34th Annual International Conference of the IEEE EMBS
San Diego, California USA, 28 August - 1 September, 2012

A Survival Prediction Model of Rats in Hemorrhagic Shock Using the Random Forest Classifier*

Joon Yul Choi, Sung Kean Kim, Wan Hyung Lee, Tae Keun Yoo, and Deok Won Kim

Abstract— **Hemorrhagic shock is the cause of one third of deaths resulting from injury in the world. Although many studies have tried to diagnose hemorrhagic shock early and accurately, such attempts were inconclusive due to compensatory mechanisms of humans. The objective of this study was to construct a survival prediction model of rats in hemorrhagic shock using a random forest (RF) model, which is a newly emerged classifier acknowledged for its performance. Heart rate (HR), mean arterial pressure (MAP), respiratory rate (RR), lactate concentration (LC), and perfusion (PF) measured in rats were used as input variables for the RF model and its performance was compared with that of a logistic regression (LR) model. Before constructing the models, we performed a 5-fold cross validation for RF variable selection and forward stepwise variable selection for the LR model to see which variables are important for the models. For the LR model, sensitivity, specificity, accuracy, and area under the receiver operating characteristic curve (ROC-AUC) were 1, 0.89, 0.94, and 0.98, respectively. For the RF models, sensitivity, specificity, accuracy, and AUC were 0.96, 1, 0.98, and 0.99, respectively. In conclusion, the RF model was superior to the LR model for survival prediction in the rat model.**

I. Introduction

Approximately 5 million people died in the world from injury in 2004 [1]. For all children aged 1 to 19 years, the first leading cause of death was unintentional injuries in 2008 [2]. By 2020, death from injury in the world will probably increase to 8 million, and the cause of one third of these deaths will result from hemorrhagic shock [3], [4]. Hemorrhagic shock is defined as circulatory dysfunction causing decreased tissue oxygenation and accumulation of oxygen debt, which can ultimately lead to multiple organ system failure if left untreated [5]. This imbalance is the most fundamental problem in all types of shock.

Machine learning has contributed much in modern day, complex clinical decision making [6]. Also, supervised learning classifiers have been recently applied to prediction models of survival or mortality in hemorrhagic shock [7]-[10]. The newly emerged random forest (RF) classifier has proven to be a highly accurate and rapid classifier [6].

Although the American College of Surgeons Advanced Trauma Life Support (ATLS) for Doctors Student Manual suggests that the severity of hemorrhagic shock can be diagnosed by traditional vital signs, such as blood pressure (BP), heart rate (HR), respiratory rate (RR), or urine output and mental status, these have been shown to be unreliable measures of acute hemorrhage due to compensatory mechanisms [5], [11], [12]. For this reason, current treatments focus on diagnosis by evidence of tissue ischemia or hypo-microcirculation, including lactate concentration (LC), and perfusion (PF) [5], [13]. However, these are also under debate as to whether they can diagnose hemorrhagic shock early and accurately [14]. Thus, the diagnosis of hemorrhagic shock may require easier and more accurate methods, rather than solely relying on the evaluation of aforementioned parameters.

The objective of this study was to diagnose hemorrhagic shock by predicting survival in rat models using RF and compare it to logistic regression (LR), which is widely used as the gold standard among medical practitioners. First, input variables were selected for model construction among proposed diagnosis indices including HR, BP, RR, LC, and PF. Second, the performance of RF and LR models were compared for sensitivity, specificity, accuracy, receiver operating characteristic (ROC) curve, and computer time of model construction .

