An Integrated Data Mining Approach to Real-time Clinical Monitoring and Deterioration Warning

Yi Mao*
Department of Computer Science and Engineering
Washington University in St. Louis, Saint Louis, USA
maoy@seas.wustl.edu

Wenlin Chen
Department of Computer Science and Engineering
Washington University in St. Louis, Saint Louis, USA
wenlinchen@wustl.edu

Yixin Chen
Department of Computer Science and Engineering
Washington University in St. Louis, Saint Louis, USA
chen@cse.wustl.edu

Chenyang Lu
Department of Computer Science and Engineering
Washington University in St. Louis, Saint Louis, USA
lu@cse.wustl.edu

Marin Kollef
Washington University School of Medicine
Washington University in St. Louis, Saint Louis, USA
mkollef@dom.wustl.edu

Thomas C.Bailey
Washington University School of Medicine
Washington University in St. Louis, Saint Louis, USA
tbailey@dom.wustl.edu

ABSTRACT

Clinical study found that early detection and intervention are essential for preventing clinical deterioration in patients, for patients both in intensive care units (ICU) as well as in general wards but under real-time data sensing (RDS). In this paper, we develop an integrated data mining approach to give early deterioration warnings for patients under real-time monitoring in ICU and RDS.

Existing work on mining real-time clinical data often focus on certain single vital sign and specific disease. In this paper, we consider an integrated data mining approach for general sudden deterioration warning. We synthesize a large feature set that includes first and second order time-series features, detrended fluctuation analysis (DFA), spectral analysis, approximative entropy, and cross-signal features. We then systematically apply and evaluate a series of established data mining methods, including forward feature selection, linear and nonlinear classification algorithms, and exploratory undersampling for class imbalance.

An extensive empirical study is conducted on real patient data collected between 2001 and 2008 from a variety of ICUs. Results show the benefit of each of the proposed techniques, and the final integrated approach significantly improves the prediction quality. The proposed clinical warning system is currently under integration with the electronic medical record system at Barnes-Jewish Hospital in preparation for a clinical trial. This work represents a promising step toward general early clinical warning which has the potential to significantly improve the quality of patient care in hospitals.

*Visiting doctoral candidate from Xidian University, China

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*Data mining*; J.3 [**Computer Applications**]: Life and medical sciences

General Terms

Experimentation, Algorithms, Performance

Keywords

Real-time clinical monitoring, deterioration warning, time-series classification, feature selection

1. INTRODUCTION

Every year, 4-17% of patients undergo cardiopulmonary or respiratory arrest while in hospitals. Lots of these patients could have been saved if warning of serious clinical events could be provided early, before its occurrence rather than when it is happening. Early prediction based on real-time electronic monitoring data has become an apparent need in many clinical areas.

At Washington University, we have carried out a NIH-funded clinical trial of a real-time patient monitoring system in a step-down unit at Barnes-Jewish Hospital, one of the largest hospitals in the United States [9]. This clinical trial uses wireless sensing devices to collect real-time vital sign data for patients not only in ICU but also in general units. The goal of this pioneering study is to show the feasibility of using data mining algorithms to give early warning of sudden deterioration, ultimately leading to the prevention of death.

However, most prior studies focus on some specific disease prediction. For example, McQuatt et al. [2] introduced a decision tree method to analyze head injury. Loforte et al. [13] found statistical indexes for detecting sepsis by investigating

the relationship between heart rate and respiration. Khosla et al. [6] applied multiple machine learning techniques for stroke prediction. There have been little study on general prediction and warning for serious clinical deterioration and death.

Data mining on clinical data has great potential to improve the treatment quality of hospitals and increase the survival rate of patients. Data-driven prediction technology strongly hinges on the data collection of patients' vital signs. In most hospitals, only intensive care units (ICUs) are equipped with real-time electronic medical sensors. In general hospital units, patients' vital signs are typically collected manually by a nurse, at a granularity of only a handful of readings per day, which confronts us with the challenged of sparseness and irregularity of data. To handle this problem, our team has proposed a real-time data sensing (RDS) system, which enables patients' vital signs data in general hospital units to be collected via wireless sensors. Currently, RDS can provide constant monitoring of patients' heart rate and oxygen saturation rate. Through a pilot study at the Barnes-Jewish Hospital, the RDS system has been installed in its step-down unit. The success of RDS will dramatically enlarge the population that have real-time monitoring data. Hence, there is a critical need for a data mining system that can effectively utilize the multi-dimensional, real-time time-series data from ICU and RDS in order to prevent deterioration and mortality in hospitalized patients.

In this paper, we develop an integrated data mining approach to identify the signs of clinical deterioration and provide early warning for possible mortality. In particular, we build classification methods to monitor real-time signal of heart rate and oxygen saturation rate of patient and issue early warning alerts before clinical deterioration/death. This system enables at-risk patients to be timely checked and treated by healthcare professionals in order to prevent potential deterioration and death.

Studies have found that real-time clinical data has certain unique features, due to the underlying dynamics of biological systems [5]. Therefore, some dedicated algorithms based on nonlinear dynamics, such as detrended fluctuation analysis (DFA) and spectral analysis, have been proposed for clinical data. However, such prior work has two downsides. First, most of them consider only a single vital sign. For example, Penzel et al. [11] used DFA and spectral analysis for sleep apnea detection leveraging on the heart rates. Mining multiple time series is a more challenging problem. Second, time series data has rich information that previous work did not make full use of. For single time series, there are first-order and second-order features, as well as some sophisticated patterns, such as DFA, spectra and entropy. For multiple time series, features such as correlation and coherence can be used. Our approach not only leverages all these information but also performs feature selection to select the most relevant and discriminative features. For classification, researchers in the clinical community routinely use simple linear classifiers [5]. Our system incorporates more robust classifiers such as SVM with RBF kernels and logistic regression. Moreover, we use an exploratory undersampling method to address the challenge of class imbalance widely observed in clinical data.

