An Intelligent Scoring System and Its Application to Cardiac Arrest Prediction

Nan Liu, Zhiping Lin, *Senior Member, IEEE*, Jiuwen Cao, Zhixiong Koh, Tongtong Zhang, Guang-Bin Huang, *Senior Member, IEEE*, Wee Ser, *Senior Member, IEEE*, and Marcus Eng Hock Ong

***Abstract*—Traditional risk score prediction is based on vital signs and clinical assessment. In this paper, we present an intelligent scoring system for the prediction of cardiac arrest within 72 h. The patient population is represented by a set of feature vectors, from which risk scores are derived based on geometric distance calculation and support vector machine. Each feature vector is a combination of heart rate variability (HRV) parameters and vital signs. Performance evaluation is conducted on the leave-one-out cross-validation framework, and receiver operating characteristic, sensitivity, specificity, positive predictive value, and negative predictive value are reported. Experimental results reveal that the proposed scoring system not only achieves satisfactory performance on determining the risk of cardiac arrest within 72 h but also has the ability to generate continuous risk scores rather than a simple binary decision by a traditional classifier. Furthermore, the proposed scoring system works well for both balanced and imbalanced datasets, and the combination of HRV parameters and vital signs shows superiority in prediction to using HRV parameters only or vital signs only.**

***Index Terms*—Cardiac arrest, heart rate variability, machine learning, scoring system.**

I. Introduction

SCORING systems have been widely used in intensive care units (ICUs) to predict clinical outcomes and assess the severity of illness [1]. An accurate outcome prediction allows critically ill patients to be considered for proper treatment as early as possible. In the past three decades, many scoring systems have been developed among which the most famous methods include acute physiology and chronic health evaluation [2], simplified acute physiology score (SAPS) [3], and mortality probability model [4]. Each scoring system has a specific purpose and its own range of applications. For example, risk of death, organ dysfunction assessment, and severity of illness are possible outcomes of some scoring systems.

The development of scoring systems usually relies on appropriate selection of variables with which prediction outcomes are associated. Most scoring systems use physiological vital signs to make clinical judgment. Vital signs are physical measures that indicate the physiological status of an individual, such as heart rate, temperature, blood pressure, pain score, Glasgow Coma Scale (GCS), and oxygen saturation. These vital signs may be observed, measured, and monitored to assess an individual's level of physical functioning. However, traditional vital signs were reported to have limitations in predicting morbidity and mortality [5]. Meanwhile, heart rate variability (HRV), a noninvasive measurement for investigating autonomic influence on the cardiovascular system [6], has been found to be closely correlated with clinical outcomes [7], [8] and is applicable in a wide range of clinical conditions like congestive heart failure [9], acute myocardial infarction [10] and dilated cardiomyopathy [11]. Therefore, HRV might hold promise to become an additional "vital sign" for clinicians [12]. Several studies have attempted to use HRV to predict mortality in the elderly [8] and for heart disease [13], but there has been limited study on correlating HRV with clinical outcomes such as cardiac arrest within 72 h and length of hospital stay.

We have previously discovered that a combination of HRV parameters and vital signs has potential to be an effective indicator of mortality [14]. Preliminary studies focused on predicting binary outcomes (death or survival) rather than using a score to grade the severity of an individual patient's condition. Furthermore, it appears that prediction of cardiac arrest within 72 h is more informative to clinicians than prediction of mortality because early identification of cardiac arrest can assist in providing a quick response to patients and minimizing the potential impact. However, to our best knowledge, there is so far no existing scoring system specifically for predicting cardiac arrest. Moreover, current scoring systems are usually not adaptable when they are required to incorporate new inputs [15], such as HRV parameters together with vital signs.

With the advancement of statistical and computational techniques, machine learning has been found to be useful to scoring systems in terms of improving predictive performance [16], handling imbalanced data [17], and enhancing system adaptability [15]. In this study, we aim at proposing an intelligent scoring system and exploring the utility of both HRV parameters and vital signs to predict cardiac arrest within 72 h. The scoring system will take advantage of a novel geometric distance-based score calculation algorithm, with which accurate scores are predicted and reliable decisions are achieved. Machine learning techniques are the pillar of the scoring system and are implemented in the process of score calculation and updating. More importantly, the intelligent scoring system is

Manuscript received January 18, 2012; revised June 18, 2012; accepted August 5, 2012. Date of publication August 8, 2012; date of current version November 16, 2012.

N. Liu, Z. Koh, T. Zhang, and M. E. H. Ong are with the Department of Emergency Medicine, Singapore General Hospital, Singapore 169608 (e-mail: marcus.ong.e.h@sgh.com.sg).

Z. Lin, J. Cao, G.-B. Huang, and W. Ser are with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798 (e-mail: ezplin@ntu.edu.sg).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TITB.2012.2212448

designed and constructed following a standard machine learning structure, making it adaptable to other medical applications in the future.

The organization of this paper is as follows. In Section II, we introduce the dataset considered in this study, and describe the proposed intelligent scoring system. In this section, we also present the method for performance evaluation. The experimental results are provided in Section III where both balanced dataset and imbalanced dataset are employed for validation. Then, discussion is given in Section IV and conclusion is drawn in Section V.

II. Methods

A. Dataset

A prospective observational clinical study was conducted at the Department of Emergency Medicine, Singapore General Hospital from November 2006 to December 2007. All patients were initially triaged by a nurse, and those with Airway, Breathing, Circulation problems, or thought to be possibly unstable and needing close monitoring are routinely put on ECG monitoring using the LIFEPAK 12 defibrillator/monitor (Physio-Control, Redmond, WA). The demographic data (for example, age, race, gender, and medical history) and selected vital signs were obtained from the hospital records. In this study, vital signs measured at the first presentation were used. Furthermore, clinical outcomes such as cardiac arrest within 72 h, death, ICU admission, and length of hospital stay were obtained from hospital charts.

