Paper presented at Chemical and Biological Sensing VIII - The International Society for Optical Engineering (SPIE)
Orlando, FL
17-21 April 2006

Estimation of confidence levels for physiology variables measured by a vital signs detection system

Jingyu Liu[a], Thomas M. McKenna[a], Andrei Gribok[a], Beth A. Beidleman[b], William T. Tharion[b], Jaques Reifman[*a]
[a]Bioinformatics Cell, U.S. Army Medical Research and Materiel Command, Fort Detrick, Maryland, 21702
[b]U.S. Army Research Institute of Environmental Medicine, Natick, Massachusetts, 01760

ABSTRACT

Quantifying the accuracy of physiological data measured by a Vital Signs Detection System (VSDS) plays a key role in making trustworthy decisions about the physiological status of a soldier. We developed an algorithm to report VSDS-measured heart and respiratory rates and their associated confidence levels. Heart and respiratory rates were measured about every 2 seconds for about 4 hours, while subjects engaged in low (e.g., sitting), medium (e.g., sit-ups), and high intensity (e.g., running) activities. The mean heart and median respiratory rates are calculated every 15 seconds by an in-house developed algorithm, and associated confidence levels for each variable are estimated simultaneously using a fuzzy-logic-based algorithm. Inputs into the algorithm are features that represent two types of information; the quality of each variable, and the relationship between the variables. Faulty data points are separated from good measures by setting a threshold. When data with pre-classified faults are tested with the confidence level threshold set at 0.5, the sensitivity and specificity of the algorithm for heart rate are 91% and 97%, respectively. For respiratory rate, because of the intrinsically noisy property of the data, the sensitivity and specificity are 87% and 93%, respectively. These preliminary results demonstrate that the fuzzy logic algorithm can accurately qualify heart and respiratory rates measured by a VSDS.

Keywords: heart rate, respiratory rate, fuzzy logic, confidence levels

1 INTRODUCTION

The Warfighter Physiological Status Monitoring (WPSM) program is a multi-institute research program that is focused on the monitoring, and interpretation, of real-time physiological data from warfighters in the field. The ultimate goal of this program is to develop a suite of wearable sensors and decision-support algorithms to provide critical physiologic information to commanders and medics to aid in evaluating the health status of individuals in the field.[1,2] The Vital Signs Detection System (VSDS) tested by the U.S. Army Research Institute of Environmental Medicine (USARIEM) is an important part of the WPSM system. Heart rate (HR), respiratory rate (RR), skin temperature, body position and motion, and ballistic impact to the body, such as might occur from a bullet, are all variables that will potentially be monitored by the VSDS. This information will be used to determine the health status of warfighters, and thereby reduce fatalities incurred by medics that attempt to aid an already dead soldier. Furthermore, the information the VSDS provides may lead to reduced morbidity and mortality of casualties by facilitating appropriate medical responses.[2,3]

Heart and respiratory rates are key physiology variables that must be accurate enough to support reliable decisions about a warfighter's status, and measures of these variables were collected during a variety of activities that are part of normal military duties.[4] In this paper, we present an algorithm to qualify HR and RR as measured by a VSDS. The data specification will be described, along with the VSDS architecture. The data qualification algorithm, including data pre-processing, estimation of representative values, feature extraction, and the fuzzy-logic-based estimation of confidence level, will be presented. Then, test datasets, derived from data measured by various sensors, will be introduced and the performance of the algorithm will be documented in terms of receiver operating characteristics (ROC) summary charts.

* jaques.reifman@us.army.mil; Tel: +1 301-619-7915

2 DATA SPECIFICATION

Here, we briefly explain the WPSM system and the VSDS system (Figure 1), and how a qualification algorithm operates on the VSDS. The VSDS consists of multiple sensors that measure physiology variables (e.g., HR and RR), and a microprocessor to process the measurements and send them to a network hub. The network hub gathers information from several devices that make up the WPSM system, including the VSDS, and uploads the information via a radio network to the medic or command staff. The diagnostic system, which resides on the hub, analyzes the information, and diagnoses the warfighter's status, with the goal of optimizing physical and mental performance, injury prevention and casualty management. The whole process, from data collection, data propagation, to decision making, is integrated within the WPSM system, of which the VSDS could be considered the most vital part. The algorithm described in this paper is to be embedded in the microprocessor to calculate physiology values and their confidence levels every 15 seconds.