In summary, this paper contains the following contributions:
1. We developed an integrated data mining approach to early deterioration warning based on real-time monitoring data. The approach is applicable to both ICU patients and patients in general hospital wards equipped with real-time sensing devices.
2. We bridge the gap between biomedical community and data mining community by incorporating popular features and methods in both areas.
3. We strengthen the early warning system by applying important data mining techniques such as feature selection, exploratory undersampling, and advanced classifiers to address the challenging multi-dimensional time-series classification problem.
4. We apply our integrated approach to a large collection of real patient data recorded from ICUs and show significant improvement over previous methods. The results also validate the effectiveness of the employed techniques.

The rest of this paper is organized as follows. Section 2 surveys the related work in detecting clinical deterioration. An overview on our early warning system is presented in Section 3. Section 4 describes our general approach and the evaluation criterion. Section 5 lists and describes all the features extracted from multiple vital sign time series. Section 6 and Section 7 describe the feature selection and prediction approach, respectively. Section 8 shows the experimental result of our real-time early warning system. Finally, we draw conclusion in Section 9.

2. RELATED WORK

Medical data mining is a key technique to extract useful clinical knowledge from medical records. A number of scoring systems exist that use medical knowledge for various medical conditions.

For example, Several Community-Acquired Pneumonia (SCAP) and Pneumonia Severity Index (PSI) were used to predict outcomes in patients with pneumonia [18]. Similarly, outcomes in patients with renal failures may be predicted using the Acute Physiology Score, Chronic Health Score, and APACHE score [17]. In [20], Zhou proposed a Multi-Task Learning Formulation for Predicting Alzheimer's Disease. The integration of heterogeneous data (neuroimages, demographic, and genetic measures) for AD prediction based on a kernel method was proposed in [19].

Detrended fluctuation analysis (DFA) and spectral analysis for heart rate variability were evaluated to classify sleep apnea and normal sleep [11]. RR (inter-beat) interval and spontaneous respiration were analyzed using approximated entropy(ApEn) and regularity index to distinguish sepsis [13]. Decision tree is introduced in predicting the outcome of head injury patients using both background (demographic) data and temporal (physiological) data [2]. Also, SVM and feature selection are employed to predict stroke [6].

However, most of these algorithms are designed for some specific diseases and to be used in some specialized hospital units. In contrast, the detection of clinical deterioration requires more general algorithms. For example, a team at the John Hopkins University developed the Modified Early

Warning Score (MEWS) [7], which uses manually-collected systolic blood pressure, pulse rate, temperature, respiratory rate, age and BMI to predict clinical deterioration. Our team has also developed a learning algorithm to identify high-risk patients based on clinical data collected by nurses [9]. However, both of these works are applied to manually collected data which has only a handful of readings per day. Here, our goal is different. We are aiming at mining real-time vital signs read by electronic devices at ICU and wireless sensors in our project. Such data is very different from manually collected data, as they are regular and have high frequency (reading gaps being minutes instead of hours).

3. REAL-TIME MONITORING IN GENERAL HOSPITAL UNITS

Our work aims to prevent clinical deterioration in patients, for both patents in ICU as well as in general hospital units equipped with sensing devices, such as our RDS system..

In our RDS pilot study at the Barnes-Jewish Hospital, patients are provided with wireless sensor network (WSN) devices which collect and stream real-time vital sign data to the learning system. If deterioration is predicted, a warning is sent to nurses on the patient's floor over the hospital's paging system. The nurses may then intervene to prevent deterioration. Figure 1 shows a WSN-based wireless pulse oximeter device developed by our team, which is capable of collecting equipped patent's real-time heart rate and oxygen saturation rate and then route the data to a wired access point using an onboard radio based on the IEEE 802.15.4 radio standard. WSN network coverage is achieved by plugging additional TelosB nodes into electrical outlets in patients' rooms and in the hallway. These new nodes will autonomously locate other nearby nodes and participate in routing sensor data from patient nodes to a wired access point, where they are entered into our data mining system.

RDS enjoys robust system reliability and lifetime. It achieved high network reliability, with a median of 99.68% of packets successfully delivered to the base station. Network outages were infrequent and had a 95%-percentile time-to-recovery of 2.4 minutes. Patient nodes achieved a lifetime of up to 69 hours from a 9V battery power source.

The establishment of the RDS represents a important step toward early clinical warning that has the potential to significantly improve the quality and outcome of patient care in hospitals. RDS shows the feasibility that, real-time physiologic data will be available for not only patients at ICU, but also patients in the general hospital units with real-time monitoring. Due to the inevitable trend of popular uses of such sensing systems in hospitals, our data mining approach is meeting a critical challenge in the healthcare industry and has the potential of very broad usage.