II. Materials and Methods

A. Experimental Protocol and Data Acquisition

Thirty six male Sprague-Dawley (S-D) rats were divided into three groups with 12 rats in each group depending on controlled blood volume loss. After anesthesia with an isoflurane inhalation system (RC2, VetEquip, Pleasanton), blood volumes of 2 mL/100 g, 2.5 mL/100 g, or 3 mL/100 g were withdrawn over 15 min for all groups through the right carotid arterial catheter. Uncontrolled hemorrhage was performed by amputation of the tail at 75 % of its length at 1 min after initiating volume controlled hemorrhage. HR, BP, and RR were measured with a sampling frequency of 1 kHz using LabChart 6 Pro (AD Instruments, Colorado Springs), as physiological parameters. The data were fed into an Analog/Digital system (PowerLab 8/30, AD Instruments). PF, as a measurement of microcirculation, was monitored using a laser Doppler perfusion monitor (PeriFlux system 5000, Perimed, Sweden) with a probe (Probe 407, Perimed), which was attached to the right front sole. The data were acquired with a sampling frequency of 32 Hz and analyzed by a

*This study was supported by student and faculty research grants of Yonsei University College of Medicine for 2011 (6-2011-0087).

J. Y. Choi is with the Brain Korea 21 Project for Medical Science, Yonsei University, Seoul, Korea (corresponding author; phone: 82-2-2228-1920; fax: 82-2-363-9923; e-mail: jychoi717@yuhs.ac).

S. K. Kim is with the Graduate Program in Biomedical Engineering, Yonsei University, Seoul, Korea (e-mail: sdm04sdm@yuhs.ac).

W. H. Lee is with Yonsei University College of Medicine, Seoul, Korea (e-mail: wanhyung@yuhs.ac).

T. K. Yoo is with Yonsei University College of Medicine, Seoul, Korea (e-mail: fawoo2@yuhs.ac).

D. W. Kim is a Professor at Dept. of Medical Engineering, Yonsei University College of Medicine, Seoul, Korea (e-mail: kdw@yuhs.ac).

program (Perisoft for window, Perimed). Blood sampling for LC was repetitively performed, then analyzed immediately by a portable blood lactate analyzer (Lactate Pro LT-1710, ARKRAY, Japan) as shown in Fig. 1. To obtain continuous LC data, linear interpolation was performed using computer software (LabVIEW 2009, National Instruments, Austin). All data were analyzed for 5 min after "Bleeding" in Fig. 1, because we simulated an emergency situation in which bleeding was stopped.

For the construction of prediction models, HR, mean arterial pressure (MAP), RR, LC, and PF were used for input data sets. MAP was used as a representative of BP. Therefore, 175 (1 set/min * 5min * 35 rats) data sets were obtained in this study (one rat was excluded from analysis since it died during the analysis period). Data sets consisted of data that was obtained for one minute of data collection. When survival and death were determined 150 min after initiation of the experiment, the numbers for the survival and death sets were 75 and 100, respectively. We used MATLAB Version 2011 (Mathworks Inc, Natick) for analysis of RF and SPSS 18.0 (SPSS Inc, Chicago) for analysis of LR.

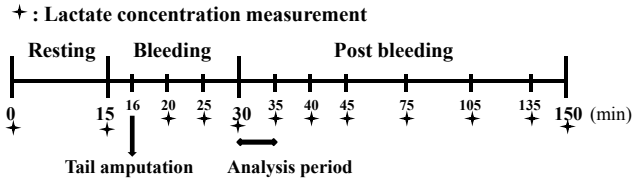

Figure 1. Experimental protocol for rats with hemorrhagic shock

B. Logistic Regression (LR)

Logistic regression is used to generate a predictive model for dichotomous response variables by fitting data to a logistic function (1), which always takes on values between zero and one:

$$P(x) = \frac{1}{1+e^{-(\alpha+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\ldots+\beta_d x_d)}} \quad (1)$$

In equation (1), $\beta_1 \sim \beta_d$ are called the regression coefficients of input variables respectively. Each of the regression coefficients describes the size of the contribution of that risk factor. LR models are usually used for comparison in machine learning studies [6]. We compared the performance of the LR models to that of RF models.

