Received: February 6, 2020

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

A Support Vector Machine Based on Kernel K-Means for Detecting the Liver Cancer Disease

Lailil Muflikhah[1,2]* **Widodo Widodo**[2] **Wayan Firdaus Mahmudy**[1] **Solimun Solimun**[2]

[1]*Faculty of Computer Science, Brawijaya University, Malang, Indonesia*
[2]*Faculty of Mathematics and Natural Science, Brawijaya University, Malang, Indonesia*
* Corresponding author's Email: lailil@ub.ac.id

Abstract: Liver cancer can be caused by hepatitis B virus (HBV) infection. This genome virus inserts genetic material into its host. Moreover, each gene has numerous HBV DNA sequences due to its high mutation rate in replication. Thus, detecting virus is a difficult task. A support vector machine (SVM) is a robust machine-learning algorithm for detecting liver cancer disease. However, a high data volume can reduce its computation speed and performance measures. Therefore, we propose data simplification using Kernel *k*-means clustering method to construct the SVM classifier model by minimizing objective function as object distance. Based on experimental results, the proposed method's performance evaluation was higher than SVM algorithm without kernel *k*-means, especially for the sensitivity significantly increased. The accuracy rate and AUC of the proposed method were around 98% and 0.95. Furthermore, the performance of proposed method is also predominant of the other machine learning: Random Forest, Naïve Bayes, Naural Network and C5.0.

Keywords: Liver cancer, DNA sequence, SVM, Kernel *k*-Means.

1. Introduction

Owing to the recent increase in the number of patients with liver cancer, this disease has attracted the attention of the government and researchers. According to the Global Cancer Observatory database, the International Agency for Research on Cancer discovered that liver cancer was the third-ranking cause of death at 781,531 cases; it is behind the other two leading cancers: lung and colorectal cancers. Geographically, the prevalence of hepatocellular carcinoma (HCC) is unevenly distributed with Asia dominating at 72.5% (609,596 cases), with a mortality rate of 72.4% (566,269 cases) [1].

The hepatitis B virus X protein *(*HB*x)* is one of the hepatitis B virus (HBV) genes whose role is to replicate for survival in the host. Insertion of the genome virus affects mutation of the deoxyribonucleic acid (DNA) sequences. Numerous previously conducted studies demonstrated the roles of *HBx* in the pathogenesis of virus-induced liver cancer [2]. Moreover, there is a relationship between putative mutation and liver cancer. Mutation of the *HBx* DNA sequence at AA 127, 130, and 131 as *HBx* deletion was implicated in liver cancer [3]. The truncation mutation at the 3' end of *HBx* had a potential role in HBV related to liver cancer [4]. The HBV *HBx* gene multi-site mutations also occurred in the clinical HCC (liver cancer) disease and were related to the development of the disease. In China, data indicated that 44 cases with a mutation pattern of *HBx* involving 60 patients who had liver cancer and were HBV-positive [5]. By applying the computational approach, the study on the patient's liver cancer that infected the HBV was conducted by profiling the DNA sequence of the HBV using the clustering method [6].

A support vector machine (SVM) is a robust classification method for predicting the liver cancer disease. Numerous studies were conducted using various types of data sets, such as image, clinical, and microarray data of gene expression, and a DNA sequence [7–10]. When SVM is applied to the large volume of data sets, its challenge is that it decreases the computational speed and the accuracy rate. Therefore, by clustering method, the variance can be reduced to simplify the SVM classifier model.

K-means is one of the methods for clustering to preprocess the large data volume with high speed in computational time and accuracy rate. Numerous previously conducted studies applied the *k*-means cluster algorithm in combination with the SVM algorithm (KSVM) but still maintained the accuracy [11]. However, the *k*-means covered only a linear separable input space. A Kernel *k*-means is an advanced *k*-means cluster algorithm that can cluster a non-separable input space by mapping a data object to the high dimension of input space, and it is applied *k*-means algorithm [12]. Therefore, this current research aims to improve the performance by applying the Support Vector Machine on Kernel *k*-Means clustering method for liver cancer detection using HBV DNA sequences. In this two-stage method, the DNA sequences were clustered by using Kernel *k*-Means algorithm, then the SVM classifier method was implemented in each clustering result.

The rest of this paper is organized as follows. In Section 2, the background of this study, the development of liver cancer and the detection of the disease in biological and computational approaches are discussed. Additionally, using the SVM algorithm based on kernel *k*-means, the proposed method is studied. Section 3 presents the results and discussion. Section 4 provides the conclusions, and Section 5 gives recommendations for further studies.

2. Related works

Many studies on the detection of liver cancer were conducted using a supervised learning method. In 2014, research on the classification of DNA sequences of the Hepatitis B virus at genotype B and C conducted by Rekha et, al. It used Non-Linear Integral and fuzzy measure method to identity liver cancer disease [13]. Another research, by using microarray data of gene expression (GSE20948) was applied Support Vector Machine with clustered network topology to identify hepatocellular carcinoma [10]. Kesh, at al. in 2015 conducted research using DNA sequence to categorize binary on unbalanced class distribution for liver cancer disease using Bayesian vs Voted perceptron [14]. Xin Bai et all (2018) conducted research on deep sequencing of HBV pre-S DNA sequence and based on the frequent pattern to classify liver cancer disease and get their association using KNN against SVM[7].

3. Material and method

As depicted in Fig. 1, the system is constructed using two main parts of the research method, including kernel *k*-means clustering and the SVM classification algorithm, as the general step for detecting liver cancer.

A. DNA sequence decomposition

The first stage of this study is preprocessing of data, including crawling and transformation data, to nucleotide composition or amino acid composition. The crawling data was utilized to obtain the desired data, such as DNA sequences and their status “carcinoma,” “HCC,” or “liver cancer” on the remark site.