4. OVERVIEW OF OUR APPROACH

In this section, we overview the proposed approach for early prediction of clinical deterioration/death based on real-time monitoring data from ICU and RDS. Specifically, in this work, we try to predict mortality (death) of hospitalized patients based on two vital signs that are most popular: heart rate and oxygen saturation rate. Our pulseox sensor

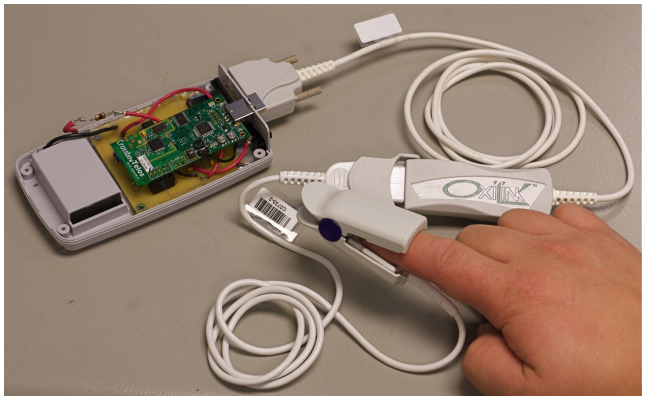

Figure 1: A wireless pulse-oximeter node developed at Washington University.

can measure these two signs. The procedure consists of the following steps:

- Preprocess the data set by removing abnormal values.
- Extract features from the patient's collected real-time vital sign time series, including heart rate and oxygen saturation rate. The extracted features contains DFA, spectral analysis, first order and second order features, and multi-sign features. All features are normalized to the range of [0,1].
- Apply feature selection techniques to select the most relevant and discriminative features.
- Apply classification algorithms to train the model, and evaluate the prediction performance by cross validation. To deal with class imbalance widely observed in clinical data, we also use an exploratory undersampling method.

The evaluation of the prediction performance is based on the following criteria: AUC (Area Under receive operating characteristic Curve), PPV (Positive Predictive Value), NPV (Negative Predictive Value), sensitivity, and specificity. In clinical community, the PPV stands for the proportion of patients who actually suffer deterioration/death, among the candidate patients who are warned by our system. A high NPV means that patients who survived are rarely misclassified. PPV/NPV and sensitivity/specificity are trade-off pairs where typically lowering one makes the other higher. PPV and NPV are also sensitive to how imbalanced the dataset is. Hence, we adopt the AUC as our main metric since it is a comprehensive measurement combining sensitivity and specificity to deal with the imbalance dataset.

5. FEATURES

In this section, we describe all the features extracted from patients' vital sign time series - heart rate and oxygen saturation rate. There are in total 34 features from these two time series data, including features within single time series and features linking the two time series.

5.1 Detrended Fluctuation Analysis (DFA)

In stochastic processes, chaos theory and time series analysis, detrended fluctuation analysis (DFA) is a method for

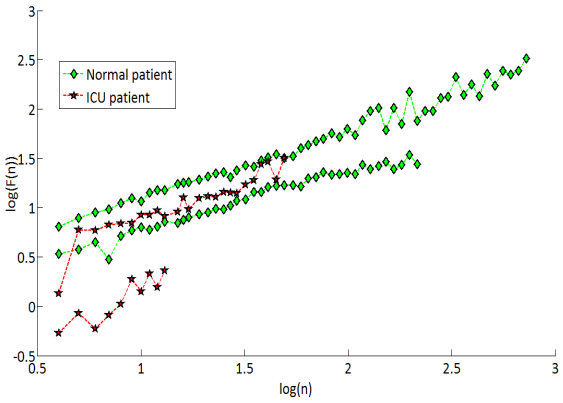

Figure 2: DFA analysis on the heart rates of two ICU and two non-ICU patients.

determining the statistical self-similarity of a signal [5]. Self-similarity is a key feature that is widely observed in natural systems, including human physiological signs. Mathematically, DFA is a scaling analysis method to reveal long-range power-law correlation exponents in noisy time series [5]. It is most suitable for non-stationary time series with slowly varying trends, such as heart rate and oxygen saturation rate. The DFA of a time series is calculated as the average fitting error over all segments in different scales. Given a time series $\{x(i)\}$,$1 \leq i \leq N$, integration is performed to convert the original time series as follows:

$$y(j) = \sum_{i=1}^{j} [x(i) - \langle x \rangle], 1 \leq j \leq N$$

where

$$\langle x \rangle = \frac{1}{N} \sum_{i=1}^{N} x(i)$$

Next, the integrated time series $y(i)$ $(i = 1, 2, ...N)$ is divided into boxes of equal length n. A polynomial function y_n is fitted to each box of length n by minimizing the least square error. Then, we detrend the integrated time series y by subtracting y_n from each box. The root-mean-square fluctuation of the detrended time series is calculated by

$$F(n) = \sqrt{\frac{1}{N} \sum_{j=1}^{N} [y(j) - y_n(j)]^2}$$

Typically,$F(n)$ increases with n and follows the power law: $F(n) \sim n^{\alpha}$. Using a log-log plot, we can get a nearly linear curve as shown in Figure 2. The scaling exponent α characterizes the self-similarity level of original time series $x(t)$, which can be calculated using the slope of this curve.

As pointed out in [10], it is important to differentiate the short-range and long-range levels of self-similarity. Hence, we divide the curve into two pieces and fit a linear function to each piece of curve, which generates two slops - α_1 and α_2. α_1 represents the slope of curve $\log(F(n))$ vs. $\log(n)$ in the range $1 \leq n \leq \phi(N)$ while α_2 represents the slope of curve in the range $\phi(N) < n \leq N$. In this way, we can see that α_1 reflects the short-range self-similarity level and α_2 reflects the long-range self-similarity level. As different patients have time series of different length, we cannot apply the same curve-segmentation method as in [11]. Instead, we select the $\phi(N)$ by a ratio γ, where $\phi(N) = \gamma * N$. Regarding the selection of γ, we first sample two groups of died patients and survived patients, respectively. Next, for a fixed γ, we calculate the α_1 for each patient as well as the sum of all α_1 of each group. γ is selected to maximize the difference between sums of α_1 of the two groups.

