Open Access **Baghdad Science Journal** **P-ISSN: 2078-8665**
2021, 18(4) Supplement: 1406-1412 **E-ISSN: 2411-7986**

DOI: https://dx.doi.org/10.21123/bsj.2021.18.4(Suppl.).1406

Recurrent Stroke Prediction using Machine Learning Algorithms with Clinical Public Datasets: An Empirical Performance Evaluation

Fadratul Hafinaz Hassan*** ***Mohd Adib Omar***

School of Computer Sciences, Universiti Sains Malaysia, 11800 Minden, Pulau Pinang, Malaysia.
*Corresponding author: fadratul@usm.my
E-mails: adib@usm.my

Received 14/10/2021, Accepted 14/11/2021, Published 20/12/2021

Abstract:

Recurrent strokes can be devastating, often resulting in severe disability or death. However, nearly 90% of the causes of recurrent stroke are modifiable, which means recurrent strokes can be averted by controlling risk factors, which are mainly behavioral and metabolic in nature. Thus, it shows that from the previous works that recurrent stroke prediction model could help in minimizing the possibility of getting recurrent stroke. Previous works have shown promising results in predicting first-time stroke cases with machine learning approaches. However, there are limited works on recurrent stroke prediction using machine learning methods. Hence, this work is proposed to perform an empirical analysis and to investigate machine learning algorithms implementation in the recurrent stroke prediction models. This research aims to investigate and compare the performance of machine learning algorithms using recurrent stroke clinical public datasets. In this study, Artificial Neural Network (ANN), Support Vector Machine (SVM) and Bayesian Rule List (BRL) are used and compared their performance in the domain of recurrent stroke prediction model. The result of the empirical experiments shows that ANN scores the highest accuracy at 80.00%, follows by BRL with 75.91% and SVM with 60.45%.

Keywords: Artificial Neural Network, Bayesian Rule List, Machine Learning, Recurrent Stroke Prediction, Support Vector Machine.

Introduction:

It is reported that stroke is one of the top five leading causes of death in Malaysia. It is happened when the brain cells stop functioning due to the blockage of blood flow to the brain [1]. The clogging of the blood may reduce the oxygen level that may further create another symptom such as loss of speech, weakness, or paralysis of one side of the body. However, these symptoms can be reduced by immediate and appropriate medical care in the early stage. For example, controlling high blood pressure and diabetes [2]. Recurrent means something that happens repeatedly. Recurrent stroke means the repeated occurrence of stroke. This repetitive occurrence causes even worst impact – more rate of death and disability due the history of the patients with first time stroke. The brain already injured by the first-time stroke may not be strong as the patient without history of stroke [3]. There are 795,000 strokes per year while 185,000 of them are recurrent strokes. The risk for another stroke can increase more than 40% within 5 years of a first stroke [4].

Besides that, this research also aims to investigate which machine learning algorithm has better performance in predicting recurrent stroke. Support Vector Machine (SVM), Artificial Neural Network (ANN) and Bayesian Rule List (BRL) have been suggested to be implemented in the recurrent stroke prediction model. The result of these machine learning algorithms will be examined and concluded at the end of this research [5].

Materials and Methods:

Recurrent Stroke Prediction Model

The amount of recurrent stroke has been increased recently which causes more people to suffer. This is because recurrent stroke leads to 40% increment of patients' death. To prevent this from happening, the doctor needs to decide the most suitable therapy for patients. For example, if the patients have a high probability of getting a recurrent stroke, the doctor should give suitable

treatment to prevent another stroke. Also, the therapy could be individualized to the patients.

To fulfill the requirement of individualizing of treatment, the doctor should be able to know which risk factors are the causes of recurrent stroke so that doctors could know the reason for patients to get a stroke. Identifying the cause that trigger the initial event of stroke is the main step in hindering recurrent stroke. At the same time, the doctor should also be able to predict the recurrent stroke correctly so that the doctor could know which type of treatment could be used in the patients. This is because different severity of patients is used a different type of medication. Hence, a personalized therapy is the best treatment in preventing the recurrent stroke [6].

In this case, the accuracy of recurrent stroke prediction models has become an important factor. This is because the doctor can make used of the model to identify the initial event and predict the risk of getting the recurrent stroke. Based on the result shows, the doctor can even use different strategies to treat their patients. At the same time, patients can know their own conditions through the recurrent stroke prediction model.