The *HBx* gene of HBV has various DNA sequences when integrated into its host. Each patient has a different DNA sequence of *HBx*. However, the similarity rate of mutation did not indicate cancer. Therefore, this study utilizes DNA sequence decomposition in the percentage of nucleotide and amino acid as a feature for predicting liver cancer.

a) Nucleotide decomposition

Moreover, the next step of preprocessing data was nucleotide decomposition employed to construct a codon in each sequence. Besides, the regions of DNA coding are representative of the nucleotide compositions at the first, second, and third positions of the codon. As an illustration, the *HBx* gene had a length of 465 sites and was constructed three basal

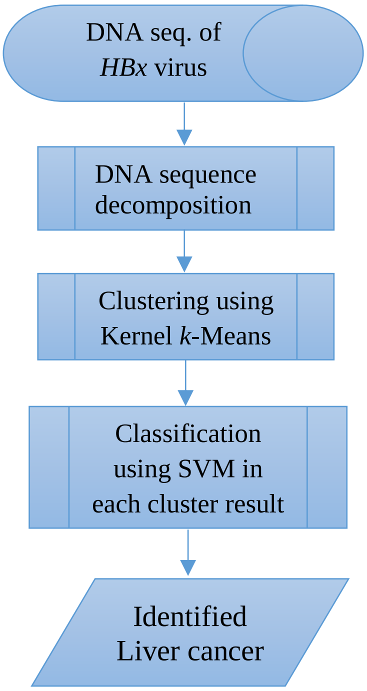
DNA seq. of
HBx virus
DNA sequence decomposition
Clustering using Kernel *k*-Means
Classification using SVM in each cluster result
Identified Liver cancer

Figure. 1 General steps for predicting liver cancer

Table 1. Standard genetic code

1st nucleotide	2nd nucleotide				3rd nucleotide
	T	**C**	**A**	**G**	
T	Phe/F	Ser/S	Tyr/ Y	Cys/ C	**T**
					C
	Leu/L		Stop	Stop	**A**
			Stop	Trp/ W	**G**
C		Pro/ P	His/ H	Arg/ R	**T**
					C
			Gln/ Q		**A**
					G
A	Ile/I	Thr/ T	Asn/ N	Ser/S	**T**
					C
			Lys/ K	Arg/ R	**A**
	Met/ M				**G**
G	Val/V	Ala/ A	Asp/ D	Gly/ G	**T**
					C
			Glu/ E		**A**
					G

nucleotides in each codon for coding a protein based on standard genetic code as shown in Table 1 [15].

The relative frequencies of four nucleotides can be computed for one specific sequence or all sequences: adenine (A), thymine (T), guanine (G), and cytosine (C). For the coding regions of DNA, additional columns are presented for the nucleotide composition at the first, second, and third codon positions [13]. In addition, the nucleotide composition in percentage was used as a feature for the learning process of the classification method. In this study, we implemented the MEGA tools of bioinformatics software to decompose DNA sequences. As demonstrated in Table 2, the DNA sequence of *HBx* was transformed to nucleotide composition in each codon.

Table 2. The nucleotide composition of the *HBx* DNA

SId	**T1**	**C1**	**A1**	**G1**	**…**	**G3**	**Status**
1	28	25.2	14.8	32.3		25.2	LC
2	26	24.5	16.8	32.3		24.5	LC
3	28	23.9	16.1	32.3		25.2	LC
4	28	23.9	14.8	33.5		25.2	N
.							.
.							.
n	28	23.9	14.8	33.5	…	25.2	N

b) Amino acid decomposition

From the DNA sequence, another data set is amino acid decomposition. The relative frequencies of the four nucleotides or the 20 amino acid residues (amino acid composition) can be computed for a particular sequence or all the sequences. The 20 amino acids include the following: alanine (Ala), arginine (Arg), asparagine (Asn), aspartic acid (Asp), cysteine (Cys), glutamic acid (Glu), glutamine (Gln), glycine (Gly), histidine (His), hydroxyproline (Hyp), isoleucine (Ile), leucine (Leu), lysine (Lys), methionine (Met), phenylalanine (Phe), proline (Pro), serine (Ser), threonine (Thr), tryptophan (Trp), tyrosine (Tyr), and valine (Val). The data set comprises amino acid sequences that can be translated into proteins [16].

Then, the amino acid composition is the number of amino acids of each type normalized with the total number of residues. It is defined in the Eq.1.

$$Comp(i) = \sum n_i x \, {}^{100}/_{N}, \qquad (1)$$

where *i* denotes the 20 amino acid residues, n_i denotes the number of residues of each type, and *N* denotes the total number of residues. The summation is taken over all the residues in the considered protein [14]. The results of the amino acid composition are presented in a tabular form, as presented in Table 3.

B. The proposed method

The large data volume requires a high computational time and decreases the performance of the SVM classifier. Therefore, we propose to cluster the data before applying the classification method. The kernel *k*-means method is *k*-means advanced by applying the following methods.

1) K-means clustering

K-means is a clustering method with high computation speed and good accuracy. Using an X-ray image as the data set, Kwang et al. conducted a study in which the *k*-means clustering was applied to develop an automatic segmentation of wrist bone

Table 3. The amino acid composition of *HBx* DNA

SId	**Ala**	**Cys**	**Asp**	**Glu**	**Phe**	**…**	**Status**
1	11.0	5.8	3.9	4.5	4.5		LC
2	9.7	5.8	3.9	4.5	5.2		LC
3	10.4	5.8	3.2	4.5	4.5		LC
4	10.4	5.8	3.9	4.5	5.2		N
.	..	…	…	…	..		…
…	…	…	…	…	…	…	…
n	.	.	.	.	.	.	.

fractures and achieved 80% accuracy [17]. This method has also been applied to genetics algorithm hybridization to reduce the computational time for feature selections [18].