5.2 Approximate entropy (ApEn)

Approximate entropy was introduced by Pincus to quantify the time series complexity closely related to entropy [12]. It is a measurement designed to quantify the degree of regularity versus unpredictability. It quantifies the unpredictability of fluctuations in a time series. A low value of the entropy indicates that the time series is deterministic while a high value means that the time series is unpredictable(randomness).

To compute the approximate entropy (ApEn) of a time series, we divide the time series into $N - m + 1$ sub-series, calculate the similarity between each other and then figure out the entropy. First, sub-series of vectors of length m, $v(n) = [x(n), x(n+1), ...x(n+m-1)]$, $n = 1, \cdots, N-m+1$, are derived from the original signal. The distance $D(i, j)$ between two vectors $v(i)$ and $v(j)$ is defined as the maximum difference in the scalar components of $v(i)$ and $v(j)$. Then $N^{m,r}(i)$, the number of vectors j such that the distance between the vectors $v(j)$ and the generic vector $v(i)$ is lower than r, is computed. The index r is a fixed parameter which sets the "tolerance" of the comparison. Then we consider $C^{m,r}(i)$, the probability to find a vector that differs from $v(i)$ by a distance less than r, as:

$$C^{m,r}(i) = N^{m,r}(i)/(N - m + 1) \quad (1)$$

and the logarithmic average over all the vectors of the $C^{m,r}(i)$ probability is calculated as follows:

$$F^{m,r} = \frac{\sum_{i=1}^{N-m+1} \ln(C^{m,r}(i)}{N - m + 1} \quad (2)$$

ApEn is given by:

$$ApEn^{m,r} = F^{m,r} - F^{m+1,r} \quad (3)$$

As recommended in [12], we use m = 2,r = 20% of the standard deviation of the time series in our analysis.

5.3 Spectral analysis

Spectral analysis is another major method for analyzing clinical time-series data [11]. For the calculation of the power spectra, the time series was resampled at 3.4 Hz using interpolation. The mean value, the standard deviation value was subtracted from the time series before applying the Fast Fourier Transformation (FFT). In case of having less than 2^N of records, zero-padding was used. The spectral analysis and the interpolation routine was implemented using Matlab 7.11.0. We calculated the component values for VLF (<=0.04HZ), LF (0.04-0.15HZ), HF (0.15-0.4HZ), and the ratio LF/HF for each time-series. Figure 3 shows an example spectral distribution of the heart rate of an ICU patient.

5.4 First order features

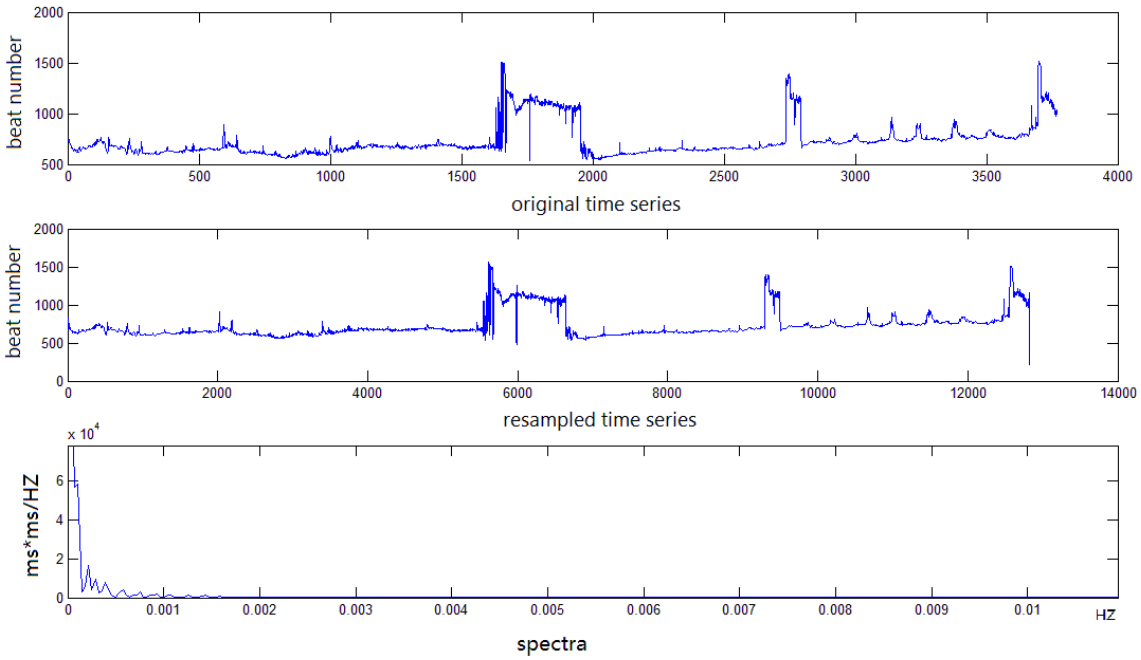

Figure 3: The calculation of the spectral features using FFT.

