Predictive Monitoring of Mobile Patients by Combining Clinical Observations With Data From Wearable Sensors

Lei Clifton, David A. Clifton, Marco A. F. Pimentel, Peter J. Watkinson, and Lionel Tarassenko

***Abstract*—The majority of patients in the hospital are ambulatory and would benefit significantly from predictive and personalized monitoring systems. Such patients are well suited to having their physiological condition monitored using low-power, minimally intrusive wearable sensors. Despite data-collection systems now being manufactured commercially, allowing physiological data to be acquired from mobile patients, little work has been undertaken on the use of the resultant data in a principled manner for robust patient care, including predictive monitoring. Most current devices generate so many false-positive alerts that devices cannot be used for routine clinical practice. This paper explores principled machine learning approaches to interpreting large quantities of continuously acquired, multivariate physiological data, using wearable patient monitors, where the goal is to provide early warning of serious physiological determination, such that a degree of predictive care may be provided. We adopt a one-class support vector machine formulation, proposing a formulation for determining the free parameters of the model using partial area under the ROC curve, a method arising from the unique requirements of performing online analysis with data from patient-worn sensors. There are few clinical evaluations of machine learning techniques in the literature, so we present results from a study at the Oxford University Hospitals NHS Trust devised to investigate the large-scale clinical use of patient-worn sensors for predictive monitoring in a ward with a high incidence of patient mortality. We show that our system can combine routine manual observations made by clinical staff with the continuous data acquired from wearable sensors. Practical considerations and recommendations based on our experiences of this clinical study are discussed, in the context of a framework for personalized monitoring.**

***Index Terms*—E-health, novelty detection, personalized monitoring, predictive monitoring.**

Manuscript received March 25, 2013; revised July 19, 2013; September 19, 2013; accepted October 24, 2013. Date of publication November 26, 2013; date of current version May 1, 2014. The work of L. Clifton was supported by the NIHR Biomedical Research Centre Programme, Oxford. The work of D. A. Clifton was supported by a Royal Academy of Engineering Research Fellowship and the Centre of Excellence in Personalised Healthcare funded by the Wellcome Trust and EPSRC under Grant WT 088877/Z/09/Z. The work of M. A. F. Pimentel was supported by the RCUK Digital Economy Program under Grant EP/G036861/1 (the Oxford Centre for Doctoral Training in Healthcare Innovation).

L. Clifton, D. A. Clifton, M. A. F. Pimentel, and L. Tarassenko are with the Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, Oxford, OX1 2JD, U.K. (e-mail: lei.clifton@eng.ox.ac.uk; david.clifton@eng.ox.ac.uk; marco.pimentel@eng.ox.ac.uk; lionel.tarassenko@eng.ox.ac.uk).

P. J. Watkinson is with the Nuffield Department of Anaesthetics, University of Oxford, Oxford, OX1 2JD, U.K. (e-mail: peter.watkinson@ouh.nhs.uk).

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

Digital Object Identifier 10.1109/JBHI.2013.2293059

I. Introduction

THE majority of patients in the hospital are ambulatory, and thus, they are well suited to be monitored using wearable sensors for the purposes of predictive care. The goal of such systems is to provide early warning of physiological deterioration such that preventative clinical action may be taken to improve patient outcomes. However, the current state of the art is not at a level suitable for wide-scale adoption, and there is a perceived "plague of pilots" in unvalidated data collection systems [1]–[3], whereby the majority of published studies are concerned with the demonstration of algorithms using small numbers of subjects, who are often not representative of actual patient groups.

Despite wearable patient monitors now being manufactured commercially, allowing the collection of continuous physiological data from ambulatory patients, the resulting quantity of data acquired each day is large, and a "data deluge" effect occurs. The workload of clinicians and healthcare workers prevents them inspecting long time-series of multivariate patient physiological data to a high degree of accuracy, and the predictive aspect to patient monitoring is lost. "Intelligent," online processing of these large datasets is, therefore, required for predictive monitoring, the results of which should then focus the limited resources of human experts to those subsets of patients who are deemed to be most at risk of being physiologically unstable, and who are in need of expert review. However, existing clinically validated devices often simply compare physiological data to heuristically determined, univariate thresholds and generate an alert if those thresholds are exceeded (e.g., "alert if heart rate (HR) exceeds 130 beats/min"). Such simplistic schemes result in large numbers of false alerts, which make these devices largely unusable in clinical practice [4]–[6]. Due to the difficulty of acquiring large datasets of patient physiology in clinical trials, there have been few attempts to investigate the large-scale clinical use of wearable patient sensors for predictive monitoring, and this area of e-health remains largely unexplored. A review of existing methods may be found in Section III.

A. Contributions of This Paper

1) We address the perceived lack of evidence for the large-scale clinical adoption of "intelligent" predictive monitoring systems by describing (in Section II) a study in which wearable sensors are used for the routine care of a large population of high-risk, ambulatory patients.
2) We adopt a machine learning approach to cope with the large quantity of vital-sign data acquired from monitoring

ambulatory patients in real time, comparing four techniques, the majority of which have not been applied to the predictive monitoring of patient data. A survey of existing methods is described to set the context of this study, given in Section III.

3) Existing methods for automatically determining the parameters of machine learning models (as required in online patient monitoring) suffer from many disadvantages; these problems, and a novel method for estimating suitable model parameters for the unique constraints involved in predictive patient monitoring, are introduced in Section III. Results are presented in Section IV.
4) A discussion and conclusions are presented in Section V, in which we describe how the work described in this paper makes a step toward the ultimate goal of *personalized* predictive monitoring.

II. Background

We undertook a clinical study approved by the local Research Ethics Committee[1] of 200 patients in a postoperative ward of the Cancer Centre, Oxford University Hospitals NHS Trust, Oxford, U.K. Patients were discharged to the ward following upper-gastrointestinal (GI) cancer surgery. This group of patients was selected for our study because of the high incidence (up to 20%) of postsurgical complications, whereby patients can deteriorate physiologically, resulting in adverse outcomes such as readmission to the intensive care unit (ICU) or death. Readmission to the ICU is prolonged and the mortality rate of such patients is high. These adverse events may occur when the physiological condition of the patient is not recognized or acted upon early enough [5], motivating the need for predictive monitoring patient vital signs (HR, measured in beats per minute, respiratory rate RR, measured in breaths per minute, blood oxygen saturation SpO_2, measured as a percentage, and systolic blood pressure SysBP, measured in mmHg). The goal of such "predictive" systems is to provide early warning of physiological deterioration, such that preventative clinical action may be taken.