Review on Machine Learning Algorithms in Stroke Prediction Model

The implementation of the machine learning algorithm in recurrent stroke prediction model is important to increase the prediction accuracy of the model. There are several algorithms that have been mentioned in the previous research paper: Bayesian Rule List (BRL), Support Vector Machine (SVM), Artificial Neural Network (ANN), Random Forest (RF) and K-nearest Neighbors (KNN). In this subtopic, machine learning algorithms will be implemented in a recurrent stroke prediction model to compare the prediction accuracy.

One previous research paper shows that machine learning algorithm could increase the performance of stroke prediction model [7]. This has been proven by using one of the research paper that researcher compare Congestive heart failure, Hypertension, Age, Diabetes mellitus, Prior Stroke ($CHADS_2$) with prediction model of Artificial Neural Network, found out that Artificial Neural Network perform better than $CHADS_2$. Thus, after reviewing the research papers, Table I has been designs to show the frequency of proposed machine learning algorithms related to stroke prediction model. From the previous research paper, there are several machine learning algorithms have been used which are ANN, BRL, RF, KNN and SVM.

Table 1 summarizes the frequency values of ML methods in the stroke prediction model from the previous works. SVM has been proposed for four times while ANN has been proposed for three times and BRL for two times. Since recurrent stroke has the similarities of variables in some aspect with stroke, this research is going to examine whether the ML algorithm that proposed in stroke prediction will have equally performance in prediction of recurrent stroke. Thus, based on the proposed frequency, this research is going to perform an experiment to investigate which machine learning algorithm has better performance in predicting recurrent stroke.

Experiments:

Datasets

Table 2 shows that when Dependent Variable (DV) is zero means that patients do not have a recurrent stroke while DV is one means that patients have a recurrent stroke. The public dataset is from the public dataset repository, Kaggle website. Eight variables in the dataset are History Anti Platelet (P_ALT), Hyperlipidemia (HPLD), Ischemic Heart Disease (IHD), Age (AGE0), Angiotensin-Converting Enzyme (P_ACEI), Smoking status (SMOKER), Angiotensin Receptor Blocker (P_ARB), and Lipid-Lowering drug (P_LL) as shown in the Table 2. P_ALT, P_ACEI, P_ARB, and P_LL are the type of drugs took by the patients. In details, P_ALT is the drug preventing platelets sticking to atherosclerotic plaques, P_ACEI is the drugs controlling blood pressure, P_ARB is the drugs helping to relax to the blood vessels and P_LL is the drugs treating the level of fats in the blood.

Algorithms:

Artificial Neural Network: The example in Table 3 below shows how ANN is used to predict stroke. The data is separated into training and test data set. The model is then build based on the setting below. It consists of five layers in the built ANN model. The learning rate of model is 0.01, while the epoch has been sets to 200. The information of the layer is defined as below.

Table 1. Comparison of frequency values with their strength and limitation of ML algorithm applied in the stroke prediction model

ML Algorithms	Frequency	Strength	Limitation
SVM	4	1. Has a regularization parameter to avoid over-fitting [8] 2. Engineering kernel helps to build an expert knowledge [8] 3. Has convex optimization problem for which there are efficient methods [8,9] 4. The estimate to a bound-on test fault percentage [10]	1. The theory only really covers the determination of the parameters [11,12] 2. Overfitting issue in optimizing parameters in modelling [12] 3. Hence, the kernel models can be oversensitive
ANN	3	1. Can handle many data sets [13] 2. Able to detect complex non-linear correlation between dependent and independent variables [14,15] 3. Able to detect all potential interactions between predictor variables [15]	"Black-box" nature of Artificial Neural Networks, can be difficult to interpret [15]
BRL	2	1. Simple in structure [16] 2. Moderately stable in classifying different sizes of training data sets and producing moderately good results of classification 3. Fastest in producing classification results	Not practical to represent the causal relationships of the training data sets due to the fact that the assumption of independence of features can be false [16]
RF	1	Has an effective method for estimating missing data and maintains accuracy when a large proportion of the data are missing [17]	Random forests have been observed to over-fit for some data sets with noisy classification or regression tasks [17]
KNN	1	Robust to noisy training data Effective if training data is large [18]	Lazy learner, it does not learn anything from the training data and simply uses the training data itself for classification [19]

