journal homepage: www.intl.elsevierhealth.com/journals/cmpb

A clinical decision-making mechanism for context-aware and patient-specific remote monitoring systems using the correlations of multiple vital signs

Abdur Rahim Mohammad Forkan [a,b,*], *Ibrahim Khalil* [a,b]

[a] *School of Science (Computer Science), RMIT University, Melbourne, Victoria 3001, Australia*
[b] *National ICT Australia (NICTA), Australia*

ARTICLE INFO

Article history:
Received 21 April 2016
Received in revised form
11 August 2016
Accepted 18 October 2016

Keywords:
Context-aware monitoring
Clinical decision support system
Multi-label classification
Personalized healthcare
Vital signs

ABSTRACT

Background and objectives: In home-based context-aware monitoring patient's real-time data of multiple vital signs (e.g. heart rate, blood pressure) are continuously generated from wearable sensors. The changes in such vital parameters are highly correlated. They are also patient-centric and can be either recurrent or can fluctuate. The objective of this study is to develop an intelligent method for personalized monitoring and clinical decision support through early estimation of patient-specific vital sign values, and prediction of anomalies using the interrelation among multiple vital signs.

Methods: In this paper, multi-label classification algorithms are applied in classifier design to forecast these values and related abnormalities. We proposed a completely new approach of patient-specific vital sign prediction system using their correlations. The developed technique can guide healthcare professionals to make accurate clinical decisions. Moreover, our model can support many patients with various clinical conditions concurrently by utilizing the power of cloud computing technology. The developed method also reduces the rate of false predictions in remote monitoring centres.

Results: In the experimental settings, the statistical features and correlations of six vital signs are formulated as multi-label classification problem. Eight multi-label classification algorithms along with three fundamental machine learning algorithms are used and tested on a public dataset of 85 patients. Different multi-label classification evaluation measures such as Hamming score, F1-micro average, and accuracy are used for interpreting the prediction performance of patient-specific situation classifications. We achieved 90–95% Hamming score values across 24 classifier combinations for 85 different patients used in our experiment. The results are compared with single-label classifiers and without considering the correlations among the vitals. The comparisons show that multi-label method is the best technique for this problem domain.

Conclusions: The evaluation results reveal that multi-label classification techniques using the correlations among multiple vitals are effective ways for early estimation of future values of those vitals. In context-aware remote monitoring this process can greatly help the doctors in quick diagnostic decision making.

* *Corresponding author.* School of Science (Computer Science), RMIT University, Melbourne, Victoria 3001, Australia.
E-mail address: abdur.forkan@rmit.edu.au (A.R.M. Forkan).

http://dx.doi.org/10.1016/j.cmpb.2016.10.018
0169-2607/

1. Introduction

A decision support system (DSS) [1] in remote patient monitoring [2] is designed to assist healthcare professionals with decision making tasks such as disease prevention and diagnosis [3]. Context-awareness [4] is an essential part of clinical decision analysis [5] in patient's healthcare [6]. In a real-time monitoring, various health-related data of a patient are collected at a fixed sampling interval and analysed continuously to discover the current health situation of the patient. Using the ongoing context information of a patient the application can send proper alerts or messages to the doctors in remote monitoring centres. Some clinical decision support systems (CDSSs) [5] are capable of sensing clinically abnormal consequences in advance based on the intelligent analysis over recently observed medical data of a patient and before they are presented to the clinicians. The traditional CDSSs primarily advise medical abnormalities by assessing data of a single vital sign [7] such as heart rate (HR), blood pressure (BP), respiratory rate (RR), oxygen saturation (SPO_2), electrocardiogram (ECG) [8] and body temperature. However, they suffer from high false alerts. Most of the serious clinical anomalies occur as a result of irregularities in multiple vital signs [9] at the same time. Therefore, an important part of a context-aware remote monitoring application is the accurate and early prediction of abnormalities [10] that occur due to the changes in multiple vital signs.

The trends in different vital signs are patient-specific. For example, the mean blood pressure (MBP) value is always high for a hypertensive patient (i.e. patient with high blood pressure). The MBP value does not contain any anomaly for this patient as single vital sign unless it raises above a specific threshold. For a normal patient, the threshold of abnormality is different, i.e. the same MBP value can be abnormal for a normal patient but deemed as normal for a hypertensive patient. Hence for many health parameters, normal is actually a relative value for the patient. This can differ subject to age, disease history, user activity, gender etc. Each vital sign and related interactions must be interpreted in the context of a patient. This necessitates the need of a personalized model for clinical decision support that can discover patient-specific anomalies independently by employing a common learning technique.

For many clinical abnormalities, a single vital sign does not contain enough information for the doctors. The symptomatic patients are likely to have several abnormal vital signs [11]. For example hypotension (low blood pressure), tachycardia (elevated heart rate) and hypothermia (decrease in body temperature) can cause *sepsis*. Hypoventilation can occur when low respiratory rate is accompanied by low oxygen saturation (SPO_2) [12]. Therefore, strong positive or negative correlations in multiple vital signs contain useful information for predicting disease symptoms or anomalies. Some examples of such correlations are shown in Fig. 1. These correlations can be repetitive in some patients and vary over time in some other patients. The remote monitoring doctors must be aware of these anomalies and must incorporate them explicitly into a decision to avoid any potentially dangerous clinical situation associated with the changes in multiple vital signs. Unfortunately, no good learning model exists that can perceive such patient-specific changes early enough to assist the physicians in real-time to make proper clinical decisions. Therefore, an enhanced CDSS is required with a fast, well-trained and adaptive learning model which is less likely to make false predictions and so the doctors can take proper diagnostic actions.

In a context-aware assisted living system [13] a patient is generally equipped with different body sensors (e.g. ECG sensor, pulse oximeter, BP sensor). These sensors have capability to collect health-related vital parameters data of the patient in a continuous manner. The wireless communication ability (bluetooth, wifi, zigbee) of these sensors simplifies the data transmission process to the cloud repositories via a mobile device (e.g. smart phone or tablet) having a high speed Internet connection. The cloud has large distributed storage and high processing capability. Thus, by applying data mining techniques inside the cloud environment over continuous batches of collected data of multiple vital signs, it is possible to induce logical models which can infer the future values of those parameters. The predicted values of these vital signs and their

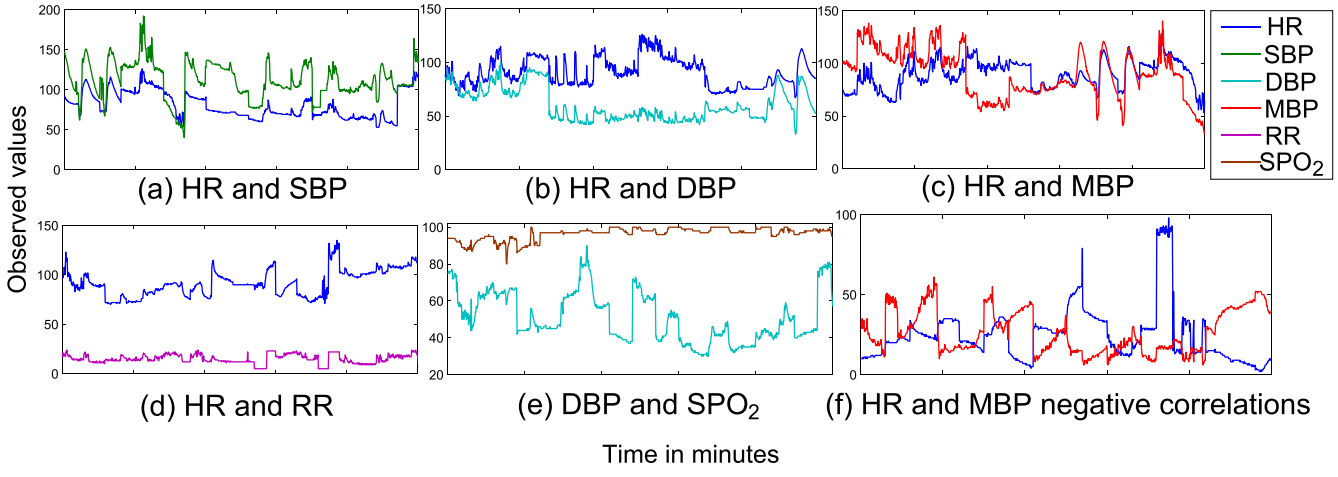

Fig. 1 – Strong correlations between a pair of vital signs. (a)–(e) examples of positive correlation, (f) example of negative correlation. Here SBP, DBP and MBP refer to Systolic, Diastolic and Mean blood pressures respectively.

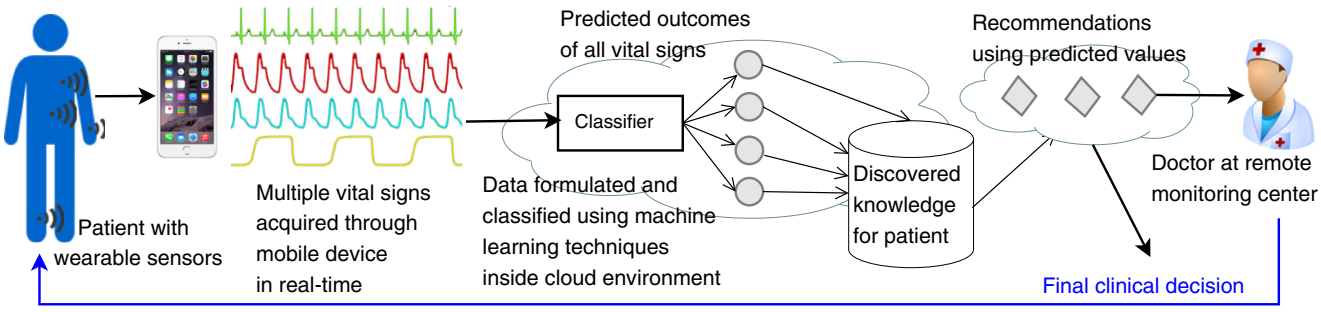

Fig. 2 – The workflow of real-time vital sign monitoring, classification, predictions and clinical decision support.

interactions are used to advise patient about the medical consequences or to recommend patient's doctors for subsequent clinical decisions. This is also known as context-aware actions by the system. This scenario is presented in Fig. 2.

