Probabilistic Novelty Detection with Support Vector Machines

Lei Clifton*, David A. Clifton*, *Member, IEEE*, Yang Zhang†, Peter Watkinson‡, Lionel Tarassenko*, and Hujun Yin§, *Senior Member, IEEE*

*Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, Oxford, UK.
†Department of Mechanical Engineering, University of Sheffield, Sheffield, UK.
‡Nuffield Department of Clinical Neurosciences, University of Oxford, Oxford, UK.
§School of Electrical and Electronic Engineering, University of Manchester, Manchester, UK.

***Abstract*—Novelty detection, or one-class classification, is of particular use in the analysis of high-integrity systems, in which examples of failure are rare in comparison with the number of examples of normal behaviour, such that a conventional multi-class classification approach cannot be taken. Support Vector Machines (SVMs) are a popular means of performing novelty detection, and it is conventional practice to use a *train-validate-test* approach, often involving cross-validation, to train the one-class SVM, and then select appropriate values for its parameters. An alternative method, used with multi-class SVMs, is to calibrate the SVM output into conditional class probabilities. A probabilistic approach offers many advantages over the conventional method, including the facility to select automatically a probabilistic novelty threshold. The contributions of this paper are (i) the development of a probabilistic calibration technique for one-class SVMs, such that on-line novelty detection may be performed in a probabilistic manner; and (ii) the demonstration of the advantages of the proposed method (in comparison to the conventional one-class SVM methodology) using case studies, in which one-class probabilistic SVMs are used to perform condition monitoring of a high-integrity industrial combustion plant, and in detecting deterioration in patient physiological condition during patient vital-sign monitoring.**

***Index Terms*—Support vector machine, novelty detection, one-class classification, calibration, condition monitoring.**

ABBREVIATIONS

EHM	Engine Health Monitoring
SVM	Support Vector Machine
MLP	Multi-Layer Perceptron
PAV	Pair-Adjacent Violators
GMM	Gaussian Mixture Model
ROC	Receiver Operating Characteristic
AUC	Area-under-the-Curve
ROC	Receiver Operating Characteristic
SDU	Step-Down Unit
ICU	Intensive Care Unit

NOTATION

$\mathbb{R}^d$	d-dimensional feature space
$\mathbb{F}$	a feature space by a non-linear transformation $\mathbf{\Phi}$: $\mathbb{R}^d \rightarrow \mathbb{F}$
$\mathbf{s}_i$	support vectors
N_s	number of support vectors $\mathbf{s}_i$; i.e, $i = 1 \ldots N_s$
k	kernel function
σ	bandwidth of a multivariate Gaussian
α_i	Lagrangian multiplier
$z(\mathbf{x})$	novelty score
$s(\mathbf{x})$	re-scaled novelty score in the range $[0, 1]$
$\mathbb{A}$	data available during classifier construction that are normal
$\mathbb{B}$	data available during classifier construction that are abnormal
r	average distance to the centroid of real normal data $\mathbf{x}_i \in \mathbb{A}$
r_a	radius in a hyper-sphere
N	number of real normal data $\mathbf{x}_i$; i.e, $i = 1 \ldots N$
$\hat{\mathbb{A}}$	artificial normal data
$\hat{\mathbb{B}}$	artificial abnormal data
b_o	offset of an optimal hyperplane
p	probability densities
P	probabilities
$g^*(\mathbf{x})$	stepwise-constant isotonic function
$M(\boldsymbol{\theta})$	model of normality with parameters $\boldsymbol{\theta}$

I. INTRODUCTION

A. Novelty Detection for On-Line Monitoring

Novelty detection, also termed one-class classification, involves construction of a model of normality using examples of normal system behaviour, and then classifies test data as either normal or abnormal with respect to that model. This technique is particularly applicable to the monitoring of high-integrity systems, such as jet engines, human patients, or manufacturing processes, in which examples of abnormal system behaviour are scarce in comparison to the number of examples of normal system behaviour, due to the reliability of such systems. In such cases, there are typically too few examples of system faults to be used to construct a robust two- or multi-class classifier, as would be used in a conventional "fault detection" approach to condition monitoring. For example, an engine health monitoring (EHM) system may have available many thousands of hours of data acquired from normal flying time with very few examples of failure throughout its operational life.

The difficulties of monitoring high-integrity systems are further compounded by inter-system variability. Different instances of the same system type (such as different patients of the same age) are often so different in their observed data that examples of abnormal system behaviour from one instance are inapplicable to the condition of other instances. For example, a heart rate of 50 beats per minute may be indicative of considerable physiological abnormality in one hospital patient, while it may be entirely normal for a healthier patient of the same age and background.

Finally, high-integrity systems typically exhibit a high de-

gree of structural complexity, and can often comprise many millions of components and sub-systems that interact in a non-linear manner. Thus, the potential space of "abnormality" is extremely large, and so the large resultant number of failure modes is often poorly understood.

Novelty detection avoids such problems by modelling the normal mode of operation of the system, which is often well-understood because most high-integrity systems function "normally" most of the time. They then seek to identify deviations from that normal model. Such condition monitoring applications are of particular interest to the techniques developed within this paper.

B. Overview

This paper considers the one-class support vector machine (SVM), which is a commonly-used method of performing novelty detection. Its formulation is briefly recapped in Section II, where its disadvantages with respect to probabilistic methods are discussed. A probabilistic extension to the one-class SVM is proposed, which requires the generation of artificial data using the available normal data. Section III describes methods for generating such data, which are then used for calibrating the output of the SVM into estimates of the class-conditional probabilities, in Section IV. The proposed probabilistic approach is illustrated using both simulated data, and real-world case studies. The latter include data from monitoring a large industrial combustion system, and from the analysis of patient vital signs, described in Sections V, and VI, respectively. Limitations of the method, and potential future extensions, are also discussed in Section VII.

II. One-Class SVMs

The SVM is an oft-employed method of performing novelty detection, and it has been applied to many such problems, including jet engine condition monitoring [1], signal segmentation [2], and fMRI analysis [3], among many others, a review of which may be found in [4].

A. Formulation

This paper considers the one-class SVM formulation proposed by [5], in which a number 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 non-linear transformation $\mathbf{\Phi}$: $\mathbb{R}^d \rightarrow \mathbb{F}$, where the data are linearly separable from the origin in $\mathbb{F}$. We note in passing that this approach is typically employed in favour of the one-class formulation proposed in [6] and [7], in which a hypersphere of minimum radius is found to enclose the data in $\mathbb{F}$, but which can be considered to be equivalent to the separation from the origin in the feature space.

We here define the output of the one-class SVM in the following manner, to be interpreted as a novelty score,

$$z(\mathbf{x}) = \rho_0 - \mathbf{w}_o \cdot \mathbf{\Phi}(\mathbf{x}) \quad (1)$$

$$= \rho_0 - \sum_{i=1}^{N_s} \alpha_i k(\mathbf{s}_i, \mathbf{x}) \quad (2)$$

with parameters

$$\mathbf{w}_o = \sum_{i=1}^{N_s} \alpha_i \mathbf{\Phi}(\mathbf{s}_i) \quad (3)$$

$$\rho_o = \frac{1}{N_s} \sum_{j=1}^{N_s} \sum_{i=1}^{N_s} \alpha_i k(\mathbf{s}_i, \mathbf{s}_j), \quad (4)$$

where $\mathbf{s}_i$ are the support vectors, of which there are N_s, and where k is a kernel function, typically a multivariate Gaussian with bandwidth σ, which provides the dot product of transformed data in $\mathbb{F}$:

$$k(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{\Phi}(\mathbf{x}_i) \cdot \mathbf{\Phi}(\mathbf{x}_j) \quad (5)$$

$$= \exp\Big[- \|\mathbf{x}_i - \mathbf{x}_j\|^2 / (2\sigma^2) \Big]. \quad (6)$$

$w_o \in \mathbb{F}$, $\rho_o \in \mathbb{R}$, and that α_i are Lagrangian multipliers used to solve the dual formulation, more details of which may be found in [5], and which are not reproduced here. Thus, abnormal data (i.e., those outside the single, normal training class) take higher values of $z(\mathbf{x})$ than those fornormal data, and hence $z(\mathbf{x})$ is a novelty score.

B. Disadvantages of the Conventional Formulation

The one-class SVM, as with the conventional multi-class SVM, has been repeatedly demonstrated to perform well at the task of separating classes within example datasets. Typically, it outperforms probabilistic methods in retrospective studies (in terms of misclassification rate), when on-line analysis is not performed [8]. However, its non-probabilistic formulation gives rise to several disadvantages with respect to probabilistic methods [9], [10].