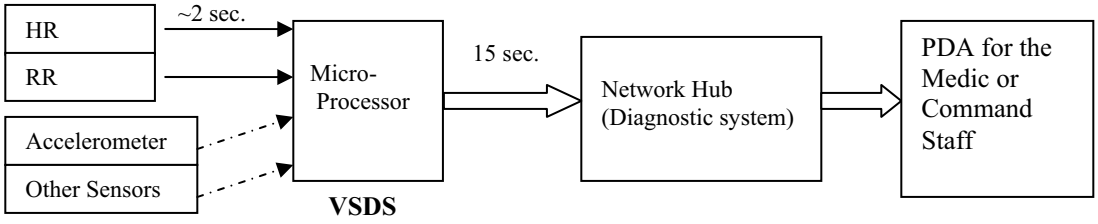

Figure 1. The architecture of the WPSM and VSDS.

The USARIEM collected datasets used to develop the algorithm consist of a total of four sensors. These included: two sensors incorporated into a VivoMetrics Lifeshirt (Ventura, CA) that simultaneously measured HR and RR, and two additional sensors that provided simultaneous, redundant measures of the HR (Schiller Cardiovit AT-6 ECG machine; Schiller Inc., Baar, Switzerland) and RR (SensorMedics Model 2900 metabolic cart; SensorMedics, Yorba Linda, CA). Eight soldiers (Mean ± SD; 21±3 yr age, 76±9 kg weight, 175±5 cm height) wore all of the sensors at the same time for approximately 4 hours each day, while engaging in low (i.e., sitting, lying, standing), medium (i.e., sit-ups, push-ups, walking), and high intensity activities (i.e., jumping jacks, running).[4] The HR and RR from the VivoMetrics Lifeshirt and data from the redundant sensors were used to design and test the data qualification algorithm. All sensors measured HR and RR at varying sample rates. The sample rate varied from 1 to 4 seconds per sample, and the average sample rate was around 2 seconds.

3 METHODOLOGY

3.1 Pre-processing

Outlier data points attributed to motion artifacts and sensor movement and malfunction were observed during measures of HR and RR. These outliers deviate greatly from true physiological values and strongly influence the calculated confidence levels. It is reasonable to remove these outliers and qualify more meaningful data. To achieve this objective, a third order median filter was used to filter the raw time series data. Since we employed a low order filter, the distortion of the dataset introduced by this nonlinear filter is minimal, but its influence on isolating outliers is dramatic and effectively eliminates them. We use the filtered time-series data in our analysis.

3.2 Estimating representative values

The representative HR and RR are values that reflect the physiological conditions within a 15-second time interval window. Based on the sampling rate of the sensors, 4 to 15 data points are reported every 15 seconds for both HR and RR. Because measures of HR are relatively more accurate than RR, every data point is equally weighted and the mean HR over the 15-second window is used to represent the HR at the end of the time window. Respiratory data, in contrast, are noisy with large absolute skewness values within a 15-second window. Therefore, we use the median value to generate the representative RR value in the 15-second window. The median is less sensitive to extreme data than the mean and this makes it a better measure for highly skewed distributions.[5]

3.3 Fuzzy logic estimation of confidence level

3.3.1 Structure of the fuzzy logic algorithm

The estimated confidence levels for HR and RR are based on two types of information: the quality of each variable, and the relationship between the variables. The method to estimate the confidence levels is, accordingly, divided into two parts (Figure 2, dotted line). The top part incorporates an evaluation of the relationships between HR and RR, termed the true relationship evaluation. A true relationship indicates that HR and RR have a physiologically reasonable relationship in terms of their ratio to each other, and also have similar trend directionality. The bottom part of the figure incorporates evaluations of HR and RR measures, termed the true measure evaluation. A true measure means that a HR or RR value is a reliable reading from a sensor based on features, such as noise and slope within the time-series data. The confidence levels are derived from both evaluation results.