The *k*-means clustering algorithm principally assigns data objects to certain clusters by defining the number of clusters as the first step in the calculation. Besides, the objective function of this algorithm is to obtain a minimum of the total intra-cluster differences. The total minimum is the squared sum of the Euclidean distances between the related objects and the cluster center [19]. However, the limitation of the *k*-means clustering is that the method covered *k*-only linear separable input space. Therefore, this study developed the *k*-means method using the Kernel *k*-means clustering.

2) Kernel k-means clustering

The lack of k-means can only detect clusters that are linearly separable, whereas kernel *k*-means can be used to detect those that are nonlinearly separable (non-convex clusters). Using the kernel function of *k*-means, the data are mainly projected onto the high-dimensional feature so that cover clusters are nonlinearly separable in the input space [12]. Moreover, the clustering algorithm replaces the Euclidean distance or similarity computation in *k*-means by the kernelized version. Let $X = \{a_1, a_2, a_3, \dots, a_n\}$ and $\pi_1, \pi_2, \dots, \pi_k$ be the set of data points and the number of *k* clusters, respectively. Principally, the kernel *k*-means clustering is applied to minimize the following objective function as distances among objects [20]:

$$D\left(\{\pi_c\}_{c=1}^{k}\right) = \sum_{c=1}^{k} \sum_{a_i \in \pi_c} \|\emptyset(a_i) - m_c\|^2, \qquad (2)$$

where

Where the centroid is noted by $m_c = \frac{\sum_{a_i \in \pi_c} \emptyset(a_i)}{|\pi_c|}$

$\|\emptyset(a_i) - m_c\|^2$ can be expanded as follows:

$$\emptyset(a_i).\emptyset(a_j) - \frac{2\sum_{a_j \in \pi_c} \emptyset(a_i).\emptyset(a_j)}{|\pi_c|} + \frac{\sum_{a_j a_i \in \pi_c} \emptyset(a_j).\emptyset(a_i)}{|\pi_c|^2}$$

and $\emptyset(a_i).\emptyset(a_j)$ is a kernel matrix where can computed distances data point and centroid.

Thus, the steps of kernel *k*-means clustering are as follows:

a. Input the *k* cluster number.
b. Randomly initialize the *c* cluster center.
c. Using Eq. (2), compute the distance of each data point and cluster center in the transformed space.
d. Assign data points to the cluster center with minimum distance.
e. Repeat to step b until there are no data points re-assigned.

3) The SVM classifier

The SVM algorithm is mainly a linear classification method that seeks the best function of the hyperplane. The function is divided into two classes of input space that are then developed into nonlinear classifiers by incorporating kernel tricks in a high-dimensional space. In addition, the data should be transformed into the vector space in a high-dimensional space. The kernel trick functions that can be utilized in nonlinear SVM classifications are a polynomial, the Gaussian radial bases function (RBF), and a sigmoid. Each label is denoted by y¬i ϵ {-1, +1}, for i = 1, 2, ..., n, where n denotes the number of data. The label is assigned +1 and −1 classes, which can be completely separated from the hyperplane using the following equation:

$$w.x + b = 0 \qquad (3)$$

where *w* is a weight vector, *x* is input vector and *b* is bias value.

The object data point x_i is assigned to -1 in the following inequality:

$$w.x_i + b \leq -1 \qquad (4)$$

It is assigned to +1 in the following inequality:

$$w.x_i + b \geq +1 \qquad (5)$$

Further, the largest margin is calculated by maximizing the distance between the hyperplane and the nearest point:

$$\frac{1}{\|w\|}. \qquad (6)$$

In principle, a nonlinear SVM concept changes the data *x* that is applied to the function Φ (*x*) in the high-dimensional vector space. Therefore, the objective function represents data in the new vector space. In the SVM, the learning process is to find support vectors by the dot product of the new vector space data.

a) Kernel trick

The kernel function aims to determine a support vector for nonlinear data in the SVM learning process [12]. It can be defined as follows:

$$K(x_i . x_j) = \Phi(x_i) . \Phi(x_j) \quad (7)$$

In this study, two kernels are applied, including polynomial and Gaussian RBF as defined in Eqs. (8) and (9):

$$K(X_i \cdot X_j) = (x_i . x_j + 1)^P \quad (8)$$

$$\text{and } K(X_i \cdot X_j) = \exp\left(-\left(\frac{\|x_i - x_j\|^2}{2\sigma^2}\right)\right) \quad (9)$$

The next step is to make predictions by implementing the Sequential SVM algorithm, as well as the calculation of the Hessian matrix. Thus, the steps are repeated to reach the maximum of the least error rate or max (| δα |) < ε. When they are complete, the bias and similarities between the testing and training data are calculated. The results are obtained in positive or negative classes. Moreover, many various parameter value combinations can be tuned in the SVM method, so that it is required to get the best parameter value.

b) Tuning the SVM parameter

The advantage of the SVM algorithm is that it can achieve a global optimum. Furthermore, the SVM algorithm can handle the problem of high dimensionality [21]. However, setting parameters and kernels can affect the performance of the learning process and the general performance of the SVM [22]. Using specified method types, such as cross-fold and bootstrapping, the combination of parameters can lead to an error rate estimation on certain data. In this study, the SVM parameters with the polynomial kernel used include gamma (ϒ) and cost (C), which require an optimum value. The cost parameter (C) of the SVM formulation controls a penalty for misclassified data training and thus the complexity of the prediction function. Hence, a high-cost value C can force the SVM to create enough complex prediction function to misclassify as few training points as possible, whereas a low-cost parameter can lead to a simpler prediction function [23]. Moreover, the gamma parameter indicates how far the influence of a single training example reached with low and high values meaning “far” and “close,” respectively. The gamma parameters can be considered as the inverse of the radius of influence of samples selected by the model as support vectors. In this study, the best parameter values on gamma (ϒ) and cost (C) are at certain values that are applied to the proposed method.

