IMPROVING PERFORMANCE FOR PREDICTION OF HEPATOCELLULAR CARCINOMA USING STACKING METHOD OF SUPPORT VECTOR MACHINE

Lailil Muflikhah, Widodo, Wayan Firdaus Mahmudy, Solimun

Brawijaya University, Indonesia, lailil@ub.ac.id

ABSTRACT

Hepatocellular Carcinoma is a serious disease that can be caused of Hepatitis B virus-infected and lead to the death. Support Vector Machine (SVM) is a robust classifier method to predict the disease. However, unbalanced class data distribution is often effect to the performance of prediction and tend to identify the high volume class. Therefore, this research aims to ensemble methods by stacking the learning model of SVM with other classifier methods to increase the performance evaluation measurement. In the proposed method, two or three algorithms, i.e. Random Forest with *k*-Nearest Neighbor, and or Generalized Linear Model were ensemble to construct as a bottom-layer classifier model and were applied to the SVM model as a top-layer classifier. As a result, the performance measure of the proposed method was higher than the conventional SVM. The proposed method the accuracy rate of 89%, sensitivity of 87.2% and specificity of 82%. It slightly increased from the SVM as 2% for accuracy and sensitivity. However, the specitivity significantly increased as 82%.

Keywords: hepatocellular carcinoma, SVM, stacking model

1. INTRODUCTION

In nowadays, Hepatocellular Carcinoma (HCC) is making the government to give extra attention due to increasing the total amount of patients each year. The HCC was the third-ranking cause of death at 8.2% (781,531 cases) based on the Global Cancer Observatory database, the International Agency for Research on Cancer (IARC) (Global Cancer Observatory, 2019.). Many studies showed that the role of *HBx* in the pathogenesis of viral-induced HCC(Ali *et al.*, 2014). By using the computational approach, the research on the patient's HCC that infected the *HBx* Hepatitis B virus was conducted by profiling the DNA sequence of the Hepatitis B Virus using the clustering method (Muflikhah *et al.*, 2019).

Support Vector Machine (SVM) is a robust classification method for HCC prediction. A lot of research was conducted using many types of datasets, such as image, clinical, microarray data of gene expression and DNA sequence (Ali *et al.*, 2014a; Bai *et al.*, 2018; Radha and Divya, 2016; Shen and Liu, 2017). However, the drawback of SVM is when applied to the huge data volume and unbalanced class. Ensemble methods are advanced techniques often used to solve complex machine learning problems. The method is a process where different and independent models as weak learners are combined to produce an outcome. The hypothesis is that combining multiple models can produce better results by decreasing generalization error. Stacking is one of the ensemble methods by combining many machine learning algorithms as a base layer of classification to predict the new data The research on the stacking machine learning model was conducted to improve accuracy in parallel computers (Gunes *et al.*, 2017). Many other studies also conducted the ensemble construction to make a robust classifier than a single classifier to improve the accuracy and ROC (Abawajy *et al.*, 2012; Buzhou Tang *et al.*, 2010). Therefore, this research aims to detect the Hepatocellular Carcinoma using the nucleotide composition of *HBx* HepB virus using stacking learning model of machine learning to SVM.

This paper is organized as follows. First, the background of this study and development of the hepatocellular carcinoma and detection of the disease in biological and computational approaches. Second, the research method including the proposed method using an ensemble method by stacking the learning model of SVM with other machine learning methods. The third section presents the results and discussion. The last section provides a conclusion and recommendations for further studies.

2. RESEARCH METHOD

In general, the research was conducted with several steps as shown in Figure 1. The nucleotide composition was a result of the transformation from DNA sequence of *HBx* Hepatitis B virus database at URL: https://hbvdb.ibcp.fr/HBVdb/. First, the dataset was applied to three machine learning algorithm i.e. Random Forest, K-Nearest Neighbor (KNN), and Generalized Linear Model (GLM) including the SVM algorithm. Second, get the correlation between SVM to other machine learning methods and make sure the correlation value is not high. Then, applied the proposed method by stacking the the machine learning algorithms to SVM algorithm.

