International Journal of Engineering & Technology, 7 (4.15) (2018) 4-6

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Predicting Sudden Deaths Following Myocardial Infarction in Malaysia Using Machine Learning Classifiers

Muhammad Hazrani Abdul Halim, Yumn Suhaylah Yusoff*, Mazlynda Md Yusuf

Fakulti Sains dan Teknologi,Universiti Sains Islam Malaysia, Bandar Baru Nilai, 71800 Nilai, Negeri Sembilan, Malaysia
**Corresponding author E-mail: suhaylah@usim.edu.my*

Abstract

Myocardial infarction (MI) is among the top causes of death in Malaysia. The mortality rate following MI was high, especially within the first 30 days after the onset. This paper study the ability of k-Nearest Neighbors (kNN) and Naïve Bayes algorithms to predict the 30-day mortality of MI patients, using. The dataset used for this study is provided by National Cardiovascular Disease Database (NCVD) which consist of 2840 MI patients from hospitals in Malaysia. The sudden death predictions made by the machine learning are based on the age, gender, year of onset, smoking habit, BMI, diabetes, hypertension and cholesterol level. The result suggests that kNN algorithm has better performance in predicting the sudden death compared to Naïve Bayes. The number of independent variables plays an important role in mortality prediction, and removing insignificant variables improve the performance.

Keywords: *Machine learning; Mortality prediction; Myocardial infarction; Sudden death.*

1. Introduction

Myocardial infarction (MI) or also known as heart attack is a very common disease, which contributed to extensive number of deaths every year in Malaysia. The risk of mortality due to MI is high, especially within the first 30 days after MI onset which is the critical period for MI patients. The 30-day mortality is also referred to as sudden death. On average, the mortality rates within the first 30 days after MI in Malaysia are at 12% [1]. Short-term mortality due to MI is somehow influenced by the patients' conditions such as age, BMI, hypertension, diabetes, and cholesterol level. Further investigation on these factors may help reducing the risk of sudden death following MI for future patients. Utilizing the data of past MI patients, the probability of 30-day mortality can be calculated. The mortality of MI patients might also be predicted using various machine learning algorithms with sufficient data set.

2. Predicting Mortality Using Machine Learning

In this era where technology is advancing and abundance of data are available for analysis, people can correctly predict the outcome of an event based on past data using machine learning. Machine learning algorithms seek functions that predict a value or an outcome from its characteristics based on sample data [2]. Today, machine learning algorithms are widely used for predictions and classifications in many fields including medical and health. Researchers adopted machine learning techniques to detect and identify the potential MI patients earlier based on their medical records [3, 4].

Aside from earlier detection of diseases, machine learning is also used as survivability or mortality predictions from various causes. Having computer to classify high-risk patients from a vast number of patients will help doctors to take appropriate actions which will reduces the rate of mortality. The most commonly used machine learning algorithms for predicting the mortality of patients are k-Nearest Neighbors (kNN), Naïve Bayes (NB), Bayesian Network (BN), Neural Network (NN), Support Vector Machine (SVM), Random Forest (RF), and Logistic Regression (LR). Table 1 shows past studies on mortality prediction using various machine learning algorithms [5-15].

Table 1: Past studies on short-term mortality prediction using machine learning

References	Mortality Prediction	Machine Learning Algorithm						
		kNN	NB	BN	NN	SVM	RF	LR
Torra et al, 2016	In-hospital mortality of patients undergoing open repair of abdominal aortic aneurysm			/	/			
Vomlel et al., 2012	30-day mortality of ST Elevation MI patients		/	/	/			
Galiatsatos et al., 2016	30-day mortality of patients after hip fracture surgery			/	/			
Wallert et al., 2017	Two-year mortality of MI patients					/	/	/
Cooper et al., 1997	60-day mortality of pneumonia patients	/		/	/			/

Motwani et al., 2017	Five-year all-cause mortality of patients with suspected coronary artery disease.							
Makar et al., 2015	Six-month mortality of elderly Medicare beneficiaries.	/	/		/	/	/	/
Helwan et al., 2017	One-year mortality of MI patients				/			
Wiemken et al., 2017	30-day mortality of community-acquired pneumonia patients		/				/	/
Taylor et al., 2015	In-hospital mortality of emergency department patients with sepsis						/	/
Watcharapasorn & Kurubanjerdjit, 2016	Mortality of patients after undergoing surgery	/						

Machine learning requires a large set of data to increase the accuracy and reliability of the mortality predictions. By utilizing the MI patient data from hospitals in Malaysia, we might be able to predict the mortality of MI patients within the first 30 days after onset using machine learning. This paper study the ability of several machine learning methods in accurately determining MI patients who needs urgent medical attention after admission.