This paper shows how multi-label classification (MLC) techniques [14–16] can be utilized to identify the changes in multiple vital signs quickly and efficiently. Considering the short-term summary statistics and correlations of all vital signs in parallel, a suitable feature vector for multi-label classifier is generated. Those features are then used to build machine learning models to make short-term predictions of vital sign threshold values. The overall system will reduce false alerts in the monitoring stations and will also help the early detection of clinical dangers. We formulate our model to a MLC problem because we want to achieve multiple targets (the range of all vital sign values) using the same information and at the same time.

1.1. *Motivations*

Chronic diseases caused by abnormalities in vital signs are major reasons of deaths in Australia and throughout the world. The long-term effects of the irregularities in multiple vitals can result in serious chronic illness. A clinical decision support system (CDSS) with early prediction capability can mitigate such problems. In our previous works we developed learning techniques for personalized knowledge discovery [17], future abnormalities prediction, and behavioural change detection [10] using various contexts of a patient in an assisted living environment. We have also shown the advantage of utilizing cloud platforms for such learning tasks and the process of context modelling and reasoning [13,17]. From the motivations of these works, in this study we mainly focus on vital sign correlations and utilize these contexts to produce a useful tool for the healthcare professionals by estimating patient-specific future trends of various vitals in advance. Our developed methodology will assist the doctors to make better decisions, diagnosis and treatment, resulting in improved healthcare service quality and less chronic disease-related deaths.

1.2. *Contributions*

The developed methods make several contributions to the biomedicine and healthcare related research. They are as follows.

1. We developed an intelligent clinical decision support system (CDSS) by adopting MLC techniques that can detect the upcoming trends in multiple vital signs at the same time using their correlated features. To make an accurate clinical decision a doctor should not only consider the occurrence of abnormality in each vital sign separately, but also take into account the effect of their correlations. In our approach, multi-label classifiers are extensively applied to detect such correlations in advance [18]. According to the literature review, this is the first attempt of employing MLC in vital sign predictions using their correlations. There are no previous studies which provide such an experimental comparison of state-of-the-art MLC algorithms on vital signs data.
2. We utilized the high resourceful cloud technologies for classifier training and decision support so that the system can work simultaneously for many patients. As a result, our innovative technique provides clinical decision support to a big community containing versatile patients by utilizing a common platform. At present no such system exists that can serve such a large number of patients. The proposed approach also reduces hospital load, because many patients can be monitored from home continuously. Ultimately, the adoption of our techniques can reduce the high cost of treatment.
3. The developed approach provides personalized and real-time clinical decision support by detecting patient-specific anomalies, disease symptoms and emergencies in advance. In addition, this can assist healthcare experts in diagnostic decision making with greater knowledge [19]. The accuracy of vital signs prediction is greatly improved by considering the patient-specific correlations of those vitals. Thus, this individualized system reduces the amount of false predictions in remote monitoring centres.

1.3. *Rest of the paper*

The outline of the paper is as follows. Section 2 presents a literature review. Section 3 describes the overall system design and concepts. Section 4 explains the theoretical methods and related implementations. Section 5 shows the experimental results and comparisons. Finally, the conclusion and future works are described in Section 6.

2. Related work

There are several studies related to discovering correlation patterns in multiple vitals such as heart rate, blood pressure,

respiration, O_2 saturation, and ECG. They mainly focus on finding future abnormalities in a specific vital sign [20–22]. However, very few attempts have been made to find future abnormalities in multiple vital signs. In biomedical area, multi-label classification techniques are mainly used in clinical text mining [23] and finding adverse effects of a patient in response to different drug events [24]. Some recent works show the advantage of multi-label classifier in clinical data analysis [25–27].

Furthermore, machine learning techniques are widely adopted in biomedical data analysis, healthcare and clinical abnormality predictions [28]. There are quite a number of CDSSs being proposed in the literature for different purposes. Various classifiers are developed using data mining methods for these CDSSs to aid healthcare providers in clinical decision-making process [29]. Classical data mining techniques such as support vector machine [30], artificial neural network [31], and naive Bayes [32] are used to detect clinical abnormalities and predict future behaviours in vital signs such as ECG [33], blood pressure [34], respiration etc. In summary, it is clear from the literature that the researchers have emphasized towards the direction of analysing a collection of continuous physiological data in real-time to extract the best knowledge of a patient situation and to find future behaviours. Most of these predictive systems are based on single goal prediction. But in this work, we intend to find values of multiple vitals at the same time.

3. The framework for clinical decision support system (CDSS)

The objective is to develop a CDSS that helps doctors to make decisions by estimating the future values and abnormalities in multiple vital signs. The framework for the proposed CDSS is presented in Fig. 3 and described as follows.

3.1. *Patient data collection*

We consider an assisted living system where a patient lives alone in his/her home. Several body sensors are attached to a patient's body that collect data of various vital signs continuously (i.e. per minute). The measured data are sent to a mobile device using wireless connection (e.g. wifi, bluetooth, zigbee). The mobile device then transmits these vital signs data to the cloud in small batches (e.g. every 1 hour) for processing. The cloud has vast storage and high processing capability. Therefore, it can store a large amount of incoming data from many patients [13,17]. The pre-processing and

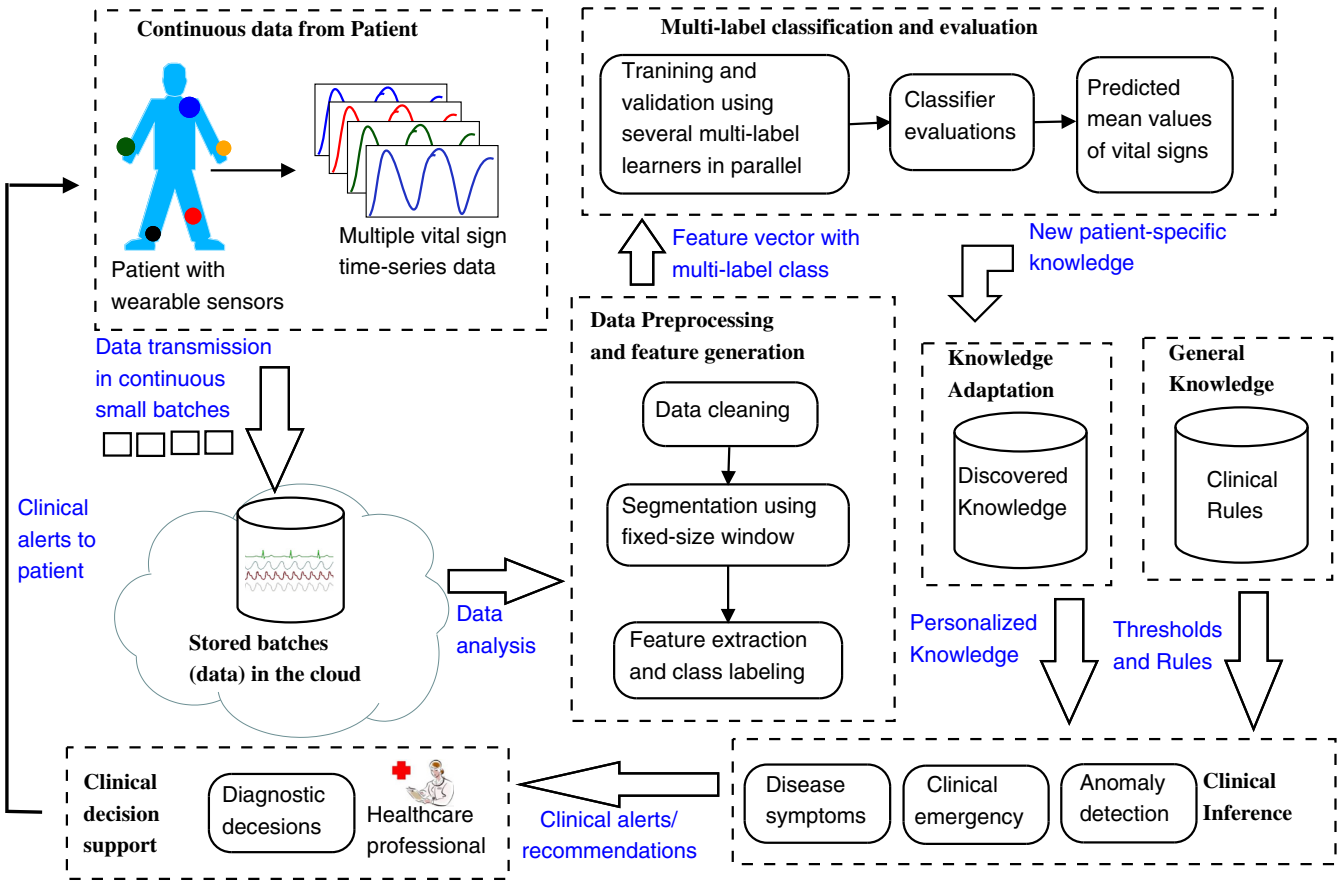

Fig. 3 – A framework of a clinical decision support system (CDSS) based on multi-label classification (MLC) using the correlations of multiple vital signs.

Table 1 – The six vital signs and their acronyms used in this study.

Bio-signal	Acronym	Unit
Heart rate	HR	Beats per minute (bpm)
Systolic blood pressure	SBP	mmHg
Diastolic blood pressure	DBP	mmHg
Mean blood pressure	MBP	mmHg
Respiratory rate	RR	Breaths per minute (bpm)
Blood O_2 saturation	SPO_2	Percentage (%)

machine learning steps are performed inside the cloud environment for parallel and fast processing.