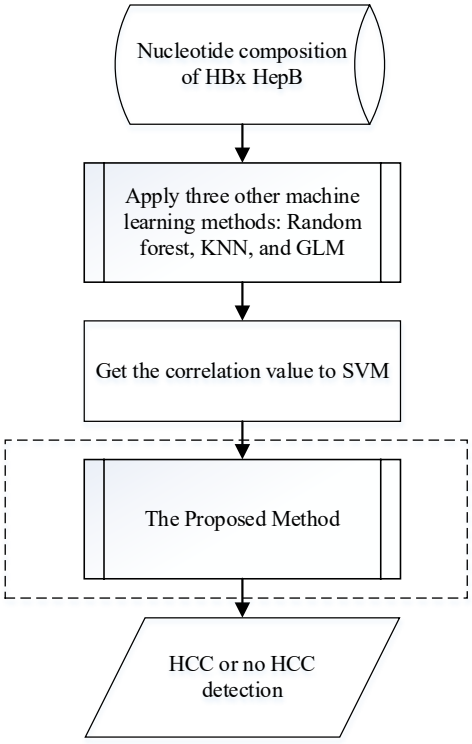

Fig. 1. General steps of this research

2.1 Random Forest Classifier

Random forest (RF) has been widely used as a robust machine learning method for classification and regression or other purposes. The classifier is built based on an ensemble of decision unpruned trees randomly during training (Breiman, 2001). The trees in the forest are grown using the CART method to maximum size without pruning. This subspace random selection scheme is resembled with bagging (resampling with replacement the training data set each time a new tree is built). It has been pointed out that the outperformance of random forests related to the good quality of each tree together with the small correlation among the trees of the forest. For the prediction of a new data sample, the classifier aggregates the outputs of all trees. The RF can deal with a large amount of data and can be used when the number of variables is much larger than the number of observations (Nguyen *et al.*, 2015).

2.2 Generalized Linear Model

Generalized Linear Models (GLM) extend the general linear model framework to address both of these issues: the range of Y is restricted (e.g. binary, count) and the variance of Y depends on the mean. A generalized linear model is made up of a linear predictor: $\eta_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}$ and two functions as this below Equation (Turner, n.d.)((Turner, n.d.):

- A link function that describes how the mean, $E(Y_i) = \mu_i$, depends on the linear predictor $g(\mu_i) = \eta_i$
- A variance function that describes how the variance, var (Y_i) depends on the mean $var\ (Y_i) = \phi V(\mu)$, where the dispersion parameter ϕ is a constant.

2.3 K-Nearest Neighbor (KNN)

K-Nearest Neighbours (KNN) algorithm is a method that uses the data points are separated into several separate classes to predict the classification of a new data point. The way in which the algorithm decides which of the points from the training set are similar enough to be considered when choosing the class to predict for a new observation is to pick the *k* closest data points to the new observation and to take the most common class among these (Sutton, 2012).

2.4 Support Vector Machine (SVM)

The Support Vector Machine (SVM) algorithm is initially a linear classification method that seeks the best function of hyperplane. The function divides two classes of input space, which are then developed into non-linear classifiers by incorporating kernel tricks in high-dimensional space. The data should be transformed into the vector space in a high dimension. The kernel trick functions that can be used in non-linear SVM classifications are Polynomial, Gaussian (RBF) and Sigmoid. Each label is denoted y¬i ϵ {-1, +1} for i = 1, 2, ..., n, where n is the number of data. The label is assigned +1 and -1 classes which can be completely separated from the hyperplane as defined in Equation (1):

$$w.x + b = 0 \quad (1)$$

The object data x_i is assigned to -1 as in Equation (2).

$$w.x_i + b \leq -1 \quad (2)$$

The object data x_i is assigned to +1 as in Equation (3).

$$w.x_i + b \geq +1 \quad (3)$$

The largest margin is calculated by maximizing the distance between the hyperplane and the nearest point

$$\frac{1}{\|w\|} \quad (4)$$

In principle, a non-linear SVM concept changes the data *x* that is applied to the function Φ (*x*) in the high dimensional vector space. The objective function represents data in the new vector space. In the SVM, the learning process is finding support vectors by dot product of the new vector space data. The kernel function aims to determine a support vector for non-linear data in the SVM learning process. The kernel function can be stated as in Equation (5)

$$K(x_i.x_j) = \Phi(x_i).\Phi(x_j) \quad (5)$$

This research used the Polynomial kernel trick as shown in Equation (6).

$$K(X_i \cdot X_j) = \exp\left(-\left(\frac{\|X_i - X_j\|^2}{2\sigma^2}\right)\right) \quad (6)$$

The next step is to make predictions by implementing the Sequential Support Vector Machine method including calculation of the Hessian matrix, iteration to reach the maximum in the least error rate or Max (| δα |) <ε. After that, the bias and similarities between the testing data and training data are calculated. As a result, it will be obtained the positive or negative classes (Vijayakumar and Wu, 1999).