For first order features, we use some traditional statistical features, such as mean (μ), standard deviation (σ), skewnes s(γ_1), and kurtosis (γ_2). In probability theory and statistics, skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable. The skewness value can be positive or negative, or even undefined. Qualitatively, a negative skew indicates that the tail on the left side of the probability density function is longer than the right side and the bulk of the values (possibly including the median) lie to the right of the mean. A positive skew indicates the opposite. A zero value indicates that the values are relatively evenly distributed on both sides of the mean, typically but not necessarily implying a symmetric distribution. Kurtosis is a measure of the "peakedness" of the probability distribution of a real-valued random variable. In a similar way to the concept of skewness, kurtosis is a descriptor of the shape of a probability distribution and, just as for skewness, there are different ways of quantifying it for a theoretical distribution and corresponding ways of estimating it from a sample from a population. The computation we used is as follows:

$$\mu = \frac{\sum_{i=1}^{N} x(i)}{N}$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N} (x(i) - \mu)}{N}}$$

$$\gamma_1 = \frac{\sum_{i=1}^{N} (x(i) - \mu)^3}{n\sigma^3}$$

$$\gamma_2 = \frac{\sum_{i=1}^{N} (x(i) - \mu)^4}{n\sigma^4} - 3$$

5.5 Second order features

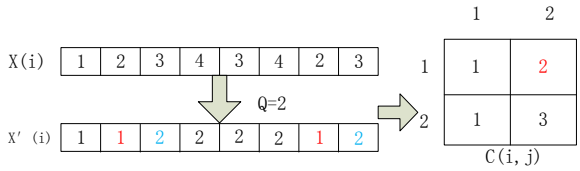

Figure 4: The process that transfers an one dimensional time series to two dimensional matrix.

Here, we employ the description which related to co-occurrence features in one dimensional time series [1]. First, the data is quantized into Q levels. Second, a two dimensional matrix $c(i, j)$ is constructed ($1 \leq i, j \leq Q$). Point (i,j) in the matrix represents the number of times that a point in the sequence with level i is followed, at a distance d_1, by a point with level j. Figure 4 shows how this process works. Finally, five co-occurrence features, including energy (E), entropy (S), correlation (COR)(ρ_x, y), inertia (F), and local homogeneity (LH), are calculated using the following equations:

$$E = \sum_{i=1}^{Q} \sum_{j=1}^{Q} c(i, j)^2$$

$$S = \sum_{i=1}^{Q} \sum_{j=1}^{Q} c(i, j) * \log(c(i, j))$$

$$\rho_{x,y} = \frac{\sum_{i=1}^{Q} \sum_{j=1}^{Q} (i - \mu_x)(j - \mu_y)c(i, j)}{\sigma_x * \sigma_y}$$

where:

$$\mu_x = \frac{\sum_{i=1}^{Q} i \sum_{j=1}^{Q} c(i,j)}{Q}$$

$$\mu_y = \frac{\sum_{j=1}^{Q} j \sum_{i=1}^{Q} c(i,j)}{Q}$$

$$\sigma_x^2 = \frac{\sum_{i=1}^{Q} (i-\mu_x)^2 \sum_{j=1}^{Q} c(i,j)}{Q}$$

$$\sigma_x^2 = \frac{\sum_{j=1}^{Q} (j-\mu_y)^2 \sum_{i=1}^{Q} c(i,j)}{Q}$$

$$F = \sum_{i=1}^{Q}\sum_{j=1}^{Q} (i-j)^2 c(i,j)$$

$$LH = \sum_{i=1}^{Q}\sum_{j=1}^{Q} \frac{1}{1+(i-j)^2} c(i,j)$$

In our experiments, we set $Q = 10$.

5.6 Cross-sign features

We also consider features that link multiple vital signs together, including linear correlation and coherence.

5.6.1 Linear Correlation

Correlation indicates the strength and direction of a linear relationship between two random variables. In general, it refers to the departure of two variables from independence and equals:

$$\gamma_{1,2} = \frac{E[(X_1(t) - E(X_1(t)))(X_2(t) - E(X_2(t)))]}{Var[X_1(t)] \cdot Var[X_2(t)]}$$

5.6.2 Coherence

Coherence provides both amplitude and phase information about the frequencies held in common between the two time series and is defined by:

$$C_{1,2} = \frac{\phi_{X_1 X_2}}{[\phi_{X_1 X_1} \cdot \phi_{X_2 X_2}]^{\frac{1}{2}}}$$

where $\phi_{x_1 x_2}$ is the cross-spectral density, and $\phi_{x_1 x_1}$ and $\phi_{x_2 x_2}$ are autospectral densities.

6. FEATURE SELECTION

Feature selection in supervised learning has been well studied, which aims to find the most relevant features and produce better classification performance than using all the features. In contrast to other dimensionality reduction techniques like those based on projection (e.g. principal component analysis) or compression (e.g. using information theory), feature selection techniques do not alter the original representation of the variables, but merely select a subset of them. Thus, besides reducing the utilization time and storage requirements, they preserve the original semantics of the variables, hence offering the advantage of interpretability to users [15].

While there are many classes of feature selection methods, here we adopt a forward feature selection algorithm which does not depend on a particular classifier. Forward feature selection is a simple search strategy to find useful features [4]. The basic idea is we starts with an empty feature subset and adds one variable each step until a predefined number of features is reached, or the approximation result does not improve any further.