C. Random Forest (RF)

Random forest is an ensemble classification algorithm that consists of many decision trees and outputs by independent trees (Fig. 2). $D_1 \sim D_t$ are training data selected randomly from the data sets with input variables to make the decision trees, $T_1 \sim T_t$ are the decision trees, and T* is a final decision tree. Each tree is built independently in combination of a bagging idea and random selection of input variables. The result is based on a majority vote of the classification of all trees [6]. Thus, the goal of random forest is to classify accurately by controlling the number of the trees. This study investigated survival prediction models arbitrarily using trees of 10 (RF10), 100 (RF100), 200 (RF200), 300 (RF300), 400 (RF400), and 500 (RF500), which are most commonly used in RF models [6], [15]-[17].

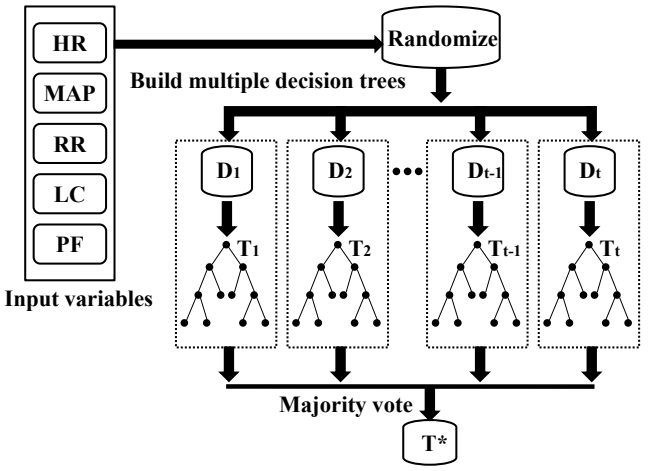

Figure 2. Features of the prediction model and their variables using random forest (RF). HR: heart rate, MAP: mean arterial pressure, RR: respiratory rate, LC: lactate concentration, PF: perfusion.

D. Variable Selection and Models Construction

For the random forest, the data were divided randomly into two mutually exclusive data sets. Among the data, approximately 70 % (n=125) were used as the training set to construct the models, and the remaining 30 % (n=50) were used in the model testing.

All the input variables, including HR, MAP, RR, LC, and PF, were used for variable selection of RF models using the training set as shown in Fig. 3. Priority of the variables was determined using the Breiman's method [18]. Then, 5-fold cross validation was repeated five times to calculate the mean accuracy of each cross validation process by progressively eliminating the least contributing ones until the most influential ones were left (backward elimination). The highest ranked variables with the best cross validated accuracy were chosen as the optimal variables. The prediction model was constructed using the optimal variables.

For logistic regression, we used the forward stepwise method for variable selection using SPSS 18.0 software. The forward stepwise method begins with no variables in a model, trying out the variables one by one, and then including them if they are statistically significant until no more variables can be added to the model.

Each of the RF and LF models designed using the optimal variables for prediction of survival in hemorrhagic shock were built using the training set. We obtained sensitivity, specificity, and accuracy to evaluate the performances of the prediction models using the remaining testing set. We also drew a ROC curve for each model and calculated the area under the ROC curve (AUC) to compare the performance of these two models. The AUCs were plotted using MATLAB software.