3.2. *Correlations of vital signs for clinical inference*

In this study we considered six vital signs collected as numerical trend data from six different bio-signals listed in Table 1. These vital signs measure different physiological functions of human body and are used to monitor patient's clinical status during patient-care. The changes in vital signs indicate the potential growth of different disease symptoms or physiological response to some treatment. Therefore, vital sign values contain useful information for clinicians to make decisions in a remote monitoring system.

Vital signs can be abnormal in virtually any disease process. Patients who are actually ill are likely to have several abnormal vital signs. Certain patterns of abnormalities develop through a strong correlation between multiple vitals. For example, a strong correlation between Hypertension (elevated SBP and DBP) and Bradycardia (low HR) can create serious clinical emergency known as *Cushing reflex*. Fever (increasing body temperature) is accompanied by tachycardia (increasing heart rate) with the general rule of thumb that the heart rate will increase by 10 beats per minute for every 1 °C increase in body temperature. The progression of such diseases can be inferred prior to emergency situation using the proposed learning model.

3.3. *Data preprocessing and feature generation*

In the pre-processing stage, noisy data are cleaned and then segmented into fixed-sized sliding window for calculating statistical and correlated features and corresponding class labels. The feature extraction process from raw data is described below.

Given the continuous value of p vital signs sampled per minute as time series $V(t)=[V_1(t), V_2(t) \ldots V_p(t)]$. $V(T_s \rightarrow T_e)$ is a batch (e.g. 24 hours) of continuous data that starts at time T_s and ends at T_e. Data between time T_s and T_e are divided into three time slices: observation time t_o, lead or forecast time t_l and prediction time t_p. To generate o number of samples from this batch t_o is divided into o windows of size w. That is, $t_o = o \times w$. A feature vector, f_j is generated from each of o windows where $1 \le j \le o$. The corresponding multi-label class vector c_j of size p is obtained from the mean values of each p vital from the prediction window of same size w located at the time after l times of w from the point where the j-th observation window t_{o_j} ends. This formulation is shown in Fig. 4.

We have numerical trend data of p vitals between time $t - t_o$ and t, the goal is to estimate future values of those p vitals at time t. That is, we want to predict an estimated average value of each vital sign between time $t + t_l$ and $t + t_l + t_p$. The prediction is not a continuous value, but instead is a vector of size p where each value is defined in one of 15 different ordinal classes. Table 2 summarizes what vital sign value for each class label stands for.

As an example, if w is 10 minutes and l is 6 then t_l is $(6 \times 10 = 60$ minutes) 1 hour. $t_p = w$, that is 10 minutes. If a batch size $(T_e - T_s)$ is 24 hours (1440 in minutes) then the size of t_o can be at most (t_o = 1440-60-10) 1370 minutes. Thus, here the number of observations o is 137 (as $o = t_o \div w$). Therefore, using $w = 10$ and $l = 6$ we can extract a total of 137 examples from 24 hour data. If we use two-thirds data for model training then we will have 92 samples for training and 45 samples for model validation.

3.4. *Multi-label classification (MLC) and evaluation*

When an adequate number of instances are available for training, they are sent to the classification engine where different MLC algorithms are applied in parallel using cloud platforms. Finally, all results are compared using various evaluation measures to evaluate their performance and the best model is determined. The model is then used to classify new unknown instances of future batches. Thus, the model can predict the normalized class value (as described in Table 2) of all vitals in advance.

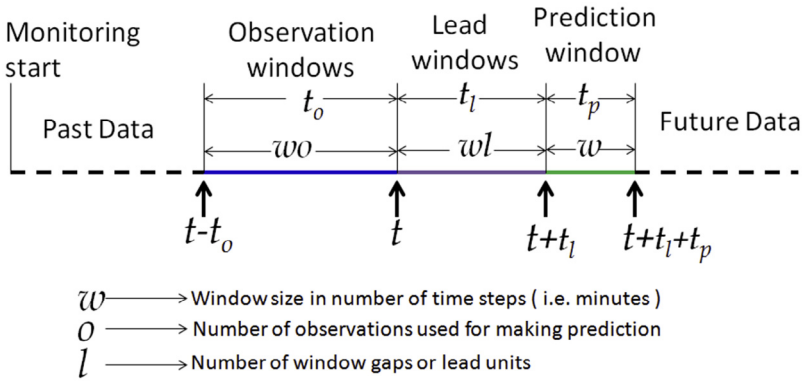

Fig. 4 – Observation, lead and prediction window times.

Table 2 – Normalized class value for different ranges of six vital signs.

Class value	HR	SBP	DBP	MBP	RR	SPO_2
–7	≤30	<50	<30	<40	<5	<83
–6	>30 and ≤40	≥50 and <60	≥30 and <35	≥40 and <50	≥5 and <6	≥83 and <85
–5	>40 and ≤45	≥60 and <70	≥35 and <40	≥50 and <55	≥6 and <7	≥85 and <87
–4	>45 and ≤50	≥70 and <80	≥40 and <45	≥55 and <60	≥7 and <8	≥87 and <89
–3	>50 and ≤55	≥80 and <90	≥45 and <50	≥60 and <65	≥8 and <9	≥89 and <91
–2	>55 and ≤60	≥90 and <100	≥50 and <55	≥65 and <70	≥9 and <10	≥91 and <93
–1	>60 and ≤70	≥100 and <110	≥55 and <60	≥70 and <80	≥10 and <12	≥93 and <95
0	>70 and ≤90	≥110 and <120	≥60 and <80	≥80 and <100	≥12 and ≤15	≥95 and ≤100
1	>90 and ≤100	≥120 and <130	≥80 and <90	≥100 and <105	>15 and ≤17	N/A
2	>100 and ≤120	≥130 and <140	≥90 and <100	≥105 and <110	>17 and ≤19	N/A
3	>120 and ≤140	≥140 and <150	≥100 and <105	≥110 and <115	>19 and ≤21	N/A
4	>140 and ≤160	≥150 and <170	≥105 and <110	≥115 and <120	>21 and ≤23	N/A
5	>160 and ≤180	≥170 and <190	≥110 and <115	≥120 and <130	>23 and ≤24	N/A
6	>180 and ≤200	≥190 and <210	≥115 and <120	≥130 and <140	>24 and ≤25	N/A
7	>200	≥210	≥120	≥140	>25	N/A

3.5. *Knowledge adaptation and decision support*

In our design, the learned model for a specific patient is also continuously adapted for new batch of data to ensure that the model is up-to-date to detect future behaviours using most recent information. The learned personalized knowledge is stored and utilized along with different clinical rules and correlations to detect anomalous situation.

3.6. *Clinical decision support*

The abnormal mean values in one or multiple vitals indicate different types of abnormalities, disease symptoms or clinical emergencies. The healthcare professionals are notified on occurrence of such anomalies. They can then further investigate the data and make diagnostic decisions. An appropriate clinical alert is sent to the patient based on the decision so that the patient gets notified before any potential danger.

4. Methodology and implementation

The process of model development and implementation is described as follows.

4.1. *Data sources*

We have used vital signs data from MIMIC [35] and MIMIC-II [36] numerical datasets of MIT Physiobank archive [37]. Data in the MIMIC/MIMIC-II database contain multi-parameter recordings, which are obtained from both bedside monitor and the medical records of the patients who stayed in ICUs. We preferred to use this dataset because it fulfilled the criteria to evaluate our implementations. Moreover, there is no public dataset available which contains multiple vital signs data of various home-monitoring patients with different correlations for a long period of time. Here we considered that home-monitoring data have similar nature when they are collected in a controlled environment and in supervision of a nurse.

MIMIC and MIMIC-II (version 2) datasets contain records of various physiological signals of about 4000 adult patients. Most of the data are sampled per minute, but some are sampled per second and they are converted to per minute sampling by taking the mean value in a minute. The data containing clean values of the six vital signs for more than 24 hours are used. Finally, 30 from MIMIC and 55 from MIMIC-II, in total 85 patient records, are used for evaluations. Patients involved in this study have a wide range of clinical problems such as sepsis, respiratory failure, congestive heart failure, pulmonary edema, myocardial infarction, cardiogenic shock and acute hypotension. Most of these clinical cases occur due to abnormalities in multiple vital signs at the same time.

4.2. *Data cleaning and preprocessing*

A few pre-processing steps are required to improve the data quality before computing the features. This is practical for real-world data. Even for a monitored patient at home, data can contain noise and outliers. The noisy or missing data occur due to sensor errors, disconnections, equipment changes, network connection interruptions and many other reasons. If all vital signs data are missing for a long period, they are considered as non-recoverable due to network interruptions or sensor errors and thus deleted. The case where one or more vital signs data are missing while clean values of others are available is considered as recoverable and imputed using median-pass [38] and k-nearest neighbour filter [39].

4.3. *Segmentation*

For each patient the model starts learning using the batch of first 24 hour data. The first batch is used as bootstrap learning for building the initial model. The incoming data are classified using this model and after every 2 hours the model is refreshed using the most recent 24-hour batch data to ensure adaptability. This process considers that the upcoming future values are vastly dependent on most recent past values (in our case 24 hours) which is also true for real life data. Thus, this incremental batch based segmentation and learning process is able to handle potentially infinite amount of data.

Here each vital sign signal is sampled per minute. Therefore, a batch of 24-hour data contains 1440 minute samples

of six vital signs. As described previously, $T_e - T_s = 1440$ minutes. According to Fig. 4 these 1440 minute data are divided into 3 time slices (t_o, t_l and t_p). The values of l and w are varied and corresponding feature vector and class labels are generated. For training we have used 66%-split, that is, the first two-thirds data are used for model training and the last one-third data are used for model validation.

4.4. *Feature extractions and class labelling*

The filtered numerical trend data of each vital sign (as in Table 1) are used to calculate the features. For an observation window size w statistical features of each of the six signals such as mean, median, minimum, maximum, standard deviation, skewness, kurtosis, percentiles and inter-percentile ranges are calculated [40]. Minimum and maximum contain the information about extreme values, mean and median represent the magnitude of each vital sign, standard deviation describes the variability and skewness is the third moment of amplitude distribution. Moreover, different percentiles (5th, 10th, 50th, 90th and 95th), IPR (inter-percentile range, which is the difference between 2 percentile values such as 95th and 5th), and kurtosis are other important statistical measures that provide a snapshot of the relation between vital parameters that vary over time.