2.5 The Proposed Method: SVM Stacking Learning Model

An ensemble is a set of classifiers that learn a target function, and their individual predictions are combined to classify the new data. Stacking is an ensemble learning technique that combines multiple classification or regression models via a meta-classifier or a meta-regressor. The base-level models are trained based on a complete training set, then the meta-model is trained on the outputs of the base level model-like features. The base-level often consists of different learning algorithms and therefore stacking ensembles are often heterogeneous. The algorithm in Figure 2 summarizes stacking (Zhou, 2012).

Input: training data $\boldsymbol{D} = \{\boldsymbol{x_i}, \boldsymbol{y_i}\}_{\boldsymbol{i=1}}^{\boldsymbol{m}}$
Output: ensemble classifier *H*
***Step* 1: learn base-level classifiers**
for *t*=1 to *T* do
    Learn $\boldsymbol{h_t}$ based on *D*
end for
***Step* 2: construct new data set of predictions**
for *i*=1 to *m* do
    $\boldsymbol{D_h} = \{\boldsymbol{x_i'}, \boldsymbol{y_i}\}$, where $\boldsymbol{x_i'} = \{\boldsymbol{h_1}(\boldsymbol{x_i}), \ldots, \boldsymbol{h_T}(\boldsymbol{x_i})\}$
end for
Step 3: learn a meta-classifier
learn *H* based on $\boldsymbol{D_h}$
return *H*

Fig. 2. General algorithm of stacking

In stacking multiple layers of machine learning models are placed one over another where each of the models passes their predictions to the model in the layer above it and the top layer model takes decisions based on the outputs of the models in layers below it. The model predictions of various individual models are not highly correlated with the predictions of other models.

This research used two-layer models as shown in Figure 3. The detail is as follows:

- The bottom-layer models (d_1, d_2, d_3) that consist of Random Forest, K-NN, and Generalized Linear Model received the original input features (x) from the nucleotide composition dataset.
- Top layer model, Support Vector Machine classifier, $f()$ which takes the output of the bottom layer models (d_1, d_2, d_3) as its input and predicts the final output.

Then, the out of fold predictions are used while predicting for the training data