4) The SVM kernel k-means clustering algorithm

This study aimed to improve the performance of the SVM classifier through the hybridization of the kernel *k*-means clustering algorithm. The algorithm details are as follows:

1. Input data set from DNA sequence decomposition, either nucleotide or amino acid composition.
2. Compute clustering data set using kernel *k*-means with the number of clusters (*k = 2*).
3. Tune the SVM parameter values (gamma (ϒ) and cost (C) to obtain optimal results, i.e., best parameter values).
4. Apply the SVM classifier method using the best parameter in each cluster.

4. Results and discussions

4.1 Data sets

In this study, the data sets were DNA sequences obtained from the HBV database (URL: https://hbvdb.ibcp.fr/HBVdb/). Two kinds of data set were decomposed from the sequences in genotypes A, B, C, and D. As presented in Table 4, there are 12 and 20 features of nucleotide composition and amino acid composition of various total data in any genotype, respectively.

4.2 Experimental results

The data sets were clustered using the kernel *k*-means method with the cluster number *k = 2.* It referred to the number of classes, i.e., Liver Cancer (LC) and normal (N). Then, using the tuning parameter of SVM with RBF or the polynomial kernel trick, this study obtained the best parameter value of the proposed method. The experimental results are presented in Tables 5 and 6.

Table 4. Data sets of nucleotide decomposition and amino acid decomposition

No	DNA decomposition	Geno-type	Status (total)	
			LC	Normal
1	Nucleotide	A	12	946
	Amino Acid			
2	Nucleotide	B	918	1542
	Amino Acid			
3	Nucleotide	C	376	2906
	Amino Acid			
4	Nucleotide	D	108	1419
	Amino Acid			

Table 5. Classification results using SVM RBF and kernel *k*-means

Data set	Prediction		Actual	
	LC	Normal	LC	Normal
Nucleo-A	11	947	12	946
Nucleo-B	506	1,954	515	1,945
Nucleo-C	360	2,922	420	2862
Nucleo-D	108	1,419	108	1,419
Amino-A	12	946	12	946
Amino-B	515	1,945	515	1,945
Amino-C	327	2,955	420	2,862
Amino-D	116	1,411	108	1,419

Table 6. Classification results using SVM polynomial and kernel *k*-means

Data set	Prediction		Actual	
	LC	Normal	LC	Normal
Nucleo-A	11	947	12	946
Nucleo-B	520	1940	515	1945
Nucleo-C	230	3052	420	2862
Nucleo-D	105	1422	108	1419
Amino-A	12	946	12	946
Amino-B	515	1945	515	1945
Amino-C	335	2947	420	2862
Amino-D	118	1409	108	1419

Based on the classification results, disease detection indicated that the error rate in average for the proposed method is 1.41 using the RBF kernel. This is lower than the proposed method using the polynomial kernel of 2.38. For liver cancer detection, the nucleotide composition data set is better to classify than the amino acid data set.

4.3 Performance result evaluation

To obtain the performance of the proposed method in classification results, it was evaluated by applying the confusion matrix. The matrix describes the performance of the classifier method on data testing in which the correct values are known as actual data. The terminology of confusion matrix is illustrated in Table 7 [24].

Remarks on Table 7:

- True positive (*tp*): The cases are predicted as carcinoma, and they are actually carcinoma.

Table 7. The confusion matrix

	Predicted: NO	Predicted YES
Actual: NO	*tn*	*fp*
Actual: YES	*fn*	*tp*

Table 8. The performance measure metrics

Measure	Formula	Definition
Accuracy	$\frac{tp+tn}{tp+fn+fp+tn}$	Correctness of a classifier
Sensitivity	$\frac{tp}{tp+fn}$	Effectiveness of a classifier to identify the positive label
Specificity	$\frac{tn}{tn+fp}$	Effectiveness of a classifier to identify the negative label
AUC	$\frac{1}{2}\left(\frac{tp}{tp+fn}+\frac{tn}{tn+fp}\right)$	The ability of the classifier to avoid false classification

- True negative (*tn*): The cases are predicted as not carcinoma (normal), and they are actually not carcinoma.
- False-positive (*fp*): The cases are predicted as carcinoma, but they are actually no carcinoma.
- False-negative (*fn*): The cases are predicted as not carcinoma, but they are actually carcinoma.

Moreover, there are various measurements for performance evaluation, including accuracy, sensitivity, specificity, and area under the curve (AUC), as presented in Table 8 [25].

4.4 Comparison of other machine learning algorithms

The most supervised learning algorithm, unbalanced data distribution effected to identify the class positive label, liver cancer, on the huge data volume. Therefore, the proposed method, SVM classification based on the Kernel *k*-Means cluster is addressed to simplify the classifier model by reducing the data variance through the similarity data characteristics in the same group.

This research was compared to other representative machine learning algorithms to detect liver cancer disease, i.e. rule based [26] or decision tree (C5.0), Probability model (Naïve Bayes)[27], Neural Network [28], and ensemble method (Random Forest)[29]. All data sets were applied using the proposed method; then, the performance of the experimental results was compared to the other machine-learning algorithms

Accuracy is a measurement to know the ability of the classifier model to predict correctly. The proposed method, either using RBF kernel or Polynomial kernel of SVM Kernel *k*-Means has the highest accuracy rate as shown in Tables 9 and 10.