We also calculated the sequential trend of each vital which is the number of increasing and decreasing values within a sliding window w [41]. Another feature is the regression slope which is calculated by fitting a linear least square regression line to the small curve of w-sized window. The slope determines sharp changes in a vital sign which can indicate a dangerous situation. Moreover, we have measured the pairwise correlation coefficients of six signals that contain the information about actual correlation between a pair of vitals. All these measures are simple and very easy to implement. Overall, a total of 123 features are computed from six vital signs.

To find the corresponding multi-label classes of each feature set, we measured the mean values of w-sized window located after $l \times w$ minutes from the observation window. Afterwards, the measured mean values of each of the six vital signs are labelled to the normalized value from –7 to 7 (for SPO_2 it is from –7 to 0) as described in Table 2. Therefore, each instance of a classifier is a set of 123 features and corresponding set of 6 classes.

The overall training examples are generated as a batch of 24 hour data. Therefore, the system generates an incremental batch of 24-hour training samples and they are fed to different multi-label classifiers.

4.5. *Multi-label classification*

Multi-label classification problem corresponds to searching for a function h that assigns to each instance. The goal is to minimize the expected prediction loss with respect to a specific loss function. An instance is represented by a vector of m features or attributes $\mathbf{X} = (X_1, X_2, X_3, \ldots, X_m)$ and a vector of d output labels $\mathbf{Y} = (Y_1, Y_2, Y_3, \ldots, Y_d)$. The h function should assign to each instance $\mathbf{X}$ that finds the most likely combination of class labels, that is, $\arg\max_{Y_1, Y_2, \ldots, Y_d} P(y_1 = Y_1, Y_2, \ldots, Y_d | \mathbf{X})$. The m-dimensional input space is represented by $\Omega_{\mathbf{X}} = \prod_{i=1}^{m} \Omega_{X_i}$. In our case all the features are numeric. So, $\Omega_{X_i} \subseteq \mathbb{R}$. A multi-label dataset with N training samples is represented by $D = \{(\mathbf{x}^{(1)}, y^{(1)}), \ldots \mathbf{x}^{(N)}, y^{(N)})\}$ where $x^{(i)} \epsilon \Omega_{\mathbf{X}}$ and $y^{(i)} \subseteq Y$ for all $i \epsilon \{1, \ldots, N\}$. As stated above, in our formulation, m is 123 and d is 6. The learning task of obtaining h is represented by $h: \Omega_{X_1} \times \Omega_{X_2} \times \ldots \times \Omega_{X_m} \rightarrow y \subseteq \mathbf{Y}$.

Multi-label classification problem can be categorized in two steps, (1) Problem transformation method and (2) Algorithm adaptation method [14]. The first one transforms the learning task into one or more single label classification tasks. They are algorithm independent. The second method extends existing machine learning algorithms (e.g. decision tree, support vector machine [42], neural network, k-nearest neighbour [43]) to handle multi-label data directly. There are various problem transformation methods such as binary relevance (BR)-based methods, label combination (LC)-based methods, pairwise methods, ranking methods via single label learning and ensemble methods.

Many variants of these problem transformation methods are described in the literature [14]. We evaluated our data with most of these methods and picked the best eight methods based on the evaluation measures and training time to interpret our results. The eight methods are binary relevance (BR), classifier chain (CC), Bayesian CC (BCC), Monte Carlo optimization of CC (MCC), Classifier Trellis (CT), Fourclass Pairwise (FW), Pruned Sets (PS) and Ranking + Threshold (RT). The classification algorithms are performed in MEKA software [44] which is a Multi-label extension of popular data mining software WEKA [45]. A short description of these algorithms is presented below.

4.5.1. *Binary relevance (BR)*

The BR method transforms the original dataset into c datasets (where c is total number of classes in a dataset), one for each class label, where each dataset includes all the instances of the original dataset and learns one binary classifier for each label independent of the rest of labels. To classify a new instance, BR outputs the union of the labels that are predicted by the c classifiers. It does not consider label relationship [46]. In our case, we have 6 classes with 15 different labels. Thus, the class labels of our problem are easily transferable to BR method.

4.5.2. *Classifier chain (CC)*

The CC method contains classifiers which are linked along a chain, where each classifier handles the binary relevance problem associated with each label. It creates a chain of classifiers $C_1, C_2, \ldots, C_L$, where L is the total number of labels. To classify a new instance, CC starts from C_1 and runs down along the chain. Each classifier determines the probability of being classified into $L_1, L_2, \ldots, L_L$. The chain method passes label information between classifiers to take into account label correlation. It combines the advantages of binary relevance and label dependency. The CC method is based on the decomposition of the conditional probability of the class vector $\mathbf{Y}$ using the product rule of probability. $p(\mathbf{Y}|\mathbf{X}) = p(Y_1|\mathbf{X}) \prod_{i=2}^{L} (Y_i|Y_1, \ldots, Y_i, \mathbf{X})$ [46].

A variant of CC is Bayesian Classifier chain (BCC). The objective of BCC is to find a joint distribution of the classes $\mathbf{Y} = Y_1, Y_2, \ldots, Y_d$ for a given feature set $\mathbf{X} = (X_1, X_2, X_3, \ldots, X_m)$, such that $p(\mathbf{Y}|\mathbf{X}) = \prod_{i=1}^{n} (Y_i|pa(Y_i)|\mathbf{X})$, where $pa(Y_i)$ represents the parents of class Y_i. A Bayesian Network (BN) is induced to

represent the joint distribution. In this setting, a classifier chain can be constructed by inducing first the classifiers that do not depend on any other class and then proceed with their predecessors. Another variant of CC is Monte Carlo classifier chain (MCC) that uses Monte Carlo search method for estimating joint probability. Another classifier chain method that approximates the probability $p(\mathbf{Y}|\mathbf{X})$ by maintaining a lattice structure graph is known as Classifier Trellis (CT) [47]. As the class labels of our formulated feature vector have high correlations most of the variants of classifier chains produced better predictions.

4.5.3. *Pruned sets (PS)*

Pruned Sets (PS) is a variant of Label Power Set or Label Combination (LC) method. LC takes into account label dependency. Label power set considers each unique occurrence of a set of labels as one class [48]. In LC each different set of labels becomes a different class in a new single-label classification task. Unfortunately, basic LC must discard any new training examples that have a labelset combination that is not one of the class labels. So, this does not suit incremental learning. Pruned sets (PS) uses pruning to focus on core combinations. It is much better suited to this incremental learning context. PS drastically reduces the number of class-labels in the transformation by pruning all examples with infrequent labelsets. It then additionally subsamples the infrequent labelsets for frequent ones so it can reintroduce examples into the data without reintroducing new labelset combinations. PS thus retains (and often improves upon) the predictive power of LC, while being up to an order of magnitude faster.

4.5.4. *Fourclass pairwise (FW) and ranking + threshold (RT)*

These are ranking-based methods. It learns $\frac{d}{d-1}$ binary models, one for each pair of labels. Each model is trained based on examples that are annotated by at least one of the labels but not both. The FW model compares each class pair Y_j, Y_k to one of the four classes 00,01,10,11 with threshold. The RT method duplicates multi-label examples into examples with one label each, trains a multi-class classifier, and uses a threshold to reconstitute a multi-label classification.

4.5.5. *Algorithm adaptation method*

In our evaluations we used three popular machine learning algorithms for result analysis. They are J48 Decision Tree (J48), Random Tree (RT) and Sequential Minimal Optimization (SMO, a simplified version of support vector machine). We have also tested some other algorithms such as Naive Bayes (NB) and Multi Layer perception (MLP) but the outcomes of those adaptation methods were not satisfactory and so not included in this study for result interpretation.

4.5.6. *Evaluation measure*

The evaluation methods for multi-label classifications are different from those used for single-label classifications. The evaluation methods can be divided into example-based measures, label-based measures, and ranking-based measures [46]. Here we have used four measures for the performance evaluations of our experiments. These are described below.

1. Hamming score: is the accuracy for each label (class) to correctly be predicted, averaged across all labels. This is the opposite of Hamming Loss which reports how many times on average the relevance of an example to a class label is incorrectly predicted. Hamming loss takes into account the prediction error (an incorrect label is predicted) and the missing error (a relevant label is not predicted), normalized over total number of classes and total number of examples.
2. Accuracy: is the ratio between the correct labels to the total number of labels for each instance, averaged across all instances.
3. F1 micro average: is the harmonic mean between precision and recall, where the recall refers to the percentage of relevant labels that are predicted, and precision refers to the percentage of predicted labels that are relevant.
4. Exact match: is the accuracy of each example where all label relevance must match exactly for an example to be correct.

4.6. *Abnormality prediction*

The classifier model we developed can detect the output label of multiple vitals at same time. In our formulation the normal ranges (according to general medical rule) of all vital signs have class label 0. We can consider that values near 0 (that is 1 or −1) are nearly normal. Other than this, high or low values in class labels are considered as abnormal, and high or low values in multiple vitals and very high or low value in one or more vitals can be treated as dangerous situation. Thus, our system has the mechanism for early prediction of such abnormal conditions well ahead of time and sends proper alert to the doctors.

4.7. *Incremental and patient-specific learning*

The best classifier is picked in terms of Hamming score, F1 measure and model building time. Once a classifier model is selected for a patient using the bootstrap batch, the new incoming instances are classified using only this classifier. If the performance of classification for new instances falls below an expected threshold value (in our case Hamming score <90%) before the model is refreshed, the classifier is re-trained using most recent batch (past 24-hour data from current) to maintain the desired performance level. Otherwise, the model is refreshed every 2 hours. It is expected that the nature of correlations of vitals will not be the same for a long duration for a specific patient. Thus, this incremental learning process is adaptive and keeps the knowledge of the model up to date with new instances. Therefore, our model can be easily used in patient-specific abnormality prediction as it maintains patient-specific knowledge.