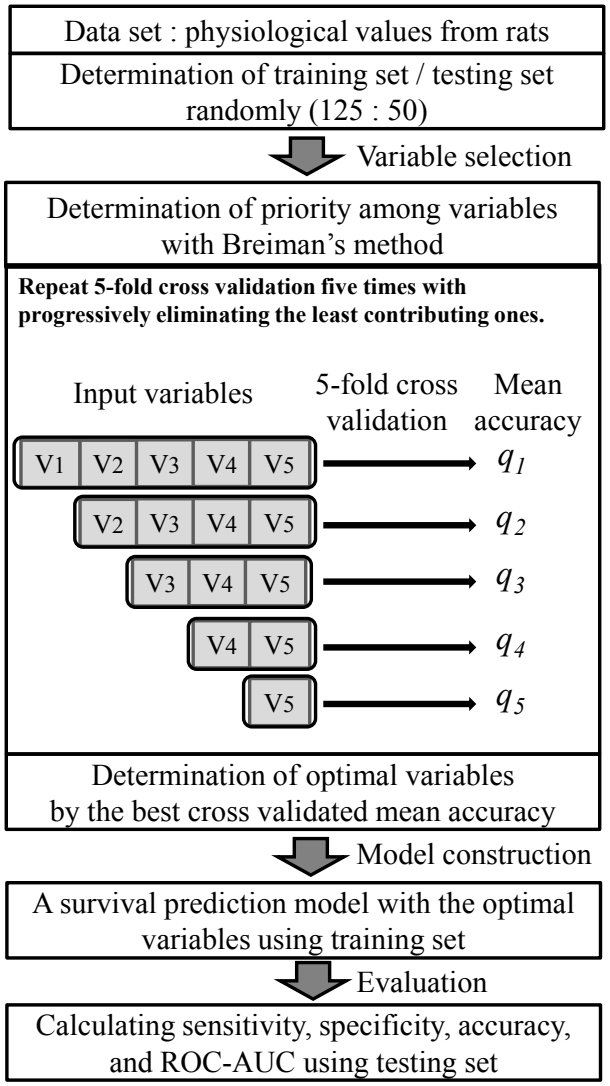

Figure 3. Flowchart of backward elimination variable selection and survival prediction model construction using random forest

III. Result

The priority of the input variables, ranked in descending order, was PF, RR, MAP, HR, and LC. Table I shows the selected variables resulting from variable selection. PF and RR were selected for the RF model as well as the LR model. Table II lists the AUC, sensitivity, specificity, accuracy, and execution time of several RF models with the various numbers of trees and LR. For the LR model, AUC, sensitivity, specificity, and accuracy were 0.98, 1, 0.89, and 0.94, respectively. For all of the RF models, AUC, sensitivity, specificity, and accuracy were 0.99, 0.96, 1, and 0.98, respectively. The execution time of the various RF models was considerably shorter than that of the LR model. Fig. 4 shows the ROC curves of the RF model with 100 trees, as a representative RF model, and the LR model.

TABLE I. Input Varibles Used in Construction of The Models

Variable	RF						LR
	The number of trees						*Forward*
	10	*100*	*200*	*300*	*400*	*500*	
HR							
MAP							
RR	*	*	*	*	*	*	*
LC							
PF	*	*	*	*	*	*	*

*Variables selected

RF: Random forest, LR: logistic regression, HR: heart rate, MAP: mean arterial pressure, RR: respiratory rate, LC: lactate concentration, PF: perfusion.

TABLE II. Performance of The RF and LR Survival Prediction Models

	RF						LR
	The number of trees						*Forward*
	10	*100*	*200*	*300*	*400*	*500*	
ROC-AUC	0.99	0.99	0.99	0.99	0.99	0.99	0.98
Sensitivity	0.96	0.96	0.96	0.96	0.96	0.96	1
Specificity	1	1	1	1	1	1	0.89
Accuracy	0.98	0.98	0.98	0.98	0.98	0.98	0.94
Time (sec)	0.06	0.07	0.08	0.08	0.09	0.10	1.23

RF: Random forest, LR: logistic regression, ROC-AUC: receiver operating characteristic - area under the curve

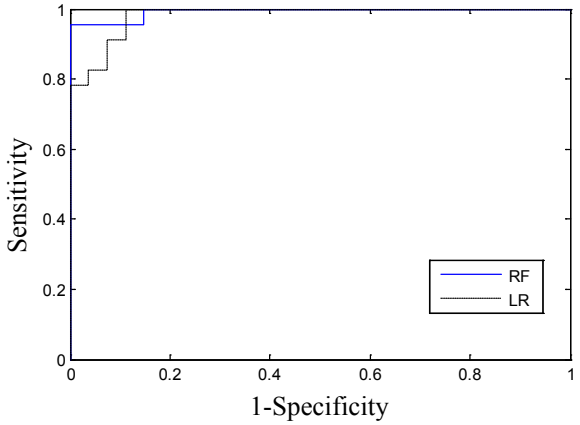

Figure 4. Receiver operating characteristic (ROC) curves of random forest (RF) with 100 trees and logistic regression (LR) for survival prediction model of rats in hemorrhagic shock