Table 9. Comparison of the accuracies of machine-learning algorithms (Nucleotide composition)

Algorithms	**Datasets**			
	N$_A$	**N$_B$**	**N$_C$**	**N$_D$**
SVM RBF	0.99	0.93	0.88	0.97
SVM Poly	0.99	0.79	0.88	0.94
SVM KkM-RBF	**0.99**	**0.98**	**0.97**	**0.99**
SVM KkM-Poly	**0.99**	**0.96**	**0.93**	**0.98**
Random Forest	0.99	0.98	0.96	0.93
Naïve Bayes	0.97	0.81	0.87	0.93
Neural Network	0.99	0.86	0.87	0.93
C5.0	0.99	0.86	0.87	0.93

Table 10. Comparison of the accuracies of machine-learning algorithms (Amino Acid composition)

Algorithms	**Datasets**			
	AA$_A$	**AA$_B$**	**AA$_C$**	**AA$_D$**
SVM RBF	0.98	0.94	0.89	0.97
SVM Poly	0.99	0.95	0.91	0.97
SVM KkM-RBF	**1**	**0.98**	**0.97**	**0.98**
SVM KkM-Poly	**1**	**0.98**	**0.97**	**0.98**
Random Forest	0.99	0.98	0.96	0.98
Naïve Bayes	0.31	0.78	0.84	0.64`
Neural Network	0.99	0.91	0.89	0.97
C5.0	0.99	0.93	0.89	0.97

Table 11. Comparison of sensitivities of machine-learning algorithms (Nucleotide composition)

Algorithms	**Datasets**			
	N$_A$	**N$_B$**	**N$_C$**	**N$_D$**
SVM RBF	0.17	0.73	0.03	0.66
SVM Poly	0.67	0.03	0.04	0.19
SVM KkM-RBF	**0.96**	**0.96**	**0.82**	**0.86**
SVM KkM-Poly	**1**	**0.94**	**0.51**	**0.89**
Random Forest	0.92	0.92	0.72	0
Naïve Bayes	0.25	0.87	0	0
Neural Network	1	0.89	1	1
C5.0	1	0.93	0.12	0.71

Then, another measurement, a sensitivity of the proposed method, especially using RBF kernel has high sensitivity, closed to 1. It can predict positive liver cancer correctly even though the related data is small size when compared to the normal data (negative class) as shown in Tables 11 and 12.

Generally, the proposed method, SVM hybrid kernel k-means, has the highest performance in terms of accuracy, sensitivity, and specificity when compared to the other representative machine learning algorithms. By contrast, the naïve Bayes method has the lowest performance evaluation measurement. The average accuracy rate for all the machine-learning algorithms is presented in Fig. 2. It showed that the proposed method is dominant in other algorithms. The SVM using kernel polynomial and RBF combined with the kernel k-means cluster method reached accuracy rates of 0.97 and 0.98, respectively.

Otherwise, specificities, the ability to predict in negative class, of all method are high except neural network algorithm. The methods can predict a negative class (normal, non-Liver Cancer), due to high data volume in this class as shown in Table 13 and Table 14.

Then, the AUC of the proposed method is approximately 0.9. It is better than that of the SVM without the clustering algorithm, which is approximately 0.7, as depicted in Fig. 3. It is also dominant in other machine-learning algorithms. The

Table 12. Comparison of sensitivities of machine-learning algorithms (Nucleotide composition)

Algorithms	**Datasets**			
	AA$_A$	**AA$_B$**	**AA$_C$**	**AA$_D$**
SVM RBF	0	0.78	0.15	0.56
SVM Poly	0.75	0.85	0.26	0.56
SVM KkM-RBF	**1**	**0.96**	**0.75**	**0.87**
SVM KkM-Poly	**1**	**0.96**	**0.76**	**0.89**
Random Forest	0.92	0.94	0.7	0.89
Naïve Bayes	1	0.78	0.17	0.89
Neural Network	1	0.94	0.89	0.99
C5.0	1	0.75	0.13	0.59

Table 13. Comparison of specificities of machine-learning algorithms (Nucleotide composition)

Algorithms	**Datasets**			
	N$_A$	**N$_B$**	**N$_C$**	**N$_D$**
SVM RBF	1	0.98	1	0.99
SVM Poly	1	1	1	1
SVM KkM-RBF	**1**	**0.99**	**0.99**	**0.99**
SVM KkM-Poly	**1**	**0.97**	**0.99**	**0.99**
Random Forest	1	0.99	0.99	0.99
Naïve Bayes	0.98	0.81	1	1
Neural Network	0	0.75	0	0
C5.0	0.99	0.93	0.99	0.99

Table 14. Comparison of specificities of machine-learning algorithms (Amino Acid composition)

Algorithms	**Datasets**			
	AA$_A$	**AA$_B$**	**AA$_C$**	**AA$_D$**
SVM RBF	1	0.98	0.99	0.99
SVM Poly	1	0.97	0.99	0.99
SVM KkM-RBF	**1**	**0.99**	**0.99**	**0.99**
SVM KkM-Poly	**1**	**0.99**	**0.99**	**0.99**
Random Forest	1	0.95	0.99	0.99
Naïve Bayes	0.3	0.78	0.94	0.62
Neural Network	0	0.80	0.78	0.69
C5.0	1	0.97	0.99	0.99