A typical fuzzy logic application includes three steps. 1) A crisp number (actual value) is transformed into a membership degree (from 0 to 1) based on a membership function. 2) A fuzzy rule (IF-THEN rule) is used to reason about the data. The truth value for the premise of each rule is computed on input membership values, and applied to the conclusion part of each rule. 3) The output membership values from all rules are combined together to form a single membership value for each output variable. See Bojadziev and Bojadziev (6) for additional information about fuzzy logic theory. In this research application, inputs to the fuzzy logic estimation algorithm are features extracted from HR and RR. These features are first transformed into membership values based on predefined membership functions. The fuzzy logic engine, composed of fuzzy rules, takes in the input membership values, and calculates the output membership values. The final output membership values from the confidence level estimation blocks are the possibility of a HR or a RR value representing a true physiological condition, which we define as the confidence level. Because the confidence levels are based on an immediate 15-second time window, we call them instantaneous confidence levels.

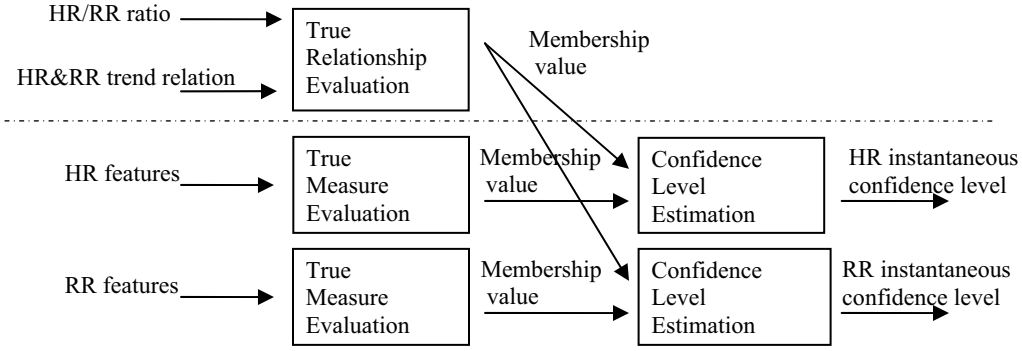

Figure 2. Structure of the fuzzy logic algorithm.

3.3.2 Feature extraction

A total of 10 features are generated for use in the HR and RR qualification algorithm. Two of the features focus on the relationship between HR and RR, while the remaining features capture characteristics of the HR and RR time-series data.

The two features representing the relationship between HR and RR are the HR/RR ratio and the HR and RR trend relation. The HR/RR ratio captures the relative conjunction between HR and RR values, when both values fall within a physiologically reasonable range. The principle for evaluating this relationship is that when HR and RR values establish an unreasonable relationship to each other, although neither of them is obviously faulty, we don't trust either one of them and assign a low membership value to the true relationship. If either measure is apparently faulty (out of the normal range), then the HR/RR ratio is set to a default value 4, disabling the ratio evaluation, and the true measure (instead of relationship) evaluation of each variable will influence the final confidence level. The HR/RR ratio is calculated as shown in (1).

$$\text{HR/RR ratio} = \begin{cases} \dfrac{\text{HR mean value}}{\text{RR median value}}; & 45 \le \text{mean HR} \le 190 \text{ beats/min} \\ & 10 \le \text{median RR} \le 70 \text{ breaths/min} \\ 4; & \text{otherwise} \end{cases} \quad (1)$$

The HR mean value is the representative HR value calculated every 15 seconds, and the RR median value is the representative RR value determined every 15 seconds.

The relationship between HR and RR trends is also evaluated. In general, it is expected that directional changes of HR and RR are correlated, taking into account time lags and a certain degree of individual manipulation of RR (e.g., 'pacing' during push-ups). When a contradiction occurs between HR and RR trends beyond a certain extent, we don't trust either of them and assign a low membership value to the true relationship. The HR and RR trend evaluation is based on HR and RR 1-minute slopes. The HR 1-minute slope is estimated by a least squares error (LSE) regression method. The RR 1-minute slope is calculated by taking the RR median value in the current 15-second window and subtracting the median value in the window of 60 to 45 seconds before the current time, and dividing by the time interval of 45 seconds.