3. Methodology

Machine learning needs a set of data to be able to predict the mortality of MI patients accurately. For that, this paper acquired and analyzed the data of MI patients form various medical institutions and hospitals in Malaysia. The dataset is randomly divided into smaller datasets, which will be used for training and validation. Using these datasets, the potential of sudden deaths is predicted using machine learning packages installed in R.

3.1. Data Description

A dataset consists of 28420 MI patients from 19 different medical institutions and hospitals in Malaysia, from year 2006 until 2013 is acquired from National Cardiovascular Disease Database (NCVD). It include variables which are known to affect the risk of MI such as age, gender, year of occurrence, smoking habit, BMI, diabetes, hypertension, and cholesterol level.

To observe the ability of machine learning to accurately classify the potential of sudden death, we need a different dataset for testing. As such, we randomly divided the dataset obtained from NCVD into 10 smaller datasets. One of the datasets is used as the testing dataset, while the other nine small datasets is used to train the machine learning algorithm.

The model used in this study include all eight independent variables and sudden death as the dependent variable. However, this model may contain both significant and insignificant variables which might affect the performance of machine learning. Thus, we regressed the model using logistic regression to determine the relationship between sudden death and each independent variable. We discovered that gender, year of MI and smoking are insignificantly associated with sudden death, thus these variables were removed from the model to form a new model with only significant variables. We are interested in observing the performance of machine learning while using different models with different variables, so both model 1 and model 2 are tested using machine learning algorithms, and the results are compared. The list of variables used in model 1 and model 2 is as follows:

Model 1: Age, gender, year of MI, smoking habit, BMI, diabetes, hypertension, cholesterol

Model 2: Age, BMI, diabetes, hypertension, cholesterol

3.2. Machine Learning Classifiers

There are a lot of machine learning methods that can be used as classifiers. Each classifier has different algorithm which will affect the ability to accurately predict the outcome based on the nature of the dataset. To identify which machine learning classifier is suitable for predicting the potential of sudden death, we train and test the data using different classifiers and compare the accuracy rates of the predictions. The machine learning classifiers and the packages used in this paper is shown in Table 2.

The k-Nearest Neighbors algorithm is a machine learning method which predict the outcome of instances by directly comparing them with the training dataset. The algorithm will choose k numbers of outcome from the training dataset which are the most similar with the instances, then assign the predominant classification to the instances [9]. The Naïve Bayes algorithm compute the conditional probability of the outcome given a set of independent variables using Bayes' Theorem. A mortality prediction can be derived given the likelihood of the variables occurring, as well as the prior probabilities of the outcome [13].

To use the machine learning classifiers as the mortality prediction, first, we train each classifier using the train dataset and placing the variable for sudden death as the dependent. Next, we use the trained classifier to predict the outcome of MI patients using the test dataset. Comparing the prediction results with the actual outcome of the patients will determine the performance of the classifier.

4. Results and Discussion

To determine which model and algorithm produced better mortality prediction, we measured the performance of the models. The performance of the machine learning prediction is usually compared based on the accuracy [5-6, 8, 12-13, 15], precision [5, 13, 15], recall [5, 8, 13, 15], and area under the ROC curve (AUC) [6, 8, 10-11, 13-14]. The performance for each model is recorded in Table 3 and compared.

The most basic performance measure for machine learning is the accuracy of the classification or prediction. It shows the percentage of instances correctly predicted out of total instances. From Table 1, it is shown that both k-Nearest neighbors and Naïve Bayes models have high prediction accuracy. Mortality prediction for model 2 shows higher accuracy compared to model 1, and the result is more noticeable for kNN algorithm. This suggests that removing insignificant variables from the model helps improving the accuracy of mortality prediction.