A. Existing Manual Monitoring

Clinical guidance in the U.K. [6] recommends the regular observational recording of vital signs, combined with the use of early warning score (EWS) systems. The latter involve the clinician applying univariate scoring criteria to each vital sign in turn (e.g., "score 3 if HR exceeds 130 beats/min"). Care is then escalated to a higher level if any of the scores assigned to individual vital signs, or the sum of all such scores, exceed some threshold.

The length-of-stay of patients in our study is shown in Fig. 1(a), where the mean length-of-stay is nine days following surgery. However, the distribution shown in the figure has a long tail, extending up to 60 days, corresponding to patients for whom earlier discharge is not possible. This is typically due to continued physiological instability of the patients, and concern on the part of the ward staff such that the patient cannot be discharged. Such patients can accumulate several hundred manual vital-sign observations during their stay on the ward; a histogram of the time between consecutive manual observations (across all patients) is shown in Fig. 1(b). The latter shows that most observations are taken at intervals of several hours, with a mean of 4.1 h between observations (but often rising to as long as eight h between observations).

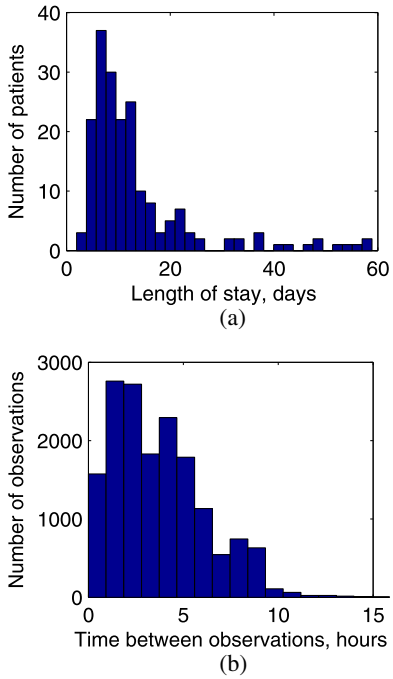

Fig. 1. (a) Histogram of the length-of-stay of 200 studied patients in the Cancer Centre. (b) Histogram of time between manual observations, over all patients.

This current standard of care for "predictive monitoring," involving manual observation, has a number of disadvantages. 1) The EWS assigned to each vital sign, and the thresholds against which the scores are compared, are typically heuristic [7]. 2) EWS systems are used with periodic observation of vital signs, which may be made as infrequently as once every 12 h in some wards. Patients may deteriorate significantly between observations. 3) There is a significant error rate associated with manual scoring, especially in the high-workload setting of a high-dependence clinical ward. 4) Each vital sign is treated independently and correlations between vital signs are not taken into account. The approach described in this paper attempts to address these disadvantages.

B. Continuous Wearable Monitoring

Patients in our study are connected to conventional bed-side monitors during the first day after their surgery. However, as is common in most hospital wards, the majority of patients are mobilized after the first day, to gain exercise by walking around the ward. This demonstrates the difficulty of monitoring the majority of patients in hospital (and at home), because they are mobile, and which therefore strongly motivates the use of wearable monitors to perform predictive monitoring.

[1]Mid & South Bucks Research Ethics Committee reference 08/H0607/79.

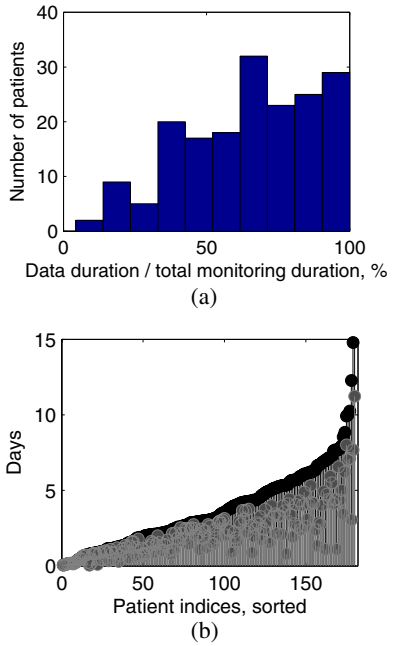

Fig. 2. (a) Histogram of continuous data completeness as a percentage of the total time that the patient was equipped with a wearable patient monitor. (b) Time that patients were equipped with a wearable patient monitor (sorted in ascending order and shown in black) with actual time of acquired data (shown in gray).

Continuous wearable monitoring devices are widely available, despite the disadvantages of high false-alarm rates described in Section I. The system deployed in the study described by this paper used mobile pulse oximeters manufactured by Nonin Medical, Inc. (for the acquisition of the photoplethysmogram or *PPG*, from which SpO_2 and HR may be derived). Mobile ECG sensors manufactured by Corscience GmbH & Co. KG. (for the acquisition of the ECG, from which HR may be derived) were also used. We note that the alarm functions of these wearable monitors were deactivated, and the devices were used only for continuous data acquisition, to which the machine learning methods described in Section III were then applied retrospectively.

These wearable devices were configured to communicate via Bluetooth to a patient-worn PDA, which collected ECG at 256 Hz and the PPG at 75 Hz. These waveforms, along with derived estimates of HR and SpO_2, were transmitted to a central server via wi-fi. The central station stored data along with anonymized patient information for later analysis.

There are few reliable methods for acquiring blood pressure in a nonintrusive continuous manner, and so manual measurements of SysBP made by the ward staff were entered into the patient PDA, along with measurements of RR. After entry into the PDA, these manual measurements were automatically transmitted to the central station, where they were then associated with the continuous data described above.

Fig. 2(a) shows a histogram of the percentage of the total monitoring time for each patient (defined to be the time for which wearable sensors were attached to the patient) for which actual data were acquired. It may be seen that the completeness of data acquisition is far below 100%, with a mean of 62%. The major causes of data incompleteness were infrequent malfunction of the wearable sensors and PDAs, failures in the hospital wi-fi network, occasional crashes of the central server, and expiration of batteries in the wearable sensors and PDAs. A team of research nurses was responsible for ensuring that patient compliance and device readiness was kept as high as possible.