We use two metrics for our forward selection, AUC and F-score. The AUC is the area under the ROC curve. F-score [3] is defined as:

$$F(i) \equiv \frac{(\overline{x_i}^{(+)} - \overline{x_i})^2 + (\overline{x_i}^{(-)} - \overline{x_i})^2}{\frac{1}{n_+ - 1}\sum_{k=1}^{n_+}(x_{k,i}^{(+)} - x_i^{(-)})^2 + \frac{1}{n_- - 1}\sum_{k=1}^{n_-}(x_{k,i}^{(-)} - x_i^{(-)})^2}$$

where $\overline{x_i}$, $\overline{x_i}^{(+)}$, $\overline{x_i}^{(-)}$ are the average of the ith feature of the whole, positive,and negative data sets, respectively; $x_{k,j}^{(+)}$ is the ith feature of the kth positive instance, and $x_{k,j}^{(-)}$ is the ith feature of the kth negative instance. The numerator indicates the discrimination between the positive and negative sets, and the denominator indicates the one within each of the two sets. The larger the F-score is, the more likely this feature is more discriminative.

7. CLASSIFICATION ALGORITHMS

In our approach, we apply Support Vector Machine (SVM) and logistic regression for prediction. In addition, to deal with the class imbalence of our dataset, we adopt the exploratory undersampling method.

7.1 Support vector machine

SVM is one of the most popular classification methods based on statistical learning theory. The key idea is to learn an optimal hyperplane that can separate the training data set with maximum margin [16]. Some previous work in DFA and spectral analysis use a linear separator [5,10,11] which corresponds to SVM with a liner kernel. Generally, by importing non-linear kernel, SVM has higher accuracy compared to other linear classifiers.

7.2 Logistic regression

Logistic regression is a model for predicting the probability of an event, which can also be used for binary classification. Logistic regression has the benefit of being able to output a numerical score to reflect the severity of the patient. Also, logistic regression allows us to control the sensitivity/specificity tradeoffs by adjusting the classification threshold.

7.3 Exploratory undersampling

Looking through the records, we have a skewed dataset. Among 772 records, 175 are from visits belong the positive (death) set. Undersampling [8] is a very popular method in dealing with the class-imbalance problem. The idea is to combine the minority class with only a subset of the majority class each time to generate a sampling set, and take the ensemble of multiple sampled models. We have tried undersampling on our data but obtained very modest improvements.

Method	Features	AUC	Specificity	Sensitivity	PPV	NPV
linear SVM	DFA of heart rate (HR)	0.5759	0.9497	0.0755	0.2550	0.7781
Logistic Regression	DFA of HR	0.4742	0.9483	0.0729	0.3181	0.7555
kernel SVM	DFA of HR	0.5897	0.9497	0.1265	0.3643	0.7879
linear SVM	DFA of oxygen saturation rate (OS)	0.4473	0.9497	0.0346	0.1300	0.7705
Logistic Regression	DFA of OS	0.4902	0.9483	0.0313	0.1667	0.7473
kernel SVM	DFA of OS	0.5016	0.9497	0.0676	0.2450	0.7768
Logistic Regression	DFA of HR+OS	0.5370	0.9483	0.0521	0.2500	0.7513
kernel SVM	DFA of HR+OS	**0.6332**	0.9497	**0.1428**	**0.4146**	**0.7911**

Table 1: The performance comparison of different classifiers using DFA in different time series and their combination.

To improve the performance further, we used a recent method called exploratory undersampling [8], which makes better use of the majority class than simple undersampling. The idea is to iteratively remove those samples that can be correctly classified by a large margin to the class boundary by the existing model.

Specifically, we fix the number of the died patients, and then randomly choose the same amount of survived patients to build the training dataset at each iteration. The main difference to simple undersampling is that, each iteration, it removes 5% in both the majority class and the minority class with the maximum classification margin. For logistic regression, we remove those died patients that are closest to 1 (the class label of death) and those survived patients that are closest to 0. For SVM, we remove those correctly classified patients with the maximum distance to the boundary.

8. EXPERIMENT RESULTS

In this section, we present the experimental result of the performance of our integrated learning algorithms. We first introduce the database of real-time vital signs used in our study. Then, we show and discuss the experimental results on the performance and advantages of various proposed techniques.

8.1 Database and experimental setup

Since our RDS system is still in a smaller scale clinical trial and does not yield enough data, we test our approach based on the MIMIC II (Multiparameter Intelligent Monitoring in Intensive Care) [14] database which contains comprehensive clinical data from tens of thousands of Intensive Care Unit (ICU) patients. Data were collected between 2001 and 2008 from a variety of ICUs (medical, surgical, coronary care, and neonatal). The database denotes the outcome of each patient while they are in ICU (died or survived). The database also includes thousands of records of continuous high-resolution physiologic waveforms and minute-by-minute numerical time series (trends) of physiologic measurements. Since our learning model is to be embedded into an early warning system that is based on heart rate and oxygen saturation rate collected by wireless sensors, we only extract the time series of heart rate and oxygen saturation rate of patients for training the model from MIMIC II. We also discover that the extracted dataset suffers the class-imbalance problem, i.e. most patients in the dataset are from one class (no deterioration). To deal with this skewed dataset, later on we will adopt explorary undersampling and demonstrate its improvement on performance.

Our experiment is based on 10-fold cross validation. That is, we divide the data set into 10 smaller datasets with 9 training sets and 1 test set. Each subset is validated once as testing set. The results are the average of all the 10 validations. In addition to listing the AUC, we also give the sensitivity, specificity, PPV and NPV. For logistic regression, since we can vary its results by changing the threshold, we always set the threshold so that the specificity close to 0.95 by choosing a proper threshold when presenting sensitivity, PPV and NPV. For the implementation of nonlinear SVM, we use a RBF kernel.