Table 2: Classifier packages used in R

Classifier	Package	Training Code	Testing Code
kNN	caret	fit <- knn3(Suddendeath ~ Age + Gender +…, data=train, k=1)	predictions <- predict(fit, data=test, type="class")
Naïve Bayes	e1071	fit <- naiveBayes(Suddendeath ~ Age + Gender + …, data=train)	predictions <- predict(fit, data=test)

Table 3: Performance of machine learning on sudden death prediction

	k-Nearest Neighbors		Naïve Bayes	
	Model 1	Model 2	Model 1	Model 2
Accuracy	86.82%	90.16%	90.83%	90.87%
AUC	54.83%	59.20%	66.46%	67.73%
Precision	14.45%	19.02%	8.33%	-
Recall	9.02%	2.39%	0.04%	-

Observing the accuracy of the prediction using when using machine learning is important, however, it is less reliable for biased datasets like the one used in this paper. Since the number of 30-day survivors is around 90%, we will have accuracy of more than 90% even if the machine learning algorithm predicts all patient as survivors. This is observed when we use Naïve Bayes to predict the mortality from model 2, and it only produce one outcome (survive). Accuracy alone not sufficient to measure the performance of machine learning algorithm, thus we also looked at precision, recall, and AUC as additional performance measures.

Recall rate is the percentage of correctly predicted sudden death out of actual sudden death cases from the test datasets. Both kNN and Naïve Bayes have low ability to correctly identify high-risked MI patients as high-risk. The rates worsen when the number of variables is reduced from eight to five. The recall rate for kNN decrease from 9.02% to 2.39%, while Naïve Bayes recalls only 0.04% for model 1 and fails to predict any sudden death potentials from for model 2. Small number of variables might be the reason why the recall rates are very low, thus increasing the number of independent variables should improve the performance of machine learning algorithms as mortality predictors.

Precision reflects the number of correctly predicted sudden death out of the total number cases predicted as sudden death. The ability of kNN and Naïve Bayes to precisely predict 30-day mortality from the dataset is also not that high. However, removing all insignificant variables from the model increases the precision rate of the machine learning prediction, as observed with kNN algorithm. As such, using only variables that have strong relationship with the dependent variables will improve the precision of the predictions.

Another criterion observed to determine the performance of the machine learning is the area under the ROC curve (AUC). ROC stands for Receiver Operator Characteristic, which show how the correctly predicted positive value vary with the incorrectly classified negative value. It is recommended to use AUC when comparing the performance of the classifiers, aside from accuracy [16]. Both machine learning algorithms used in this study show acceptable AUC value of more than 50%, and Naïve Bayes has better AUC compared to kNN. AUC for model 2, which consist of only significant variables, has better AUC than model 1.

Overall, k-Nearest Neighbors shows better prediction ability compared to Naïve Bayes in predicting the potential sudden death cases based on existing dataset. Eliminating insignificant independent variables from the model helps improving the performance of the machine learning, although doing that will reduce the recall rates.

5. Conclusion

This study suggests that k-Nearest Neighbors can predict the sudden death following MI better compared to Naïve Bayes. However, the predictive ability of machine learning is limited due to biasness of the dataset and small number of significant variables used in the model. Adding more independent variables into the model may help improving the ability of machine learning classifiers to predict the 30-day mortality after MI. Filtering out insignificant variables from the model using logistic regression will enhance the performance of machine learning as mortality predictor.

Acknowledgement

The authors gratefully acknowledge all the supports and funds provided by Ministry of Education Malaysia under a research grant to Faculty of Science and Technology, Universiti Sains Islam Malaysia with a grant number USIM/RAGS/FST/36/50115. The authors would also like to express their gratitude to National Cardiovascular Disease Database (NCVD) for providing the data used in this research. We also acknowledge all participating hospitals and institutions; SGH Heart Centre, Penang Hospital, University Malaya Medical Centre, National Heart Institute, Tengku Ampuan Rahimah Hospital, Sultanah Aminah Hospital, Tuanku Ja'afar Hospital, Sultanah Bahiyah Hospital, Serdang Hospital, Tengku Ampuan Afzan Hospital, Raja Permaisuri Bainun Hospital, Malacca Hospital, Kuala Lumpur Hospital, Sabah Heart Centre, Queen Elizabeth Hospital, Sultanah Nur Zahirah Hospital, Ampang Hospital, Raja Perempuan Zainab II Hospital, and Tuanku Fauziah Hospital, for contributing the data.

References

[1] Ahmad W A W and Sim K H. Annual Report of the NCVD-ACS Registry 2009 & 2010. *National Cardiovascular Disease Database*, 2013.

[2] Mullainathan S and Spiess J. Machine Learning: An Applied Economic Approach. Journal of Economic Perspectives, 2017, 31(2):87-106.