Table 2. Sample data set used to predict recurrent stroke

P_APLT	HPLD	IHD	AGE0	P_ACEI	SMOKER	P_ARB	P_LL	DV
0	0	0	58.76	1	1	1	0	0
0	1	0	72.942	1	1	1	0	0
0	0	0	49.297	1	1	1	0	0
0	0	0	55.997	1	1	1	0	1
0	0	1	80.395	1	1	1	0	0
1	0	1	68.735	1	0	1	1	0
0	0	0	73.538	1	1	1	1	0
1	1	0	70.614	1	0	1	1	0
0	0	0	73.877	1	1	1	0	0
0	0	0	52.031	1	0	1	0	0
1	0	0	55.428	1	1	1	1	1
0	1	0	52.476	1	0	1	0	0
1	1	0	60.639	1	1	1	0	0
1	0	0	67.791	1	0	1	1	0
0	0	1	57.772	1	0	1	1	0
0	0	0	59.8	1	0	1	0	0
0	1	0	64.916	1	1	1	0	0
0	1	0	59.381	1	1	1	1	0
1	1	0	51.242	1	1	1	1	0
0	0	0	81.439	1	0	1	0	0

Table 3. Layer setting in ANN model

Number of layer	Neuron in the layer
First	128
Second	64
Third	64
Fourth	64
Fifth	10

• **Support Vector Machine:** the example in Table 4 below shows how SVM is used for prediction of stroke. The first step is to prepare the training data used for differentiating between patients having recurrent stroke and non-recurrent stroke. Next, is to plot the graph to see the distribution of data. After plotting the graph as in the Figure 1, a suitable line or hyper-lane to separate between data of stroke and recurrent stroke is selected. The graph is used to differentiate whether the patients will get recurrent stroke or not if plotting point is closer to the group that have recurrent stroke means that the patients will have recurrent stroke.

Table 4. Sample of score risk factors values to plot graph in SVM

Total score of other risk factors	Age0	DV
3	58.760	0
4	72.942	0
3	49.297	0
3	55.997	1

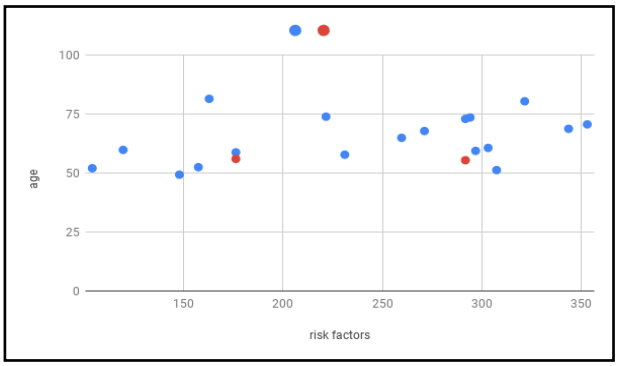

Figure 1. Distribution of data using SVM.

• **Bayesian Rule List:** The example below shows the process of predicting recurrent stroke based on Bayesian Rule List model. When the model has been given 3 variables which are p_ACEI, smoker and p_ARB to predict recurrent stroke. First, we need to consider the training data that have been shown in Table 2. A likelihood table needs to be prepared based on the number of variables, if there are three variables used for prediction, then three likelihood tables need to be prepared.

Table 5. Likelihood table for the p_ACEI variable

Likelihood table		Recurrent stroke		
		Yes	No	
p_ACEI	Yes	2/2	18/18	20/20
	No	0/2	0/18	0/20
		2/20	18/20	20

Table 6. Likelihood table for smoker variable

Likelihood table		Recurrent stroke		
		Yes	No	
SMOKER	Yes	2/2	10/18	12/20
	No	0/2	8/18	8/20
		2/20	18/20	20

Table 7. Likelihood table for the p_ARB variable

Likelihood table		Recurrent stroke		
		Yes	No	
P_ARB	Yes	2/2	18/18	20/20
	No	0/2	0/18	0/20
		2/20	18/20	20

In this experiment, three likelihood tables are shown in the Table 5, Table 6 and Table 7. After preparing the likelihood table for each variable, the model will calculate on the probability whether will get a recurrent stroke or not. Thus, a patient that has p_ACEI, p_ARB and is a smoker who has higher probability to get recurrent stroke based on BRL.