8.2 Comparison of linear and nonlinear separation using DFA

Prior study [11] in clinical community only considers one single time series with linear separation. We evaluated a number combinations of classifiers and features listed in Table 1 to demonstrate that classifying multiple time series with nonlinear separation will improve the prediction performance for clinical data. For comparison, we employed different classifers with DFA features on heart rate only, oxygen saturation rate only, and both.

From Table 1, we could observe the following facts. First, on single time series, we discover that kernel SVM outperforms linear SVM, showing that clinical data with sparse features is better separated by a nonlinear classfier. Second, the combination of both time series greatly improve the performance of each classifier, which justifies our motivation of combining multiple time series.

8.3 Features combination

In this section, we use a series of increasingly large feature set on two classifiers to see the performance of multiple features, as shown in Table 2. From the table, we can discover that as we add more features, the performance is improved. This is in contrast with Table 1, in which kernel SVM outperforms logistic regression when there are only a small number of features. It shows that all of the newly introduced features contribute to the improvement of performance. We can also observe that the logistic regression classifer outperforms SVM with RBF kernel in our dataset, which means that nonlinear SVM may suffer overfitting when the number of features becomes larger.

8.4 Features selection and exploratory undersampling

Algorithms	Features	AUC
RBF Kernel SVM	DFA	0.6332
	DFA+Cross Feature	0.6565
	DFA+Cross Feature+ApEn	0.6753
	All Features	0.7090
Logistic Regression	DFA	0.5370
	DFA+Cross Feature	0.5731
	DFA+Cross Feature+ApEn	0.5974
	All Features	0.7402

Table 2: The performance comparison of different features.

To avoid overfitting and improve the generalization ability, we perform the forward feature selection (FFS) technique based on SVM and logistic regression to select the most relevant features. We adopt two kinds of selection criterion: F-Score and AUC. Table 3 shows the performance comparison of feature selection using SVM and logistic regression. We can observe that by using feature selection, both SVM and logistic regression get significant improvement on their performance. And, logistic regression outperforms SVM, which makes logistic regression our first choice for further experiments of our learning system. In addition, feature selection based on AUC outperforms the one based on F-Score, for both logistic regression and SVM. Finally, we can see that the number of features selected by logistic regression is larger than the one by SVM, which further demonstrates our conclusion that too many features tend to cause overfitting in SVM with RBF kernel.

Regarding the class-imbalance problem of our dataset, we apply an exploratory undersampling method whose result is shown in Table 4. We can see that, with all features used, logistic regression with exploratory undersampling trained on all features (AUC = 0.7767) outperforms the one without exploratory undersampling (AUC=0.7402). In addition, the combination of logistic regression, exploratory undersampling and forward feature selection with AUC further improves the performance (AUC = 0.8082), which forms the current learning model for our early warning system. Comparing the first row of Table 3 against the last row of Table 4, we can also see that exploratory undersampling also improves performance in case feature selection is used on logistic regression.

8.5 Identifying leading risk factors

Other than just achieving better result on prediction performance, with our learning system, we also provide some insights on the leading risk factors for patient deterioration and mortality. Table 5 lists the first dozen of features by our feature selection method with AUC score. We can see that all kinds of proposed features are selected, including DFA, spectral analysis, first order, second order, ApEn and cross-sign features.

Table 6 provides the 10 most significant features of our final logistic regression model, ordered by the magnitude of their coefficients in the model. Note that all feature values are normalize to between 0 and 1, so the coefficient denotes the sensitivity of the model to a feature. With this table, we can identify some factors that are highly related to deterioration. For example, if the local homogenity of patent's heart rate (top ranked in Table 6) is low, then this patient is likely to suffer sudden clinical deterioration and death. The standard deviation of oxygen saturation rate is also a significant factor for clinical deterioration. It is confirmed by the clinical observation that high variation in a physiologic index strongly indicates the low stability of patients' health status.

Feature
standard deviation of heart rate
Apen of heart rate
Energy of oxygen saturation
LF of oxygen saturation in SPA
LF of heart rate in SPA
DFA of oxygen saturation
Mean of heart rate
HF of heart rate in SPA
Inertia of heart rate
Homogeneity of heart rate
Energy of heart rate
linear correlation

Table 5: The first 12 selected features in logistic regression using forward feature selection with AUC.

Feature	Coefficient
local homogenity of heart rate	-14.50
standard deviation of oxygen saturation	10.20
entropy of oxygen saturation	10.17
low frequency of heart rate	8.62
local homogenity of oxygen saturation	7.77
LF/HF of oxygen saturation	4.53
inertia of heart rate	3.86
entropy of heart rate	2.97
low frequency of oxygen saturation	-2.89
mean of oxygen saturation	-2.86

Table 6: The 10 highest-weighted variables of our final logistical regression model (with exploratory undersampling and feature selection based on AUC).

9. CONCLUSION

Preventing clinical deterioration and death in hospital patients by mining electronic medical records is a promising and important trend in US hospitals. We have developed a predictive system that can provide early alert of deterioration for patients under real-time monitoring in ICUs and in general hospital units, as enabled by wireless sensing systems such as the RDS system we developed in Barnes-Jewish Hospital. Our approach integrated features from a diversity of fields including chaos theory (DFA), signal processing (spectral analysis and entropy), and machine learning (time-series features). We showed that the combined feature set gives significant performance improvement. We also showed that robust classifiers such as kernel SVM and logistic regression outperform previously used linear classifiers. Moreover, we showed that using established data mining methods, including feature selection and exploratory undersampling, can also improve the performance. With a AUC over 0.8 and sensitivity near 0.5 at 0.95 specificity, our final logistic regression model is approved by medical experts for a clinical

Method	Number of Selected features	AUC	Specificity	Sensitivity	PPV	NPV
Logistic Regression (AUC)	23	**0.7844**	0.9483	0.5208	0.7692	0.8567
Logistic Regression (F-score)	26	0.7592	0.9483	0.5104	0.7656	0.8540
SVM (AUC)	5	0.7752	0.9654	0.4852	0.8041	0.8651
SVM (F-score)	4	0.7736	0.9497	0.4833	0.7163	0.8652

Table 3: The performance comparison of SVM and logistic regression combined with feature selection based on AUC score and F-score.