- Uncertainty in classification is not modelled. This is of importance in the monitoring of high-integrity systems, for example, because it may be advantageous for a decision to be made only if the certainty in classification is sufficiently high. Monitoring systems that provide "don't know" outputs can often be more accepted in practice [11], due to their appropriate handling of uncertainty, which can reduce false-positive classifications (at the expense of reduced sensitivity to system abnormality).
- Not explicitly modelling the uncertainty in the classification makes it difficult to obtain "error bars" on the output, as would be required, for example, in a decision support system [12].
- Cross-validation is required to set model parameters, such as the novelty threshold on the novelty score $z(\mathbf{x})$. This is computationally expensive, and can require the "holding out" of data, such that the quantity of data available for training is decreased. In many condition monitoring applications, the rate of data acquisition is low, and hence, if a model is being learned on-line, all available data must be used for training if the model is to be able to sufficiently characterise normal system behaviour. For example, a common application is the monitoring of jet engines, in which a very small number of data are broadcast from the aircraft to a ground-based monitoring station via limited-bandwidth satellite-

or airport-based communications systems [13], [14]. This limited quantity of data must be used to learn a model of normality on-line, such that monitoring can take place after completion of a small number of flights. Similarly, in the home monitoring of patients with chronic illnesses, a typical "m-health" monitoring system may acquire measurements of vital signs (such as blood pressure and temperature) daily or twice-daily, from which meaningful models of normality must be constructed on-line [15], [16].

- Finally, probabilistic approaches allow the use of peripheral classification techniques, such as probabilistic feature selection [17] and probabilistic combination-of-classifiers [18]–[20].

Several probabilistic extensions have been proposed for multi-class SVMs, which will be reviewed in Section IV, while other techniques such as Relevance Vector Machines (RVMs) have sought to embed probabilistic, sometimes Bayesian, methodologies into the multi-class kernel machine framework [10], [21], [22]. This paper develops probabilistic calibration techniques for one-class SVMs, suitable for performing on-line novelty detection, which will be described in Section IV. Prior to that, we introduce the notion of artificial data generation in Section III, which will be required by our probabilistic calibration method.

III. Generating Artificial Calibration Data

Suppose the range of novelty scores $z(\mathbf{x})$ produced by a one-class classifier is $[-a, a]$, then $z(\mathbf{x})$ may be mapped onto the range $[0, 1]$ through a simple linear re-scaling

$$s(\mathbf{x}) = \frac{z(\mathbf{x}) + a}{2a}, \tag{7}$$

where $s(\mathbf{x})$ is the re-scaled score in the range $[0, 1]$. However, $s(\mathbf{x})$ tends to be poorly calibrated because the novelty score $z(\mathbf{x})$ may not be proportional to the actual probability of the sample being abnormal [23]. Different approaches to calibration must therefore be considered.

A. Justification of the Need for Artificial Data

In the following, we here denote sets $\mathbb{A}$ and $\mathbb{B}$ to refer to those data available during classifier construction that are normal and abnormal, respectively. Typically, in novelty detection application, $\mathbb{B}$ is empty or under-represented.

Previous studies of two-class SVMs have shown that calibration of output into probabilities requires a separate validation set [24], which comprises a suitable number of normal and abnormal examples (i.e., $\mathbb{A}$ and $\mathbb{B}$ of suitable size). However, as noted in Section I, novelty detection applications typically have very few examples in $\mathbb{B}$.

Furthermore, for small datasets such as those that may occur in some low-bandwidth condition monitoring applications, as discussed in the previous section, one cannot afford to further split the training set of normal examples $\mathbb{A}$ to form a validation set suitable for calibration.

Thus, with an under-represented (or frequently empty) abnormal set $\mathbb{B}$, and a potentially limited normal set $\mathbb{A}$, an alternative approach must be taken to probabilistic calibration.

In this section, we provide a possible solution to the problem by generating artificial samples to be used as a dataset suitable for calibration of the novelty scores.

We will describe the generation of artificial normal data $\hat{\mathbb{A}}$, and artificial abnormal data $\hat{\mathbb{B}}$. These sets must be based entirely on the real, available normal data $\mathbb{A}$ acquired from the system, as we cannot assume the existence of a non-empty $\mathbb{B}$, nor can we assume that what limited examples we may have in $\mathbb{B}$ provide a complete understanding of abnormal conditions for our system. (This latter effect is due to inter-system variability, and the large numbers of failure modes that may occur for a complex system, as described in Section I.)

B. Generating Normal Data $\hat{\mathbb{A}}$, and Abnormal Data $\hat{\mathbb{B}}$

Generating artificial abnormal data has been used in novelty detection problems where only normal data are available [25], [26]. Generating a compact set of artificial abnormal data can avoid dealing with excessive quantities of empty feature space around the normal data [26]. In [25], uniformly-distributed artificial abnormal data were generated to surround a hypersphere of normal data, where the latter were assumed to have a multivariate, unimodal Gaussian distribution. A similar approach was employed in [26], whereby artificial abnormal data generated around normal data were used to form closed decision boundaries for a multi-layer perceptron (MLP). We note that, in all such work, the artificial data have been used merely to ensure that a decision boundary can be constructed around the single, known class of normal data; hence, no strong assumption is made about the abnormal class, other than that it lies outside the locus of normal data.

Let r be the average Euclidean distance of the real normal data $\mathbf{x}_i \in \mathbb{A}$ to their centroid $\mathbf{c}$, and let N be the number of those real normal data; i.e, $i = 1 \ldots N$. The procedure for generating artificial data consists of the following four steps, which have been adapted from [25], [26] to include the generation of artificial normal data $\hat{\mathbb{A}}$.

STEP 1. Generate uniformly distributed artificial data *inside* and *around* real normal data $\mathbb{A}$ in a hyper-sphere of radius r_a. We have chosen $r_a = 2r$ to generate a compact set of data outside the real normal data, as required in step 4 below. Results from experiments not shown here indicate that $r_a > 2r$ does not improve the performance of the calibration using the one-class SVM classifier, because this simply involves the generation of data further out into the already "abnormal" regions of the data space.

STEP 2. For each real normal $\mathbf{x}_i \in \mathbb{A}$ define a local average distance quantity Δ_l to be the average of the Euclidean distance of its k-nearest neighbours, $k = \sqrt{N}$. The global average distance Δ_g is found by averaging Δ_l over all the real normal data $\mathbb{A}$.

STEP 3. For each artificial data-point, find the distance Δ_a from its nearest neighbour among the real normal data $\mathbb{A}$.

STEP 4a. Those artificial data with $\Delta_a \geq \Delta_g(1 + d_a)$ lie outside the locus of real normal data $\mathbb{A}$, and are thus used to form the artificial abnormal set $\hat{\mathbb{B}}$.

The parameter $d_a \in \mathbb{R}^+$ controls the boundary space between the real normal data $\mathbb{A}$ and the artificial abnormal

data $\hat{\mathbb{B}}$; a greater value of d_a increases the separation between the two sets.

STEP 4b. Those artificial data with $\Delta_a < \Delta_g(1 - d_n)$ are used to form the artificial normal set $\hat{\mathbb{A}}$. The parameter $d_n \in [0, 1]$ is usually a small positive value, which determines how close to the boundary of the locus of real normal data $\mathbb{A}$ that the artificial normal data $\hat{\mathbb{A}}$ are permitted to approach. As $d_n \to 0$, the artificial "normal" data $\hat{\mathbb{A}}$ reach the edge of the locus of the real "normal" data $\mathbb{A}$.

The values of parameters d_n and d_a are additional quantities that one may optimise during the modelling process, and which may be determined with prior knowledge of the domain or with knowledge of any abnormal examples that may exist; in the case studies considered later, we follow the analogous method described in [26] by setting $d_n = 0.1$, and $d_a = 0.5$.

We note in passing that there is no strong link between the method used to generate artificial data and the algorithm used ultimately to perform novelty detection. It is feasible, for example, that we might use the above algorithm to generate artificial data and then re-use parts of the algorithm to form a k-nearest neighbour analysis for eventual novelty detection. However, the focus of the work described by this paper is to demonstrate the benefit of probabilistic output for one-class SVMs, due to the traditionally superior classification performance of the latter with respect to many other methods [8].

IV. Calibrating Novelty Scores into Probabilities

In this section, we investigate several methods of calibrating novelty scores $z(\mathbf{x})$ into class-conditional probabilities. The most popular methods in the literature for achieving this calibration (with multi-class classifiers) include *sigmoid fitting*, *binning*, and *isotonic regression*. The first two are briefly recapped in subsections IV-A, and IV-B, respectively.

A. Sigmoid Fitting

In [27], and recently refined in [28], a sigmoid function is used to map the output of a two-class SVM classifier onto probabilities. With training data labelled according to $\{\mathbf{x}_i, y_i\}$, $i = 1, \ldots, n$, $y_i \in \{-1, 1\}$, $\mathbf{x}_i \in \mathbb{R}^d$, a typical unthresholded output $z(\mathbf{x})$ of a two-class SVM classifier [29] is

$$z(\mathbf{x}) = \sum_{i=1}^{N_s} \alpha_i y_i k(\mathbf{s}_i, \mathbf{x}) + b_o, \tag{8}$$

where b_o is the offset of an optimal hyperplane.