A plot of total monitoring times (sorted into ascending order) is shown in Fig. 2(b), where the actual monitoring time for each patient is also shown. Comparison of this figure with Fig. 1(a) shows that patients were typically connected to the wearable patient monitors for a proportion of their stay on the ward, with a maximum total monitoring time of approximately 25 days (compared with a maximum length-of-stay of approximately 60 days). There was a mean total monitoring time of approximately 5.2 days (compared with an average length-of-stay of approximately ten days).

Much of the difference between total stay on the ward and total monitoring time is due to the patient compliance; the ECG sensors were particularly unpopular with patients, despite their small size, probably due to their positioning on the chest following upper-GI surgery. The pulse oximeters were tolerated much better by patients, being attached to the fingertip. However, patients typically removed the pulse oximeters prior to eating or showering and often failed to replace the devices afterward. This was particularly evident during weekends, when research nurses were unavailable to check the connectivity of each patient. Due to the perceived discomfort of the ECG sensors, they were discontinued from use after 52 patients had been continuously monitored.

The total quantity of continuous data acquired for all 200 patients was 63.8 GB, and subsequently used for investigating our machine learning approach to analyzing the data for demonstration that predictive monitoring could be performed by early identification of deterioration.

III. Methods

Monitoring complex, high-integrity systems (such as patients in the hospital or at home) can be confounded by the variability between individual systems of the same system type. In our case, patients of similar demographic backgrounds can exhibit significantly different "normal" physiology. The few examples of "abnormal" behavior (e.g., physiological deterioration) that may exist for some population are, therefore, often inapplicable to the analysis of previously unseen individuals. For example, an HR of 50 beats/min may be indicative of considerable physiological abnormality in one hospital patient, while it may be entirely normal for a fitter patient of the same age and background.

Furthermore, high-integrity systems also typically exhibit a high degree of structural complexity and can often comprise many subsystems that interact in a nonlinear manner. Thus, the potential space of "abnormality" is extremely large, and so the large resultant number of failure modes is often poorly understood. For example, the exact response of a particular human's physiology to a given failure mode (such as deterioration leading to myocardial infarction) will vary significantly between

patients. Those data that do exist are typically insufficient for constructing accurate models of these failure states, because the data are usually obtained from a small number of patients, with differing comorbidities, lifestyles, etc. We have demonstrated in the previous section some of the difficulties that arise in collecting large datasets of physiological data from patients.

A. Existing Work

Much existing work has focused on the development of communications infrastructures, platforms and protocols for data transfer, and decision support frameworks, extended reviews of which may be found in [2], [3], [8], and [9]. The application of machine learning techniques to the predictive monitoring of patient physiological data at large scale is limited; reviews may be found in [10] and [11].

Much existing work takes a "novelty detection" approach. This method attempts to avoid the problems described earlier by modeling the "normal" mode of operation of the system, which is often well understood because most high-integrity systems function "normally" most of the time. The classifier then looks for deviations from that normal model, which are classified "abnormal." This approach is appropriate for the predictive monitoring of physiological condition in patients, because sufficient data exist from "stable" patients such that a model of the well-understood "normal" state of these patients may be constructed. Physiological deterioration may then be detected as being corresponding departures in the vital signs from that "normal" state. The use of novelty detection for predictive monitoring of patients is particular appropriate, because the manual EWS systems described earlier (the use of which is standard clinical practice) are essentially novelty detection schemes, where the EWS may be directly interpreted as a novelty score that increases as patient physiology deviates from "normality."

While the field of novelty detection is well explored in jet engine condition monitoring [12], signal segmentation [13], and FMRI analysis [14], among many others (a review of which may be found in [15]), its use for tracking patient physiological condition remains largely unexplored, possibly due to the difficulty of acquiring and labeling physiological data. Key papers include the use of kernel estimates with patient vital-sign data [16]: a low-dimensional approach based on Kalman filtering for neonatal ICU patients [17], a support vector machine (SVM) [18], neural networks in univariate sleep analysis [19], and univariate Gaussian processes (GPs) for denoising HR data [20].

This paper compares four methods of performing novelty detection: two discriminative methods (using one-class SVMs and one-class GPs) and two generative methods (using Gaussian mixture models, or GMMs, and a kernel density estimate). We describe a novel parameter selection technique for the SVM-based approach, suitable for training the model for novelty detection with patient physiological data.

B. One-Class SVMs

We briefly recap the formulation of the one-class SVM to introduce our notation, and refer the reader to the original formulation [21] for further details.

A quantity l of d-dimensional data $\{\mathbf{x}_1, \ldots, \mathbf{x}_l\} \in \mathbb{R}^d$ are mapped into a (potentially infinite-dimensional) feature space $\mathbb{F}$ by some nonlinear transformation $\mathbf{\Phi}$: $\mathbb{R}^d \rightarrow \mathbb{F}$. A kernel function k provides the dot product between pairs of transformed data in $\mathbb{F}$, such that $k(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{\Phi}(\mathbf{x}_i) \cdot \mathbf{\Phi}(\mathbf{x}_j)$. A Gaussian kernel allows a point to be separated from the origin in $\mathbb{F}$ [22], hence is chosen for us in the work described by this paper: $k(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\|\mathbf{x}_i - \mathbf{x}_j\|^2 / 2\sigma^2\right)$, where σ is the width parameter associated with the Gaussian kernel.

The decision boundary between "normal" and "abnormal" subspaces in $\mathbb{F}$ is $z(\mathbf{x}) = w_o \cdot \mathbf{\Phi}(\mathbf{x}) - \rho_0$, with parameters

$$w_o = \sum_{i=1}^{N_s} \alpha_i \mathbf{\Phi}(\mathbf{s}_i) \tag{1}$$

$$\rho_o = \frac{1}{N_s} \sum_{j=1}^{N_s} \sum_{i=1}^{N_s} \alpha_i k(\mathbf{s}_i, \mathbf{s}_j) \tag{2}$$

where $\mathbf{s}_i$ are the support vectors, of which there are N_s, and where k is the Gaussian kernel. Here, $w_o \in \mathbb{F}$, $\rho_o \in \mathbb{R}$, and that α_i are Lagrangian multipliers used to solve the dual formulation, more details of which may be found in [22] and which are not reproduced here. Test data $\mathbf{x}$ are classified as being either "normal" or "abnormal" according to the sign of $z(\mathbf{x})$.