[3] Weiss J C, Natarajan S N, Peissig P L, McCarty C A and Page D. Machine Learning for Personalized Medicine: Predicting Primary Myocardial Infarction from Electronic Health Records. Association for the Advancement of Artificial Intelligence, 2012, 33(4):33-45.

[4] Seenivasagam V and Chitra R. Myocardial Infarction Detection using Intelligent Algorithms. Neural Network World, 2016, 26(1):91-110.

[5] Torra A M, Fernandez D R, Alonso O M, Paya O S, Mackenzie J C and Jaimes M C. Using Machine Learning Methods for Predicting Inhospital Mortality in Patients undergoing Open Repair of Abdominal Aortic Aneurysm. *Journal of Biomedical Informatics*, 2016, 62:195-201.

[6] Vomlel J, Kruzik H, Tuma P, Precek J and Hutyra M. Machine Learning Methods for Mortality Prediction in Patients with ST Elevation Myocardial Infarction. *Proceedings of the WUPES*, 2012, pp. 204-213.

[7] Galiatsatos D, Anastassopoulos G, Drosos G, Ververidis A, Tilkeridis K and Kazakos K. Prediction of 30-day Mortality after a Hip Fracture Surgery using Neural and Bayesian Networks. *Proceedings of the IFIP International Conference on Artificial Intelligence Applications and Innovations*, 2014, pp. 566-575.

[8] Wallert J, Tomasoni M, Madison G and Held C. Predicting Two-year Survival versus non-Survival after First Myocardial Infarction using Machine Learning and Swedish National Register Data. *BMC Medical Informatics and Decision Making*, 2017, 17(1):99.

[9] Cooper G F, Aliferis C F, Ambrosino R, Aronis J, Buchanon B J, Caruana R, Fine M J, Glymour C, Gordon G, Hanusa B H, Janosky J E, Meek C, Mitchell T, Richardson T, and Spirtes P. An Evaluation of Machine-Learning Methods for Predicting Pneumonia Mortality. *Artificial Intelligence in Medicine*, 1997, 9(2):107-138.

[10] Motwani M, Dey D, Berman D S, Germano G, Achenbach S, Al-Mallah M H, Andreini D, Budoff M J, Cademartiri F, Callister T Q, Chang H J, Chinnaiyan K, Chow B J W, Cury R C, Delago A, Gomez M, Gransar H, Hadamitzky M, Hausleiter J, Hindoyan N, Feuchtner G, Kaufmann P A, Kim Y J, Leipsic J, Lin F Y, Maffei E, Marques H, Pontone G, Raff G, Rubinshtein R, Shaw L J, Stehli J, Villines T C, Dunning A, Min J K and Slomka P J. Machine Learning for Prediction of All-Cause Mortality in Patients with Suspected Coronary Artery Disease: a 5-year Multicentre Prospective Registry Analysis. *European Heart Journal*, 2017, 38(7):500-507.

[11] Makar M, Ghassemi M, Cutler D M and Obermeyer Z. Short-term Mortality Prediction for Elderly Patients using Medicare Claims Data. *International Journal of Machine Learning and Computing*, 2015, 5(3):192-197.

[12] Helwan A, Ozsahim D U, Abiyev R and Bush J. One-year Survival Prediction of Myocardial Infarction. *International Journal of Advance Computer Science and Applications*, 2017, 8(6):173-178.

[13] Wiemken T L, Furmanek S P, Mattingly W A, Guinn B E, Cavallazzi R, Botran R F, Wolf L A, English C L and Ramirez J A. Predicting 30-day Mortality in Hospitalized Patients with Community-Acquired Pneumonia using Statistical and Machine Learning Approaches. *University of Louisville Journal of Respiratory Infections*, 2017, 1(3):50-56.

[14] Taylor R A, Pare J R, Venkatesh A K, Mowafi H, Melnick E R, Fleischman W and Hall M K. Prediction of In-Hospital Emergency Department Patients with Sepsis: A Local Big Data-Driven, *Machine Learning Approach. Academic Emergency Medicine*, 2016, 23(3):269-278.

[15] Watcharapasorn P and Kurubanjerdjit N. The Surgical Patient Mortality Rate Prediction by Machine Learning Algorithms. *Proceedings of the International Joint Conference on Computer Science and Software Engineering*, 2016, pp. 1-5.

[16] Provost F, Fawcett T and Kohavi R. The Case against Accuracy Estimation for Comparing Induction Algorithms. *Proceedings of the Fifteenth International Conference on Machine Learning*, 1998, pp. 445-453.