If we are to apply the method to one-class SVM classifiers for the purposes of novelty detection, this technique requires further evaluation. It is reported in [23] that, for some data sets, the sigmoid shape does not appear to fit two-class naïve Bayes scores as closely as it fits two-class SVM scores. Furthermore, the sigmoid-fitting method is based on the assumption that class-conditional densities $p(z \mid y = \pm 1)$ are exponential, motivated by the empirical distribution of many data sets [23], [27]. For different data sets, the suitability of a calibration method needs to be validated using a reliability diagram, which visualises the calibration of a classifier [30]. For each score $s \in [0, 1]$, the empirical probability $P(C|s(\mathbf{x}) = s)$ can be calculated as being the number of data with score s that belongs to class C divided by the number of all samples with score s. If the classifier is well-calibrated, the plot of s versus $P(C|s(\mathbf{x}) = s)$ will be the line $y = x$, meaning that the scores are equal to the empirical probabilities. However, for many of the novelty detection datasets that we have considered, such as those described later in this paper, the reliability diagram shows that the above assumption is not reasonable [8].

B. Binning

A histogram method, also termed *binning*, can be used to obtain calibrated probabilities from a naïve Bayesian classifier [31]. The training samples are first sorted according to their scores; the sorted set is then divided into b subsets of equal size, called *bins*. For each bin B_i ($i = 1, \ldots, b$), the lower and upper boundary of scores within that bin are calculated. Given a test example $\mathbf{x}$ from B_i, the probability that $\mathbf{x}$ belongs to class C is estimated as the fraction of all samples in B_i that belong to class C.

The number of bins b determines the number of different probability estimates, and must be small enough to reduce the variance of the probability estimates [31]. For instance, $b = 10$ was used in [31]. Compared with uncalibrated scores, binning improves the accuracy of probability estimates by reducing both variance and bias, at the price of reduced resolution of probability estimates.

The binning method is a non-parametric method which does not make any assumption about the mapping function between the output scores and probabilities, but it has several disadvantages [23]. First, the number of bins b is normally chosen by cross-validation. However, cross-validation often fails to indicate the optimal value of b if the training set is too small or unbalanced, as may be the case in novelty detection applications. Second, the size of the bins is fixed, and the position of boundaries are calculated accordingly. If two examples from the same bin are required to have different probability estimates, the binning method may fail to produce meaningfully calibrated novelty scores. We have chosen not to use the binning method for our calibration task due to these shortcomings.

C. Isotonic Regression

In [23], and further examined in [32], it was proposed to use an intermediary approach between sigmoid fitting and binning, termed isotonic regression. The latter is a non-parametric form of regression, with the restriction that the mapping from novelty scores $z(\mathbf{x})$ into probabilities is isotonic (i.e., non-decreasing). Essentially, we wish to ensure that data are ranked correctly: a higher novelty score should always correspond to a higher probability of belonging to the abnormal class, and vice versa.

The pool- or pair-adjacent violators (PAV) algorithm is employed to perform the isotonic regression, which finds the stepwise-constant isotonic function $g^*(\mathbf{x})$ that fits the data according to a mean-squared-error criterion. This algorithm allows us to identify those cases in which correct ranking is not

taking place; that is, the PAV algorithms ensures that higher novelty scores always correspond to higher probabilities of being "abnormal".

Let $\mathbf{x}_i$ ($i = 1, \ldots, n$) be the training examples from normal and abnormal classes $\{C_0, C_1\}$, let $g(\mathbf{x}_i)$ be the value of the function to be learned for each training sample, and let $g^*(\mathbf{x})$ be the isotonic function obtained from isotonic regression. The PAV algorithm works as follows.

STEP 1. Sort the examples $\mathbf{x}_i$ according to their novelty scores $z(\mathbf{x}_i)$, in ascending order. Initialise $g(\mathbf{x}_i) = 0$ if $\mathbf{x}_i \in C_0$, or $g(\mathbf{x}_i) = 1$ if $\mathbf{x}_i \in C_1$. It is most likely that at this stage there will be several cases in which higher novelty scores do *not* correspond to higher probabilities of being "abnormal", and vice versa.

STEP 2. If $g(\mathbf{x}_i)$ is isotonic, then return $g^*(\mathbf{x}_i) = g(\mathbf{x}_i)$. Else, proceed to STEP 3.

STEP 3. Find a subscript i such that $g(\mathbf{x}_i) > g(\mathbf{x}_{i+1})$. The examples $\mathbf{x}_i$, and $\mathbf{x}_{i+1}$ are called pair-adjacent violators. These are pairs that violate our requirement that increasing novelty scores correspond to increasing probabilities of abnormality. We then replace $g(\mathbf{x}_i)$ and $g(\mathbf{x}_{i+1})$ with their average

$$g^*(\mathbf{x}_i) = g^*(\mathbf{x}_{i+1}) = \left[g(\mathbf{x}_i) + g(\mathbf{x}_{i+1})\right]/2. \tag{9}$$

This replacement removes the conflict, by smoothing the cdf (by introducing a quantisation in the probability, according to the average described above).

STEP 4. Set $g(\mathbf{x}_i)$ as the new $g^*(\mathbf{x}_i)$. Proceed to STEP 2.

Thus, $g^*(\mathbf{x})$ is a step-wise constant function which consists of horizontal intervals, and may be interpreted as $P(C_1|\mathbf{x})$, the probability that sample $\mathbf{x}$ is abnormal. For a test example $\mathbf{x}$, we first find the interval to which its score $z(\mathbf{x})$ belongs. Then we set the value of $g^*(\mathbf{x})$ in this interval to be $P(C_1|\mathbf{x})$, the probability estimate of C_1 given $\mathbf{x}$.

If the scores rank all examples correctly, then all class C_0 examples will appear before all class C_1 examples in the sorted data set in STEP 1. The calibrated probability estimate $g^*(\mathbf{x}) = 0$ for class C_0, and $g^*(\mathbf{x}) = 1$ for class C_1. Conversely, if the scores do not provide any information, $g^*(\mathbf{x})$ will be a constant function, taking the value of the average score over all examples in class C_1.

The PAV algorithm used in isotonic regression may be viewed as a binning algorithm, in which the position and the size of the bins are chosen according to how well the classifier ranks the samples [23]. Therefore, the isotonic regression overcomes the previously-described drawbacks of the binning method.

D. Obtaining and Using the Calibration Function

Isotonic regression may now be used with the artificial data generated as described in Section III, using the following procedure.

STEP 1. Generate artificial normal $\hat{\mathbb{A}}$ and abnormal $\hat{\mathbb{B}}$ data, using the method in Section III.

STEP 2. Construct a one-class SVM classifier following the method in Section II, using a training set comprising all available real normal data $\mathbb{A}$.

STEP 3. Calibrate the novelty scores $z(\mathbf{x})$ from the one-class SVM classifier into probabilities $P(C_n|x)$, using $\{\hat{\mathbb{A}}, \hat{\mathbb{B}}\}$. This calibration results in a PAV isotonic regression function $g^*(\mathbf{x})$.

STEP 4. Calibrate novelty scores $z(\mathbf{x})$ of all real data into probabilities according to $g^*(\mathbf{x})$ obtained in STEP 3.

A novelty threshold κ may be set by taking advantage of the probabilistic nature of the output, at $P(C_1|\mathbf{x}) > \kappa = 0.5$, on the calibrated novelty scores; i.e., a test sample $\mathbf{x}$ is classified abnormal if $P(C_1|\mathbf{x}) > 0.5$, and normal otherwise. We use the threshold $\kappa = 0.5$ following [21], [33] for problems in which we have uniform class priors. Case studies presented in subsequent chapters will demonstrate the suitability of this choice in application to datasets acquired during the monitoring of example high-integrity systems. We will also show results of area-under-the-ROC, for which all values of the threshold are considered.

V. Case Study I: Industrial Combustion Monitoring

A. Introduction

To demonstrate the proposed novelty detection method, this section considers the condition monitoring of an industrial combustion system, the Typhoon G30 combustor (Siemans Industrial Turbomachinery Ltd.).

Combustion instability, caused by the resonant coupling between combustive heat and acoustic pressure, is a major problem in the operation of jet engines and power generators. Early warning of combustion instability is required to prevent catastrophic system failure, and detection of a deviation away from normal operating status is important for being able to perform pre-emptive maintenance, such that hazards (and associated costs) can be avoided.

B. Methodology

As described in Section I, it is often desirable to perform system-specific novelty detection, in which a model of normality $M(\boldsymbol{\theta})$, with parameters $\boldsymbol{\theta}$, is constructed on-line using data acquired from an individual system during its service life. It is typically assumed that the first training interval comprises normal data $\mathbb{A}$, such that a model can be constructed. (This assumption can be validated on-line by performing an initial comparison of data acquired during the system's first period of operation with a population-generic model, trained using data acquired from a population of systems of the same type.) That model $M(\boldsymbol{\theta})$ is then used for testing further data acquired from the system, such that new data are compared for novelty with respect to the previous, assumed normal, operation of that same system [34]. This model may be periodically re-trained as further normal data are acquired; for example, $M(\boldsymbol{\theta})$ may be re-trained at the end of each flight that is deemed to be normal compared with the existing model, in the case of aerospace EHM [35]. In the case of human vital-sign monitoring, the model could be re-trained after every N-hour period of normal patient physiology [36], [37].