C. Proposed Parameter Optimization for a One-Class SVM

For the case of a Gaussian kernel $k(\mathbf{x}_i, \mathbf{x}_j)$, it is important to choose an appropriate value for the bandwidth parameter σ. Larger values of σ result in smoother decision boundaries, which therefore tend to exhibit lower variance at the expense of increased bias (using the standard terminology from probabilistic modeling). Conversely, smaller values of σ provide decreased bias, but at the expense of increased variance. The "optimal" value for σ will depend on the distribution of the particular dataset under consideration, and it is not usually obvious how one should choose the value of σ. For a Gaussian kernel $k(\mathbf{x}_i, \mathbf{x}_j)$, the quantity $-\log k(\mathbf{x}_i, \mathbf{x}_j)$ is the Euclidean distance between two observations scaled by a factor $1/2\sigma^2$. Based on this link between σ and Euclidean distance, we propose the following three-step method to determine an appropriate value for σ, estimated directly from the available training data. The following is an SVM-based extension of the popular method proposed by Bishop [23], originally for use with multilayer perceptrons.

A1: Calculate the local average Euclidean distance Δ_i of K nearest neighbors from each observation in the training set, where $K = \sqrt{l}$, $\Delta_i = \frac{1}{K} \sum_{j \in \mathcal{D}} \|\mathbf{x}_i, \mathbf{x}_j\|$, $\forall i = 1 \ldots l$, and where $\mathcal{D}$ is the set of K nearest neighbors for $\mathbf{x}_i$.

A2: Calculate the global average distance Δ_G by averaging Δ_i over all the training data, $\Delta_G = l^{-1} \sum_i \Delta_i$.

A3: Δ_G provides a guide for the range of σ, where we define $\sigma = \kappa \times \Delta_G$, and where κ is a linking constant between the value of σ and the global average distance Δ_G of any dataset. Therefore, κ provides a guide for the appropriate value of σ, which is independent of the size of the dataset l. Once an appropriate value of κ is chosen for one dataset, it provides a

good starting point for another dataset with similar dynamics (e.g., for another patient vital-sign dataset), allowing the value of κ to be reused from previous analyses, when the dataset has changed. This is of particular importance for the online predictive monitoring of patients, in which such prior information gained from previous studies can be useful in parameter optimization for new patient-monitoring studies.

The other parameter to optimize in a one-class SVM is ν, defined below. The support vector constraints in terms of the SVM penalty parameter (typically denoted C in the literature) are $\sum_i \alpha_i = 1,\ 0 \leq \alpha_i \leq C$, allowing us to state[2] that $1/l \ \leq C \ \leq 1$. We may equivalently write $C = 1/\nu l$ [21], so we have $1/l \leq \nu \leq 1$. Therefore, ν and C take values in the same range.

The parameter ν serves as an upper bound on the proportion of training observations that lie on the "wrong" side of the hyperplane, and is also a lower bound on the fraction of support vectors among normal training data [22], i.e., $\nu \leq N_s/l$. Parameter ν is used in this investigation instead of C, due to its clear meaning, as described above; the value of C can be easily recovered using $C = 1/\nu l$.

We, therefore, need to optimize SVM parameters (κ, ν) and propose the following novel method to do so, which exploits the nature of the physiological datasets typically acquired during patient monitoring applications:

B1: Choose a pair of parameter values (κ, ν).

B2: Use the chosen (κ, ν) to train a one-class SVM, which is dependent on a training set of "normal" data.

B3: Use the resulting SVM to classify a validation dataset, which comprises both "normal" and "abnormal" data in equal quantity.

B4: Compute *partial AUC*, defined below, using the validation results obtained in the previous step.

B5: Repeat *B1–B4* using different values of (κ, ν), typically using a grid search. Choose the (κ, ν) with the maximum partial AUC, where the latter is defined below.

The performance of a two-class decision rule can be summarized in a receiver operating characteristic (ROC) curve, which plots the true-positive rate on the vertical axis against the false-positive rate (FPR) on the horizontal axis, as the decision threshold varies. One possible comparison of different ROC curves is to consider the area-under-the-ROC-curve (AUC), which integrates the FPR over varying thresholds. AUC is independent of a fixed decision threshold and is invariant to prior class probabilities [24]. AUC represents the probability that a randomly chosen positive observation is correctly classified, and therefore, a higher value of AUC indicates better separation between the two classes. Most practical novelty detection systems require low FPRs, and so we are primarily interested in the ROC curve for low values of FPR when evaluating the performance of a novelty detector. (Its performance at higher FPRs is irrelevant, and possibly confounding, because these represent choices of decision threshold that would never be used in practice.) We, therefore, consider *partial AUC* in our proposed algorithm above, to restrict evaluation of the classifier to those ranges of decision threshold that are likely to be used in practice. Partial AUC is defined as the integral area between two FPRs [25]. Unlike AUC, whose maximum value is always 1, partial AUC depends on the two chosen FPRs, over which the ROC curve is integrated.

Note that our proposed method exploits the typical case encountered in physiological monitoring and assumes the presence of some examples of "abnormal" behavior, which are placed within the validation set for the purposes of parameter optimization. However, as noted previously, these are likely to be small in quantity compared with the number of "normal" observations, and hence, the training set is entirely comprised of "normal" data, and a one-class approach is taken.

A commonly employed alternative which uses only "normal" data [21], [26] is to vary the SVM parameters until some fixed value of the false-positive classification rate α is achieved (e.g., $\alpha = 0.05$) when presented with the training set of "normal" examples. However, as demonstrated in [12], the overall expected performance of the one-class SVM can be improved by setting parameters by taking into account any available examples of "abnormal" data that may be available, even if they are few in comparison to the number of "normal" training data. Therefore, we adopt our proposed approach and include any available "abnormal" data in our validation set. A comparison with the conventional one-class method of [21], [26] is provided in the next section.