During the study, a total number of 1386 critically ill patients were monitored and their ECG tracings were collected. Out of these, 361 patients were excluded due to high percentage of artifacts, nonsinus beats, and ectopics combined together. As a result, 1025 ECG records were used for analysis. Lead II ECGs sampled at 125 Hz were extracted as text files for HRV analysis using CODE-STAT Suite data review software (version 5.0, Physio-Control) and proprietary ECG extraction software (Physio-Control). Cases with ECG recordings were prospectively identified and had identity confirmed by querying hospital charts and records. Charts were included for review if they had an ECG recording showing sinus rhythm and were excluded if they were in nonsinus rhythm (ventricular or supraventricular arrhythmias). Table I presents the characteristics of total 1025 critically ill patients in this study. There are 52 cases of cardiac arrest within 72 h (5.1%) in the dataset.

TABLE I
Characteristics of 1025 Patients in the Dataset

Characteristics	(N=1025)
Mean age (SD)	61.74 (15.86)
Male (%)	622 (60.68)
Race (%)	
Chinese	689 (67.2)
Malay	153 (14.9)
Indian	126 (12.3)
Other	57 (5.6)
Signs and symptoms (%)	
Breathing difficulty	496 (48.4)
Chest Pain	433 (42.2)
Cough	182 (17.8)
Fever	141 (13.8)
Others	620 (60.5)
ED outcome (%)	
Admitted to ICU	201 (19.6)
Admitted to Ward Transferred	868 (84.7)
Cardiac arrest within 72hrs	52 (5.1)
Died	100 (9.8)
Hospital outcome (SD)	
Mean length of stay (days)	6.81 (10.61)
Duration in ICU (days)	0.78 (4.49)
Duration in Ward (days)	5.05 (7.14)

Algorithm: Geometric distance based score prediction
Input
- HRV parameters and vital signs of N training patients and one testing patient $\mathbf{x}_t$ where each patient is a sample in the database.
- Patients' hospital records and characteristics.

1) Variable selection
- Select HRV parameters and vital signs to form a feature vector to represent each patient's health condition.
- Transform both training and testing feature vectors into the interval $[-1, 1]$ by min-max normalization.

2) Initial score calculation
- Obtain the cluster center C_p of the positive class (patients with cardiac arrest within 72 hours) in the Euclidean feature space.
- Calculate distance D_p, which is the distance between C_p and one positive training sample nearest to C_p.
- Calculate distance D_n, which is the distance between C_p and one negative training sample farthest to C_p.
- Calculate distance D_t between the testing sample $\mathbf{x}_t$ and C_p and compute the initial score based on D_p and D_n.

3) Classification based score updating
- Predict binary outcome with SVM classifier and obtain the number of positive samples N_p within K neighbors.
- Implement predefined rules for score updating.

Output
- Predictive risk score on the clinical outcome.

Fig. 1. Proposed geometric distance-based risk score prediction method.

B. Proposed Scoring System

An intelligent prediction model is proposed to compute a risk score on a patient's clinical outcome, utilizing both HRV parameters and vital signs. The scoring system is built based on the calculation of geometric distances among a set of feature vectors obtained from the records of multiple patients. The proposed score prediction algorithm is summarized in Fig. 1 and the details are elaborated in the following.

1) Variable Selection: We select 24 variables (16 HRV parameters and 8 vital signs) to study their possible association with the risk of cardiac arrest within 72 h. Vital signs in this study are heart rate, temperature, systolic blood pressure, diastolic blood pressure, pain score, GCS, respiratory rate, and oxygen saturation. A package developed in MATLAB (R2009a, The Mathworks, Natick, MA) is used to process ECG record and to calculate 16 time domain and frequency domain HRV parameters as shown in Table II following the widely used HRV

TABLE II
LIST OF HRV PARAMETERS USED IN THIS STUDY

HRV parameters	Description
aRR (s)	Mean of the RR intervals
STD (s)	Standard deviation of the RR intervals
Mean HR (bpm)	Mean of the instantaneous heart rate
SD of HR (bpm)	Standard deviation of the instantaneous heart rate
RMSSD (s)	Root mean square of differences between adjacent RR intervals
NN50 (count)	Number of consecutive RR intervals differing by more than 50ms
pNN50 (%)	Number and percentage of consecutive RR intervals differing by more than 50ms
HRV triangular index	Total number of all RR intervals divided by the height of the histogram of intervals
TINN	Baseline width of a triangle fit into the RR interval histogram using a least squares
TP (ms^2)	Total power
VLF (ms^2)	Power in the very low frequency range ($\leq$ 0.04 Hz)
LF (ms^2)	Power in the low frequency range (0.04-0.15 Hz)
HF (ms^2)	Power in the high frequency range (0.15-0.40 Hz)
LF norm (nu)	LF power in normalized units: LF/(TP-VLF)$\times$100
HF norm (nu)	HF power in normalized units: HF/(TP-VLF)$\times$100
LF/HF	Ratio of LF power to HF power

analysis standard recommended by Task Force of the European Society of Cardiology the North American Society of Pacing Electrophysiology [18]. In the proposed scoring system, HRV parameters and vital signs are combined to form 24-dimensional feature vectors to represent patients.

Prior to risk score calculation, the feature set is transformed into the interval $[-1, 1]$ by performing min-max normalization on the original data [19]. Given a dataset $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N]$ where each $\mathbf{x}$ represents a patient, let $\min_A$ and $\max_A$ denote the minimum and maximum values of an attribute vector $A = [\mathbf{x}_1(m), \ldots, \mathbf{x}_N(m)]$ where $m = 1, 2, \ldots, 24$, and N is the total number of samples. Min–max normalization maps a value, v, of A to v' in the range $[\min'_A, \max'_A]$ by computing

$$v' = \frac{v - \min_A}{\max_A - \min_A}(\max'_A - \min'_A) + \min'_A. \qquad (1)$$

The normalization process is able to preserve the relationships among the original data values, therefore facilitates geometric distance calculation as well as risk prediction.