Method	AUC	Specificity	Sensitivity	PPV	NPV
Logistic Regression + all features	0.7402	0.9483	0.3646	0.7000	0.8185
Logistic Regression+all features+exploratory undersampling	0.7767	0.9500	0.4615	0.9000	0.6440
Logistic Regression+exploratory undersampling+feature selection	**0.8082**	0.9473	0.4865	0.9000	0.6546

Table 4: The performance of logistic regression using exploratory undersampling. Feature selection is based on the AUC.

trial at a major hospital.

10. REFERENCES

[1] R. J. Alcock and Y.Manolopoulos. Time-series similarity queries employing a fearture-based approach. pages 27–29, 1999.

[2] A.McQuatt, P.J.D.Andrews, D.Sleeman, V.Corruble, and P.A.Jones. The analysis of head injury data using decision tree techniques. *Artifical Intelligent in Medicine*, 1620:336–345, 1999.

[3] Y. Chen and C. Lin. Combining svms with various feature selection strategies. In *Taiwan University*, 2005.

[4] I. Guyon, S. Gunn, M. Nikravesh, and L. A.Zadeh, editors. *Feature Extraction: Foundations and Applications (Studies in Fuzziness and Soft Computing)*. 2006.

[5] K. Hu, P. Ch.Ivanov, Z. Chen, P. Carpena, and H. E. Stanley. Effect of trends on detrended fluctuation analysis. *Physical Review*, 2001.

[6] A. Khosla, Y. Cao, C. Lin, H. Chiu, J. Hu, and H. Lee. An integrated machine learning approach to stroke prediction. In *Prodeedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 183–192, 2010.

[7] J. Ko, J. H. Lim, Y. Chen, R. Musvaloiu-E, A. Terzis, G. M. Masson, T. Gao, W. Destler, and L. S. R. P. Dutton. Medisn: Medical emergency detection in sensor networks. In *Prodeedings of the 6th ACM coference on Embedded network sensor systems*, pages 361–362, 2010.

[8] X. Liu, J. Wu, and Z. Zhou. Exploratory under-sampling for class-imbalance learning. *Sixth IEEE International Conference on Data Mining*, 2006.

[9] Y. Mao, Y. Chen, G. Hackmann, M. Chen, C. Lu, M. Kollef, and T. Bailey. Early deterioration warning for hospitalized patients by mining clinical data. *International Journal of Knowledge Discovery in Bioinformatics*, 2(3):1–20, 2012.

[10] C. Peng, S. Havlin, S. H. Eugene, and G. A. L. Quantification of scaling exponents and crossover phenomena in nonstationary heartbeat time series. *Chaos*, 5:82, 1995.

[11] T. Penzel, J. W.Kantelhardt, L. Grote, J.-H. Peter, and A. Bunde. Comparison of detrended fluctuation analysis and spectral analysis for heart rate variability in sleep and sleep apnea. *IEEE Transactions on Biomedical Engineering*, 50:10, 2003.

[12] S. Pincus. Approximate entropy as a measure of system complexity. volume 88, pages 2297–2301, 1991.

[13] R.Loforte, G.Carrault, L.Mainardi, and A.Beuche. Heart rate and respiration relationship as a diagnostic tool for late onset sepsis in sick preterm infants. *Computers in Cardiology*, 33:737–740, 2006.

[14] M. Saeed, M. Villarroel, A. T. Reisner, G. Clifford, L.-W. Lehman, G. Moody, T. Heldt, T. H. Kyaw, B. Moody, and R. G. Mark. Multiparameter intelligent monitoring in intensive care ii: a public-access intensive care unit database. *Critical Care Medicine*, 39(5), 2011.

[15] Y. Saeys, I. Inza, and P. Larranaga. A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23:2507–2517, 2007.

[16] W. Vladimir N. Vapnik. *Statistical Learning Theory*. 1998.

[17] W.A.Knaus, E.A.Draper, D.P.Wagner, and J.E.Zimmerman. Apache ii: a severity of disease classification system. *Critical Care Medicine*, 13(10):818–829, 1985.

[18] Yandiola, P. P. Espana, A. Capelastegui, J. Quintana, D. Rosa, G. Inmaculada, B. Amaia, Z. Rafael, M. Rosario, and T. Antonio. Prospective comparison of severity scores for predicting clinically relevant outcomes for patients hospitalized with community-acquired pneumonia. 135(6):1572–9, 2009.

[19] J. Ye, K. Chen, T. Wu, J. Li, Z. Zhao, R. Patel, M. Bae, R. Janardan, H. Liu, G. Alexander, and E. Reiman. Heterogeneous data fusion for alzheimer's disease study. pages 1025–1033, 2008.

[20] J. Zhou, Y. Lei, J. Liu, and J. Ye. A multi-task learning formulation for predicting disease progression. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 814–822, 2011.