D. Other Novelty Detection Schemes

We compare results obtained with the SVM, and its proposed training scheme, to three probabilistic methods.

The GMM is a semiparametric technique [27] and is defined by the pdf $p(\mathbf{x}) = \sum_{i=1}^{M} \pi_i\ p(\mathbf{x}|\boldsymbol{\theta}_i)$, which is comprised of M component distributions, each of which has a prior probability π_i and a likelihood $p(\mathbf{x}|\boldsymbol{\theta}_i) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, where $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ have their usual meanings of the center and covariance matrix for multivariate Gaussian i, respectively. The maximum likelihood estimates of the model parameters were determined using expectation maximization [24].

The kernel density estimate is a nonparametric method that has been used previously for vital-sign monitoring [16], which is essentially a GMM with a kernel placed on each of the training data, and where each kernel has the same (isotropic) covariance, σ.

The one-class GP is that proposed by Kemmler *et al.* [28], details of which will not be replicated here due to the limitations of space. This method uses the familiar GP classification framework [29].

E. Classifier Training Methodology

All four candidate approaches will, therefore, be trained using 4-D inputs, corresponding to HR, SpO_2, RR, and SysBP, where the former two are collected from wearable sensors. Manual observations include measurement of all four variables, although SpO_2 was measured using the pulse oximeter because no manual method exists for estimating this vital sign. Input vectors of the

[2]where the lower constraint arises because, in the worst case, we have all training data as support vectors and $N_s = l$, and therefore $C \geq 1/l$ in order for $\sum_i \alpha_i = 1$. The upper constraint arises because $\alpha_i \leq C$.

absolute values of the vital signs (after zero-mean, unit-variance normalization, using coefficients derived from the training set) were provided to the classifiers by updating the inputs whenever new data were available. This approach directly replicates the use of manual EWS systems, which perform a heuristic version of novelty detection as noted previously. Additionally, members of the clinical staff are encouraged to measure HR, RR, and SysBP using manual methods (counting pulses, counting movements of the chest wall, and use of a sphygmomanometer, respectively). For those patients with both ECG and PPG measurements, HR was estimated using the pulse oximeter to allow fair comparison with those patients who had no ECG measurements.

Thirty-seven patients were deemed by clinicians to be sufficiently "abnormal" that the patient would require clinical review. This labeling occurred retrospectively, with clinicians reviewing all manually acquired patient data, but not those data acquired from the wearable sensors. The remaining patients were thus classified as being "normal." The available "abnormal" data are insufficient to train a multiclass classifier, being small in comparison with the number of "normal" data, and therefore, the novelty detection approach is justified for this application.

The available examples of abnormality must be split between the validation set (to enable parameter optimization, as described in Section III-C) and the test set (to allow out-of-sample evaluation of the results). However, it is important that each of the 37 "abnormal" patients contributes to *either* the validation set or the test set, but not both. If one patient contributed data to both sets, the test set would no longer be independent of the training and validation sets, due to the dependence between observations for a single patient. Results could, therefore, be unfairly skewed in favor of correct classification, and any poor performance of the classifier would not be discovered until it is applied to classifying truly independent test data, from further patients. Therefore, the 37 "abnormal" patients are split equally between validation and test sets, where the partition of the "abnormal" patients into two disjoint subsets is random, giving $\{\text{validation}\} \cap \{\text{test}\} = \emptyset$ as required.

Similar numbers of "normal" data are required for each of the validation and test sets; again, no "normal" patient should contribute data to more than one set, similarly giving $\{\text{training}\} \cap \{\text{validation}\} \cap \{\text{test}\} = \emptyset$.

Table I shows how patients were assigned to each of the training, validation, and test sets. The split between the training, validation, and test sets was performed randomly. In order to test the variability of the results to this random partitioning, 50 experiments were performed, each experiment containing a different random partition of patients between the training, validation, and test sets. Each experiment, therefore, included retraining of the classifier, revalidation, and retesting, in order to obtain fully independent results for each experiment. Partial AUC was determined over the range $\text{FPR} = [0,\ 0.15]$.

F. Classifier Evaluation Methodology

There is no "gold standard" for the labeling of time-series physiological data, which makes the application of machine learning techniques to such datasets a particular challenge. For this study, retrospective clinical review of the manual observations and patient case-notes resulted in 1-h intervals that were identified as being indicative of patient deterioration, which occurred within the 37 "abnormal" patient time-series, as described previously. These 1-h intervals are, therefore, the "positive" cases that the candidate classifiers will attempt to identify. We subsequently partitioned data from the remaining 163 "normal" patients into 1-h intervals which will be treated as "negative" cases.

TABLE I
DATASET PARTITIONS, ACROSS 200 PATIENTS (COMPRISING 163 NORMAL, 37 ABNORMAL)

	Train	Validate	Test
Normal	126	18	19
Abnormal	0	18	19

All available data, both manual observations and those from patient-worn sensors when available, are provided to each of the candidate algorithms. Where data are missing or incomplete, missing channels are not provided to the classifiers, but replaced by the mean of that channel.

Note that each of the 50 experiments results in model retraining and revalidation, and the models therefore have different "optimal" novelty detection thresholds for each experiment, according to which threshold provided the best performance on the validation set for that experiment. Results on the test set for each experiment are reported in the next section. We follow previous work in this area [16] in deeming a novelty detection to have occurred if a novelty threshold is exceeded for four or more minutes in any 5-min window of data.

Defining true-positive, true-negative, false-positive, and false-negative to be TP, TN, FP, and FN, respectively, a TP will occur if a 1-h "positive" interval contains a novelty detection, or FN otherwise. Similarly, a TN will occur if a 1-h "negative" interval contains no novelty detection, or FP otherwise.

We will consider *accuracy*, defined to be (TP + TN)/(TP + TN + FP + FN), *sensitivity* as being TP/(TP + FN), and *specificity* as being TN/(TN + FP).

IV. RESULTS

A. Classifier Performance

Table II shows the overall results after 50 experiments, at the "optimal" threshold for each experiment (that threshold determined from the validation set in each of the 50 experiments). Here, we have included the results for conventional SVM parameter optimization [21], [26], referred to as "SVM-0" in the table, for comparison with results obtained using the proposed parameter optimization technique exploiting partial AUC, referred to as "SVM" in the table. The SVM using the proposed optimization method achieves the highest accuracy and partial AUC in comparison to the other methods when evaluated using the independent test data. This is confirmed by the ROC plots shown in Fig. 3, in which it may be seen that the (mean) ROC curve for the SVM is higher than that for comparator methods throughout most of the interval on the horizontal axis.