5. Experimental evaluations and results

To evaluate the performance of the proposed approach different MLC algorithms are applied over our experimental data. We have performed the testing for multiple patients individually and concurrently inside cloud environment. We have used multiple m3.2xlarge Elastic Compute Cloud (EC2) instances

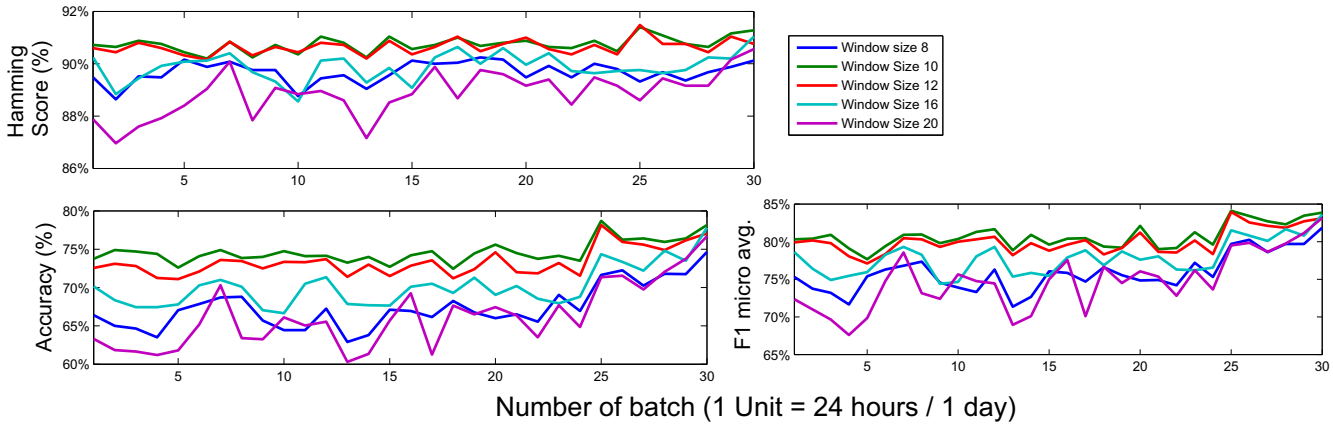

Fig. 5 – Selection of optimal window size for calculating features. Three evaluation measures from 30 day monitoring data of patient a40493n using different window sizes. Here the multi-label classifier is BR and learning algorithm is Random Tree. Window size 10 achieved the best Hamming score, accuracy and F1 value compared to other window sizes.

provided by Amazon Web Service (AWS) for data mining process using MEKA software tool. We have also used Amazon Simple Storage Service (S3) for storing patient data as incremental batch. All the results are presented below.

5.1. *Selection of window size*

As described in the implementation we have used fixed-length sliding window for calculating the feature vectors of all vital signs. Thus, the selection of optimal window size w is important in our analysis. A short window size (e.g. 4 minutes) does not contain enough information for calculating statistics and correlations. On the other hand, a long window size (e.g. 30 minutes) is also not suitable as there is high possibility that physiological signals can fluctuate in this long duration.

We have evaluated window size w as 8, 10, 12, 16, 20 minutes for multiple patients using BR classifier. The values are selected in such a way that they can be divided by 60 minutes (1 hour). The evaluations of window size for patient a40493n are shown in Fig. 5.

For selecting the window size we performed a statistical significance test using 8 problem transformation methods and 3 algorithm adaptation methods (in total 24 combinations) across 85 patients with different window sizes. We performed a paired t-test for each window combination and at 5% significance level using Hamming score. The obtained average p-values across 24 algorithms are presented in Table 3. The results show that using Hamming score of different classifiers window 10 is significantly better than the others. Therefore, we used window size 10 for further evaluations.

Table 3 – Statistical tests results (paired t-test at 5 percent significance level) for different window sizes. Here only mean p-values are presented across 24 algorithms and 85 patients.

Window size	10	12	16	20
8	0.02	0.36	0.24	0.43
10		0.005	0.02	0.03
12			0.04	0.05
16				0.25

5.2. *Selection of forecast period*

If l is unit for lead time and w is the window size then the forecast period (or lead period) $t_l = l \times w$. For example, for $w = 10$ and $l = 6$ lead period is *60 minutes* or *1 hour*. That is, the prediction window is located after 1 hour and our system will predict the mean value of 10 minute prediction window 1 hour before.

We varied the l value from 1 to 7 (i.e., forecast window from 10 to 70 minutes). The shorter forecast window has better prediction accuracy. Normally accuracy will degrade when lead time becomes longer. 1 hour preceding prediction window is considered fair enough for a doctor to make clinical decisions. Therefore, we have used l value up to 7 for our experiment. Once again, this evaluation was performed for all 85 patients with different classifiers. The evaluation for patient a40493n using lead time 1–7 is presented in Fig. 6. For all cases the achieved Hamming score was satisfactory (≥85%).

5.3. *Classifier performance evaluation*

As stated above, to evaluate the performance of all classifier combinations we have tested it with 8 problem transformation methods and 3 algorithm adaptation methods (in total 24 combinations) for 85 patients.

5.3.1. *Performance for a single patient*

Table 4 shows the observed result for a single patient on random 24 hour data where first two-thirds data are used for training and the rest one-third for validation. Different classifier combinations perform well in different patients. As in Table 4, we can see for all combinations that we have Hamming >86% (at least), accuracy >60%, exact match >31% and F1 score >70%. For this particular patient SMO has better accuracy than J48 and Random Tree for most of the problem transformation methods. Here, we found that FW and Random Tree

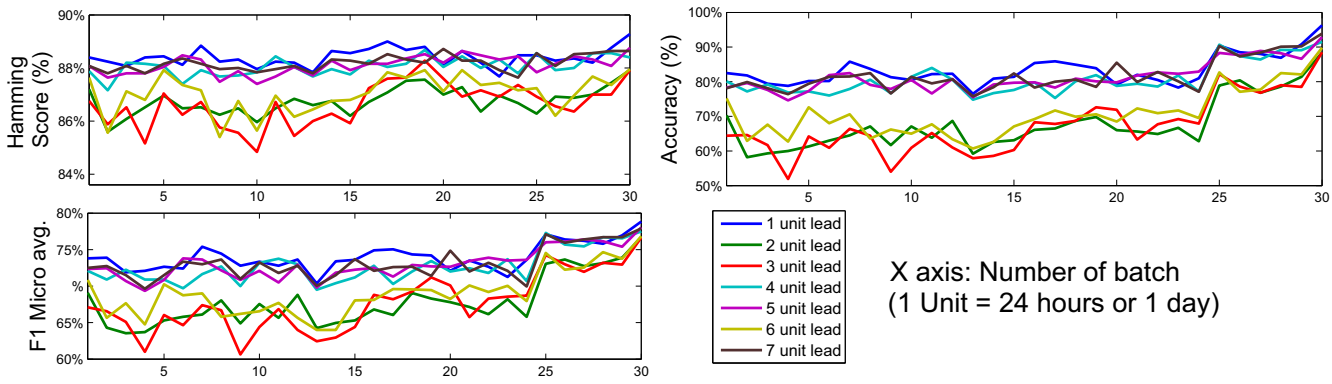

Fig. 6 – Performance for single patient for different lead time using $w = 10$. Three evaluation measures from 30 day monitoring data of patient a40493n using l 1–7. Here the multi-label classifier is BR and learning algorithm is Random Tree. For all 7 cases the classification has good Hamming score value (≥85%).

Table 4 – Performance of different multi-label classifiers using the features of random 24 hour data of patient a40215n. Here $w = 10$ and $l = 7$.

	Hamming score (%)			Accuracy (%)			Exact match (%)			F1 micro average (%)		
	J48	RT	SMO	J48	RT	SMO	J48	RT	SMO	J48	RT	SMO
BR	92.4	92.9	93.4	71	73.4	72	34.7	55.1	26.5	79.3	80.5	81.4
BCC	92.3	92.6	93.4	70.8	72.3	72.4	38.8	55.1	32.7	79.1	79.4	81.1
CC	91.8	92.4	93.6	68.4	73.4	74.9	30.6	55.1	46.9	77.4	79.5	82.3
MCC	89.7	92.4	93.6	63.5	73.4	74.9	22.4	55.1	46.9	73.7	79.5	82.3
FW	93.1	94.1	93.8	74.3	77.7	74.8	53.1	57.1	36.7	81.1	84	82.8
CT	92.9	93.3	93.5	72.7	74.4	71.9	44.9	55.1	26.5	80.3	81.2	81.3
PS	89.1	91.3	93.7	60.5	70.5	77	34.7	57.1	61.2	70.1	76.2	82.7
RT	86.3	92	93.8	68.3	71.8	75.6	31.2	57.1	46.9	72.8	77.9	83

combination produces the best result which has Hamming score of 94.1% and accuracy of 77.7% (shown in bold face).

According to our problem formulation the class labels are mean values of vital sign within small ranges. Therefore, it is very common for a classifier to predict something that is not exactly accurate but near the class label (e.g. HR class predicted as 2 where original class label is 1). As the metric Hamming loss actually represents the average distance from the correct class labels, a smaller class deviation that is a small value of Hamming loss (or in reverse large value of Hamming score) indicates a better classifier performance. Fig. 7 shows the boxplot of the Hamming score values across all 24 classifier combinations for 85 different patients used in our experiment. We can see that most of the mean values are in the range of 90–95%.

