real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests. *BMC Reseach Notes*, 4(299), 1-14.
46. Ali, S.; and Smith, K.A. (2003). Automatic parameter selection for polynomial kernel. *Proceedings of Fifth IEEE Workshop on Mobile Computing Systems and Applications*. Las Vegas, USA, 243-249.