TABLE II
NOVELTY DETECTION PERFORMANCE, ± ONE STANDARD DEVIATION

Classifier	Accuracy	Partial AUC	Sensitivity	Specificity
GMM	0.90 ± 0.02	0.24 ± 0.02	0.97 ± 0.02	0.84 ± 0.05
Kernel	0.91 ± 0.02	0.26 ± 0.01	0.94 ± 0.04	0.87 ± 0.04
GP	0.90 ± 0.02	0.26 ± 0.01	0.91 ± 0.05	0.89 ± 0.04
SVM-0	0.90 ± 0.01	0.26 ± 0.02	0.92 ± 0.03	0.87 ± 0.04
SVM	0.94 ± 0.01	0.28 ± 0.03	0.96 ± 0.02	0.93 ± 0.02

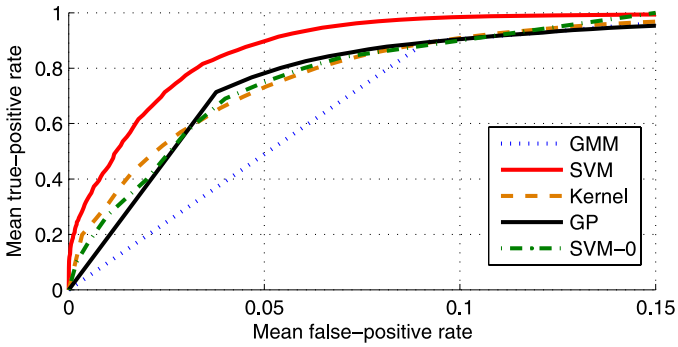

Fig. 3. ROC curve for novelty detection results. The mean of 50 experiments has been shown at each point on the ROC curve.

B. Case Studies

We now demonstrate the performance of the generative and discriminative approaches to novelty detection for predictive monitoring with case studies from "abnormal" patients who were known to deteriorate, ending with ICU readmission, and, in some cases, death. As described previously, the goal is to identify this deterioration as early as possible, to provide maximum opportunity for preventative action to be taken in advance of subsequent emergency conditions.

An example of the application of the techniques to patient vital-sign data is shown in Figs. 4 and 5. The first example shows an "abnormal," deteriorating patient for whom manual observations were taken throughout the patient stay. Only the fifth set of observations (indicated by the black box) caused the conventional EWS system to alert. Excursions of abnormally high HR peaking at 130 beats/min prior to this were not observed by staff (the abnormality falls between the third and fourth manual observations, shortly after 18.00 hours). However, this deterioration is clearly represented by increases in novelty scores for both the SVM and GMM. It may be seen that the scores for the kernel estimate and GP are constantly above threshold for large periods of the interval shown.

The remainder of the manual observations for this patient were deemed "normal" by the manual EWS system, but increasingly frequent desaturations in SpO_2 may be seen throughout the time-series (decreasing as low as 84%, which is highly abnormal), while periods of tachycardia (elevated HR) increasing to approximately 130 beats/min were not observed by the manual method. The patient was immediately admitted to the ICU under emergency conditions after the period shown in the figure. While the majority of time for this patient was considered "normal" by the conventional EWS system, frequent corresponding increases of the novelty scores of the SVM and GMM

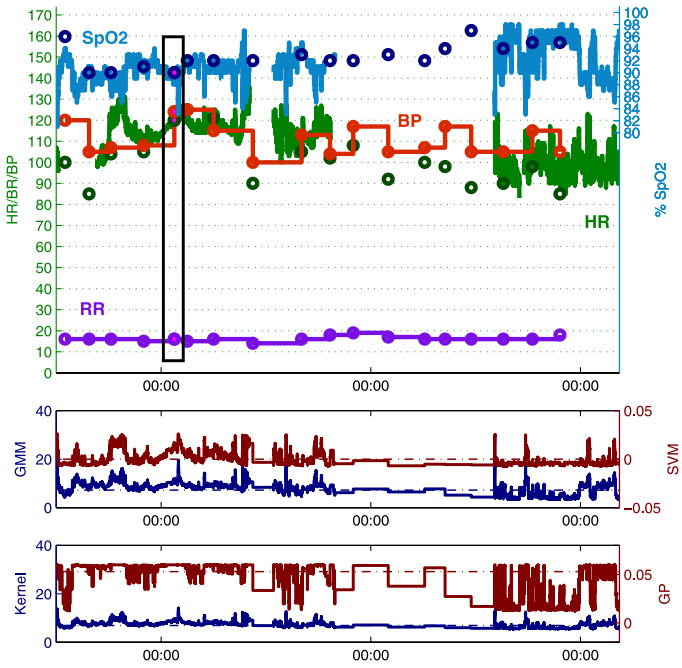

Fig. 4. Upper plot shows time-series of vital signs for an exemplar patient, showing HR, RR, SpO_2, and BP in green, purple, blue, and red, respectively, with time (in hours, with midnights of successive days marked as 00:00) shown on the horizontal axis. The lower plots show novelty scores derived from GMM and kernel density outputs $-\log p(\mathbf{x})$, SVM output $z(\mathbf{x})$, and GP output on the same time-base. Horizontal lines in the lower plots show the decision thresholds for each classifier. Manual observations are shown using circles. (Note that all RR and SysBP data are manually observed, while the time-series of HR and SpO_2 are continuous data from wearable sensors.)