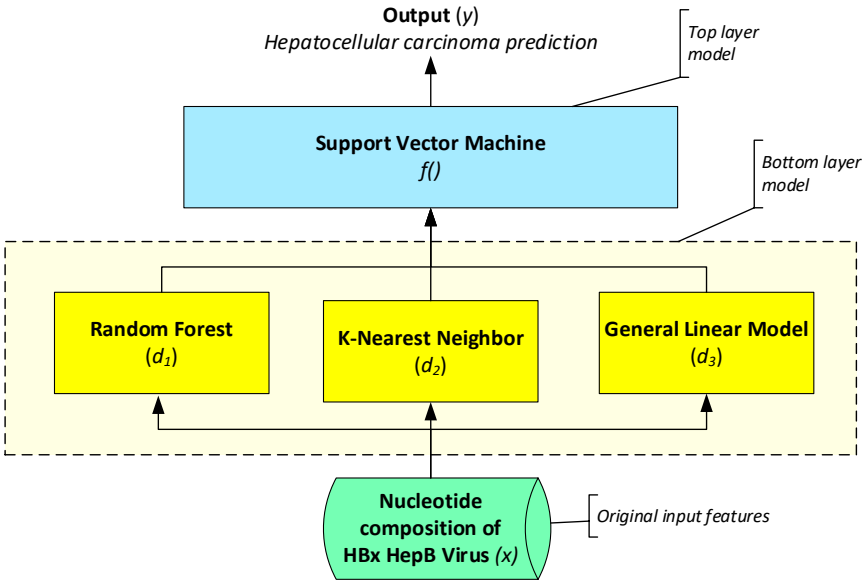

Fig. 3. The proposed method of stacking multi-layer machine learning model

3. RESULT AND DISCUSSION

3.1 Data Sets

This research used the nucleotide composition of DNA sequence - *HBx* Hepatitis B virus in genotype-C as data set. The regions of DNA sequence are representative of the nucleotide compositions such as Thymine (T), Cytocine (C), Adenine (A), Guanine (G) at the first, second, and third positions of codon. The relative frequencies of the four nucleotides can be computed for one specific sequence or for all sequences. For the coding regions of DNA, additional columns are presented for the nucleotide compositions at the first, second, and third codon positions (Kumar *et al.*, 2016). In this research, we use the MEGA tools of bioinformatics software that was transformed to nucleotide composition in percentage of codon as it was showed in Tabel 1.

Table 1. Nucleotide composition of HBx DNA sequence.

SId	T1	C1	A1	G1	T2	C2	A2	G2	T3	C3	A3	G3	Status
1	28	25.2	14.8	32.3	26	32.9	19.4	21.3	23	33.5	18.7	25.2	HCC
2	26	24.5	16.8	32.3	27	32.9	18.7	21.3	23	33.5	19.4	24.5	HCC
3	28	23.9	16.1	32.3	26	32.3	20	21.3	23	32.9	18.7	25.2	HCC
4	28	23.9	14.8	33.5	27	32.9	18.7	21.3	24	32.9	18.1	25.2	N
.													.
n	28	23.9	14.8	33.5	27	32.3	19.4	21.3	23	32.9	20	25.2	N

The composition of data set used is unbalanced class distribution, with 420 of HCC and 2862 of non-HCC as illustrated in Figure 4. The number of normal cases (non-HCC) is more than the number of carcinoma cases (HCC).

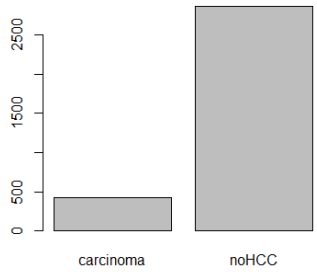

Fig. 4. The data set composition

3.2 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 2 ("Simple guide to confusion matrix terminology," 2014).

Table 2. The confusion matrix.

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

Remarks on Table 2:

- True positive (*tp*): The cases are predicted as carcinoma, and they are actually carcinoma.
- 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 3 ("Evaluating a Classification Model," 2019).

Table 3. 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

The data set was applied to RF, KNN, GLM, and SVM. As a result, the accuracy rate was shown in Table 4 and the kappa value was shown in Figure 5. The accuracy mean of SVM using Polynomial kernel achieved of 87.26%. However, the Random Forest can achieve the highest accuracy.

Table 4. Accuracy rate of machine learning algorithms sub models (RF, KNN, GLM, and SVM).

	Min.	1st Quartile	Median	Mean	3rd Quartile	Max.
Rf	0.878049	0.884909	0.887195	0.887569	0.892531	0.893617
Glm	0.871951	0.871951	0.871951	0.872029	0.871951	0.87234
Knn	0.859756	0.864329	0.870427	0.871111	0.878327	0.884146
svmPoly	0.871951	0.871951	0.871951	0.872638	0.872243	0.87538