2) Initial Score Calculation: The calculation of geometric distance-based risk score is described as follows. First, the cluster centers of positive and negative samples are calculated in the Euclidean space, where positive samples are patients with cardiac arrest within 72 h as outcome and negative samples are patients without cardiac arrest. As shown in Fig. 2, two geometric distances D_p and D_n are computed where C_p and C_n are the cluster centers of the positive class and the negative class, respectively, given by $C_p = (1/N_1)\sum_{\mathbf{x}_i \in \omega_1} \mathbf{x}_i$ and $C_n = (1/N_0)\sum_{\mathbf{x}_i \in \omega_0} \mathbf{x}_i$. N_i is the number of samples in class ω_i where $i = 1$ indicates the positive class and $i = 0$ indicates the negative class. D_p is the distance between C_p and one positive training sample nearest to C_p, and D_n is the distance between C_p and one negative training sample farthest from C_p. Note that the proposed scoring system is based on 24-D features whereas the example in Fig. 2 uses 2-D features for the convenience of illustration.

Given a new testing sample $\mathbf{x}_t$, indicated as the triangle pattern in Fig. 2, the Euclidean distance D_t between $\mathbf{x}_t$ and C_p is

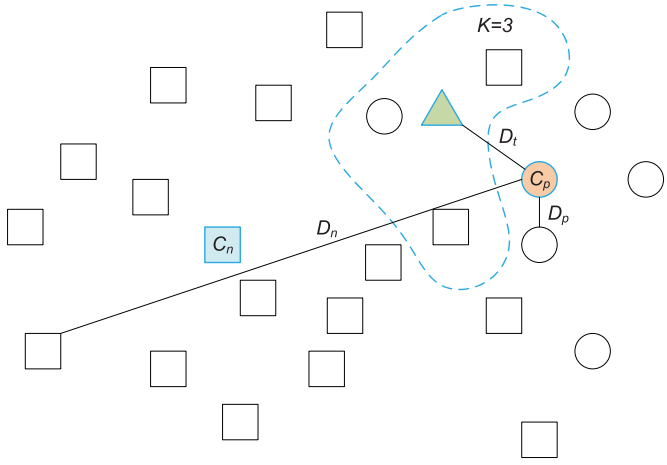

Fig. 2. Illustration of initial score calculation and score updating in the proposed score prediction algorithm. Circle patterns indicate patients with positive outcomes (cardiac arrest within 72 h) and square patterns indicate patients with negative outcomes. Triangle pattern is the testing data point. Note that C_p and C_n are virtual center points calculated from positive samples and negative samples, respectively.

calculated. If D_t is less than D_p, an initial score is assigned to 100 due to the fact that $\mathbf{x}_t$ is closer to C_p than any other positive samples. Similarly, the risk score is assigned to 0 if D_t is larger than D_n as the testing sample is further away to C_p than any other negative samples, that is, the testing sample is unlikely to have a positive outcome. Except for the earlier two cases, the risk score S is computed as

$$S = \frac{W_n}{W_n + W_p} \times 100 \qquad (2)$$

where $W_p = \exp(|D_t - D_p|)$ and $W_n = \exp(|D_t - D_n|)$ are weights representing how close pattern $\mathbf{x}_t$ is connected with the positive class and the negative class, respectively. The exponential function is adopted to enhance the discriminatory power of the weights. If the difference between D_t and D_p is small, weight W_p becomes small thus producing a high risk score. In

other words, geometric similarity between patterns is able to correlate with predictive score by means of (2). In general, the risk score reflects a possibility on the prediction of a positive outcome, i.e., cardiac arrest in our study.

3) Classification-Based Score Updating: In many conditions, samples from different classes are usually cluttered together as shown in Fig. 2. It is therefore difficult to distinguish one category from the other by simply relying on the geometric distance-based score predictor. To overcome this difficulty, a score updating method is proposed, which consists of two major components: classification and neighborhood check.

The output of a conventional classifier is a binary prediction on cardiac arrest within 72 h. Although the predictive outcomes are difficult for clinicians to accurately assess a patient's condition, such a binary decision could also be helpful in enhancing risk score calculation. Given its merit in classification performance, support vector machine (SVM) [20] is employed as a classifier in this study. SVM is a supervised learning method based on a novel statistical learning theory and has been successfully implemented in many biomedical systems [21], [22]. The LIBSVM package [23] is used for algorithm implementation and a brief introduction to SVM is presented as follows.

Assume that the training set is given as $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)$ and a hyperplane $\langle \mathbf{w}, \mathbf{x} \rangle + b = 0$ is constructed where $\mathbf{x}_i \in \mathbf{X}$, $y_i \in \{\pm 1\}$. The set of patterns is said to be optimally separated by the hyperplane if it is separated without errors and the margin is maximal. A canonical hyperplane has the constraint for parameters $\mathbf{w}$ and b: $\min_{\mathbf{x}_i} y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) = 1$. A separating hyperplane in canonical form must satisfy the following constraint:

$$y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq 1, \quad i = 1, 2, \ldots, N. \tag{3}$$

Then quadratic programming is used for solving the constraint optimization problem in order to find the optimal hyperplane. The optimization criterion is the width of the margin between the class. For a new pattern $\mathbf{x}_t$, the hyperplane decision function can be written as (4) where the kernel trick may be used to extend the classifier to be nonlinear

$$f(x) = \operatorname{sgn}\left(\sum_{i=1}^{N} \alpha_i y_i k(\mathbf{x}_t, \mathbf{x}_i) + b\right). \tag{4}$$

Several kernel functions are available for use, such as linear kernel $k(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i \cdot \mathbf{x}_j$, sigmoid kernel $k(\mathbf{x}_i, \mathbf{x}_j) = \tanh(\alpha \mathbf{x}_i \cdot \mathbf{x}_j + \gamma)$, and radial basis function (RBF) kernel $k(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2 / 2\sigma^2)$ where σ is the width of RBF function. The linear kernel is chosen in this study due to its simplicity in optimization and high training efficiency.