Moreover, our class distribution is uneven. Thus, the accuracy metric does not indicate the true performance of the classifiers. That's why we also measure F1 score because it represents a classifier performance in case of an uneven class distribution in terms of precision and recall performance metrics. However, we also kept the accuracy and exact match measure for our evaluations to interpret the performance of

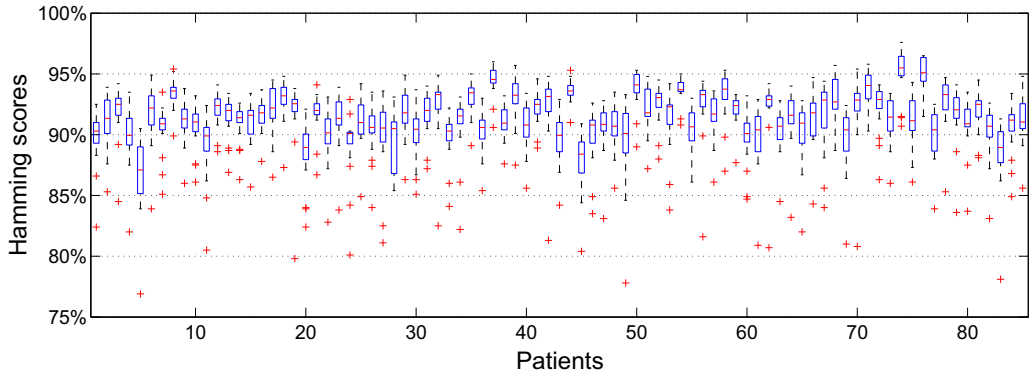

Fig. 7 – Hamming scores of 85 patients separated using the statistics from 24 classifier combinations.

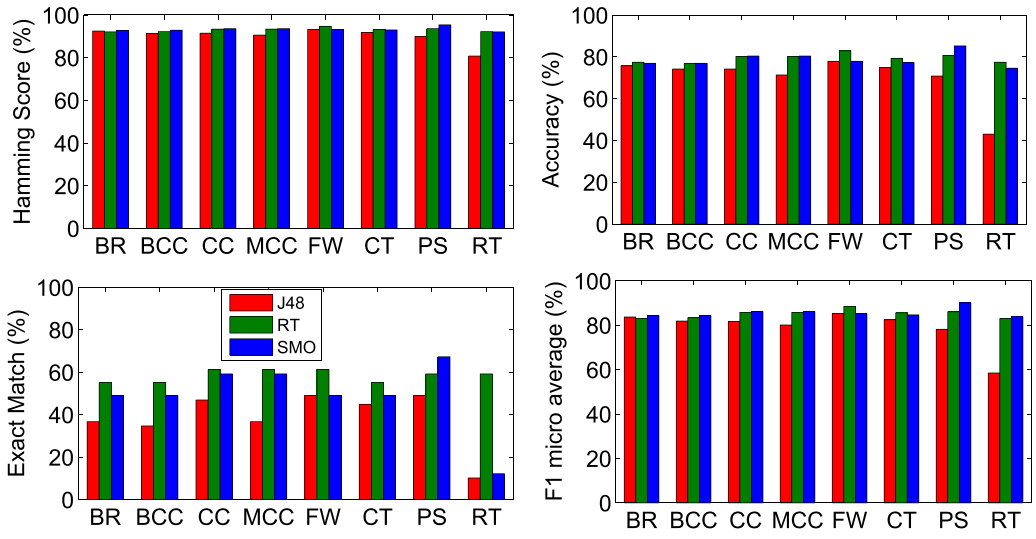

Fig. 8 – Performance measure of different classifier combination using random 24 hour data of patient a42141n.

the classifier with Hamming score and F1 score. But here we mostly decide depending on Hamming score as this is the best accuracy measure to describe our problem.

Fig. 8 shows graphical interpretation of the classification performance of another patient in terms of 4 measures we considered. Once again, we find that for most of the combinations we have Hamming score >90%. For this patient also FW and Random Tree combination achieved the best Hamming score but SMO performs well as base learner in most transformation methods. In the analysis, other than FW we can see that PS and CT also have better performance.

5.3.2. *Performance for multiple patients*

To understand the overall average performance of all classifiers for all patients we measured the average Hamming score of 85 patients for different classifier combinations using random 24 hour data. The results are presented in Fig. 9. We found that the mean Hamming score is still over 90%. The best average mean is obtained for FW and Random Tree combination. To validate the statistical significance of this statement we performed a paired t-test that compares the performance score (e.g. Hamming scores) of each pair of algorithms across 85 patients at 5% significance level [49]. The statistical test compared the best results with the results of the other learning methods, in a two-by-two basis. Due to the space limit of the paper, we only present the test based on the Hamming score which has a normal distribution across 85 patients.

The null hypothesis is that the algorithms being compared are equally good and the alternative hypothesis is that FW and Random Tree (RT) combination is better than others. The combination of algorithm number is shown in Table 5 and the p-values of 24 combinations are presented in Table 6.

Note that in Table 6 small values for the p-values indicate that a method is significantly better than the other and high p-value (>5%) indicates there are no statistically significant differences in the results accomplished by a pair. We can observe that for algorithm (14) (FW and Random Tree) all p-values are 0 that is significantly small. This justifies our previous statement that the best results are obtained from this combination. Moreover, we can see that for many cases there are

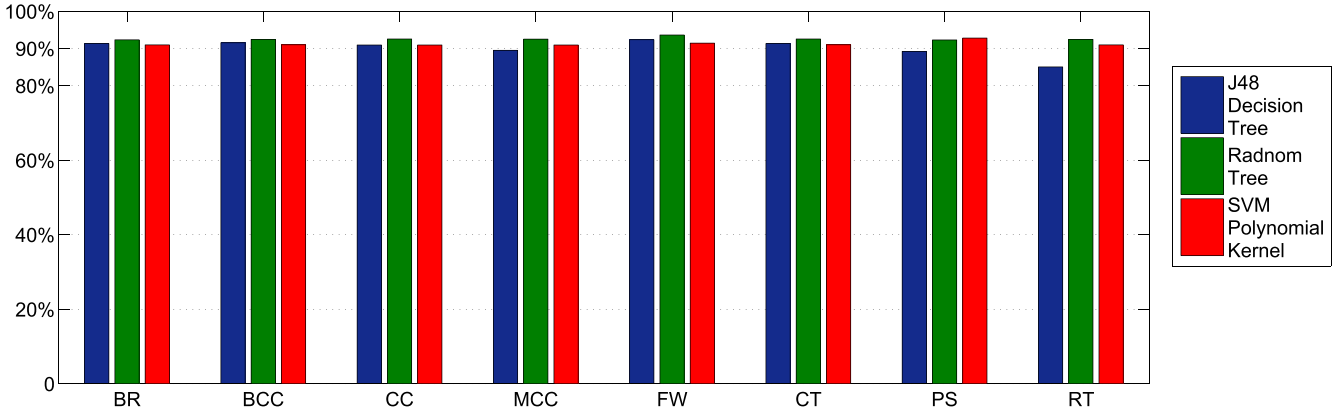

Fig. 9 – Average performance of different combinations of multi-label classifiers and learning algorithms in terms of Hamming score. These average values are calculated from random 24 hour data of each of the 85 patients using $w = 10$ and $l = 7$. Here, for most cases Hamming score is above 90%. The performance of Random Tree and SMO is better than J48.

Table 5 – The numbering map of 24 combinations of 8 problem transformations and 3 algorithm adaptation methods.

	J48	RT	SMO
BR	1	2	3
BCC	4	5	6
CC	7	8	9
MCC	10	11	12
FW	13	14	15
CT	16	17	18
PS	19	20	21
RT	22	23	24

no statistically significant (p value >5%) differences between a pair of algorithms.

The above analysis proves the efficiency of prediction for our model. That is, it can still perform well if class output label varies across multiple patients. Fig. 10 shows another average performance chart using boxplot. Here the results of Hamming score of 85 patients are summarized in terms of Hamming score of 24 classifier combinations. As we also see here most of the values are between 90 and 95%.

5.3.3. *Performance in terms of model building time*

In a clinical decision support system it is also important to take a quick decision. Therefore, we need to measure the efficiency of our model in terms of learning time. Fig. 11 graphically shows the average model training time for 85 patients. We can see that for most cases the average model building time is less than a second using 24-hour data. The FW method is a slow learner. SMO is slow learner than Random Tree and J48. However CT, PS and RT can learn fast using our data. PS and Random Tree combination has the minimum building time. Therefore, when we consider both Hamming score and building time we can say that CT, PS and RT problem transformation methods along with Random Tree adaptation method are the best classifiers for our techniques.

5.4. *Performance for using correlations*

To understand the importance of correlation coefficients in MLC performance we have tested the same dataset that was used for MLC. Here we used our 3 adaptation methods (J48, Random Tree and SMO) as single-label classifier. That is, the objective is to classify the output label of each vital individually with the following three settings.

1. Considering all 123 features generated from 6 vitals to predict the output label of each vital individually.
2. Excluding correlation coefficients as features for all and considering only the statistical features of corresponding vital sign.
3. Using selected features by a feature selection algorithm. As feature selection algorithm we have used correlation-based feature selection algorithm that evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them. As search function we have used best first search that searches the space of attribute subsets by greedy hillclimbing augmented with a backtracking facility [50].

The results of this comparison for a single patient are presented in Table 7. The results are generated using WEKA software. We can see that the prediction accuracy is higher when we consider all the correlated features for all types of classifier. In few cases (e.g. HR and RR) the prediction accuracy with selected attributes performed better than using all classifiers, but for most other cases it is not as correlation feature selection process is biased by picking the attributes related to the desired vital sign (e.g. HR) only. Thus, in individualized classification we can get the prediction performance of some vital signs better than others but when they are combined with correlations we have a stable prediction model. This proves the importance of using correlation in MLC. The correlated features have high impact on predicting other vitals.

5.5. *Discussions*

From the above observations, we can conclude the MLC is a better option than single-label classifiers for the developed CDSS. Because for estimating values of 6 vitals we need to build 6 learning models using the same set of features. We can see from Table 7 that the vitals such as SPO_2 which have low variations have higher accuracy than the vitals with high variations and this makes the prediction inconsistent. In case of multi-label classifier we can estimate the values of all vitals using just one training model which have high Hamming score. That is, using MLC we can achieve high prediction rate and low model training time which are essential for real-time patient monitoring.