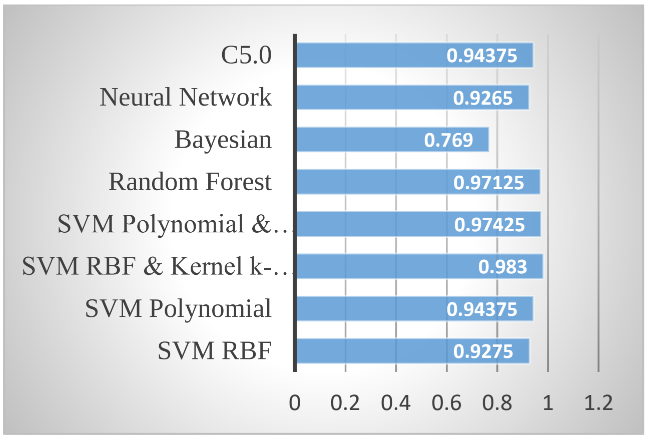

Figure. 2 Comparison of accuracy rates of machine-learning algorithms

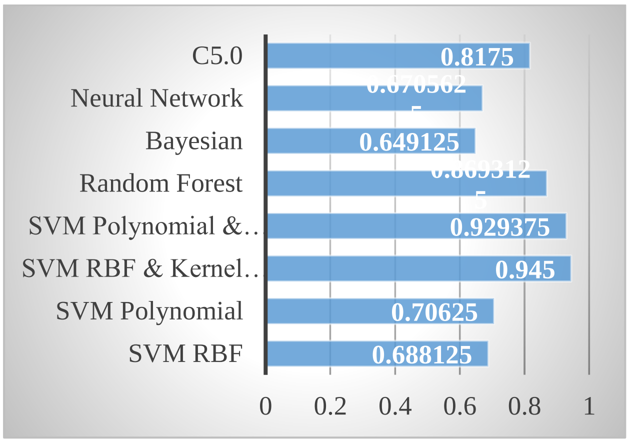

Figure. 3 Comparison of the AUC for machine-learning algorithms

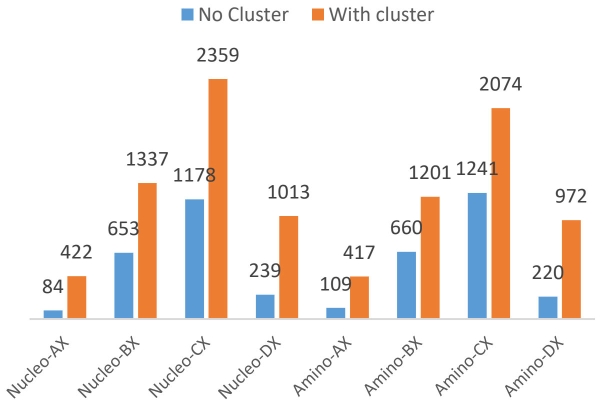

Figure. 4 Comparison of the number of support vector

AUC value of the proposed method is approximately 1. This implies that the method can well classify the DNA sequences.

When compared to the conventional SVM, the proposed method is effect to number of support vector as hyperplane of dataset as a classifier model as shown in Fig.4. The number of support vector of proposed method using cluster based is more than the number of support vector of the conventional SVM without cluster.

Futhermore, the quality of cluster is also effect to the performance result of detection. In the proposed method, the data set was clustered using Kernel *k*-Means before applied the SVM algorithm. The cluster was evaluated for the purity, entropy, shilouette to know the quality of the cluster and then, relationship of between the quality and performance of classification.

The purity and entropy are evaluation measurement to know the ability of a clustering method to recover the suitable classes [30]. Suppose we are given *l* categories, while the clustering method generates *k* clusters. The purity of the clustering with respect to the known categories is given by:

$$Purity = \frac{1}{n}\sum_{q=1}^{k} \max_{1\leq j\leq 1} n_q^j \qquad (11)$$

Where:

- n is the number of object data
- n_q^j is the number of object data in cluster q that belongs to original class *j* (1≤ *j* ≤ *l)*

The purity has range value in [0, 1]. The larger the purity, the better the performance of clustering.

Another evaluation is entropy of the clustering with respect to the known categories as the formulation given by:

$$Entropy = -\frac{1}{n\log_2 l}\sum_{q=1}^{k}\sum_{j=1}^{l} n_q^j \log_2 \frac{n_q^j}{n_q} \qquad (12)$$

Where:

- n is the number of object data
- n_q is the number of object data in cluster *q* (1≤ *q* ≤ *k)*
- n_q^j is the number of object data in cluster q that belongs to original class *j* (1≤ *j* ≤ *l)*

The smaller the entropy, the better the performance of clustering [30]. Then, the quality of cluster for all data sets can be shown in Fig 5. The nucleo-AX has the highest performance of cluster, i.e. silhouette, purity, and entropy. The silhouette and purity is the highest but the entropy of clustering is lowest when was compared to other datasets. The higher silhouette and purity of clustering, the higher accuracy, sensitivity and specificity of classification.

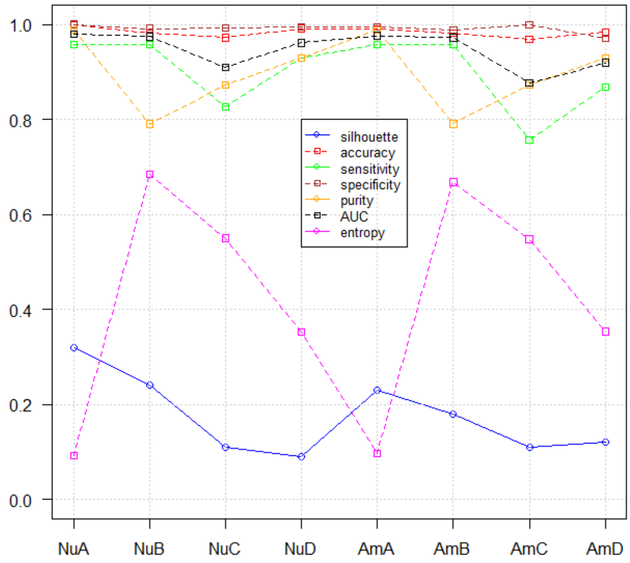

Figure. 5 Comparison of the evaluation performance for Kernel *k*-Means Clustering