The output of the SVM classifier is either positive or negative. To update the risk score, we need to further determine the characteristics of K-nearest neighbors of the testing sample. An example is shown in Fig. 2, in which K is 3 and there are two negative samples and one positive sample in the area defined by the dotted line. The majority class in the neighborhood provides the evidence that its member patterns might share similar characteristics with the testing sample. That is to say, the more positive samples are found in the nearest neighbors, the higher risk score the testing sample may have. Define N_p as the total number of positive samples in K-nearest neighbors and an integer ϵ as the threshold, we have the following two rules for risk score updating.

1) The risk score increases to two times of its original value if the predictive outcome of the SVM classifier is positive and N_p is larger than threshold ϵ.
2) The risk score decreases to half of its original value if the predictive outcome of the SVM classifier is negative and N_p is smaller than threshold ϵ.

C. Performance Evaluation

Evaluation of the scoring system is based on the leave-one-out cross-validation framework. In a dataset of N samples, N iterations are required for algorithm validation. Within each iteration, one sample is used as the testing sample while the rest samples are used for training. The proposed score prediction process needs to repeat N times so that each sample can be tested individually. Having the risk scores for the entire dataset, a proper threshold is derived to report sensitivity and specificity

$$\text{Sensitivity} = \frac{\text{TP}}{(\text{TP} + \text{FN})} \tag{5}$$

$$\text{Specificity} = \frac{\text{TN}}{(\text{TN} + \text{FP})}. \tag{6}$$

Furthermore, the receiver operating characteristic (ROC) curve, the positive predictive value (PPV), and the negative predictive value (NPV) are also used to present system performance, where PPV and NPV are defined as

$$\text{PPV} = \frac{\text{TP}}{(\text{TP} + \text{FP})} \tag{7}$$

$$\text{NPV} = \frac{\text{TN}}{(\text{TN} + \text{FN})}. \tag{8}$$

In these parameters, true positive (TP) indicates patients with cardiac arrest within 72 h correctly predicted as cardiac arrest within 72 h; false positive (FP) indicates healthy patients incorrectly predicted as cardiac arrest within 72 h; true negative (TN) indicates healthy patients correctly predicted as healthy; and false negative (FN) indicates patients with cardiac arrest within 72 h incorrectly predicted as healthy. In general, high sensitivity, specificity, PPV, and NPV are desired for a scoring system.

III. Experiments

The geometric distance-based scoring system was used to correlate HRV parameters and vital signs to cardiac arrest within 72 h, and ROC analysis was adopted to investigate prediction performance. In the evaluation results, sensitivity, specificity, PPV, and NPV were reported.

A. Balanced Data Versus Imbalanced Data

In the proposed scoring system, classification plays an important role in score updating. An investigation into the dataset is necessary because highly imbalanced training data would make

normal classifiers fail in prediction. As shown in Table I, our dataset consists of a majority group of 973 negative samples (94.9%) and a minority group of 52 positive samples with cardiac arrest within 72 h as outcome (5.1%). When the SVM classifier is applied to such an imbalanced dataset, the majority class will dominate model learning which eventually leads to poor generalization performance on new samples from the minority class. A solution of handling imbalanced data is to create a decision ensemble [24]. We proposed using an undersampling strategy to partition the majority class into M nonoverlapped groups where each group had the same sample size as minority class did. By joining each group of majority class samples with minority class samples separately, M balanced datasets were created, on which an ensemble of M individual prediction models was trained for pattern classification.

In the experiments, we used one balanced dataset with 52 positive samples and 52 negative samples and one imbalanced dataset with 52 positive samples and 973 negative samples for system evaluation. Note that the balanced dataset was only a subset of the imbalanced dataset. The same algorithm depicted in Fig. 1 was implemented on both datasets. However, prediction methods in score updating were different, i.e., an individual SVM classifier was used on the balanced data while an ensemble of M classifiers was used on the imbalanced data. Moreover, implementation of the scoring system on the imbalanced data required another change in neighborhood check. When the balanced data were used, K was assigned to 5 and the threshold ϵ was assigned to 3; when the imbalanced data were used, the threshold remained unchanged and the number of neighbors became $K' = \alpha \times K$ where α is an integer defining the imbalance ratio of the number of negative samples and the number of positive samples, i.e., $\alpha = \lfloor 973/52 \rfloor$ in this study. This ratio compensates the bias in data distribution to provide an effective neighborhood check.

B. Results on the Balanced Dataset

Table III presents the performances on predicting cardiac arrest within 72 h with the proposed scoring system where both HRV parameters and vital signs were used for pattern representation. Cutoff indicates the threshold applied to the ROC curve, from which desired sensitivity or specificity can be achieved. By adjusting the cutoff from 40.0 to 65.0, sensitivity dropped from 84.6% to 59.6% and specificity increased from 53.8% to 82.7%. PPV and NPV also changed accordingly to reflect the prediction performances. The best performance in the ROC analysis was obtained when 49.8 was the cutoff score. In medical applications, we usually expect high sensitivity and PPV values while maintaining favorable specificity and NPV because misclassifying unhealthy patients is risky.