IV. Discussion and Conclusion

We constructed survival prediction models of rats in hemorrhagic shock using RF and LR after selection of optimal variables in this study. The results showed that the RF models were more accurate and faster than LR. Archer and Albert et al. reported that the number of trees must be chosen large enough to get stable estimates of variable importance [15], [16]. On the other hand, Chen and Ham et al. reported that a larger number of trees did not provide improved performance [17], [19]. In this study, the number of trees did not demonstrate improved performance.

One of the advantages of the RF model was the relative ease with constructing the models. It was easy because only determining the number of trees is needed [6]. However, The RF model is not allowed to examine the individual trees separately [20]. In this study, the execution time of the RF models was approximately 10 times faster than that of the LR model. The execution time would be very crucial for evaluation of a large amount of data, such as in the field of bioinformatics. Albert J et al. showed only 1 min to construct the RF model using a training set of 10,000 [16]. Very few studies utilized RF models for prediction of hemorrhagic shock, and our study showed its potential in this field. Our study also applied a validation process to minimize data bias in sampling data for training and testing.

Variable selection is an important process to constructing efficient models with a minimum number of input variables. In this study, we used HR, MAP, RR, LC, and PF as the initial input variables for survival prediction of rats in hemorrhagic shock. After 5-fold cross validation, two variables, RR and PF, were selected, and as a result the RF model was more accurate than the LR model. Consequently, we suggest that variable selection should be applied to construct diagnosis models for many input variables in the field of bioinformatics, in particular gene analysis.

Even though perfusion was shown to be the greatest contributing input variable, it is quite sensitive to motion artifacts, provides only relative values [21], and is too expensive to obtain, requiring the use of laser Doppler flowmetry, in emergency rooms. In conclusion, it was shown that the random forest method was more accurate and faster than logistic regression method for predicting survival of rats in hemorrhagic shock. It would be useful to give preferential emergency treatment to patients who are more in danger.

REFERENCES

[1] World Health Organization, World health statistics 2010, *World Health Organization Press*, 2010, pp. 62–70.

[2] T. Mathews, A. M. Minino, M. J. K. Osterman, D. M. Strobino, and B. Guyer, "Annual summary of vital statistics: 2008," *Pediatrics,* peds. 2010-3175, pp. 146-157, Dec. 2010.

[3] C. J. Murray and A. D. Lopez, "Alternative projections of mortality and disability by cause 1990-2020: Global Burden of Disease Study," *The Lancet*, vol. 349, no. 9064, pp. 1498-1504, May. 1997.

[4] C. D. Deakin and I. R. Hicks, "AB or ABC: pre-hospital fluid management in major trauma," *Journal of Accident and Emergency Medicine*, vol. 11, no. 3, pp. 154-157, Mar. 1994.

[5] M. Wilson, D. P. Davis, and R. Coimbra, "Diagnosis and monitoring of hemorrhagic shock during the initial resuscitation of multiple trauma patients: a review," *Journal of Emergency Medicine,* vol. 24, no. 4, pp. 413-422, May 2003.

[6] C. H. Hsieh, R. H. Lu, N. H. Lee, W. T. Chiu, and M. H. Hsu et al., "Novel solutions for an old disease: Diagnosis of acute appendicitis with random forest, support vector machines, and artificial neural networks," *Surgery*, vol. 149, no. 1, pp. 87-93, Mar. 2010.