In the case study described in this section, the initial normal period of system operation was simulated by operating the Typhoon G30 combustor in a stable manner at atmospheric pressure. Normal data $\mathbb{A}$ were acquired from the system, as

described below, from which a model of normality $M(\boldsymbol{\theta})$ was constructed using both (i) the proposed method, in which probabilistic calibration was performed, and (ii) the conventional method, in which probabilistic calibration was not performed. Subsequently-acquired data were then tested against this model. To examine the capability of the two novelty detection systems, the combustor then simulated a fault by being deliberately operated in a manner that promoted unstable combustion.

This abnormal operation was achieved by increasing fuel flow-rates above some threshold, while maintaining a constant air flow-rate, which provided abnormal test data $\mathbb{B}$ with which to evaluate the performance of the system-specific model of normality constructed previously.

C. Datasets

Two combustion datasets $\{\mathcal{D}_1, \mathcal{D}_2\}$ were acquired from a Typhoon G30 combustor, as described in the previous section. Each combustion dataset consists of measurements from three channels $\{X_1, X_2, X_3\}$, with sampling frequencies of 1 kHz. Channel X_1 is the gas pressure of the fuel methane (CH_4) in the main burner. For stable combustion, the swirl air flow rate was 0.039 kgs^{-1}; the fuel supplied to the main, and the pilot burners was fixed at flow rates 22.61×10^{-4} kgs^{-1}, and 10.20×10^{-4} kgs^{-1}, respectively. To initiate combustion instabilities, the flow rates of fuel supplied to the main and pilot burners were increased to 26.18×10^{-4} kgs^{-1}, and decreased to 4.37×10^{-4} kgs^{-1}, respectively. Channels X_2 and X_3 are luminosity measurements recorded within the combustion chamber. A bundle of fine optical fibres was mounted at the rear focal point of a Nikon 35 mm camera, such that all light passing through the front lens was collected. The flame luminosity from the combustion chamber was measured using this system. The fibre optic bundle was bifurcated, each channel connected to a photomultiplier (ORIEL model 70704). This design allowed the measurement of chemiluminescent emitters of C_2 radicals (visible at light wavelength 513 nm), and the global intensity of unfiltered light, corresponding to the second and third channels in the datasets, respectively.

Combustion flame images from a high-speed camera have been investigated to predict instability [38], in which a Gaussian mixture model (GMM) was constructed to identify novel flame patterns. A novelty detection method using SVMs [39] was able to achieve earlier identification of combustion instability, and greater distinction between stable and unstable classes than the conventional GMM method. The optical measurements methods described above have been used to study the flame dynamics of unstable combustion [40].

The two triple-channel combustion datasets $\{\mathcal{D}_1, \mathcal{D}_2\}$ contain 5,700, and 7,400 data-points, respectively, which were divided into non-overlapping windows of length $L = 64$. This design resulted in 89, and 115 windows of data for datasets $\mathcal{D}_1$, and $\mathcal{D}_2$, respectively. Wavelet analysis [41], [42] was used, with the Daubechies-3 wavelet function, to obtain wavelet coefficients for each window. Following [43], the mean value of the first-level approximation coefficients λ_1 and the energy of the first-level detail coefficients γ_1 were obtained for each window, to provide a bivariate data space. We note that the case study described in this section is bivariate, such that the data space can be plotted to illustrate the proposed novelty detection procedure; a multivariate example using patient vital-sign data is considered in the next section.

TABLE I
DATASET INDICES FOR EXAMPLE REAL AND SYNTHETIC COMBUSTION DATASETS.

Dataset	$\mathbb{A}$	$\mathbb{B}$	Total size
$\mathcal{D}_1$	1…41	42…89	89
$\mathcal{D}_2$	1…56	57…115	115
$\mathcal{G}$	1…200	201…400	400

This procedure yielded 41 bivariate feature vectors of normal data $\mathbb{A}$, and 48 bivariate feature vectors of abnormal data $\mathbb{B}$ for example combustion dataset $\mathcal{D}_1$. 56, and 59 feature vectors for normal $\mathbb{A}$, and abnormal $\mathbb{B}$ were obtained for dataset $\mathcal{D}_2$, respectively.

For comparison, a synthetic combustion dataset $\mathcal{G}$ was also generated, consisting of 200 normal examples $\mathbb{A}$, from which to construct a model of normality, and 200 abnormal examples $\mathbb{B}$ with which to examine the performance of the resulting novelty detection system. These examples were generated from a bivariate, three-component GMM, with full covariance matrices. Table I shows the components of each dataset used by the case study in this section.

A key point to note is that these datasets are used solely for the purposes of retrospective evaluation of the proposed novelty detection technique, and hence we have collected example abnormal test data $\mathbb{B}$ with which to determine the effectiveness of the one-class SVMs under consideration. When novelty detection is performed in practice, as noted previously, only data assumed to be normal $\mathbb{A}$ are available, from which the model of normality is constructed. Indeed, due to the rarity of abnormal events in most high-integrity systems, it is usually the case that most such systems run "normally" for the great majority of their operational lives. If sufficient abnormal data were available at the training stage such that all possible fault conditions could be represented, then one may consider taking a conventional multi-class approach in which each abnormal condition is explicitly modelled. Such a system could be expected to out-perform a one-class classifier, due to the inclusion of prior knowledge of non-normal classes [21]. Performance of a system using unlabelled data, with unlabelled abnormal examples mixed with the known normal examples, could also be expected to out-perform a novelty detection system [44]–[46]. However, we will here confine ourselves to evaluating the proposed extension to the one-class SVM method, which represents the "standard" condition monitoring case in which only normal data are available at the time of model construction.

D. Model Construction

The conventional one-class SVM requires use of a validation set to determine the threshold κ on its novelty-score output $z(\mathbf{x})$. Hence, we illustrate this case for the conventional

method by using 80% of the available normal data $\mathbb{A}$ for construction of a model of normality, and the remaining 20% of the normal data $\mathbb{A}$ for setting of the novelty threshold, such that $\kappa = \max[\ z(\mathbf{X}_v)\ ]$, where $\mathbf{X}_v$ is the validation set.

The SVM has two key parameters, the values of which need to be determined. The validation set is typically used to determine these values. The first parameter is the kernel bandwidth σ, as defined in (5), which corresponds to the width of the Gaussian kernel. The second parameter is typically termed $\mathcal{C}$, which defines the complexity of the decision boundary:

$$\min_{w\in F,\ \xi\in\mathbb{R}^l,\ \rho\in\mathbb{R}} \quad \frac{1}{2}\|\mathbf{w}\|^2 + \mathcal{C}\sum_{i=1}^{l}\xi_i - \rho \tag{10}$$

$$\text{subject to} \quad \Big(\mathbf{w}\cdot\mathbf{\Phi}(\mathbf{x}_i)\Big) \geq \rho - \xi_i, \quad \xi_i \geq 0, \tag{11}$$

where ξ_i are the individual errors [47]. We note the notational distinction between the SVM parameter $\mathcal{C}$, and the class labels $\{C_0, C_1\}$ previously used to describe the "normal," and "abnormal" classes, respectively.

More intuitively, the value of the $\mathcal{C}$ parameter determines the flexibility of the decision boundary; if $\mathcal{C}$ takes large values in the above, then misclassifications are penalised more significantly, resulting in a decision boundary that is more flexible, attemping to include all training data. Conversely, if $\mathcal{C}$ takes small values in the above, then misclassifications are penalised less, and therefore more "misclassifications" are allowed to occur. This latter case results in a smoother decision boundary [48]. A grid search is typically performed to obtain suitable values for $\{\mathcal{C}, \sigma\}$. We will term the conventional one-class SVM method *SVM-1*.

As noted in the previous section, the proposed probabilistic approach allows automatic selection of a novelty threshold at $P(C_1|\mathbf{x}) = 0.5$, and hence all available normal data $\mathbb{A}$ at the time of model construction may be used as the training set for the one-class SVM. The data in the training set $\mathbb{A}$ were then used to generate $N = 200$ artificial $\hat{\mathbb{A}}$ data, and $N = 200$ artificial $\hat{\mathbb{B}}$ data, as described in Section III, for the purposes of calibrating the SVM output into probabilities. The choice of σ in (5) may also be performed automatically, without the need for validation, noting that, 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 any two samples scaled by a factor $1/(2\sigma^2)$. Based on this close link between σ and Euclidean distance, we propose the following method to determine an appropriate value for σ, based on a similar method proposed by [49] for selecting σ in a density estimator.

First, as described before, we calculate the local average Euclidean distance Δ_l of k nearest neighbours from each sample in the training data, where k is set to be the square root of the number of normal training samples $\mathbb{A}$. Next, the global average distance Δ_g is found by averaging Δ_l over all the training data. The value of Δ_g provides a guide for the range of σ, such that $\sigma = d_\Delta * \Delta_g$. Experiments not shown here (using each dataset considered in this article) indicate that the results obtained are insensitive to the value of d_Δ, and we have selected $d_\Delta = 1.5$. These experiments varied $d_\Delta \in [1\ 2]$, where ROC values were similar to 1 d.p. for values of d_Δ over that range, using each dataset considered in this paper. This insensitivity to the value of d_Δ was previously observed by other authors [49]. We note that this value may be inappropriate for data spaces of particularly high dimension (e.g., 20-dimensional spaces), which this paper does not consider, where a larger value of d_Δ may be appropriate.