The number of nearest neighbors K serves as a vital factor in controlling score updating, it is therefore worth investigating the impact of K on prediction performance. Various values of K were evaluated and the results are illustrated in Fig. 3. As mentioned in Section II-B, the threshold ϵ in neighborhood checking was assigned to 3 in score prediction. If sensitivity was kept as a constant value at 80.8%, performances degraded

TABLE III
PREDICTION RESULTS WITH THE PROPOSED INTELLIGENT SCORING SYSTEM ON THE BALANCED DATASET

Cutoff	Sensitivity	Specificity	PPV	NPV
40.0	84.6%	53.8%	64.7%	77.8%
45.0	80.8%	71.2%	73.7%	78.7%
49.8	**78.8%**	**80.8%**	**80.4%**	**79.2%**
50.0	75.0%	80.8%	79.6%	76.4%
55.0	69.2%	82.7%	80.0%	72.9%
60.0	67.3%	82.7%	79.5%	71.7%
65.0	59.6%	82.7%	77.5%	67.2%

The best results in ROC analysis are highlighted in bold.

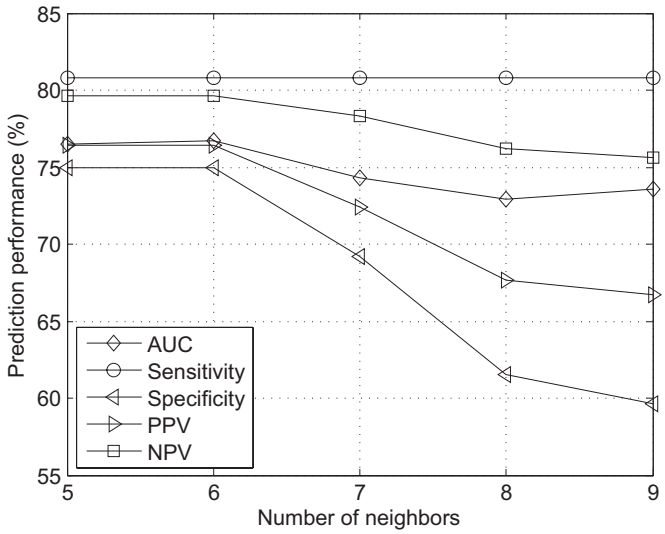

Fig. 3. Prediction results with different number of nearest neighbors on the balanced dataset where five performance indicators are used.

dramatically with the increase of K, particularly on specificity and PPV. It appears that selecting five samples in the neighborhood of the testing sample was the most suitable setting in the balanced dataset.

Then we continued investigating the correlation between measurement type and clinical outcome. Fig. 4 depicts that the combined features of HRV parameters and vital signs outperformed HRV parameters only and vital signs only in predicting cardiac arrest within 72 h. In the ROC analysis, the closer the curves approaches to the upper-left corner, the better the prediction performance is and the larger the value of area under curve (AUC) is. Fig. 4 presents that HRV parameters get larger AUC value than vital signs, which shows evidence that HRV is also an effective indicator to cardiac arrest within 72 h. Overall, the combined features were superior to either HRV parameters only or vital signs only in terms of the prediction performance.

Furthermore, a comparison between the proposed intelligent scoring system and several classical classifiers was conducted on the prediction of cardiac arrest within 72 h. These classifiers are SVM with linear kernel (SVM-LIN), SVM with RBF kernel (SVM-RBF), and generalized linear model (GLM), all of which are able to produce probability outputs. To evaluate the prediction performance, both the ROC analysis and several widely used performance measures were used and the

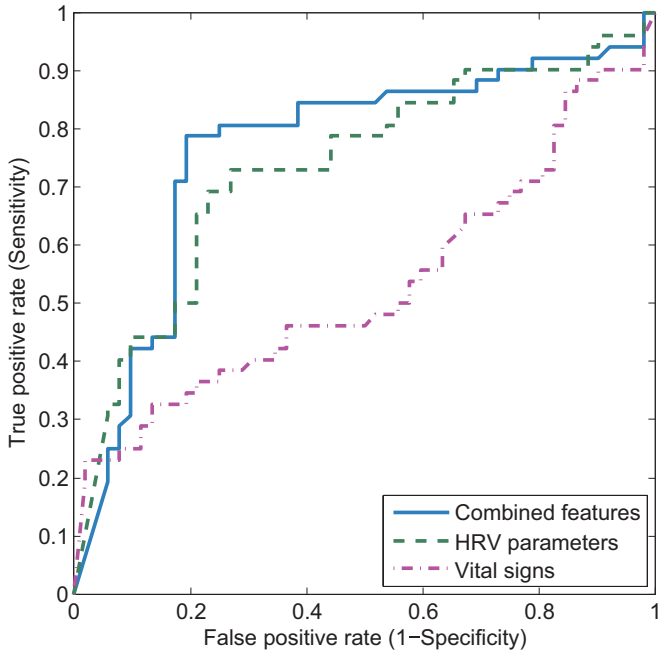

Fig. 4. ROC curves generated on the balanced dataset with three different types of features: the combined features, HRV parameters, and vital signs.

TABLE IV
PREDICTION RESULTS WITH THE PROPOSED SCORING SYSTEM AND CLASSICAL CLASSIFICATION METHODS ON THE BALANCED DATASET

Measure	Proposed	SVM-LIN	SVM-RBF	GLM
Sensitivity	**78.8**%	73.1%	61.5%	63.5%
Specificity	**80.8**%	80.8%	80.8%	80.8%
PPV	**80.4**%	79.2%	76.2%	76.7%
NPV	**79.2**%	75.0%	67.7%	68.9%
MAE	**0.373**	0.632	0.591	0.551
MSE	**0.206**	0.439	0.378	0.396
LogL	**1.252**	1.759	1.430	1.776
MAPR	**0.627**	0.368	0.409	0.449
MPR	**0.627**	0.368	0.409	0.449

comparison results are shown in Table IV. Obviously, the proposed scoring system achieved the best prediction performance in terms of sensitivity, specificity, PPV, and NPV by performing the ROC analysis. Moreover, five probability-based measures, namely mean absolute error (MAE), mean squared error (MSE), LogLoss (LogL), macro average mean probability rate (MAPR), and mean probability rate (MPR), were employed to assess the reliability of the classifiers [25]. In general, the proposed scoring system performed the best among all classifiers, achieving the smallest errors (MAE, MSE, LogL) and the highest probability rates (MAPR, MPR) in the prediction. MAPR was the same as MPR because both the positive class and the negative class had the same number of examples on the balanced dataset. It is mentioned by Ferri *et al.* [25] that MSE has a strong correlation with classification accuracy and this is reflected in Table IV that the proposed scoring system outperformed other classical classifiers with the highest sensitivity, specificity, PPV, and NPV while maintained the smallest MSE.