Four features are extracted from the HR time series to evaluate the HR measures. These features include: 1) mean value, 2) 15-second slope, 3) noise, and 4) constant signal interval. The mean value is the representative HR value as described above. The 15-second slope indicates the magnitude of HR change within the current 15-second window, and it is calculated by the LSE regression within the current 15-second window. HR noise is an index of the influence of noise on the reliability of the measures of HR. To quantify noise, we first calculate a baseline HR variance (derived from high-quality measures from the Schiller Cardiovit AT-6 ECG machine) as representing physiologically true HR variance. Then, we calculate the residual variance of HR in each 15-second window as described in the following procedure, and divide the residual variance by the baseline variance to determine the magnitude of noise. This feature is calculated as follows:

1. Estimate an LSE regression line in the current 15-second window. This also defines the HR 15-second slope.
2. Create residuals by subtracting the regression line from the measurements.
3. Compute the variance of the residuals, assuming the mean is zero.
4. Normalize the residual variance by dividing by the predefined HR baseline variance to yield the noise feature.

The fourth feature, the HR constant signal interval, is the time interval during which HR measures are unchanging; it is a feature to determine whether a sensor has failed, stuck at a given value.

Similarly, four features are used to evaluate true measures of RR. These features include: 1) median value, 2) 15-second change, 3) noise, and 4) constant signal interval. The RR median value is the representative RR value as described above. The RR 15-second change is calculated by subtracting the previous median RR value from the current median RR value. The RR noise is calculated in a fashion similar to HR noise; a baseline standard deviation (SD) is calculated, based on high-quality SensorMedics data, as the true respiratory rate SD, the residual SD is calculated in the current 15-second window, and the residual SD is then divided by the baseline SD to determine the magnitude of noise.

3.3.3 Membership function design

A membership function maps a feature value to a membership value (degree of membership between 0 and 1). Only a subset of the whole input space for each feature is of concern; this is the set of all reasonable elements, called the "possible" set. Features, such as HR mean value or HR 15-second slope, have physiologically-based limits. The membership functions for the "possible" subset are accordingly defined based on these limits. Some of the limits are well known from the physiology literature, and some are not. For those not well defined, we define them, in our study, by setting a cut-off range (limit) at 0.5% of the population of features derived from high-quality data from the Schiller Cardiovit AT-6 ECG machine and the SensorMedics sensor. Data inside this range are considered possible with a degree of 1, while data outside the range are considered possible with a decreasing degree, as they get farther away from the cut-off range. In this work, we employ, among others, trapezoidal membership functions. For example, Figure 3 shows the membership function for HR mean value.

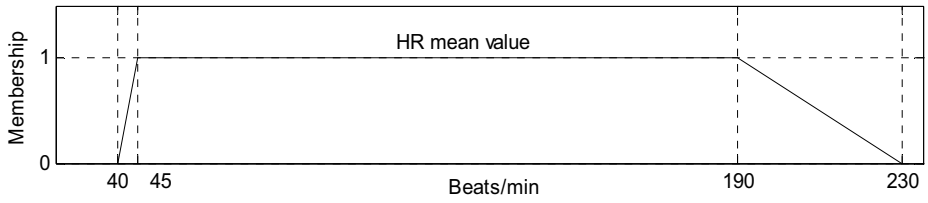

Figure 3. Membership function for HR mean value.

Other features are strongly affected by factors, such as the sensor quality, or environment-induced recording artifacts, rather than physiological limits. For these types of features, which include HR and RR noise and trend relationship, the

membership functions are derived from data distribution through a transformation based on mass assignment theory.[7] Mass assignment, a set-based probability function, builds a bridge between the probability density function and the fuzzy set membership function. The relative frequency for a feature is first computed based on its distribution. Then, a fuzzy set membership function transform is performed via mass assignment, assuming the membership of a feature element with the largest frequency is 1. Additional information about such transformations can be found in the work of Shanahan (7). The following figure shows the data-driven membership function for HR noise as an example.