Furthermore, it is common for the SVM $\mathcal{C}$ parameter (the flexibility of the decision boundary) to be written in terms of a different parameter ν:

$$\mathcal{C} = \frac{1}{\nu l}. \tag{12}$$

The support vector constraints [47]

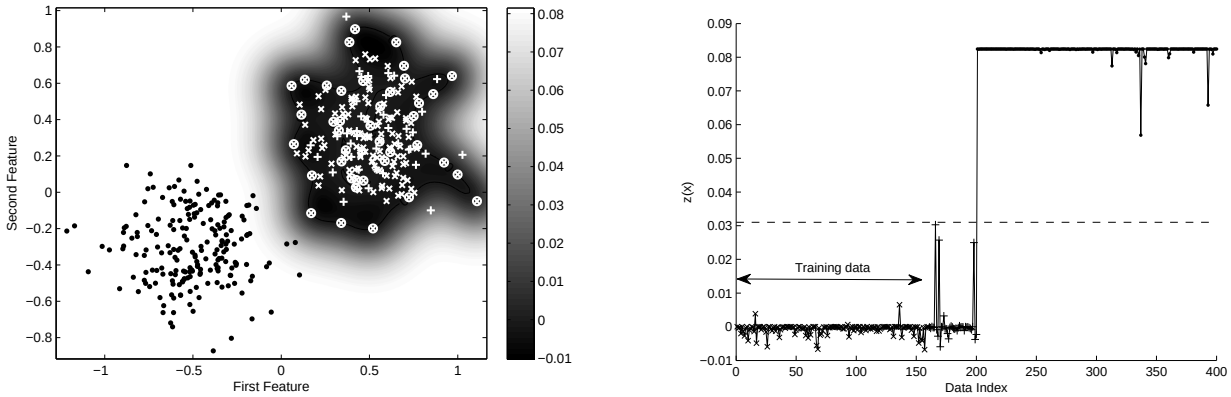
$$\sum_i \alpha_i = 1, \quad 0 \leq \alpha_i \leq \mathcal{C}. \tag{13}$$

imply that the range of the $\mathcal{C}$ parameter is $[1/l \leq \mathcal{C} \leq 1]$; and so, from the above, we have the range for ν which is $[1/l \leq \nu \leq 1]$. This result shows that the parameters ν and $\mathcal{C}$ have the same range. The parameter ν was introduced in the SVM literature because it serves as an upper bound on the number of training samples that lie on the wrong side of the hyperplane (i.e., it is the maximum mis-classification rate). It is also a lower bound on the fraction of support vectors among normal training data [5]. We therefore use the parameterisation involving ν instead of $\mathcal{C}$, due to its clear meaning, as described above. If we wish, the value of $\mathcal{C}$ can be easily recovered by (12).

We will name the proposed probabilistic one-class SVM method *SVM-1P*.

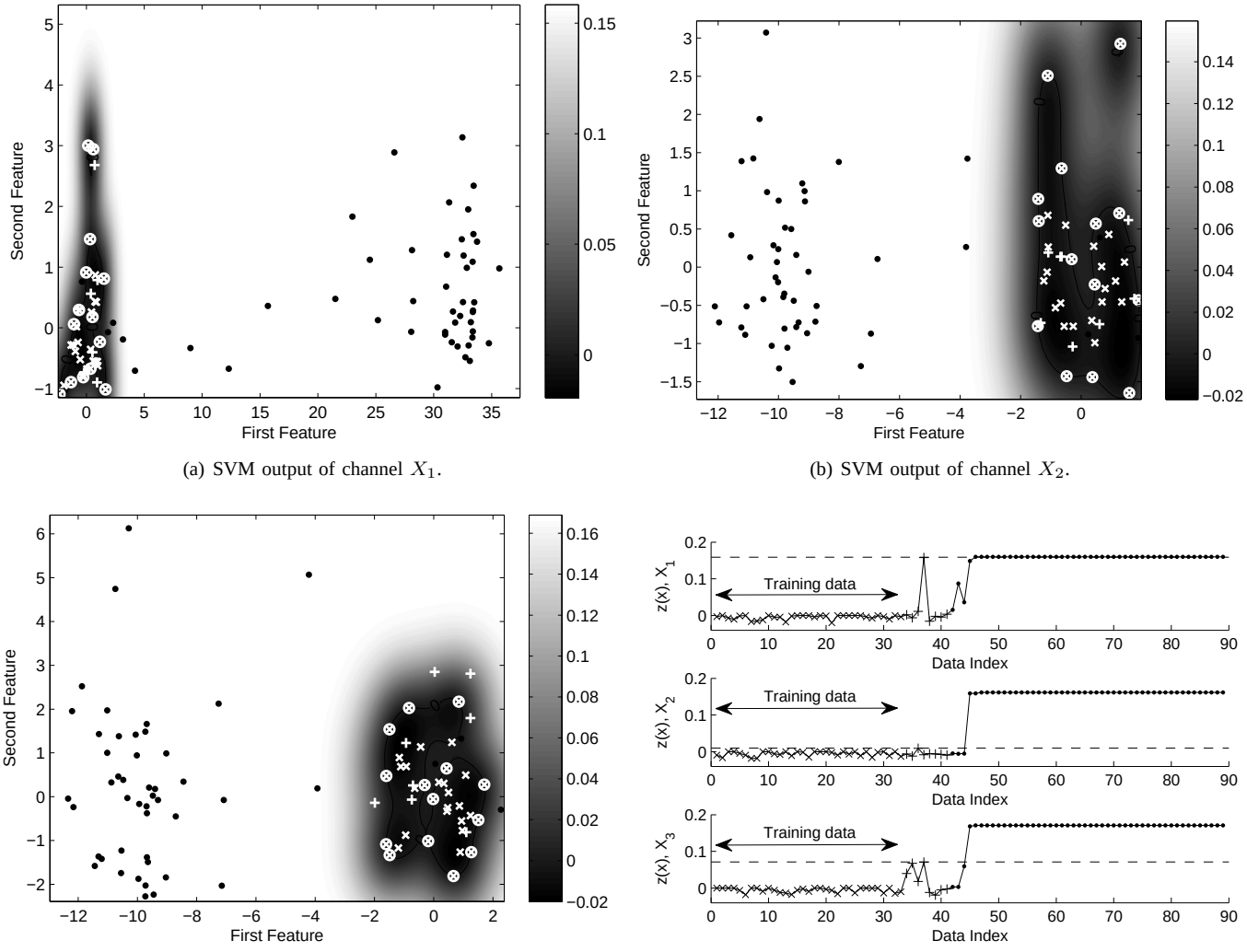
E. Results - Conventional Methods

Fig. 1 shows the results obtained when applying conventional method *SVM-1* to the synthetic dataset $\mathcal{G}$. Part (a) shows the contour plot of SVM output in the bivariate data space. Normal training data $\mathbb{A}$ are shown by white $\{\times\}$. "Normal," and abnormal test data are shown by white $\{+\}$, and black $\{\cdot\}$, respectively. Support vectors are circled. Part (b) shows the novelty score of each of the data. $\mathcal{G}$ consists of 200 normal data, simulating the initial period of normal system operation, and 200 abnormal data here generated to evaluate the novelty detection system, and to simulate a system abnormality (which would be rare, in practice). In procedure *SVM-1*, 80% of the real normal data (data indices 1...160) are used for training the one-class SVM, with the remainder of the normal data (data indices 161...200) used as the validation set. The threshold is represented by a horizontal line, and is set to be the maximum score $z(\mathbf{x})$ assigned to data in the validation set. The normal data $\mathbb{A}$ occupy a region in the upper-right quadrant of the data space, and the conventional SVM has correctly delimited the locus of "normality", as shown in Fig. 1(a). Data acquired from a conventional high-integrity system would spend the majority (if not the entirety) of their time contained within this locus, and hence would be classified normal, because the corresponding conventional SVM novelty scores $z(\mathbf{x})$ fall beneath the novelty threshold κ, as illustrated in Fig. 1(b). In this synthetic example, the simulated fault, occurring at data index 201 and persisting



(a) SVM output. (b) SVM novelty scores.

Fig. 1. Novelty detection results using procedure *SVM-1*, applied to synthetic data set $\mathcal{G}$.



(a) SVM output of channel X_1. (b) SVM output of channel X_2.

(c) SVM output of channel X_3. (d) SVM novelty scores.

Fig. 2. Novelty detection results using procedure *SVM-1*, applied to combustion data set $\mathcal{D}_1$.

for the remainder of the dataset, results in data that lie well-separated from the locus of normal data. Note that normal data in the validation set take significantly larger novelty scores than the training data, as may be seen for $z(\mathbf{x}_{166})$, $z(\mathbf{x}_{169})$, and $z(\mathbf{x}_{198})$. Hence, the selection of a suitable novelty threshold κ requires care when using the conventional one-class SVM formulation, *SVM-1*.