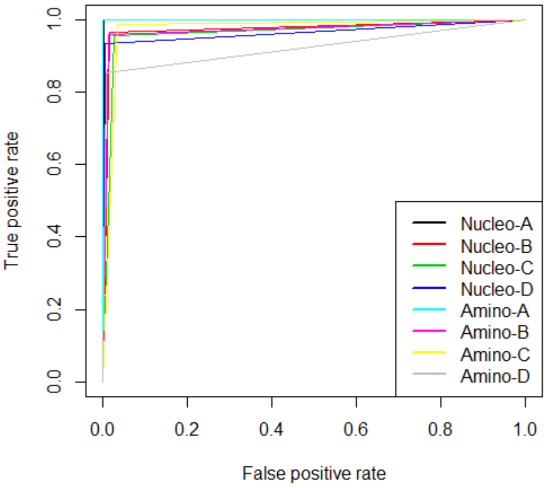

Figure. 6 The ROC of prediction using kernel *k*-means SVM (RBF)

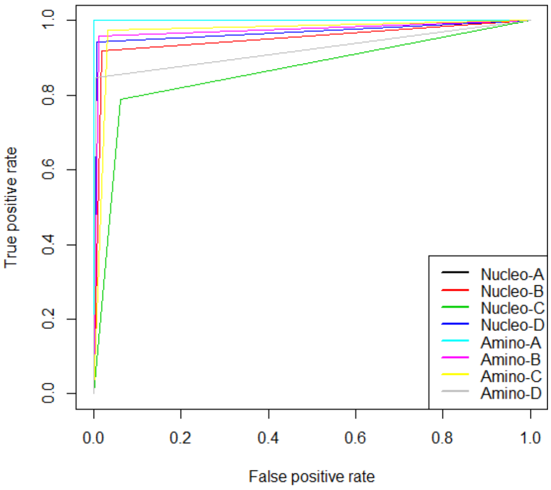

Figure. 7 ROC of prediction using kernel *k*-means SVM (polynomial)

Otherwise, the lower entropy of clustering, the higher accuracy, sensitivity and specificity of classification.

Finally, the receiver operating characteristic (ROC) curve is a tool for predicting the probability of a binary outcome. It is a plot of the false-positive rate (x-axis) versus the true positive rate (y-axis) for several candidate threshold values between 0.0 and 1.0 [31]. As illustrated in Figs. 6 and 7, the proposed method achieved a ROC of approximately 1.0. This means that the data sets can be well classified. In other words, based on the nucleotide or amino acid compositions of *HBx* HBV, the proposed method can be used to detect liver cancer disease.

5. Conclusion

Based on the kernel *k*-means clustering algorithm, this study demonstrated that the SVM algorithm can be used to detect liver cancer disease. Two kinds of data sets, including nucleotide and amino acid compositions, were extracted from the DNA sequences of *HBx* HBV in genotypes A–D. In the first stage, the DNA decompositions were clustered into two groups using the kernel *k*-means algorithm. Then, the SVM classifier was applied to each cluster by utilizing the kernel trick, including RBF or polynomial.

When clustered the data sets, each group has high similarity of the characteristic. The classification based on cluster has more the number of support vector than classification without cluster for detection. The quality of cluster has also effect to the performance of detection.The higher quality of cluster, the higher quality of detection.

In general, performance of the proposed method was higher than that of the conventional SVM method without the cluster. Moreover, it also appeared to be superior to the other machine-learning algorithms.

6. Future study

In this study, the proposed method was applied to kernel *k*-means using $k = 2$. In the future, we hope to develop a method for finding the best cluster number, *k*, to achieve the optimum prediction results.

References

[1] *Global Cancer Observatory*. [Online]. Available: http://gco.iarc.fr/. [Accessed: 07-Oct-2019].

[2] A. Ali, H.A.Hafiz, M.Suhail, A. A.Mars, M. K. Zakaria, K. Fatima, S.Ahmad, E.Azhar, A. Chaudhary, and I. Qadri,, “Hepatitis B virus, HBx mutants and their role in hepatocellular

carcinoma", *World J. Gastroenterol. WJG*, Vol. 20, No. 30, pp. 10238–10248, 2014.

[3] M. Iavarone, J. Trabut, O. Delpuech, F. Carnot, M. Colombo, D. Kremsdorf, C. Bréchot, and V. Thiers, "Characterisation of hepatitis B virus X protein mutants in tumour and non-tumour liver cells using laser capture microdissection", *J. Hepatol.*, Vol. 39, No. 2, pp. 253–261, 2003.

[4] K. Poussin , H. Dienes, H.Sirma , S.Urban, M. Beaugrand, D. Franco, P. Schirmacher, C. Bréchot, and P. P. Bréchot, "Expression of mutated hepatitis B virus X genes in human hepatocellular carcinomas", *Int. J. Cancer*, Vol. 80, No. 4, pp. 497–505, 1999.

[5] Q. Wang, T. Zhang, L. Ye, W. Wang, and X. Zhang, "Analysis of hepatitis B virus X gene (HBx) mutants in tissues of patients suffered from hepatocellular carcinoma in China", *Cancer Epidemiol.*, Vol. 36, No. 4, pp. 369–374, 2012.

[6] *ICETAS 2019 – 6th IEEE Conference*. [Online]. Available: https://icetas.etssm.org/. [Accessed: 01-Nov-2019].

[7] X. Bai, J. Jia, M. Fang, S. Chen, X. Liang, S. Zhu, S. Zhang, J.Feng, F. Sun, and C. Gao, "Deep sequencing of HBV pre-S region reveals high heterogeneity of HBV genotypes and associations of word pattern frequencies with HCC", *PLOS Genet.*, Vol. 14, No. 2, p. e1007206, 2018.