TABLE V
PREDICTION RESULTS OF THE PROPOSED SCORING SYSTEM WITH THREE FEATURE EXTRACTION METHODS ON THE BALANCED DATASET

Method	Sensitivity	Specificity	PPV	NPV
PCA	**78.8**%	**80.8**%	**80.4**%	**79.2**%
LDA	76.9%	80.8%	80.0%	77.8%
KPCA	75.0%	80.8%	79.6%	76.4%

TABLE VI
PREDICTION RESULTS WITH THE PROPOSED INTELLIGENT SCORING SYSTEM ON THE IMBALANCED DATASET

Cutoff	Sensitivity	Specificity	PPV	NPV
40.0	92.3%	13.8%	5.4%	97.1%
45.0	86.5%	38.2%	7.0%	98.2%
50.0	78.8%	61.7%	9.9%	98.2%
55.0	78.8%	61.8%	9.9%	98.2%
60.0	78.8%	62.1%	10.0%	98.2%
61.0	**78.8**%	**62.3**%	**10.0**%	**98.2**%
65.0	75.0%	62.7%	9.7%	97.9%

The best results in ROC analysis are highlighted in bold.

As a machine learning system for medical application, the proposed scoring system is possible to be further improved by integrating existing learning algorithms. We have implemented three feature extraction methods, namely principal component analysis (PCA), linear discriminant analysis (LDA), and kernel PCA (KPCA) [26], and evaluated their contributions to the prediction performance. By applying feature extraction algorithms, feature dimensions became less than 24 and the new feature vector was either the linear combination of the original features in PCA and LDA or the nonlinear combination of the original features in KPCA. The best prediction results with reduced feature dimensions are summarized in Table V. Apparently, these feature extractors cannot help to achieve better performance compared to the original scoring system. Because the feature vectors in our database only have 24 dimensions and there is no much redundant information to remove, sophisticated feature extraction methods may not be helpful for improving the prediction performance with features of low dimensions.

C. Results on the Imbalanced Dataset

To predict cardiac arrest within 72 h on the imbalanced dataset, an decision ensemble was constructed to replace a single SVM classifier in risk score updating. Parameter K was empirically chosen as 3 for evaluations, which meant $K' = \alpha \times 3$ nearest neighbors were used. Table VI shows sensitivity, specificity, PPV, and NPV with different cutoff scores. Similar to the performance obtained on the balanced dataset, sensitivity dropped and specificity increased with the increment of cutoff value. According to the ROC analysis, a cutoff score of 61.0 achieved the best performance on predicting cardiac arrest within 72 h.

When compared with the results in Table III, we noticed that lower cutoff score generated better performance on the balanced dataset but higher cutoff score performed well on the imbalanced dataset. Moreover, PPV and NPV values in Table VI were extremely low and extremely high, respectively. This observation indicated difficulties of prediction on the imbalanced data where

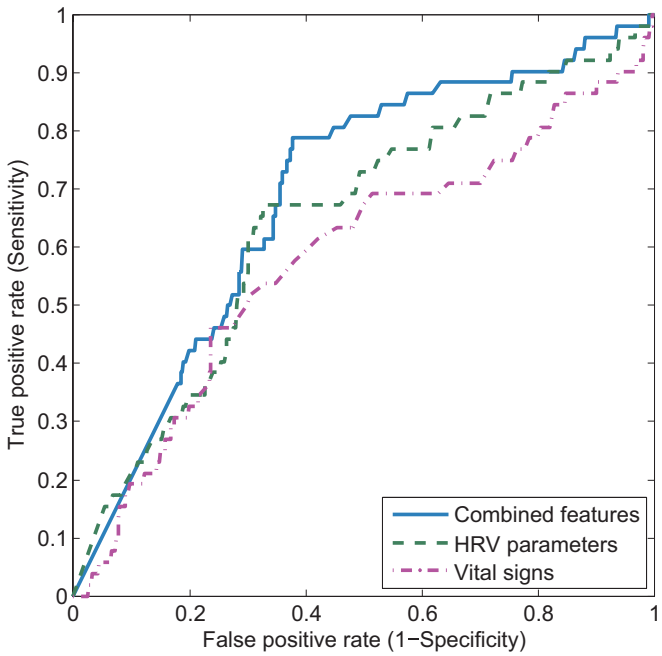

Fig. 5. ROC curves generated on the imbalanced dataset with three different types of features: the combined features, HRV parameters, and vital signs.

majority class dominated the entire learning and decision making process. For example, when cutoff was 60.0, sensitivity was 78.8% and specificity was 62.1%. Sensitivity in percentage was good but the absolute number of correctly classified positive samples was quite low given that there were only 52 such samples. This explains why PPV was low and NPV was high.

Fig. 5 gives a performance comparison among the combined features, HRV parameters and vital signs based on the imbalanced dataset. According to the ROC analysis, the combine features performed the best and HRV parameters outperformed vital signs. Compared with the ROC curves on the balanced dataset in Fig. 4, vital signs appeared to have better performance in terms of achieving less difference to HRV parameters and the combined features.

IV. Discussion

Conventional classification models usually predict labels on new testing samples. The outcome is not that helpful to a clinician, as the level of severity of a patient cannot be retrieved from the label. As a result, many scoring systems have been proposed where the outputs are risk scores rather than predictive labels [1]. In most popular scoring systems such as the modified early warning score (MEWS) [27], the CAPE triage score [28], and the SAPS [3], traditional vital signs are used for risk score prediction. However, these physiological measures may not be sufficient for reliable risk assessment, thus extra information is needed to enhance predictive abilities of scoring systems.