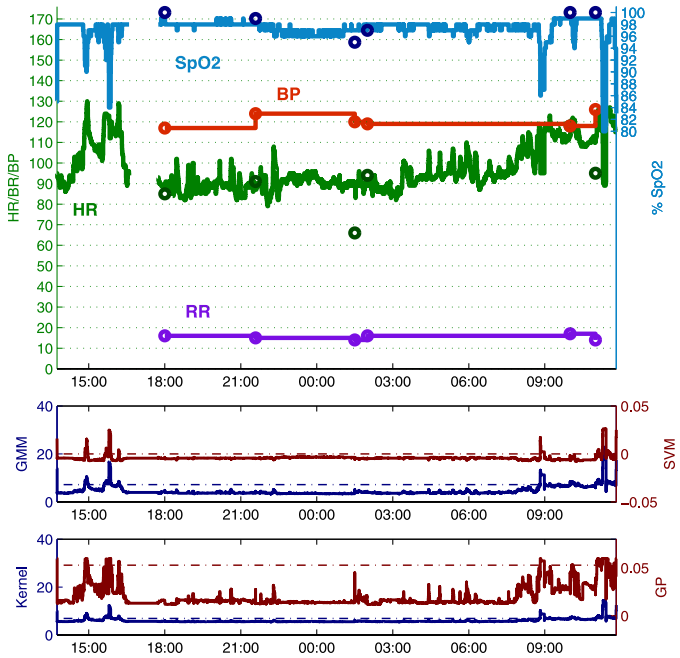

Fig. 5. Upper plot shows time-series of vital signs for a second exemplar patient, showing vital signs and novelty detection output as in the first example.

may be seen throughout the time-series, indicating that these periods of deterioration were successfully identified by the classifiers acting on the continuous data acquired from wearable sensors.

We observe in passing that the similarity of the GMM, kernel estimate, and SVM output is not accidental, as the $-\log p(\mathbf{x})$ scaling of the GMM and kernel density output makes it a comparable score to the SVM $z(\mathbf{x})$, because the latter asymptotically approaches the level sets on the pdf in its tails [30].

The second example (see Fig. 5) shows a patient who is similarly unstable at the start of their admission to the Cancer Centre ward, following surgery. This patient exhibits immediate desaturations in SpO_2, decreasing to approximately 85%, and sustained tachycardia increasing to approximately 130 beats/min. However, the first manual observation for this patient does not occur until four hours into the period shown, and these physiologically abnormalities are not observed by the manual method.

All of the manual observations made for this patient were deemed to be "normal" by the conventional EWS system. However, this patient died immediately after the period shown in the figure. Both the initial deterioration at the start of the time-series and the elevated HR and desaturations at the end of the time-series were correctly identified by all four novelty detection methods, as indicated by the increase of their outputs over their corresponding decision thresholds. In both examples, the novelty detection methods used to classify the continuously acquired data from wearable sensors identify deterioration in abnormal patients, which is not identified by existing manual methods. This demonstrates that predictive monitoring is feasible using mobile sensors and offers significant advantages to manual observation of the patient, which is the current standard of care in many hospitals.

V. Conclusions and Discussion

Advances in principled approaches to predictive patient monitoring have been limited by the difficulty of collecting physiological data from a mobile population of patients. This has been demonstrated in the context of our study by the technological and clinical (and, in the U.K., ethical) obstacles that must be overcome. For the 200 patients that were studied, with an average length-of-stay of nine days, the average time that wearable health monitors were worn by was five days. Patient compliance was generally high, with patients being informed of the potential benefits of wearing their sensors, in terms of identifying any deterioration in their condition. Even so, ECG sensors were deemed to be unacceptably uncomfortable for prolonged wear, such that the sensors had to be removed from the study. While finger-mounted pulse oximeters were more acceptable to patients, the devices were frequently removed and often not returned to the finger.

Data dropout was a significant challenge, mainly due to infrastructure problems (interruptions in the hospital wi-fi service) or expired batteries. The ECG sensor had the bare minimum battery life required for use on the ward (at approximately 24 h), such that nurses could change the device once per day. Any shorter battery life would require several changes per day, which is deemed unrealistic for clinical practice. However, the actual quantity of data ultimately collected was large.

We note that we have used manually observed estimates of blood pressure and RR. On-going work aims to provide robust methods for determining the latter from the ECG and PPG waveforms acquired from the ECG sensors and pulse oximeter, respectively. Work exists in this area [31], but trial implementations have demonstrated that resulting RR estimates are not robust, and cannot yet be used in clinical practice without further improvement of the estimation algorithms.

We have demonstrated that automated methods can be used to identify patient deterioration, fulfilling the aim of predictive monitoring, and automatically parse the large quantities of data acquired from the trial. We have shown that such methods accurately identify "abnormal" physiological data, arising due to patient deterioration, which makes mobile approaches to predictive monitoring more realistic. We have proposed a parameter-estimation method for the SVM that takes advantage of the type of data encountered in patient vital-sign monitoring, exploiting the notion that the classifier performance is only relevant within a subset of the AUC curve conventionally used for parameter selection, and which has been demonstrated to outperform other methods over the large quantity of clinical data that we have acquired.

The results of automated novelty detection show that an FPR ($1 -$ specificity) between 7% and 16% per patient-hour. These results compare favorably with those of, for example, a candidate manual EWS system for national adoption in the U.K., which has an FPR of approximately 20% [32]. We note that, as with EWS systems, the availability of clinical resources would allow a different "operating point" to be adopted by changing the novelty threshold—that is, each system could be made more or less sensitive by adjusting its novelty threshold, as is performed by changing the threshold score in EWS systems.

The on-going next phase of the clinical study will result in further data on which to confirm these preliminary findings, and aims to determine if patient outcomes are improved by revealing the output of the machine learning process to ward nurses, online, during the patient stay on the ward.

This next phase of the work makes possible the extension of the predictive monitoring described in this article to *personalized* predictive monitoring, whereby novelty detection may be performed using models constructed from the patient's own physiology. This approach is of particular interest in the high-risk group of mobile patients described in this study, while they are recovering from upper GI surgery, and where the response of each patient to surgery is likely to differ significantly between individuals. However, the construction of models of normality requires significant quantities of data, and it may be that a suitable approach to take is one in which prior models of patient condition are used initially (when few examples of patient-specific data have been collected), which are then used as the basis for posterior models that take into account the subsequently observed patient data. It is anticipated that the models constructed using data from the predictive monitoring study described in this paper could form the basis for such prior models in the personalized setting.

ACKNOWLEDGMENT

The authors wish to thank S. Vollam, D. Evans, and T. Saunders for the collection of clinical data used in this investigation.