Moreover, this technique can easily be used for patient-specific vital sign predictions. The learning models for each patient are obtained by training different models using bootstrap data (first 24 hours). The best learning for $w = 10$ and $l = 7$ is obtained using the highest Hamming score and lowest model building time while being trained in m3.2xlarge Amazon EC2 instance using MEKA classifier. The best model for each patient is stored in Amazon S3. Then, the future values are predicted for subsequent data. The abnormality alarm is when multiple vitals have high or low values, or one vital has very high/low value. The prediction is verified using Hamming score and the same model is re-trained when Hamming score goes below 90%.

6. Conclusion and future work

In this work, we propose a model for a CDSS to predict multiple vital sign values of a home-monitoring patient using their correlations. This also helps to find patient-specific anomalies in advance. The proposed technique takes the advantage of multi-label classification. Numerical trend dataset of multiple vital signs is prepared for multi-label classification engine. Different multi-label classification methods are performed over data of many patients using MEKA software and their performance was evaluated in order to extract patient-specific knowledge and predict future abnormality.

Table 6 – The observed *p*-values in paired t-tests to compare the algorithms as pair.

A	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)	(21)	(22)	(23)	(24)
(1)	0	0.22	0.31	0	0.28	0.16	0	0.18	0	0	0.18	0	0	0.57	0.91	0	0.26	0	0	0	0	0	0.19
(2)		0	0	0.61	0	0	0.39	0	0	0.4	0	0.9	0	0	0	0.36	0	0	0.89	0.05	0	0.62	0
(3)			0.04	0	0.9	0.89	0	0.91	0	0	0.91	0	0	0.1	0.21	0	0.91	0	0	0	0	0	0.96
(4)				0	0.05	0.02	0	0.03	0	0	0.03	0	0	0.72	0.4	0	0.05	0	0	0	0	0	0.03
(5)					0	0	0.68	0	0	0.7	0	0.7	0	0	0	0.65	0	0	0.53	0.12	0	0.99	0
(6)						0.79	0	0.81	0	0	0.81	0	0	0.13	0.26	0	0.98	0	0	0	0	0	0.85
(7)							0	0.99	0	0	0.99	0	0	0.07	0.15	0	0.81	0	0	0	0	0	0.94
(8)								0	0	0.98	0	0.44	0	0	0	0.98	0	0	0.33	0.28	0	0.71	0
(9)									0	0	1	0	0	0.08	0.17	0	0.82	0	0	0	0	0	0.95
(10)										0	0	0	0	0	0	0	0	0.44	0	0	0	0	0
(11)											0	0.45	0	0	0	0.96	0	0	0.34	0.26	0	0.72	0
(12)												0	0	0.08	0.17	0	0.82	0	0	0	0	0	0.95
(13)													0	0	0	0.41	0	0	0.79	0.06	0	0.7	0
(14)														0	0	0	0	0	0	0	0	0	0
(15)															0.66	0	0.12	0	0	0	0	0	0.09
(16)																0	0.24	0	0	0	0	0	0.18
(17)																	0	0	0.3	0.27	0	0.68	0
(18)																		0	0	0	0	0	0.87
(19)																			0	0	0	0	0
(20)																				0.05	0	0.54	0
(21)																					0	0.14	0
(22)																						0	0
(23)																							0

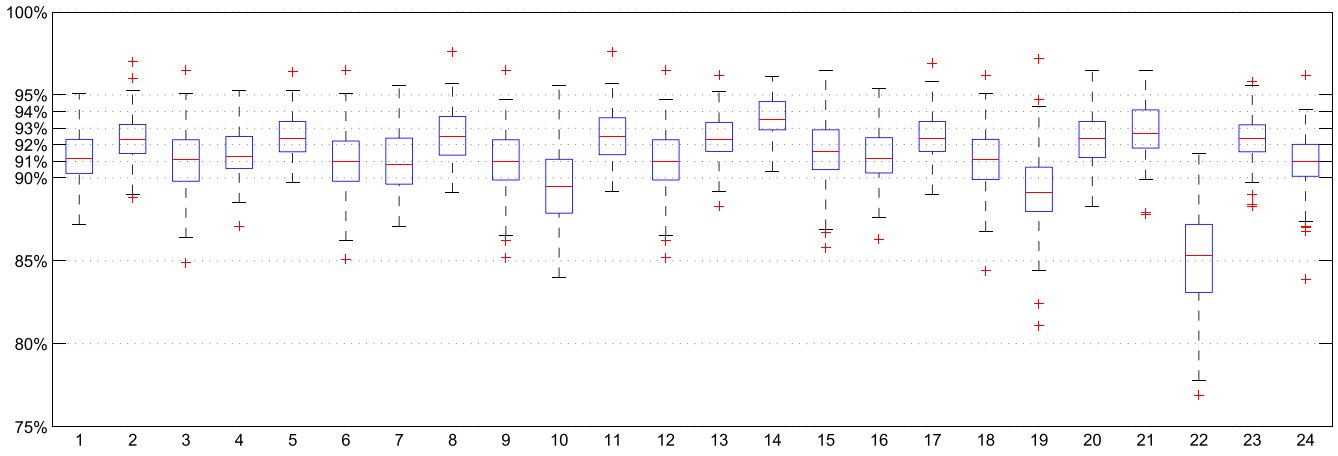

Fig. 10 – Hamming scores of 24 different classifier combinations using the statistics of 85 patients.

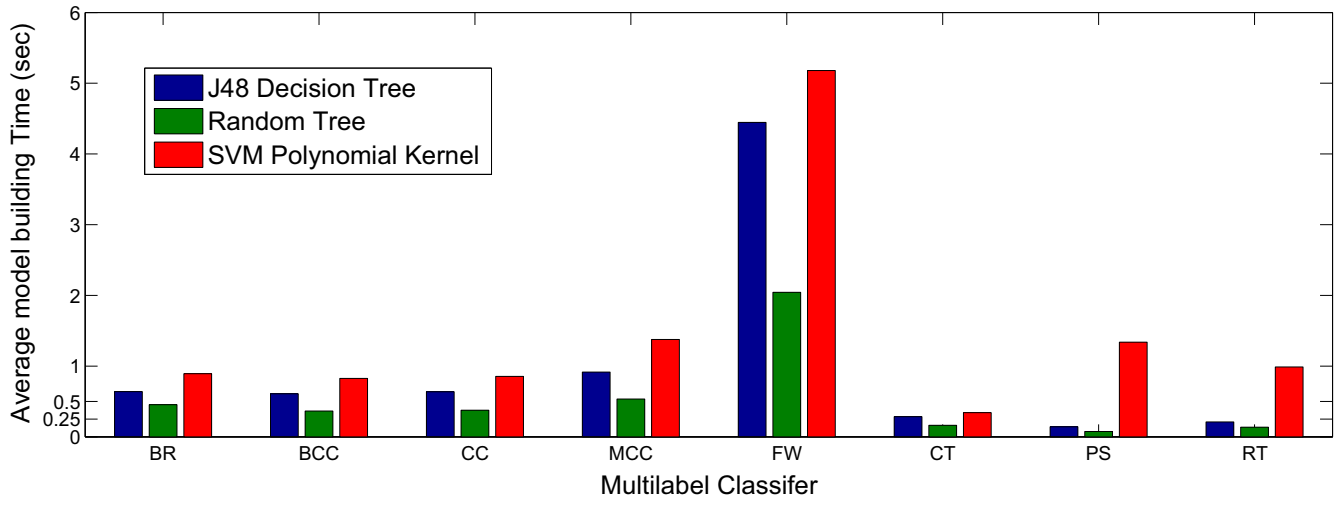

Fig. 11 – Average building time for different classifier using random 24 hour data of 85 patients. PS-Random Tree is the fastest process. SMO takes more time than Random Tree and J48. FW has very high building time.

Using the experimental evaluations we showed that multi-label classification outperformed single label classification for this type of decision support system. Based on Hamming score and model building time we obtained the best classifier for each patient individually. The model is also made adaptive and situation-aware using incremental learning process. Thus, patient-specific clinical decision can be made using the predicted outcome produced by the multi-label classifier in short time.

The major contribution of this research is the investigation of multi-label classification methods to forecast the future value of multiple vital signs and at the same time using their

Table 7 – A comparison of prediction accuracies for 6 vital signs as individual using 3 classification algorithms. This is using the same dataset and settings as in Table 4. Here, WCF means with all correlated features of all vitals, OF means using only features of corresponding vital and SF using subset of features selected by a feature selection algorithm from all features.

	J48 Decision tree			Random tree			SVM polynomial kernel		
	WCF	OF	SF	WCF	OF	SF	WCF	OF	SF
HR	89.36	72.34	85.1	85.1	80.85	89.36	82.98	80.85	87.23
SBP	72.08	60.28	70.21	69.57	57.44	68.08	72.34	51.06	64.68
DBP	70.21	61.7	68.08	74.46	61.7	70.21	70.21	55.31	68.08
MBP	72.34	63.82	63.82	74.46	65.18	68.08	70.21	59.57	55.31
RR	69.7	65.95	72.34	74.46	70.21	76.59	67.44	59.57	64.68
SPO_2	95.74	93.61	95.74	95.74	95.74	97.87	97.87	93.61	93.61

correlated features. Therefore, the outcome of multi-label classification can assist the healthcare professionals in decision making through the CDSS and thus help to detect instances when patient would be in serious clinical danger. Our model is extendible for other vital signs (e.g. body temperature) and in future, we want to include more bio-signals and clinical data and their correlation in our investigations.

Acknowledgement

The authors wish to acknowledge the support of National ICT Australia (NICTA) and RMIT University for funding the research work presented in this paper.