Previous studies have shown that HRV was potential to serve as a predictor of mortality [7], [8], myocardial infarction [13], sleep apnea syndrome [29], and congestive heart failure [9]. However, there has been limited study of the correlation of HRV with cardiac arrest within 72 h in a large prospective clinical series. Moreover, HRV cannot be incorporated into existing scoring systems due to their low adaptability on new feature integration [15]. Toward this end, we proposed in this study an intelligent scoring system and demonstrated that a combination of HRV parameters and vital signs had a strong association with cardiac arrest within 72 h. Different from a typical classification method, the proposed scoring system is able to generate a risk score between 0 and 100 as an indicator of cardiac arrest within 72 h. Having the human readable score, clinicians can easily make clinical decisions. Furthermore, the system possesses high adaptability in medical applications because it can be extended for predicting other clinical outcomes by simply changing input variables.

The proposed intelligent system consists of three major components, namely variable selection, initial score calculation, and classification-based score updating, to deliver a robust risk score analysis. Our study showed that a combination of HRV parameters and vital signs performed well in predicting cardiac arrest within 72 h on both balanced dataset and imbalanced dataset. Each patient was represented as a feature vector in the database, with which its corresponding outcome (cardiac arrest within 72 h or not) was associated. We selected a set of 16 HRV parameters and 8 vital signs for feature representation and linear SVM as the classifier to learn intrinsic characteristics from training samples. In practice, the choice of a classifier depends on the requirements of applications and the linear SVM classifier in this study makes a tradeoff between performance and speed [20].

To show its effectiveness in score prediction, the proposed method was compared to several classical classifiers that can produce probability or score outputs. The comparison results in Table IV demonstrated that the proposed scoring system was superior to other classifiers in terms of ROC analysis and five performance measures. In this study, we did not compare our method with existing scoring systems because they were not adaptable when new input features such as HRV parameters together with vital signs were required to be incorporated. It is worth noting that the performance of score prediction is data dependent as well as disease dependent [1]. Also, according to some published scoring methods [2], [15], both sensitivity and specificity in the range of 60% to 80% are usually observed in hospital setting. Therefore, the performance of the proposed scoring system is generally satisfactory in predicting cardiac arrest within 72 h.

Many databases have seen biased data distribution, particularly in medical applications where positive samples belong to the minority class. In such a case, a vast number of negative samples bring difficulties in predicting abnormal data. That is to say, when applying normal classification methods on the imbalanced dataset, majority class will dominate the learning process and therefore results in poor prediction performance. As described in Section III-A, we adopted an undersampling strategy to handle data imbalance [30], which was proved simple yet effective. Recently, many algorithms have been presented for learning from imbalanced data [17], [30]. We intend to examine various learning strategies to improve the performance of our scoring system on highly imbalanced data.

As mentioned in [1], there are several limitations in existing scoring systems. One significant problem is that prediction models can only achieve their best performance when the characteristics of new samples exactly match those of the training samples in the development population. This is also a major challenge to most machine learning algorithms that require consistency within samples between training and testing datasets. Transfer learning [31] could be of help to solving this problem, in which training data and testing data may come from different feature spaces. Other major limitations include dependency of input quality, inherent bias of the derived model, and automatic patient data management, etc. Vincent and Moreno [1] suggested that scoring systems need to update according to changes of patient population and advancement of prognostic techniques. Moreover, combining multiple scoring system could also improve the prediction performance [1], [24].

V. Conclusion

In this paper, we have proposed a novel risk score prediction system with HRV parameters and vital signs, in which geometric distance serves as the key component. The intelligent scoring system has demonstrated its ability to generate human understandable risk scores, and has shown its effectiveness to being a powerful predictor of cardiac arrest within 72 h. We foresee a potential on extending the scoring system to predict other clinical outcomes. The experimental validations on a balanced dataset and an imbalanced dataset show that the proposed system is able to achieve satisfactory prediction results. To create a stable and effective scoring system, it is worth discovering discriminatory variable subsets and investigating suitable learning strategies to handle imbalanced data, which will be a topic of future research.

References

[1] J. L. Vincent and R. Moreno, "Clinical review: Scoring systems in the critically ill," *Crit. Care*, vol. 14, p. 207, 2010.

[2] W. A. Knaus, E. A. Draper, D. P. Wagner, and J. E. Zimmerman, "APACHE II: A severity of disease classification system," *Crit. Care Med.*, vol. 13, pp. 818–829, 1985.

[3] J. R. Le Gall, S. Lemeshow, and F. Saulnier, "A new simplified acute physiology score (SAPS II) based on a European/North American multicenter study," *J. Amer. Med. Assoc.*, vol. 270, pp. 2957–2963, 1993.

[4] S. Lemeshow, D. Teres, J. Klar, J. S. Avrunin, S. H. Gehlbach, and J. Rapoport, "Mortality probability models (MPM II) based on an international cohort of intensive care unit patients," *J. Amer. Med. Assoc.*, vol. 270, pp. 2478–2486, 1993.

[5] K. M. Hargarten, C. Aprahamian, H. Stueven, D. W. Olson, T. P. Aufderheide, and J. R. Mateer., "Limitations of prehospital predictors of acute myocardial infarction and unstable angina," *Ann. Emerg. Med.*, vol. 16, pp. 1325–1329, 1987.

[6] H. V. Huikuri, T. Mäkikallio, K. E. Juhani Airaksinen, R. Mitrani, A. Castellanos, and R. J. Myerburg, "Measurement of heart rate variability: A clinical tool or a research toy?" *J. Amer. College Cardiol.*, vol. 34, pp. 1878–1883, 1999.