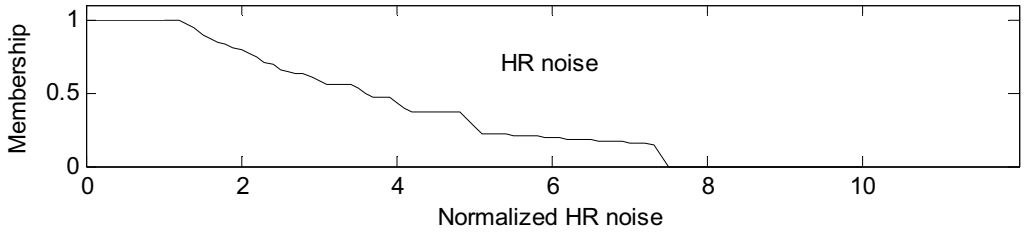

Figure 4. Membership function for HR noise.

The membership function for a constant signal interval feature is defined as decreasing linearly after 30 seconds of constant signal, with 0 membership value after 60 seconds.

In the fuzzy logic estimation structure (Figure 2), there are five output variables that indicate the relationship between HR and RR, the measures for HR and RR, and the confidence for HR and RR. The membership function for them being true is an impulse function placed at 1. This type of output membership function characterizes the fuzzy inference in this project as a Sugeno type fuzzy inference (6), which is composed of Sugeno fuzzy rules. The output membership value in each rule is weighted based on the premise of the rule. Additional information about the Sugeno type fuzzy inference can be found in Bojadziev and Bojadziev (6).

3.3.4 Fuzzy rules

The fuzzy rules are straightforward. One rule is used to evaluate the HR and RR relationship. Two rules are used to evaluate the measures of HR and RR. Two rules are used to estimate the confidence levels for HR and RR. The rules are:

1. IF the HR/RR ratio is *possible*, AND the HR and RR trend relation is *possible*, THEN the relationship is *true*.
2. IF the HR mean value is *possible*, AND the HR 15-second slope is *possible*, AND the HR noise is *possible,* AND the HR constant signal interval is *possible,* THEN the measure for HR is *true*.
3. IF the RR median value is *possible*, AND the RR 15-second change is *possible*, AND the RR noise is *possible,* AND the RR constant signal interval is *possible,* THEN the measure for RR is *true*.
4. IF the relationship is *true*, AND the measure for HR is *true*, THEN the confidence for HR is *true*.
5. IF the relationship is *true*, AND the measure for RR is *true*, THEN the confidence for RR is *true*.

The AND operation in this study is chosen as a minimum function. The membership value of confidence is assigned to the corresponding variable every 15 seconds, providing instantaneous confidence levels.

4 TEST DATASETS

We generated three datasets in order to evaluate the algorithm. One, incorporating HR and RR data measured by the VivoMetrics Lifeshirt, is used to evaluate the performance of the algorithm on raw data collected from the subjects. The others, which are derived from data measured by the Schiller Cardiovit AT-6 or the SensorMedics sensor, are used to evaluate the performance of the algorithm on simulated faults. Algorithm performance is assessed by calculating the sensitivity and specificity of fault detection. Therefore, faults need to be identified in advance. Because no information is available to identify real faults in the time-series measurements, we classified data points as presumptive faults based on comparing *redundant measures* of HR or RR from the VivoMetrics Lifeshirt and the Schiller Cardiovit AT-6 or the SensorMedics sensor. This classification results in binary fault indications (yes/no) every 15 seconds for the corresponding VivoMetrics, Schiller Cardiovit AT-6, and SensorMedics-collected data. Two hypotheses form the foundation for fault identification:

Hypothesis 1: Two time series, simultaneously measuring the same physiology variable on the same volunteer, should have the same normal distribution. This infers that differences between them should fall into a normal distribution with zero mean, and a small standard deviation. If differences do not belong to this distribution, then they more likely fall into a distribution with larger absolute values in mean and standard deviation, and at least one time series is wrong.

Hypothesis 2: If one time series in a current 15-second time window is wrong, then the one with the larger noise variance will be the one at fault. The noise variance is the variance of residuals, as previously described for the noise feature extraction. A simplifying assumption is that only a single fault may occur.

4.1 Test datasets

The VivoMetrics test dataset was used to test the ability of the qualification algorithm to identify points previously classified as faults by the redundant measurements procedure described above and illustrated in Figure 5. This dataset contains a total of 4015 HR data points with 243 of them identified as faults, and 4005 RR points with 408 identified as faults.