[8] L. Ali, A.Hussain, J. Li, A. Shah, U. Sdhakr, M. Mahmud, U. Zakir, X. Yan, B. Luo, and M. Rajak, "Intelligent image processing techniques for cancer progression detection, recognition and prediction in the human liver", In: *Proc. of 2014 IEEE Symposium on Computational Intelligence in Healthcare and e-health (CICARE)*, pp. 25–31, 2014.

[9] P. Radha and R. Divya, "Multiple time series clinical data with frequency measurement and feature selection", In: *Proc of 2016 IEEE International Conference on Advances in Computer Applications (ICACA)*, pp. 250–254, 2016.

[10] C. Shen and Z. Liu, "Identifying module biomarkers of hepatocellular carcinoma from gene expression data", In: *Proc. of 2017 Chinese Automation Congress (CAC)*, pp. 5404–5407, 2017.

[11] J. Wang, X. Wu, and C. Zhang, "Support vector machines based on K-means clustering for real-time business intelligence systems", *IJBIDM*, Vol. 1, pp. 54–64, 2005.

[12] B. Schölkopf, A. Smola, and K.-R. Müller, "Nonlinear Component Analysis as a Kernel Eigenvalue Problem", *Neural Comput.*, Vol. 10, No. 5, pp. 1299–1319, 1998.

[13] H. S. Rekha and P. Vijaya Lakshmi, "Classification on DNA Sequences of Hepatitis B Virus," In: *ICT and Critical Infrastructure: Proceedings of the 48th Annual Convention of Computer Society of India- Vol II*, pp. 431–443, 2014.

[14] V. Sruthi, N. T. Kesh, R. Priyanka, and S. G. Jacob, "Binary categorization of DNA data with unbalanced class distribution for prediction of hepatocellular carcinoma", In: *Proc. of 2015 International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT)*, pp. 490–494, 2015.

[15] *The Genetic Code*. [Online]. Available: https://web.archive.org/web/20161208155439/http://www.biology-pages.info/C/Codons.html. [Accessed: 04-Mar-2020].

[16] S. Kumar, G. Stecher, and K. Tamura, "MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets", *Mol. Biol. Evol.*, Vol. 33, No. 7, pp. 1870–1874, 2016.

[17] K. B. Kim, D. H. Song, and S.-S. Yun, "Automatic segmentation of wrist bone fracture area by K-means pixel clustering from X-ray image", *Int. J. Electr. Comput. Eng.*, Vol. 9, No. 6, pp. 5205–5210, 2019.

[18] A. N. Alfiyatin, W. F. Mahmudy, and Y. P. Anggodo, "K-Means Clustering and Genetic Algorithm to Solve Vehicle Routing Problem with Time Windows Problem", *Indones. J. Electr. Eng. Comput. Sci.*, Vol. 11, No. 2, pp. 462–468, 2018.

[19] Y. Cheung, "k*-Means: A new generalized k-means clustering algorithm", *Pattern Recognit. Lett.*, Vol. 24, pp. 2883–2893.

[20] I. Dhillon, Y. Guan, and B. Kulis, "A Unified View of Kernel k-means, Spectral Clustering and Graph Cuts", *UTCS Technical Report #TR-04-25*, p. 20. 2005.

[21] V. D. Sánchez A, "Advanced support vector machines and kernel methods", *Neurocomputing*, Vol. 55, No. 1–2, pp. 5–20, 2003.

[22] C. Sudheer, M. Rathinasamy, B. K. Panigrahi, and S. Mathur, "A hybrid SVM-PSO model for forecasting monthly streamflow", *Neural Comput. Appl.*, Vol. 24, No. 6, pp. 1381–1389, 2014.

[23] A. Karatzoglou, D. Meyer, and K. Hornik, "Support Vector Machines in R", *J. Stat. Softw.*, Vol. 15, No. 1, pp. 1–28, 2006.

[24] *Simple guide to confusion matrix terminology*, *Data School*, 26-Mar-2014. [Online]. Available:

https://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/. [Accessed: 30-Oct-2019].

[25] *Evaluating a Classification Model*, *ritchieng.github.io*. [Online]. Available: http://www.ritchieng.com/machine-learning-evaluate-classification-model/. [Accessed: 01-Nov-2019].

[26] T. A. Gameel, S. Rady, K. A. El-Bahnasy, and S. M. Kamal, "Prediction of liver cancer development risk in genotype 4 hepatitis C patients using knowledge discovery modeling", In: *Proc. of 2017 Eighth International Conference on Intelligent Computing and Information Systems (ICICIS)*, pp. 410–415, 2017.

[27] H. Ogihara, N. Iizuka, and Y. Hamamoto, "Prediction of Early Recurrence of Liver Cancer by a Novel Discrete Bayes Decision Rule for Personalized Medicine", *BioMed Research International*, 2016. [Online]. Available: https://www.hindawi.com/journals/bmri/2016/8567479/. [Accessed: 22-Nov-2019].

[28] A. M. E. Sherbini, M. M. Hagras, H. Farag, and M. R. M. Rizk, "Diagnosis and Classification of Liver Cancer using LIBS Technique and Artificial Neural Network", *International Journal of Science and Research*, pp. 1153-1158. 2015.

[29] Y. Li, Y. Dong, Z. Huang, Q. Kuang, Y. Wu, Y. Li, and M. Li, "Computational identifying and characterizing circular RNAs and their associated genes in hepatocellular carcinoma", *PLOS ONE*, Vol. 12, No. 3, p. e0174436, 2017.

[30] H. Kim and H. Park, "Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis", *Bioinforma. Oxf. Engl.*, Vol. 23, No. 12, pp. 1495–1502, 2007.

[31] T. Fawcett, "An introduction to ROC analysis", *Pattern Recognit. Lett.*, Vol. 27, No. 8, pp. 861–874, 2006.