REFERENCES

[1] S. Martin, G. Kelly, W. Kernohan, B. McCreight, and C. Nugent, "Smart home technologies for health and social care support," *Cochrane Database Syst. Rev.*, vol. 4, pp. 1–11, 2008.

[2] G. Clifford and D. Clifton, "Annual review: Wireless technology in disease state management and medicine," *Annu. Rev. Med.*, vol. 63, pp. 479–492, 2012.

[3] L. Tarassenko and D. Clifton, "Semiconductor wireless technology for chronic disease management," *Electron. Lett.*, vol. S30, pp. 30–32, 2011.

[4] C. Tsien and J. Fackler, "Poor prognosis for existing monitors in the intensive care unit," *Crit. Care Med.*, vol. 25, no. 4, pp. 614–619, 1997.

[5] National Patient Safety Association, "Safer care for acutely ill patients: Learning from serious accidents," Tech. Rep., 2007.

[6] National Institute for Clinical Excellence, "Recognition of and response to acute illness in adults in hospital," Tech. Rep., 2007.

[7] L. Tarassenko, D. Clifton, M. Pinsky, M. Hravnak, J. Woods, and P. Watkinson, "Centile-based early warning scores derived from statistical distributions of vital signs," *Resuscitation*, vol. 82, no. 8, pp. 1013–1018, 2011.

[8] A. Pantelopoulos and N. Bourbakis, "A survey on wearable sensor-based systems for health monitoring and prognosis," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 40, no. 1, pp. 1–12, Jan. 2010.

[9] J. Lahteenmaki, J. Leppanen, A. Orsama, V. Salaspuro, J. Pinnen, M. Sormunen, H. Kaijanranta, and M. Ermes, "Remote patient monitoring system with decision support," in *Proc. 8th IASTED Int. Conf. Biomed. Eng.*, 2011, pp. 491–495.

[10] S. Meystre, "The current state of telemonitoring: A comment on the literature," *Telemed. e-Health*, vol. 11, no. 1, pp. 63–69, 2005.

[11] V. Nangalia, D. Prytherch, and G. Smith, "Health technology assessment review: Remote monitoring of vital signs—current status and future challenges," *Crit. Care*, vol. 14, no. 5, pp. 1–8, 2010.

[12] P. Hayton, L. Tarassenko, B. Schölkopf, and P. Anuzis, "Support vector novelty detection applied to jet engine vibration spectra," in *Proc. Adv. Neural Inf. Process. Syst.*, London, U.K., 2000, pp. 946–952.

[13] A. Gretton and F. Desobry, "On-line one-class support vector machines: An application to signal segmentation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Hong Kong, 2003, pp. 709–712.

[14] D. R. Hardoon and L. M. Manevitz, "fMRI analysis via one-class machine learning techniques," in *Proc. 19th Int. Joint Conf. Aritif. Intell.*, Edinburgh, U.K., 2005, pp. 1604–1605.

[15] M. Markou and S. Singh, "Novelty detection: A review—Part 2: Neural network based approaches," *Signal Process.*, vol. 83, no. 12, pp. 2499–2521, 2003.

[16] A. Hann, "Multi-parameter monitoring for early warning of patient deterioration" Ph.D. dissertation, Univ. Oxford, Oxford, U.K., 2008.

[17] J. Quinn, C. Williams, and N. McIntosh, "Factorial switching linear dynamical systems applied to physiological condition monitoring," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 9, pp. 1537–1551, Sep. 2009.

[18] L. Clifton, D. Clifton, P. Watkinson, and L. Tarassenko, "Identification of patient deterioration in vital-sign data using one-class support vector machines," in *Proc. Comput. Sci. Inf. Syst.*, 2011, pp. 125–131.

[19] J. Marcos, R. Hornero, D. Alvarez, I. Nabney, F. del Campo, and C. Zamarron, "The classification of oximetry signals using Bayesian neural networks to assist in the detection of obstructive sleep apnoea syndrome," *Physiol. Meas.*, vol. 31, pp. 375–394, 2010.

[20] O. Stegle, S. Fallert, D. MacKay, and S. Brage, "Gaussian process robust regression for noisy heart rate data," *IEEE Trans. Biomed. Eng.*, vol. 55, no. 9, pp. 2143–2151, Sep. 2008.

[21] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*, 1st ed. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[22] B. Schölkopf, J. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the support of a high-dimensional distribution," *Neural Comput.*, vol. 13, no. 7, pp. 1443–1471, 2001.

[23] C. M. Bishop, "Novelty detection and neural network validation," *Proc. IEE Conf. Vision Image Signal Process.*, vol. 141, no. 4, pp. 217–222, 1994.

[24] C. M. Bishop, *Pattern Recognition and Machine Learning*. Berlin, Germany: Springer-Verlag, 2006.

[25] S. H. Park, J. M. Goo, and C. H. Jo, "Receiver operating characteristic (ROC) curve: Practical review for radiologists," *Korean J. Radiol.*, vol. 5, no. 1, pp. 11–18, 2004.

[26] B. Schölkopf and A. Smola, *Learning with Kernels*. Cambridge, MA, USA: MIT Press, 2002.

[27] I. Nabney, *Netlab: Algorithms for Pattern Recognition*, 1st ed. London, U.K.: Springer-Verlag, 2002.

[28] M. Kemmler, E. Rodner, and J. Denzler, "One-class classification with Gaussian processes," in *Proc. 10th Asian Conf. Comput. Vision*, 2011, pp. 489–500.

[29] C. Rasmussen and C. Williams, *Gaussian Processes for Machine Learning*. Cambridge, MA, USA: MIT Press, 2006.

[30] R. Vert and J. Vert, "Consistency and convergence rates of one-class SVMs and related algorithms," *J. Mach. Learn. Res.*, vol. 7, pp. 817–854, 2006.

[31] C. Orphanidou, D. Clifton, M. Smith, J. Feldmar, and L. Tarassenko, "Telemetry-based vital-sign monitoring for ambulatory hospital patients," in *Proc. IEEE Eng. Med. Biol. Conf.*, Minneapolis, MN, USA, 2009, pp. 4650–4653.

[32] G. Smith, D. Prytherch, P. Schmidt, and P. Featherstone, "Review and performance evaluation of aggregate "track and trigger" systems," *Resuscitation*, vol. 77, pp. 170–179, 2008.

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