Experimental Results:

The number of TP, TN, FP and FN will be applied in the formula of sensitivity, specificity, accuracy, precision and F1 Score to measure the performance of the machine learning algorithm. In this empirical experiment, the sensitivity value helps to detect ill patients who have a recurrent stroke condition. High specificity value measures the percentage of disease free patients who are correctly diagnosed as being disease free.

Baghdad Science Journal
P-ISSN: 2078-8665
E-ISSN: 2411-7986

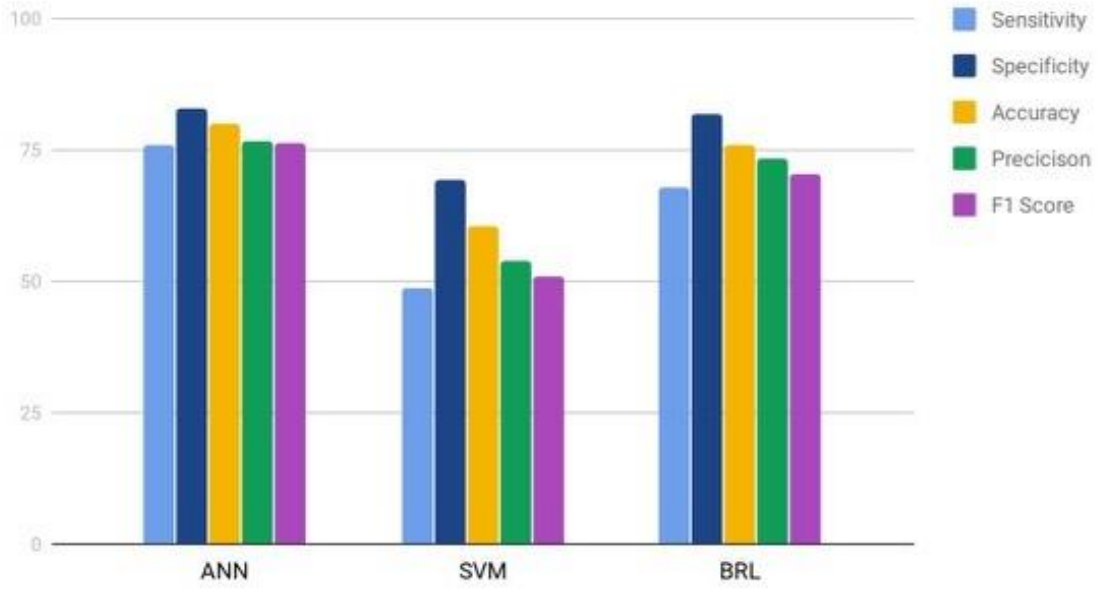

Figure 2. Performance comparison of ANN, SVM and BRL based on the sensitivity, specificity, accuracy, precision and F1 Score.

Table 8 shows that ANN model has the lowest false negative values of 67 compares to SVM and BRL. SVM scores the highest false negative values of 144 and the lowest total values of true positive of 136. It can be concluded that ANN approach is the most optimal model with the highest true positive values of 213 and the lowest false negative and false positive values.

Table 8. Total values of TP, TN, FP, FN in each ML models with clinical public datasets

Criteria	ANN	SVM	BRL
TP	**213**	**136**	190
TN	315	263	311
FP	65	117	69
FN	**67**	**144**	90

Table 9. Summary performance of ANN, SVM and BRL with clinical public datasets

Evaluation criteria	ANN (%)	SVM (%)	BRL (%)
Sensitivity	76.07	48.57	67.86
Specificity	82.89	69.21	81.84
Accuracy	**80.00**	**60.45**	75.91
Precision	76.62	53.75	73.35
F1 Score	76.34	51.03	70.50

The results from the empirical experiments show that ANN perform well and scores the highest values for all the criteria, as depicted in Table 9 and Figure 2. ANN model generated the value of 76.07% for sensitivity, 82.89% of specificity, 80.00% of accuracy, 76.62% of precision and 76.34% of F1 Score. The overall comparison of accuracy values among ANN, and SVM and BRL shows that SVM has the lowest accuracy value at 60.45% when compare to BRL with 75.91% accuracy rate and ANN with 80.00% accuracy. Thus, ANN has a better performance compare to SVM and BRL approaches when testing using clinical public dataset from Kaggle.