In contrast, the Schiller Cardiovit AT-6 and SensorMedics test datasets were used to test the robustness of the algorithm in detecting different types of simulated faults. These datasets were adjusted by modifying the faults identified by the redundant measurements procedure into acceptable values, to get a 'clean' dataset. Then, three types of simulated faults were randomly superimposed on the data (Figure 5). The three types of faults are spikes, contradictory trends, and abnormal slopes. Spikes are 2-data-points in length, 15 beats·min^{-1} in amplitude for HR, and 12 breaths·min^{-1} for RR. Their amplitude is set to the maximum 2×SD of HR or RR observed in 15-second windows from high-quality measures of the appropriate sensor. Contradictory trends are pairs of HR and RR trends with reasonable slopes, but of opposite direction (1.1(beats·min^{-1}/sec for HR, and 0.4(breaths·min^{-1}/sec for RR), lasting for 1 minute. Abnormal HR or RR slopes increase or decrease at physiologically unreasonable rates (5.5(beats·min^{-1}/sec for HR, and 1.8(breaths·min^{-1}/sec for RR) within 15-second windows. A fault of different type has different data points in length, and a given number of faults for each type were superimposed to influence the same number of 15-second windows in the datasets. There was no overlap between the different types of faults.

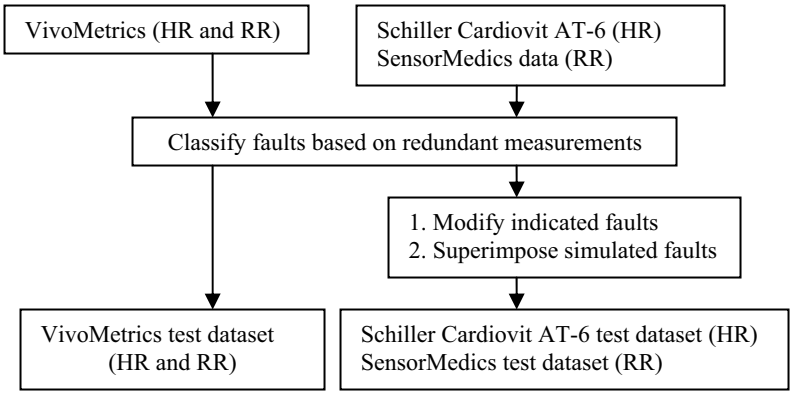

Figure 5. Generation of three test datasets.

5 EVALUATION AND RESULTS

Since the output of this algorithm is representative HR or RR values as well as their instantaneous confidence levels, a threshold for the confidence level is needed to separate faults from good measures. Representative values with confidence levels less than the threshold are assumed to be faulty; otherwise, they are assumed to be good values. Sensitivity and specificity parameters can be calculated by comparing the ability of the algorithm to identify the pre-classified or superimposed faults in the test datasets. By varying the confidence level threshold, a set of sensitivity and specificity values are derived, which are used to construct ROC curves.

5.1 Feature selection

All possible combinations of the 10 features were examined, by means of ROC curves, on the test datasets, and some representative combinations are plotted in Figures 6 and 7. Three features, the HR/RR ratio, HR mean value, and HR noise, are necessary to generate a good ROC curve based on detecting pre-identified faults in VivoMetrics HR data (Figure 6a). More features do not improve performance. The algorithm provides very similar ROC performance for VivoMetrics RR, as long as HR/RR ratio and RR noise are included (Figure 6b). However, three features, HR/RR ratio, RR median value and RR noise, are selected as the best combination because of the largest area under the ROC curve.

The robustness of the algorithm was examined using the Schiller Cardiovit AT-6 and SensorMedics test datasets, where defined faults are superimposed. The three features selected from the pre-identified HR fault test above, do not yield the best ROC performance. Addition of HR and RR trend relation, and HR 15-second slope, do improve performance (Figure 7a). For RR, as long as the three features selected from the pre-identified RR fault test above are included, different combinations with additional features yield similar ROC curves, with a marginally-greater ROC area reflecting incorporation of two more features, the HR and RR trend relation, and the RR 15-second change (Figure 7b). If we accept that sensors can malfunction and generate a constant signal, then it is reasonable to include the constant signal interval feature in the best combination. Overall, when robustness is a top priority, all features should be included.