[7] B. Eftekhar, K. Mohammad, H. Ardebili, M. Ghodsi, and E. Ketabchi, "Comparison of artificial neural network and logistic regression models for prediction of mortality in head trauma based on initial clinical data," *BMC Medical Informatics and Decision Making,* vol. 5, no. 1, p. 3, Feb. 2005.

[8] D. A. Roberts, J. B. Holcomb, B. E. Parker Jr, J. L. Sondeen, and A. E. Pusateri et al., "The use of polynomial neural networks for mortality prediction in uncontrolled venous and arterial hemorrhage," *The Journal of trauma,* vol. 52, no. 1, pp. 130-135, Jan. 2002.

[9] D. W. Kim, J. L. Choi and Y. S. Park, "Survival prediction in rats with fixed-volume hemorrhage using a logistic regression equation," *Shock*, vol.33, Suppl 1, pp. 14, Jun. 2010.

[10] K. H. Jang, T. K. Yoo, J. Y. Choi, K. C. Nam, and J. L. Choi et al., "Comparison of survival predictions for rats with hemorrhagic shocks using an artificial neural network and support vector machine," in *Proc. IEEE Eng. Med. Biol. Soc.*, Boston, 2011, pp.91-94.

[11] M. J. Vandromme, R. L. Griffin, J. A. Weinberg, L. W. Rue, and J. D. Kerby, "Lactate Is a Better Predictor than Systolic Blood Pressure for Determining Blood Requirement and Mortality: Could Prehospital Measures Improve Trauma Triage?," *Journal of the American College of Surgeons,* vol. 210, no. 5, pp. 861-867, April 2010.

[12] H. Guly, O. Bouamra, M. Spiers, P. Dark, and T. Coats et al, "Vital signs and estimated blood loss in patients with major trauma: Testing the validity of the ATLS classification of hypovolaemic shock," *Resuscitation,* vol. 82, no. 5, pp. 556-559, May 2011.

[13] R. P. Dutton, "Current concepts in hemorrhagic shock," *Anesthesiology clinics,* vol. 25, no. 1, pp. 23-34, May 2007.

[14] G. J. Pestel, K. Fukui, O. Kimberger, H. Hager, and A. Kurz et al., "Hemodynamic parameters change earlier than tissue oxygen tension in hemorrhage," *Journal of Surgical Research,* vol. 160, no. 2, pp. 288-293, May 2010.

[15] K. J. Archer and R. V. Kimes, "Empirical characterization of random forest variable importance measures," *Computational Statistics & Data Analysis,* vol. 52, no. 4, pp. 2249-2260, Jan. 2008.

[16] J. Albert, E. Aliu, H. Anderhub, P. Antoranz, and A. Armada et al., "Implementation of the random forest method for the imaging atmospheric Cherenkov telescope MAGIC," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment,* vol. 588, no. 3, pp. 424-432, April 2008.

[17] X. W. Chen and M. Liu, "Prediction of protein-protein interactions using random decision forest framework," *Bioinformatics,* vol. 21, no. 24, pp. 4394-4400, Oct. 2005.

[18] L. Breiman, "Random forests," *Machine learning,* vol. 45, no. 1, pp. 5-32, April 2001.

[19] J. Ham, Y. Chen, M. M. Crawford, and J. Ghosh, "Investigation of the random forest framework for classification of hyperspectral data," *Geoscience and Remote Sensing, IEEE Transactions on,* vol. 43, no. 3, pp. 492-501, Mar. 2005.

[20] A. M. Prasad, L. R. Iverson, and A. Liaw, "Newer classification and regression tree techniques: bagging and random forests for ecological prediction," *Ecosystems*, vol. 9, no. 2, pp. 181-199, Mar. 2006.

[21] M. L. Kaiser, A. P. Kong, E. Steward, M. Whealon, M. Patel, D. B. Hoyt, and M. E. Cinat, "Laser Doppler Imaging for Early Detection of Hemorrhage," *The Journal of trauma,* vol. 71, no. 2, pp. 401-406, Aug. 2011.