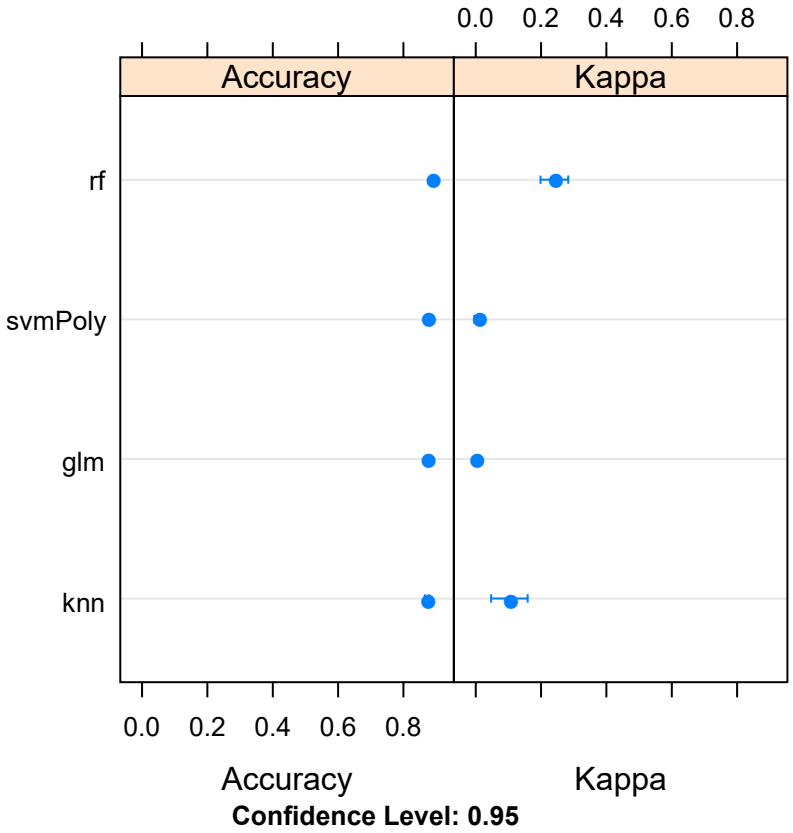

Fig. 5. Comparison of sub-models (Random Forest, SVM, GLM, and KNN) for stacking ensemble

Then, the correlation among the machine learning algorithms as sub models was evaluated to define as base-layer which is not high as shown in Table 5 and Figure 6. The highest correlation between svmPoly and Logistic Regression (GLM) at 0.475 is not high, so that recommended as sub-model layer of classifier.

Table 5. Correlation value among machine learning algorithms (RF, KNN, GLM, and SVM).

	Rf	glm	Knn	svmPoly
Rf	1	-0.00306	-0.02711	-0.1663
Glm	-0.00306	1	0.449073	0.475957
Knn	-0.02711	0.449073	1	0.024706
svmPoly	-0.1663	0.475957	0.024706	1

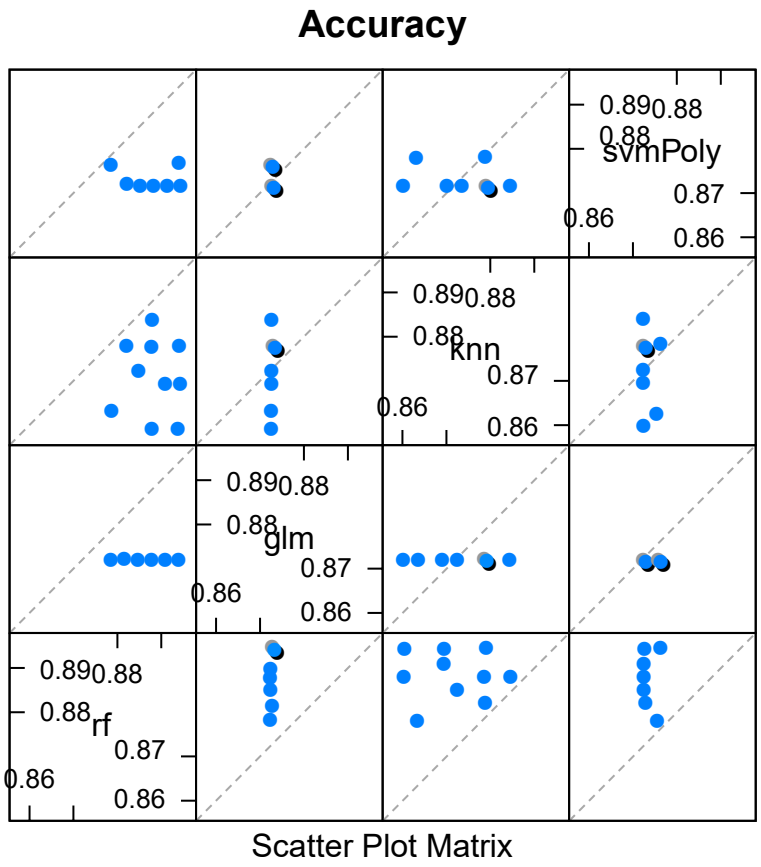

Fig. 6. Correlation between predictions made by sub-models in stacking ensemble

In this research, there are four scenarios of stacking ensemble methods as bottom layer, details as follows:

- Stack-1: Random Forest, Linear Regression and KNN
- Stack-2: Random Forest, Linear Regression
- STACK-3: Linear Regression, KNN
- STACK-4: Random Forest, KNN

By using the stacking method as bottom layer for predictor and SVM with polynomial kernel for top layer, then the comparison of accuracy rates were shown in Figure 7. The accuracy is a measurement to know the ability of the classifier model to predict correctly. The proposed method, stack-1, stack-2, and stack-4 had the higher accuracy rate than the conventional SVM. The ensemble with Random Forest increased the accuracy of SVM algorithm.

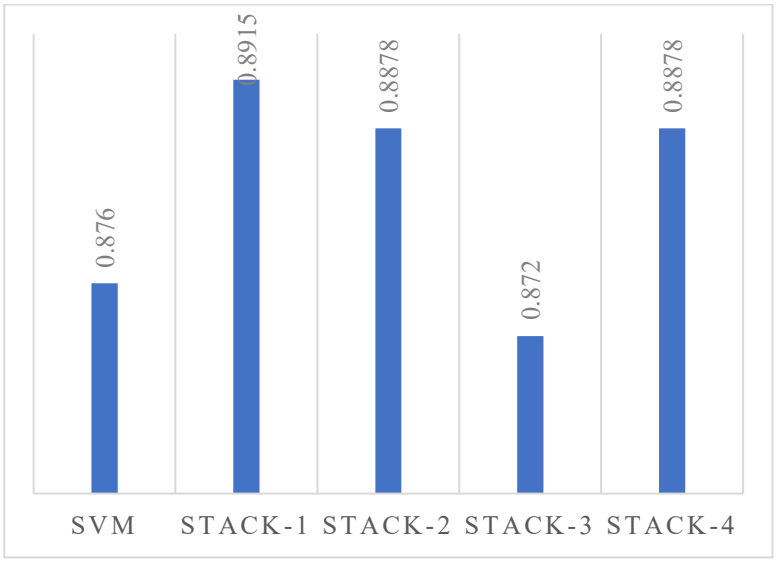

Fig. 7. Accuracy rate comparison of SVM against stacking ensemble methods

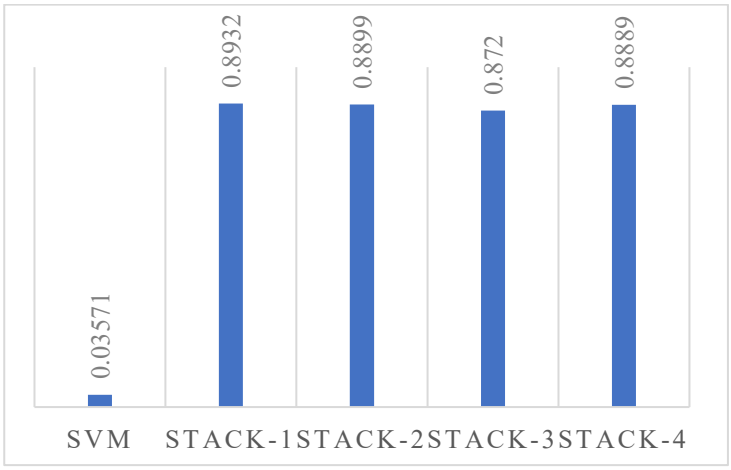

Fig. 8. Sensitivity comparison of SVM against stacking ensemble methods

Sensitivity is a measurement to predict positive carcinoma correctly even though the related data is small size when compared to the normal data (negative class). The proposed method, stacking by ensemble method showed the sensitivity is higher than the conventional SVM method as shown in Figure 8. It means that it was not depend on the class distribution.