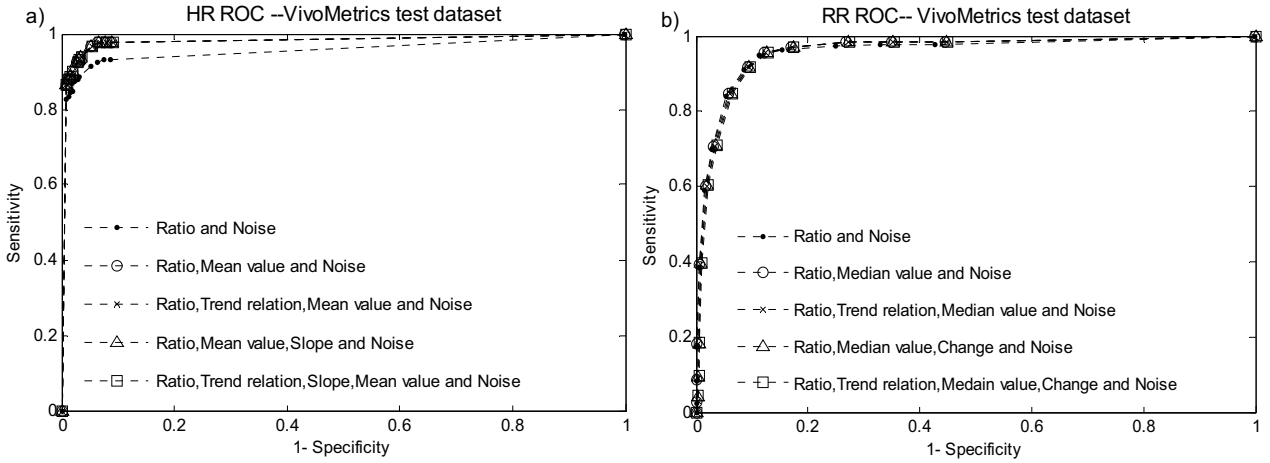

Figure 6. ROC curves derived from HR (left) and RR (right) feature combinations tested on the VivoMetrics dataset.

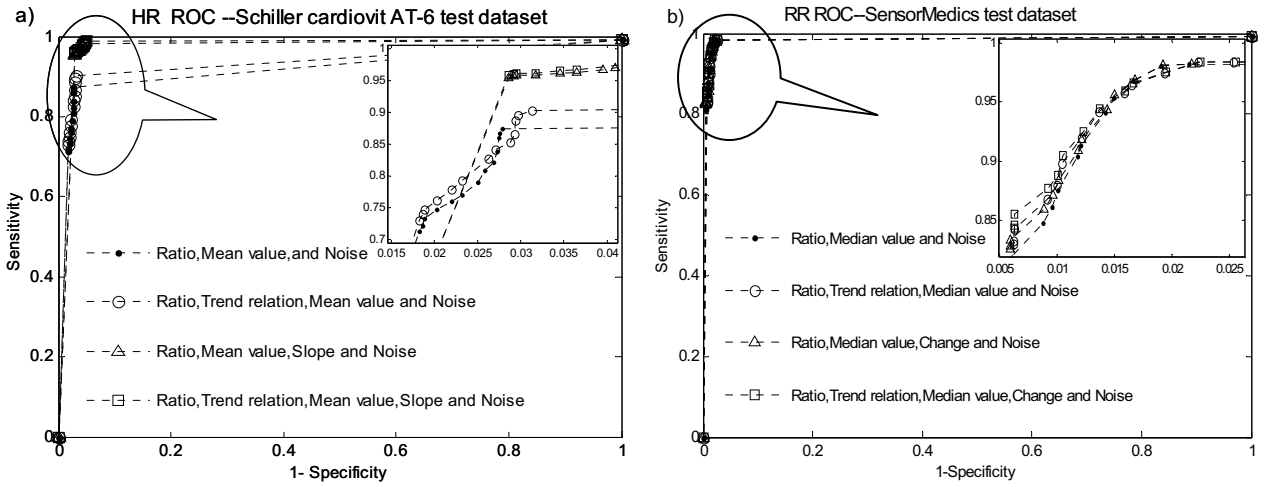

Figure 7. ROC curves derived from HR (left) and RR (right) feature combinations tested on the Schiller cardiovit AT-6 and SensorMedics test datasets.