Fig. 2 shows the results obtained when applying the conventional *SVM-1* to the example bivariate dataset, $\mathcal{D}_1$. Parts (a), (b), and (c) show contour plots of SVM novelty scores $z(\mathbf{x})$ in the bivariate data space of each channel, using the same notation as previously shown. Part (d) shows the novelty scores $z(\mathbf{x})$ of channels $\{X_1, X_2, X_3\}$, from upper to lower figures, respectively. Using *SVM-1*, normal data $\mathbb{A}$ indices $1 \dots 33$ have been used as the training data, with the remainder of normal data $34 \dots 41$ used as validation data. The seeded fault data, used to test the novelty detection system, comprise the abnormal data $\mathbb{B}$ indices $42 \dots 89$. While the separation of fault data from normal data is generally large, it may be seen that more overlap exists between the two classes than with the synthetic dataset $\mathcal{G}$. This overlap adversely affects the ability of the conventional *SVM-1* method to set the novelty threshold κ on the novelty score such that the classifier is suitably sensitive to the subsequently-acquired abnormal data $\mathbb{B}$ from the seeded fault. This effect is most evident for the SVM that is trained using data from channel X_1, in which one of the validation data $\mathbf{x}_{37}$ takes a high novelty score, which significantly reduces the sensitivity of the novelty detector to the fault data. Similarly, the SVM trained for channel X_3 results in constantly high novelty scores $z(\mathbf{x})$ throughout the normal validation dataset.

The earliest detection of the fault is at $\mathbf{x}_{47}$ for channel X_1, at $\mathbf{x}_{45}$ for channel X_2, and at $\mathbf{x}_{45}$ for channel X_3, despite the onset of the fault occurring at $\mathbf{x}_{42}$ in each case. Thus, the conventional one-class method *SVM-1* has resulted in decreased sensitivity with respect to fault data, with a number of false-negative classifications of the earlier fault data. This result is of particular significance in the monitoring of high-integrity systems, as will be discussed in Section VII.

We consider also the use of a GMM cross-validated using the same procedure as used to set the parameters of method *SVM-1*. We note that the GMM typically performs less well than the SVM, in terms of both earliest warning of abnormality and the overall area-under-the-curve (AUC), where the curve is the receiver operating characteristic (ROC) curve.

F. Results - Proposed Method SVM-1P

Fig. 3 illustrates the proposed calibration method *SVM-1P* applied to synthetic dataset $\mathcal{G}$. The upper figure shows real $\{\mathbb{A}, \mathbb{B}\}$, and artificial $\{\hat{\mathbb{A}}, \hat{\mathbb{B}}\}$ data generated around $\mathbb{A}$; it may be seen that the locus of artificial abnormal data $\hat{\mathbb{B}}$ is a torus around the normal data $\mathbb{A}$ from which they were generated. The lower figure shows the corresponding isotonic function g^* obtained using the artificial data $\{\hat{\mathbb{A}}, \hat{\mathbb{B}}\}$, on which the calibrated probabilities obtained from all data are marked. The locus of data $\mathbb{B}$ occupied by the simulated fault is confined to the lower-left quadrant of the bivariate data space, which is typical for novelty detection applications. The fault is not a full description of "abnormality", which motivates the use of the one-class approach for monitoring high-integrity systems, where only "normality" is typically well-understood. Fig. 3 (lower) shows the step-wise linear nature of the $g^*(\mathbf{x})$ isotonic function. The figure shows that $g^*(\mathbf{x})$ takes all values in the range $[0\ 1]$.

TABLE II
DATA INDEX OF THE FIRST ABNORMAL CLASSIFICATION, ACCORDING TO EACH CLASSIFIER.

	$\mathcal{D}_1$			$\mathcal{D}_2$			$\mathcal{G}$
	X_1	X_2	X_3	X_1	X_2	X_3	-
SVM-1	47	45	45	61	60	45	201
SVM-1P	43	45	44	57	57	58	201
GMM	47	46	46	62	60	59	201

Results obtained by applying the calibration method to exemplar dataset $\mathcal{D}_1$ are shown in Fig. 4. Part (a) illustrates calibration results for channel X_1, showing (upper) the generation of artificial data $\{\hat{\mathbb{A}}, \hat{\mathbb{B}}\}$ from normal dataset $\mathbb{A}$, and the corresponding isotonic $g^*(\mathbf{x})$ function. Part (b) shows output estimated probabilities for each of the three channels $\{X_1, X_2, X_3\}$. Fig. 4 shows that the locus of data space occupied by the data acquired during fault conditions $\mathbb{B}$ lies significantly far from the locus of normal data $\mathbb{A}$, and from the toroidal structure of the artificial data $\hat{\mathbb{B}}$ that were used for calibration. The corresponding calibrated probabilities for this well-separated data space are therefore extremal, taking values close to 0 and 1. The *a priori* selection of the novelty threshold κ at $P(C_1|\mathbf{x}) = 0.5$ is shown in Fig. 4(b), which shows that the estimated probabilities have been forced into the extrema of the range $P = [0\ 1]$. The onset of the fault conditions is detected earlier than with the coventional method *SVM-1*, as shown in Table II. Recall from Table I that the onset of abnormality in dataset $\mathcal{D}_1$ occurs at data index $\mathbf{x}_{42}$, which therefore represents the point at which earilest detection could occur. The proposed method identifies the deterioration in the second abnormal window, $\mathbf{x}_{43}$, which is earlier than the conventional one-class SVM and the GMM.

G. Discussion

Results obtained for dataset $\mathcal{D}_2$ are shown in Table II, where it may be seen that the differences between methods *SVM-1*, *GMM*, and *SVM-1P* are similar to those illustrated in more detail for dataset $\mathcal{D}_1$, above. Recall from Table I that the onset of seeded fault conditions occurred at data index $\mathbf{x}_{57}$ for dataset $\mathcal{D}_2$, which indicates that the proposed method *SVM-1P* is sufficiently sensitive to detect the onset of unstable combustion, while the conventional method *SVM-1* is less sensitive, and, in the case of channel X_3, generates a false alarm by incorrectly classifying normal data index $\mathbf{x}_{45}$ as being abnormal. Such false alarms are one of the principal reasons that conventional monitoring systems are ignored in practice, as discussed in Section VII. Again, the SVM-based methods outperform the GMM. The (specificity, sensitivity)

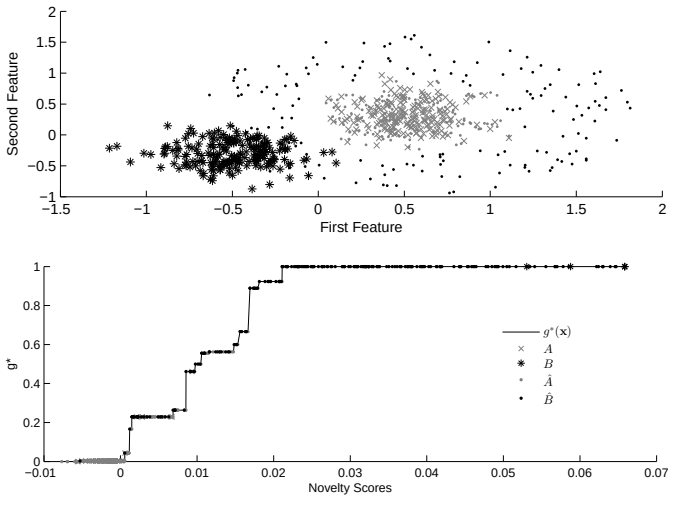
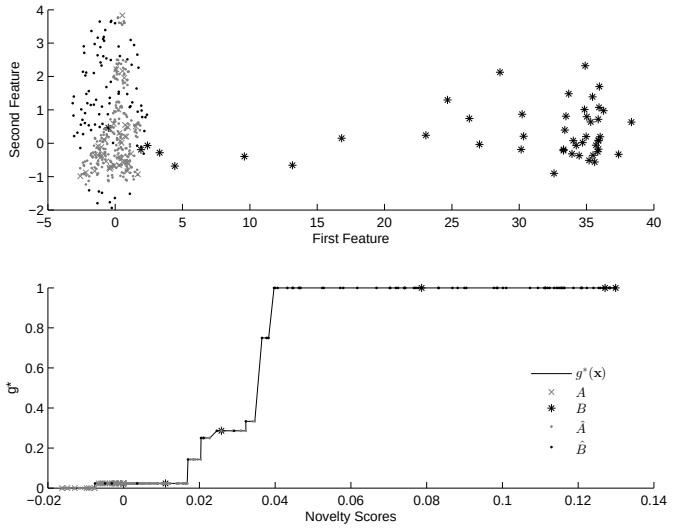
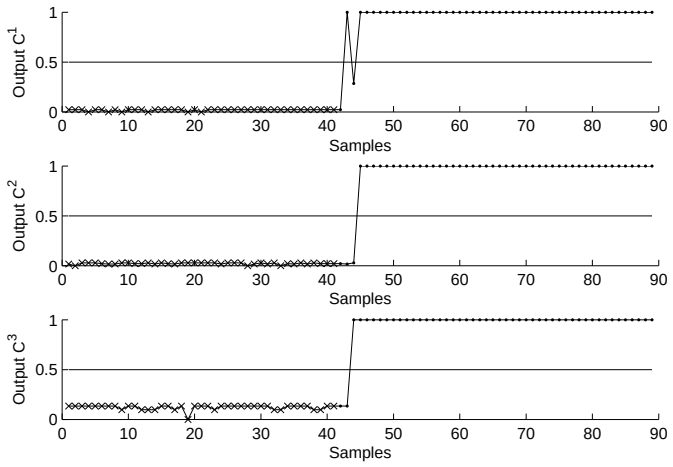


Fig. 3. Isotonic regression results using procedure *SVM-1P*, applied to synthetic data set $\mathcal{G}$.