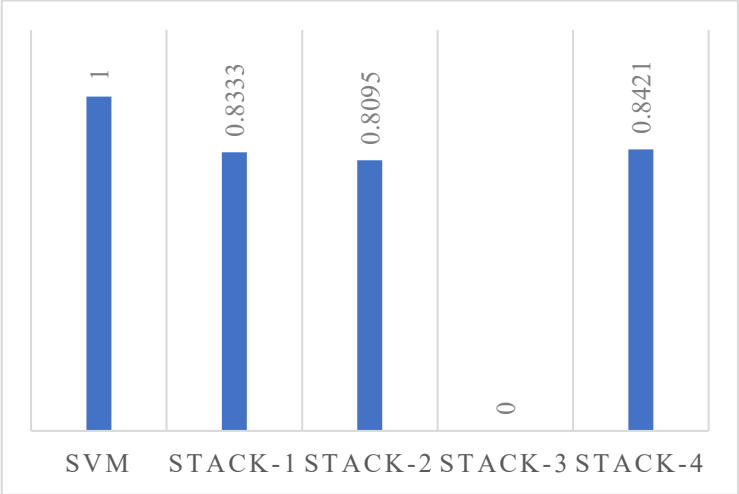

Fig. 9. Specificity comparison of SVM against stacking ensemble methods

Furthermore, the specificity, the ability to predict in negative class, of SVM is highest as shown in Figure 9. The SVM methods can predict a negative class (normal, non-HCC) as maximum, due to high data volume in this class.

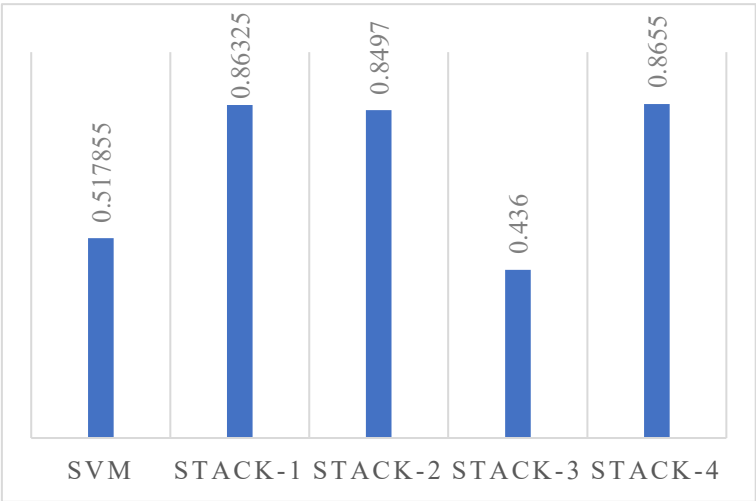

Fig. 10. The comparison of Area Under the Curve (AUC) for SVM against stacking ensemble methods

Finally, the AUC of the proposed method (Stack-1, Stack-2, and Stack-4) is higher than the conventional SVM using Polynomial kernel as shown in Fig. 10. It implies that the method can well classify to predict hepatocellular carcinoma disease based on the DNA sequences of *HBx* HepB.

4. CONCLUSION

Prediction of Hepatocellular Carcinoma disease was applied using the proposed method by stacking the learning model of Random Forest with *k*-Nearest Neighbor, and or Generalized Linear Model as a base-layer classifier model to SVM as a top-layer classifier. Because Random Forest has higher accuracy than other methods, then in stacking ensemble method, the prediction is based on the highest accuracy. In general, the performance evaluation result of the proposed method is higher than the Support Vector Machine in the single classifier model.

5. FUTURE WORK

The proposed method needs high computation for the learning model due to the ensemble from many algorithms. In the future, it is possible to develop the selected algorithm with fast computation for ensemble learning model.

ACKNOWLEDGMENTS

This research was financially supported by the Ministry of Research and Technology /National Agency for Research and Innovation (RISTEK/ DIKTI) in a program of Doctoral Dissertation Research Grant.

REFERENCES

Abawajy, J., and Kelarev, A. (2012). A Multi-tier Ensemble Construction of Classifiers for Phishing Email Detection and Filtering, In: *Cyberspace Safety and Security, Lecture Notes in Computer Science*. Xiang, Y., Lopez, J., Kuo, C.-C.J., and Zhou, W. (Eds.), 48-56. Springer: Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-35362-8_5

Ali, A., Abdel-Hafiz, H., Suhail, M., Al-Mars, A., Zakaria, M.K., Fatima, K., Ahmad, S., Azhar, E., Chaudhary, A., and Qadri, I. (2014). Hepatitis B virus, HBx mutants and their role in hepatocellular carcinoma. *World J. Gastroenterol.* WJG 20, 10238–10248. https://doi.org/10.3748/wjg.v20.i30.10238