Conclusions:

In this research paper, SVM, BLR, and ANN have been proposed to predict recurrent stroke. The performance of the algorithms will be evaluated based on the sensitivity, specificity, accuracy, precision and F1 score using the clinical public dataset. It can be concluded that:

- ANN is the most optimal model with the lowest FP and FN values of 65 and 67, respectively. In other words, ANN model with higher sensitivity will generate accurate prediction of recurrent stroke compared to other models.
- ANN reached 80.00% of accuracy and has and overall better performance in term of sensitivity, specificity, accuracy, precision and F1 score when compared to BRL and SVM approaches.

Thus, this preliminary result of empirical performance evaluation plays a remarkable role as the benchmark for further analysis of ML methods in the recurrent stroke domain. Further study would include testing the models with different datasets and compare with real datasets. It is also worth to explore and compare the accuracy values with other mathematical models.

Acknowledgment:

The "Ministry of Higher Education Malaysia" provided support for this research under Fundamental Research Grant Scheme with Project Code: FRGS/1/2018/ICT02/USM/02/10".

Authors' declaration:

- Conflicts of Interest: None.
- We hereby confirm that all the Figures and Tables in the manuscript are mine ours. Besides, the Figures and images, which are not mine ours, have been given the permission for re-publication attached with the manuscript.
- The author has signed an animal welfare statement.
- Ethical Clearance: The project was approved by the local ethical committee in University of Sains Malaysia**.**

Authors' contributions:

F.H.H and M.H.O contributed to the design and implementation of the research, to the analysis of the results and to the writing of the manuscript

References:

1. Guide- Miss. B. M. Gund, Mrs. P. N. Jagtap, Mr. V. B. Ingale, and Dr. R. Y. Patil, "Stroke: A Brain Attack," IOSR Journal of Pharmacy, vol. 3, pp. 1-23, 2013.
2. American Heart Association. (2015, August 30). Prevention and Treatment of Diabetes. Available: https://www.heart.org/en/health-topics/diabetes/prevention--treatment-of-diabetes
3. R. Murugappan, "Protecting Against Stroke," in The Star, ed. Malaysia: Star Media Group Berhad 2017.
4. J. Burn, M. Dennis, J. Bamford, P. Sandercock, D. Wade, and C. Warlow, "Long-term risk of recurrent stroke after a first-ever stroke. The Oxfordshire Community Stroke Project," Stroke, vol. 25, pp. 333-337, 1994.
5. M. Awad and R. Khanna, Efficient learning machines: theories, concepts, and applications for engineers and system designers: Apress, 2015.
6. J. D. Spence, "Recent advances in preventing stroke recurrence," F1000Research, vol. 6, 2017.
7. H. Asadi, R. Dowling, B. Yan, and P. Mitchell, "Machine learning for outcome prediction of acute ischemic stroke post intra-arterial therapy," PloS one, vol. 9, p. e88225, 2014.
8. Arslan, A.K., Colak, C. and Sarihan, M.E., "Different medical data mining approaches based prediction of ischemic stroke," Computer methods and programs in biomedicine, 130, pp.87-92, 2016.
9. Jeena, R.S. and Kumar, S., "Stroke prediction using SVM". In 2016 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT) (pp. 600-602). IEEE, 2016
10. Hung, C.Y., Chen, W.C., Lai, P.T., Lin, C.H. and Lee, C.C., "Comparing deep neural network and other machine learning algorithms for stroke prediction in a large-scale population-based electronic medical claims database", In 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (pp. 3110-3113). IEEE, 2017.
11. Weng, S.F., Reps, J., Kai, J., Garibaldi, J.M. and Qureshi, N.,"Can machine-learning improve cardiovascular risk prediction using routine clinical data?", PloS one, 12(4), p.e0174944, 2017.
12. Monteiro, M., Fonseca, A.C., Freitas, A.T., e Melo, T.P., Francisco, A.P., Ferro, J.M. and Oliveira, A.L., "Using machine learning to improve the prediction of functional outcome in ischemic stroke patients", IEEE/ACM transactions on computational biology and bioinformatics, 15(6), pp.1953-1959, 2018.
13. Asadi, H., Dowling, R., Yan, B. and Mitchell, P., "Machine learning for outcome prediction of acute ischemic stroke post intra-arterial therapy", PloS one, 9(2), p.e88225, 2014.
14. Colak, C., Karaman, E. and Turtay, M.G., "Application of knowledge discovery process on the prediction of stroke. Computer methods and programs in biomedicine", 119(3), pp.181-185, 2015
15. Purusothaman, G. and Krishnakumari, P., "A survey of data mining techniques on risk prediction: Heart disease", Indian Journal of Science and Technology, 8(12), p.1., 2015.
16. Sung, S.F., Hsieh, C.Y., Yang, Y.H.K., Lin, H.J., Chen, C.H., Chen, Y.W. and Hu, Y.H., "Developing a stroke severity index based on administrative data was feasible using data mining techniques", Journal of clinical epidemiology, 68(11), pp.1292-1300, 2015.
17. Kirasich, K., Smith, T. and Sadler, B., "Random forest vs logistic regression: binary classification for heterogeneous datasets". SMU Data Science Review, 1(3), p.9, 2018.
18. Okut, H., "Bayesian regularized neural networks for small n big p data", Artificial neural networks-models and applications, pp.28-48, 2016.
19. Kim, J. and Canny, J., "Explainable Deep Driving by Visualizing Causal Attention", In Explainable and Interpretable Models in Computer Vision and Machine Learning (pp. 173-193). Springer, Cham, 2018.