5.2 Algorithm sensitivity and specificity

The algorithm, incorporating the full set of features, was applied to the dataset with pre-identified faults (VivoMetrics) to generate 15-second representative values for HR and RR, and their associated instantaneous confidence levels. The algorithm-identified faults were identified by varying the threshold of the instantaneous confidence level; the faults were then compared to those pre-identified by the redundant measurements method to evaluate the algorithm's sensitivity and specificity. For example, as observed in Table 1, when the confidence level threshold is set to 0.5, the sensitivity and specificity of the algorithm to identify HR faults are 91% and 97%, respectively. For RR faults, the sensitivity and specificity are 87% and 93%, respectively.

Table 1. Sensitivity and specificity of HR and RR qualification algorithm.

Threshold of Confidence level	VivoMetrics dataset(HR)		VivoMetrics dataset(RR)	
	Sensitivity	Specificity	Sensitivity	Specificity
0.9	0.9630	0.9308	0.9804	0.7976
0.8	0.9424	0.9515	0.9681	0.8560
0.7	0.9300	0.9597	0.9559	0.8907
0.6	0.9177	0.9693	0.9338	0.9041
0.5	0.9054	0.9748	0.8676	0.9330
0.4	0.9012	0.9799	0.8064	0.9475
0.3	0.8971	0.9841	0.6495	0.9728
0.2	0.8930	0.9873	0.5368	0.9833
0.1	0.8683	0.9899	0.4240	0.9894

Although the sensitivity and specificity for HR and RR qualification cannot be categorized as excellent, considering the inherent uncertainty in identifying the pre-identified faults, these results are reasonable. The RR qualification is not as good as the HR qualification, due to the intrinsically noisy property of this data.

6 CONCLUSION

The algorithm presented in this paper satisfies the requirement to determine point-by-point data quality for a VSDS application. It reports the physiology measures, and their associated confidence levels, with acceptable sensitivity and specificity. This study builds a framework for the estimation of confidence levels for sensor-measured physiology variables. The algorithm can be adjusted to incorporate additional sensors and physiology variables. However, because of the small time windows used to calculate the confidence levels, they can fluctuate strongly over small time intervals. To use the confidence levels in practice, it will be necessary that they change smoothly over time. Future work will be required to optimize the relationship between the smoothness of the confidence level results and the sensitivity and specificity of the algorithm.

REFERENCES

1. Warfighter Physiological Status Monitoring Program. <http://www.usariem.army.mil/wpsm/>.
2. R.W. Hoyt, J. Reifman, T.S. Coster, and M.J. Buller, "Combat medical informatics: present and future", *Proceedings of the AMIA 2002 Annual Symposium*, 335-339, San Antonio, Texas, 2002.
3. N. Tatbul, M.J. Buller, R.W. Hoyt, S. Mullen, and S. Zdonik, "Confidence-based Data Management for Personal Area Sensor Networks", *Proceedings of the first workshop on data management for sensor networks,* 17-23, Toronto, Canada, 2003.
4. B.A. Beidleman, W.J. Tharion, M.J. Buller, R.W. Hoyt, and B.J. Freund, "Reliability and validity of devices of a life sign detection system", Natick, MA: U.S. Army Research Institute of Environmental Medicine Technical Report, 2004.
5. D.S. Moore, *Statistics: Concepts and Controversies*, W H Freeman & Co, 4th edition, New York, 1996.
6. G. Bojadziev, M. Bojadziev, *Fuzzy Sets, Fuzzy Logic, Applications,* World Scientific Publishing Company, Singapore, 1996.
7. J.G. Shanahan, *Soft computing for knowledge discovery: introducing Cartesian granule features*, Chapter 5, Kluwer Academic Publishers, Boston MA, 2000.