Ali, L., Hussain, A., Li, J., Shah, A., Sudhakr, U., Mahmud, M., Zakir, U., Yan, X., Luo, B., and Rajak, M. (2014a). Intelligent image processing techniques for cancer progression detection, recognition and prediction in the human liver, In: *2014 IEEE Symposium on Computational Intelligence in Healthcare and E-Health (CICARE)*, 25-31. IEEE, Orlando, FL, USA. https://doi.org/10.1109/CICARE.2014.7007830

Bai, X., Jia, J., Fang, M., Chen, S., Liang, X., Zhu, S., Zhang, S., Feng, J., Sun, F., and Gao, C. (2018). Deep sequencing of HBV pre-S region reveals high heterogeneity of HBV genotypes and associations of word pattern frequencies with HCC. *PLOS Genet.* 14, e1007206. https://doi.org/10.1371/journal.pgen.1007206

Breiman, L. (2001). Random Forests. *Mach. Learn.* 45, 5–32. https://doi.org/10.1023/A:1010933404324

Buzhou, T., Qingcai, C., Xuan, W., and Xiaolong, W. (2010). Reranking for Stacking Ensemble Learning, In: *International Conference on Neural Information Processing ICONIP 2010: Neural Information Processing. Theory and Algorithms*, 575-584. LNCS 6443. Springer: Berlin/Heidelberg.

Evaluating a Classification Model. (2019). http://www.ritchieng.com/machine-learning-evaluate-classification-model/ (last accessed on November 01, 2019).

Global Cancer Observatory (2019). http://gco.iarc.fr/ (last accessed on July 10, 2019).

Gunes, F., Wolfinger, R., and Tan, P.-Y. (2017). Stacked Ensemble Models for Improved Prediction Accuracy, In: *SAS Global Forum 2017*, 1-19.

Kumar, S., Stecher, G., and Tamura, K. (2016). MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Mol. Biol. Evol.*, 33, 1870–1874. https://doi.org/10.1093/molbev/msw054

Muflikhah, L., Widodo, Mahmudy, W.F., Solimun (2019). DNA Sequence of Hepatitis B Virus Clustering Using Hierarchical k-Means Algorithm. *Presented at 6th IEEE International Conference on Engineering Technologies and Applied Sciences (ICETAS)*, December 20-21, 2019, Kuala Lumpur, Malaysia. https://icetas.etssm.org/ (last accessed on October 07, 2019).

Nguyen, T.-T., Huang, J.Z., Wu, Q., Nguyen, T.T., and Li, M.J. (2015). Genome-wide association data classification and SNPs selection using two-stage quality-based Random Forests. *BMC Genomics*, 16, S5. https://doi.org/10.1186/1471-2164-16-S2-S5

Radha, P., and Divya, R. (2016). Multiple time series clinical data with frequency measurement and feature selection, In: *2016 IEEE International Conference on Advances in Computer Applications (ICACA)*, 250–254. https://doi.org/10.1109/ICACA.2016.7887960

Shen, C., and Liu, Z. (2017). Identifying module biomarkers of hepatocellular carcinoma from gene expression data, In: *2017 Chinese Automation Congress (CAC)*, 5404–5407. IEEE: New York NY. https://doi.org/10.1109/CAC.2017.8243741

Simple guide to confusion matrix terminology (2014). Data Sch. URL https://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/ (last accessed on October 30, 2019).

Sutton, O. (2012). Introduction to k Nearest Neighbour Classification and Condensed Nearest Neighbour Data Reduction 10. (accessible at: http://www.math.le.ac.uk/people/ag153/homepage/KNN/OliverKNN_Talk.pdf)

Turner, H. (2008). Introduction to Generalized Linear Models 52. (accessible at: https://statmath.wu.ac.at/courses/heather_turner/glmCourse_001.pdf)

Vijayakumar, S., and Wu, S. (1999). Sequential Support Vector Classifiers and Regression, In: *Proc. Int. Conf. on Soft Computing*, 5, 610-619.

Zhou, Z.-H. (2012). *Ensemble Methods: Foundations and Algorithms*. Chapman & Hall/CRC.