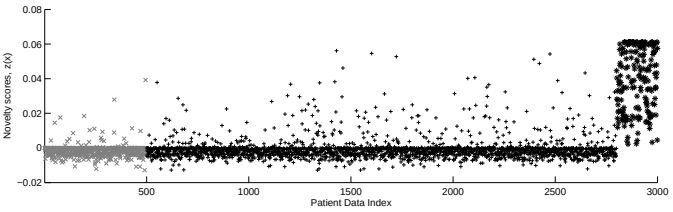


(a) g^*

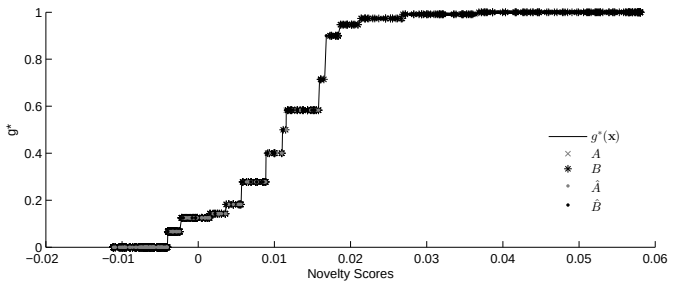


(b) Output

Fig. 4. Results obtained using procedure *SVM-1P* applied to combustion data set $\mathcal{D}_1$.



(a) *SVM-1* vital-sign output



(b) *SVM-1P* vital-sign output

Fig. 5. Results obtained using procedures *SVM-1*, and *SVM-1P*, in (a), and (b), respectively, applied to patient vital-sign data set $\mathcal{D}_3$.

TABLE III
AUC VALUES FOR EACH METHOD.

	$\mathcal{D}_1$			$\mathcal{D}_2$			$\mathcal{G}$
	X_1	X_2	X_3	X_1	X_2	X_3	-
SVM-1	0.88	0.92	0.85	0.85	0.91	0.88	0.98
SVM-1P	0.95	0.96	0.94	0.94	0.96	0.92	0.99
GMM	0.81	0.85	0.84	0.82	0.87	0.87	0.96

for *SVM-1*, *GMM*, and *SVM-1P* were (0.92, 0.91), (0.88, 0.87), and (0.95, 0.92), respectively, for dataset $\mathcal{D}_1$, using cross-validation thresholds for *SVM-1* and *GMM*, and a probabilistic threshold of $P = 0.5$ for *SVM-1P*.

AUC results for both methods, for all datasets, are reported in Table III, where it may be seen that the proposed method performs similarly using the synthetic dataset $\mathcal{G}$, due to the separability of the simulated abnormal conditions from the normal conditions; however, the proposed method provides a noticeable increase in AUC for the exemplar combustion datasets $\{\mathcal{D}_1, \mathcal{D}_2\}$, indicating that the increase in sensitivity to fault conditions, illustrated in the previous subsections, has not come at the expense of decreased specificity (i.e., the false-alarm rate is kept sufficiently low as to be usable in practice).

VI. Case Study II: Patient Vital-Sign Monitoring

This section reports results obtained from evaluating both *SVM-1* and *SVM-1P* using dataset $\mathcal{D}_3$, which is an example of novelty detection for patient vital-sign monitoring. Whereas the previous section evaluated the performance of the algorithm using bivariate data, such that the data space could be plotted to illustrate the SVM output and the probabilistic calibration method, this section examines the use of the algorithm for higher-dimensional data.

A. Dataset

This section considers vital-sign data acquired from patients in a step-down unit (SDU), which is a level of acuity lower than that of the intensive care unit (ICU). There is a need for effective novelty detection systems in such wards, because patient deterioration can go unnoticed by clinical staff, leading to adverse patient outcomes [50]. Existing patient monitors generated univariate alarms whenever vital signs exceed some pre-defined threshold, and often go unheeded due to the high false-positive rate of such alarms, where [51] reported results of a study in which it was deemed that 84% of alarms were false.

The dataset used for the work described by this section comprises measurements of heart rate, breathing rate, blood oxygen saturation, and systolic blood pressure. Data were acquired once every four hours by ward staff (as is common practice in most SDU-level wards in the UK and the US) at the Oxford Cancer Hospital, Oxford, UK. 3,000 such vectors $\mathbf{x}_i \in \mathbb{R}^4$ were acquired from 40 patients.

B. Methodology

As in Section V, procedures *SVM-1* and *SVM-1P* were applied, using the conventional and proposed methodologies, respectively. A novelty detection approach is particularly suitable to the analysis of vital-sign data from hospital patients, because it is unlikely that a full description of abnormal classes could be obtained, given the range of potential physiological deteriorations that a human may undergo, and the variation in responses to abnormal conditions between patients. Hence, given a large set of normal data $\mathbb{A}$, a model of normality can be constructed that is then used to detect abnormal physiological variation with respect to that model.

203 abnormal data $\mathbb{B}$ were acquired, deemed to be so by existing, manual clinical systems that are used to determine if a patient requires review by senior medical personnel. Of the 2,797 remaining normal data, 500 were provided to methods *SVM-1* and *SVM-1P* from which to learn a model of normality, with the remaining 2,297 normal data being used as test data, to evaluate the false-positive rate of both methods.

We note that the investigation described in this section is a "population generic" approach, collecting the data for multiple patients to construct a single model of "normal" patient physiology. This approach is directly comparable to standard clinical practice, in which population-generic sets of heuristic scores are manually applied to determine if a patient's vital signs are normal or abnormal.

C. Results

Results obtained using *SVM-1*, and *SVM-1P* are shown in Fig. 5(a), and (b), respectively. Each procedure was given the first 500 normal vital signs on which to train (and, in the case of the conventional *SVM-1* method, validate) a model of normality, which are shown in grey $\{\times\}$. The final 203 data in $\mathcal{D}_3$ were classified abnormal and representative of patient deterioration, by clinicians, which are shown in black $\{*\}$. The remainder of the data are normal, and used for testing, and are shown in black $\{+\}$. The conventional method *SVM-1* misclassifes significantly more normal data with respect to its model of normality (including some training data) than does the proposed method *SVM-1P*, due to the poor separation of *SVM-1* novelty scores $z(\mathbf{x})$ between the two classes.

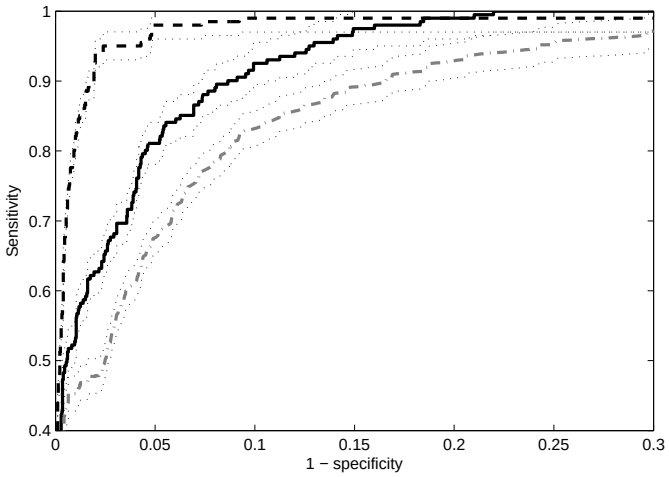


Fig. 6. ROC curves for $N = 20$ statistically independent evaluations.

ROC curves for each method, evaluated over $N = 20$ statistically independent experiments, are shown in Fig. 6, where methods *SVM-1P*, *SVM-1*, and *GMM* are shown in black dashed, black solid, and grey dashed-and-dotted lines, respectively. The mean ROC of the N experiments is shown by the thick lines, where the curves for *SVM-1P* may be seen to be closer to the upper-left corner of the ROC plot. Error bars at $\mu \pm 3\sigma$ are shown using thin dotted lines. It may be seen that the proposed method *SVM-1P* achieves both a higher sensitivity and specificity than the conventional methods over all experiments. The (specificity, sensitivity) for *SVM-1*, *GMM*, and *SVM-1P* were (0.90, 0.92), (0.82, 0.91), and (0.94, 0.97), respectively, using cross-validation thresholds for *SVM-1* and *GMM*, and a probabilistic threshold of P = 0.5 for *SVM-1P*.

VII. Conclusions, and Discussion

The conventional method of using a one-class SVM is well-understood in the literature, and has been evaluated here in comparison with a proposed method that (i) calibrates the novelty scores output by the one-class SVM into estimated posterior class probabilities, where special care was required due to the one-class formulation, (ii) utilises the probabilistic nature of the result to define a novelty threshold without the need for the conventional validation set, and (iii) proposes a procedure for determining other SVM parameters.