Open Access **Baghdad Science Journal** **P-ISSN: 2078-8665**
2021, 18(4) Supplement: 1406-1412 **E-ISSN: 2411-7986**

التنبؤ بالسكتة الدماغية المتكررة باستخدام خوارزميات التعلم الآلي مع مجموعات البيانات السريرية العامة: تقييم أداء تجريبي

فضراتول هافيناز حسن * **محمد أديب عمر**

كلية علوم الحاسبات ، يونيفرسيتي سينز ماليزيا ، 11800 ميندن ، بولاو بينانج ، ماليزيا.

الخلاصة:
غالبًا ما تكون السكتة الدماغية المتكررة مدمرة وقادرة على التسبب في إعاقة شديدة أو الوفاة. ومع ذلك ، فإن ما يقرب من 90 ٪ من أسباب السكتة الدماغية المتكررة قابلة للتغير ، مما يعني أنه يمكن تجنب السكتات الدماغية المتكررة عن طريق التحكم في عوامل الخطر ، والتي هي في الأساس سلوكية واستقلابية بطبيعتها. وبالتالي ، يتضح من الأعمال السابقة أن نموذج التنبؤ بالسكتة الدماغية المتكررة يمكن أن يساعد في تقليل احتمالية الإصابة بسكتة دماغية متكررة. أظهرت الأعمال السابقة نتائج واعدة في التنبؤ بحالات السكتة الدماغية لأول مرة باستخدام أساليب التعلم الآلي. ومع ذلك ، هناك أعمال محدودة للتنبؤ بالسكتة الدماغية المتكررة باستخدام أساليب التعلم الآلي. ومن ثم ، تم اقتراح هذا العمل لإجراء تحليل تجريبي والتحقيق في خوارزميات التعلم الآلي المطبقة في نماذج التنبؤ بالسكتة الدماغية المتكررة. يهدف هذا البحث إلى التحقيق في أداء خوارزميات التعلم الآلي ومقارنتها باستخدام مجموعات البيانات السريرية العامة للسكتة الدماغية المتكررة. في هذه الدراسة ، تم استخدام الشبكة العصبية الاصطناعية (ANN) وآلة المتجهات الداعمة (SVM) وقائمة قواعد بايزي (BRL) ومقارنة أدائها في مجال نموذج التنبؤ بالسكتة الدماغية المتكررة. تظهر نتيجة التجارب التجريبية أن ANN سجلت أعلى دقة عند 80.00٪ ، تليها BRL بنسبة 75.91٪ و SVM بنسبة 60.45٪.

الكلمات المفتاحية: الشبكة العصبية الاصطناعية ، قائمة قاعدة بايزي ، التعلم الآلي ، التنبؤ بالسكتة الدماغية المتكررة ، دعم الة المتجهات.