REFERENCES

[1] P. Nykänen, S. Chowdhury, O. Wigertz, Evaluation of decision support systems in medicine, Comput. Methods Programs Biomed. 34 (1991) 229–238.
[2] H. Xia, I. Asif, X. Zhao, Cloud-ECG for real time ECG monitoring and analysis, Comput. Methods Programs Biomed. 110 (2013) 253–259.
[3] F. Sufi, Q. Fang, I. Khalil, A. Mahmoud, Novel methods of faster cardiovascular diagnosis in wireless telecardiology, IEEE 27 (2009) 537–552.
[4] A.K. Dey, Providing Architectural Support for Building Context-Aware Applications, Ph.D. thesis, Georgia Institute of Technology, 2000.
[5] M.A. Musen, B. Middleton, R.A. Greenes, Clinical decision-support systems, in: Biomedical Informatics, Springer, 2014, pp. 643–674.
[6] C.L. Meli, I. Khalil, Z. Tari, Load-sensitive dynamic workflow re-orchestration and optimisation for faster patient healthcare, Comput. Methods Programs Biomed. 113 (2014) 1–14.
[7] T. Klingeberg, M. Schilling, Mobile wearable device for long term monitoring of vital signs, Comput. Methods Programs Biomed. 106 (2012) 89–96.
[8] K.A. Sidek, I. Khalil, Enhancement of low sampling frequency recordings for ECG biometric matching using interpolation, Comput. Methods Programs Biomed. 109 (2013) 13–25.
[9] V.S. Tseng, L.-C. Chen, C.-H. Lee, J.-S. Wu, Y.-C. Hsu, Development of a vital sign data mining system for chronic patient monitoring, in: International Conference on Complex, Intelligent and Software Intensive Systems. CISIS 2008, IEEE, pp. 649–654, (2008).
[10] A.R.M. Forkan, I. Khalil, Z. Tari, S. Foufou, A. Bouras, A context-aware approach for long-term behavioural change detection and abnormality prediction in ambient assisted living, Pattern Recognit. 48 (2015) 628–641.
[11] H. Shao, G. Li, G. Liu, Y. Wang, Symptom selection for multi-label data of inquiry diagnosis in traditional Chinese medicine, Sci. China Inf. Sci. 56 (2013) 1–13.
[12] L.W. Wilkins, ECG Interpretation Made Incredibly Easy, fifth ed., Lippincott Williams & Wilkins, 2011.
[13] A. Forkan, I. Khalil, Z. Tari, CoCaMAAL: a cloud-oriented context-aware middleware in ambient assisted living, Future Generation Comput. Syst. 35 (2014) 114–127.
[14] G. Tsoumakas, I. Katakis, Multi-label classification: an overview, Dept. of Informatics, Aristotle University of Thessaloniki, Greece (2006).
[15] M.R. Boutell, J. Luo, X. Shen, C.M. Brown, Learning multi-label scene classification, Pattern Recognit. 37 (2004) 1757–1771.
[16] K. Dembczyński, W. Waegeman, W. Cheng, E. Hüllermeier, On label dependence and loss minimization in multi-label classification, Mach. Learn. 88 (2012) 5–45.
[17] A. Forkan, I. Khalil, A. Ibaida, Z. Tari, BDCaM: big data for context-aware monitoring—a personalized knowledge discovery framework for assisted healthcare, IEEE Transactions on Cloud Computing (2015).
[18] F. Kang, R. Jin, R. Sukthankar, Correlated label propagation with application to multi-label learning, in: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), volume 2, IEEE, pp. 1719–1726 (2015).
[19] K. Glinka, A. Wosiak, D. Zakrzewska, Improving children diagnostics by efficient multi-label classification method, in: Information Technologies in Medicine, Springer, 2016, pp. 253–266.
[20] O. Salem, Y. Liu, A. Mehaoua, R. Boutaba, IEEE 18 (2014) 1541–1551.
[21] C.M. Ennett, K. Lee, L.J. Eshelman, B. Gross, L. Nielsen, J.J. Frassica, et al., Predicting respiratory instability in the ICU, in: 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. EMBS 2008., IEEE, pp. 2848–2851, (2008).
[22] A. Aboukhalil, L. Nielsen, M. Saeed, R.G. Mark, G.D. Clifford, Reducing false alarm rates for critical arrhythmias using the arterial blood pressure waveform, J. Biomed. Inform. 41 (2008) 442–451.
[23] R.-W. Zhao, G.-Z. Li, J.-M. Liu, X. Wang, Clinical multi-label free text classification by exploiting disease label relation, in: Bioinformatics and Biomedicine (BIBM), 2013 IEEE International Conference on, IEEE, pp. 311–315, (2013).
[24] M. Mammadov, A. Rubinov, J. Yearwood, An optimization approach to identify the relationship between features and output of a multi-label classifier, in: Data Mining in Biomedicine, Springer, 2007, pp. 141–167.
[25] D. Zufferey, T. Hofer, J. Hennebert, M. Schumacher, R. Ingold, S. Bromuri, Performance comparison of multi-label learning algorithms on clinical data for chronic diseases, Comput. Biol. Med. 65 (2015) 34–43.
[26] S. Bromuri, D. Zufferey, J. Hennebert, M. Schumacher, Multi-label classification of chronically ill patients with bag of words and supervised dimensionality reduction algorithms, J. Biomed. Inform. 51 (2014) 165–175.
[27] H. Banaee, M.U. Ahmed, A. Loutfi, Data mining for wearable sensors in health monitoring systems: a review of recent trends and challenges, Sensors (Basel) 13 (2013) 17472–17500.
[28] I. Yoo, P. Alafaireet, M. Marinov, K. Pena-Hernandez, R. Gopidi, J.-F. Chang, et al., Data mining in healthcare and biomedicine: a survey of the literature, J. Med. Syst. 36 (2012) 2431–2448.
[29] J.M. Hardin, D.C. Chhieng, Data mining and clinical decision support systems, in: Clinical Decision Support Systems, Springer, 2007, pp. 44–63.
[30] L. Clifton, D.A. Clifton, P.J. Watkinson, L. Tarassenko, Identification of patient deterioration in vital-sign data using one-class support vector machines, in: Federated Conference on Computer Science and Information Systems (FedCSIS), IEEE, pp. 125–131, (2011).
[31] J. Henriques, T. Rocha, Prediction of acute hypotensive episodes using neural network multi-models, in: Computers in Cardiology, IEEE, pp. 549–552, (2009).
[32] M. Seera, C.P. Lim, W.S. Liew, E. Lim, C.K. Loo, Classification of electrocardiogram and auscultatory blood pressure signals using machine learning models, Expert Syst. Appl. 42 (2015) 3643–3652.

[33] F. Sufi, I. Khalil, Diagnosis of cardiovascular abnormalities from compressed ECG: a data mining-based approach, IEEE 15 (2011) 33–39.
[34] A. Copetti, J.C. Leite, O. Loques, M.F. Neves, A decision-making mechanism for context inference in pervasive healthcare environments, Decis. Support Syst. 55 (2013) 528–537.
[35] G.B. Moody, R.G. Mark, A database to support development and evaluation of intelligent intensive care monitoring, in: Computers in Cardiology, IEEE, pp. 657–660, (1996).
[36] M. Saeed, M. Villarroel, A.T. Reisner, G. Clifford, L.-W. Lehman, G. Moody, et al., Multiparameter intelligent monitoring in intensive care ii (mimic-ii): a public-access intensive care unit database, Crit. Care Med. 39 (2011) 952.
[37] A.L. Goldberger, L.A. Amaral, L. Glass, J.M. Hausdorff, P.C. Ivanov, R.G. Mark, et al., Physiobank, physiotoolkit, and physionet components of a new research resource for complex physiologic signals, Circulation 101 (2000) e215–e220.
[38] H. Cao, D.E. Lake, J.E. Ferguson, C.A. Chisholm, M.P. Griffin, J.R. Moorman, et al., Toward quantitative fetal heart rate monitoring, IEEE 53 (2006) 111–118.
[39] F. Gallegos-Funes, V. Ponomaryov, S. Sadovnychiy, L. Nino-de Rivera, Median M-type K-nearest neighbour (MM-KNN) filter to remove impulse noise from corrupted images, Electron. Lett. 38 (2002) 786–787.
[40] A. Janghorbani, A. Arasteh, M.H. Moradi, Prediction of acute hypotension episodes using logistic regression model and support vector machine: a comparative study, in: Electrical Engineering (ICEE), 19th Iranian Conference on, IEEE, pp. 1–4, (2011).
[41] B.M. Asl, S.K. Setarehdan, M. Mohebbi, Support vector machine-based arrhythmia classification using reduced features of heart rate variability signal, Artif. Intell. Med. 44 (2008) 51–64.
[42] A. Elisseeff, J. Weston, A kernel method for multi-labelled classification, in: Advances in Neural Information Processing Systems, pp. 681–687 (2008).
[43] M.-L. Zhang, Z.-H. Zhou, ML-KNN: a lazy learning approach to multi-label learning, Pattern Recognit. 40 (2007) 2038–2048.
[44] J. Reed, P. Reutemann, MEKA: a multi-label extension to WEKA, 2015. http://meka.sourceforge.net/ (accessed 12.11.2015).
[45] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I.H. Witten, The WEKA data mining software: an update, ACM SIGKDD Explor. Newsl. 11 (2009) 10–18.
[46] J. Read, A. Bifet, G. Holmes, B. Pfahringer, Scalable and efficient multi-label classification for evolving data streams, Mach. Learn. 88 (2012) 243–272.
[47] J. Read, L. Martino, P.M. Olmos, D. Luengo, Scalable multi-output label prediction: from classifier chains to classifier trellises, Pattern Recognit. 48 (2015) 2096–2109.
[48] G. Tsoumakas, I. Katakis, I. Vlahavas, Mining multi-label data, in: Data Mining and Knowledge Discovery Handbook, Springer, 2010, pp. 667–685.
[49] R. Grodzicki, J. Mańdziuk, L. Wang, Improved multilabel classification with neural networks, in: International Conference on Parallel Problem Solving from Nature, Springer, pp. 409–416 (2015).
[50] M.A. Hall, Correlation-based feature selection of discrete and numeric class machine learning (2000).