[7] M. E. H. Ong, P. Padmanabhan, Y. H. Chan, Z. Lin, J. Overton, K. R. Ward, and D. Y. Fei, "An observational, prospective study exploring the use of heart rate variability as a predictor of clinical outcomes in pre-hospital ambulance patients," *Resuscitation*, vol. 78, pp. 289–297, 2008.

[8] H. Tsuji, F. J. Venditti Jr., E. S. Manders, J. C. Evans, M. G. Larson, C. L. Feldman, and D. Levy, "Reduced heart rate variability and mortality risk in an elderly cohort. The Framingham Heart Study," *Circulation*, vol. 90, pp. 878–883, 1994.

[9] B. M. Szabó, D. J. van Veldhuisen, N. van der Veer, J. Brouwer, P. A. De Graeff, and H. J. Crijns, "Prognostic value of heart rate variability in chronic congestive heart failure secondary to idiopathic or ischemic dilated cardiomyopathy," *Amer. J. Cardiol.*, vol. 79, pp. 978–980, 1997.

[10] L. Fei, X. Copie, M. Malik, and A. J. Camm, "Short- and long-term assessment of heart rate variability for risk stratification after acute myocardial infarction," *Amer. J. Cardiol.*, vol. 77, pp. 681–684, 1996.

[11] P. Ponikowski, S. D. Anker, T. P. Chua, R. Szelemej, M. Piepoli, S. Adamopoulos, K. Webb-Peploe, D. Harrington, W. Banasiak, K. Wrabec, and A. J. Coats, "Depressed heart rate variability as an independent predictor of death in chronic congestive heart failure secondary to ischemic or idiopathic dilated cardiomyopathy," *Amer. J. Cardiol.*, vol. 79, pp. 1645–1650, 1997.

[12] B. Goldstein and M. S. Ellenby, "Heart rate variability and critical illness: Potential and problems," *Crit. Care Med.*, vol. 28, pp. 3939–3940, 2000.

[13] C. Carpeggiani, A. L'Abbate, P. Landi, C. Michelassi, M. Raciti, A. Macerata, and M. Emdin, "Early assessment of heart rate variability is predictive of in-hospital death and major complications after acute myocardial infarction," *Int. J. Cardiol.*, vol. 96, pp. 361–368, 2004.

[14] N. Liu, Z. Lin, Z. X. Koh, G.-B. Huang, W. Ser, and M. E. H. Ong, "Patient outcome prediction with heart rate variability and vital signs," *J. Signal Process. Syst.*, vol. 64, pp. 265–278, 2011.

[15] C. B. Pearce, S. R. Gunn, A. Ahmed, and C. D. Johnson, "Machine learning can improve prediction of severity in acute pancreatitis using admission values of APACHE II score and C-reactive protein," *Pancreatology*, vol. 6, pp. 123–131, 2006.

[16] G. Cevenini and P. Barbini, "A bootstrap approach for assessing the uncertainty of outcome probabilities when using a scoring system," *BMC Med. Informat. Decis. Making*, vol. 10, p. 45, 2010.

[17] M. Khalilia, S. Chakraborty, and M. Popescu, "Predicting disease risks from highly imbalanced data using random forest," *BMC Med. Informat. Decis. Making*, vol. 11, p. 51, 2011.

[18] Task Force of the European Society of Cardiology the North American Society of Pacing Electrophysiology, "Heart rate variability: Standards of measurement, physiological interpretation, and clinical use," *Circulation*, vol. 93, pp. 1043–1065, 1996.

[19] J. W. Han and M. Kamber, *Data Mining: Concepts and Techniques*. San Mateo, CA: Morgan Kaufmann, 2006.

[20] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining Knowl. Discov.*, vol. 2, pp. 121–167, 1998.

[21] A. Kampouraki, G. Manis, and C. Nikou, "Heartbeat time series classification with support vector machines," *IEEE Trans. Inf. Technol. Biomed.*, vol. 13, no. 4, pp. 512–518, Jul. 2009.

[22] A. Temko, C. Nadeu, W. Marnane, G. B. Boylan, and G. Lightbody, "EEG signal description with spectral-envelope-based speech recognition features for detection of neonatal seizures," *IEEE Trans. Inf. Technol. Biomed.*, vol. 15, no. 6, pp. 839–847, Nov. 2011.

[23] C. C. Chang and C. J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, pp. 1–27, 2011.

[24] R. Polikar, "Ensemble based systems in decision making," *IEEE Circuits and Systems Magazine*, vol. 6, pp. 21–45, 2006.

[25] C. Ferri, J. Hernández-Orallo, and R. Modroiu, "An experimental comparison of performance measures for classification," *Pattern Recognit. Lett.*, vol. 30, pp. 27–38, 2009.

[26] A. K. Jain, R. P. W. Duin, and J. C. Mao, "Statistical pattern recognition: A review," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 1, pp. 4–37, Jan. 2000.

[27] C. P. Subbe, R. G. Davies, E. Williams, P. Rutherford, and L. Gemmell, "Effect of introducing the modified early warning score on clinical outcomes, cardio-pulmonary arrests and intensive care utilisation in acute medical admissions," *Anaesthesia*, vol. 58, pp. 797–802, 2003.

[28] S. B. Gottschalk, D. Wood, S. DeVries, L. A. Wallis, and S. Bruijns, "The cape triage score: A new triage system South Africa. Proposal from the cape triage group," *Emerg. Med. J.*, vol. 23, pp. 149–153, 2006.

[29] A. H. Khandoker, M. Palaniswami, and C. K. Karmakar, "Support vector machines for automated recognition of obstructive sleep apnea syndrome from ECG recordings," *IEEE Trans. Inf. Technol. Biomed.*, vol. 13, no. 1, pp. 37–48, Jan. 2009.

[30] H. B. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009.

[31] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

Authors' photographs and biographies not available at the time of publication.