These proposals have been illustrated using lower-dimensional data from a large-scale combustor, whereby the generation of artificial data (as is required by the calibration step) and the data space itself may be visualised. A higher-dimensional evaluation was performed using patient vital-sign data. In both cases, the proposed method achieved better overall sensitivity and specificity than the conventional technique. We note that the application of the method to datasets of very high dimension (e.g., spaces of dimension greater than 20) is not considered by the work described in this paper, in which

we restrict ourselves to considering the applications typically encountered in condition monitoring.

In practice, the acceptance of most "intelligent" methods for monitoring high-integrity systems is determined by their false-positive rates. While a high false-positive rate may be acceptable in, for example, the screening of cancer patients (in which the priority is to detect all cancers, and where the false-positive cost is relatively low), monitoring systems must seldom generate false alerts. In the case of industrial systems, such alerts could result in the premature landing of an aeroplane, or the halting of a power-generation engine; in the case of human vital-signs monitoring, such alerts would result in senior clinicians being brought to the patient bed-side to review patients that are "normal." In all cases, the cost of false-positive alerts far exceeds the cost of false-negative examples. Such monitoring methods are typically part of a "redundant" network of systems, and so failure to detect all abnormalities exhibited by the system-under-test is relatively less costly. We have demonstrated that the proposed probabilistic approach can result in significantly fewer false-alerts (i.e., it has higher specificity) than the conventional method, while still remaining acceptably sensitive to the detection of abnormal examples.

A second requirement of monitoring systems is that abnormal conditions are detected as early as possible, such that preventative action may be taken to avoid system damage from continued abnormal operation (whether it be machine maintenance, or patient review by a clinician). This is particularly important in low-bandwidth monitoring systems, in which data are acquired at a low frequency, such as in the case of aircraft monitoring, where a summary of engine performance may be downloaded at the end of each flight [13], or, as in the second case study considered by this paper, when patient vital signs are observed every four hours. We have demonstrated that the proposed probabilistic method achieves earlier warning of unstable combustion than is provided by the conventional method, which would enable the system to be shut down at the onset of unstable combustion conditions, thus avoiding further system risk. Further improvement could be gained by training the model with a sequentially-updating on-line training algorithm.

We note that in cases where the quantity of normal data acquired is particularly large, as would be the case if high sampling-rate sensors were used, then the differences between the two procedures would be smaller than is considered here, due to both systems being able to form complete models of normality from the sufficient quantity of training data. However, even in such cases, the proposed method could be exploited to provide rapid re-training of new system-specific models of normality in the case of system maintenance, whereby models of the system's pre-maintenance behaviour must be discarded, and new models have to be learned.

Acknowledgements

LC was supported by the Overseas Research Students Award Scheme, provided by the UK Government, and later by the NIHR Biomedical Research Centre Programme, Oxford. DAC was funded 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 number WT 088877/Z/09/Z.

References

[1] P. Hayton, L. Tarassenko, B. Scholkopf, and P. Anuzis, "Support vector novelty detection applied to jet engine vibration spectra," *Proc. NIPS*, Denver, US, 2000, pp. 946–952.

[2] A. Gretton and F. Desobry, "On-line one-class support vector machines. an application to signal segmentation." *Proc. IEEE ICASSP*, Hong-Kong, China, 2003.

[3] D. R. Hardoon and L. M. Manevitz, "fMRI analysis via one-class machine learning techniques," *Proc. 19th International Joint Conference on Aritifical Intelligence (IJCAI)*, Edinburgh, UK, 2005, pp. 1604–1605.

[4] M. Markou and S. Singh, "Novelty detection: A review - part 2: Neural network based approaches," *Signal Processing*, vol. 83, no. 12, pp. 2499–2521, 2003.

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

[6] D. M. J. Tax and R. P. W. Duin, "Data domain description using support vectors," *Proc. ESAN99*, Brussels, 1999, pp. 251–256.

[7] D. Tax and R. Duin, "Support vector domain description," *Pattern Recognition Letters*, vol. 20, pp. 1191–1199, 1999.

[8] L. Clifton, "Multi-channel novelty detection and classifier combination," Ph.D. dissertation, Electrical and Electronic Engineering, University of Manchester, 2007.

[9] J. Drish, "Obtaining calibrated probability estimates from support vector machines," University of California, San Diego, Tech. Rep., 2001.

[10] H. Chen, P. Tino, and X. Yao, "Probabilistic classification vector machines," *IEEE Transactions on Neural Networks*, vol. 20, pp. 901–914, 2009.

[11] D. A. Clifton, "Novelty detection with extreme value theory in jet engine vibration ddata," Ph.D. dissertation, University of Oxford, 2009.

[12] P. Sollich, "Probabilistic methods for support vector machines," *Advances in Neural Information Processing Systems*, vol. 12, pp. 349–355, 2000.

[13] D. Clifton, N. McGrogan, L. Tarassenko, S. King, P. Anuzis, and D. King, "Bayesian extreme value statistics for novelty detection in gas-turbine engines," *Proceedings of IEEE Aerospace, Montana, USA*, 2008, pp. 1–11.

[14] S. King, P. Bannister, D. Clifton, and L. Tarassenko, "Probabilistic approaches to condition monitoring of aerospace engines," *IMechE Part G: Journal of Aerospace Engineering*, vol. 223, no. 5, pp. 553–541, 2009.

[15] A. Fleury, M. Vacher, and N. Noury, "SVM-based multi-modal classification of activities of daily living in health smart homes," *IEEE Transactions on Information Technology in Biomedicine*, vol. 14, no. 2, pp. 274–283, 2009.

[16] M. Jaana, G. Pare, and C. Sicotte, "Home telemonitoring for respiratory conditions: A systematic review," *The American Journal of Managed Care*, vol. 15, no. 5, pp. 313–320, 2009.

[17] K. Shen, C. Ong, X. Li, and E. Wilder-Smith, "Feature selection via sensitivity analysis of SVM probabilistic outputs," *Machine Learning*, vol. 70, no. 1, pp. 1–20, 2008.

[18] J. Kittler, M. Hatef, R. Duin, and J. Matas, "On combining classifiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 3, pp. 226–239, 1998.

[19] D. M. J. Tax and R. P. W. Duin, "Combining one-class classifiers," *Proc. Multiple Classifier Systems*, 2001, pp. 299–308.

[20] L. I. Kuncheva, "A theoretical study on six classifier fusion stragies," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 2, pp. 281–286, 2002.

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

[22] Y. Grandvalet, J. Marithoz, and S. Bengio, "A probabilistic interpretation of SVMs with an application to unbalanced classification," *Advances in Neural Information Processing Systems 18*. MIT Press, 2006, pp. 467–474.

[23] B. Zadrozny and C. Elkan, "Transforming classifier scores into accurate multiclass probability estimates," *Proc. of ACM SIGKDD*, 2002, pp. 694–699.

[24] A. Niculescu-Mizil and R. Caruana, "Predicting good probabilities with supervised learning," *Proc. of ICML*, 2005, pp. 625–632.

[25] D. M. J. Tax and R. P. W. Duin, "Uniform object generation for optimizing one-class classifiers," *Journal of Machine Learning Research*, vol. 22, pp. 155–173, 2001.

[26] M. Markou and S. Singh, "A neural network-based novelty detector for image sequence analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 10, pp. 1664–1677, 2006.

[27] J. C. Platt, "Probabilistic outputs for support vector machines and comparison to regularized likelihood methods," *Advances in Large Margin Classifiers*, 1999, pp. 61–74.

[28] H. Lin, C. Lin, and R. Weng, "A note on Platt's probabilistic outputs for support vector machines," *Machine Learning*, vol. 68, no. 3, pp. 267–276, 2007.

[29] V. Vapnik, *The Nature of Statistical Learning Theory*, 2nd ed. Berlin: Springer-Verlag, 2000.

[30] M. H. DeGroot and S. E. Fienberg, "The comparison and evaluation of forecasters," *Statistician*, vol. 32, no. 1, pp. 12–22, 1982.

[31] B. Zadrozny and C. Elkan, "Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers," *Proc. of ICML*, 2001, pp. 609–616.

[32] S. Ruping, "Robust probabilistic calibration," *Proc. of ECML*, 2006, pp. 743–750.

[33] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. New York: John Wiley and Sons, 2001.

[34] I. Nabney, *Netlab: Algorithms for Pattern Recognition*, 1st ed. London: Springer, 2002.

[35] D. Clifton, S. Hugueny, and L. Tarassenko, "Novelty detection with multivariate extreme value statistics," *Journal of Signal Processing Systems, in press*, 2010.

[36] D. Clifton, L. Clifton, and L. Tarassenko, "Patient-specifi biomedical condition monitoring for post-operative cancer patients," *Proc. Condition Monitoring, Dublin, Ireland*, 2009, pp. 424–433.

[37] S. Hugueny, D. Clifton, and L. Tarassenko, "Probabilistic patient monitoring using extreme value theory," *Proc. Biomedical Systems and Technologies, Valencia, Spain*, 2010, pp. 5–12.

[38] L. Wang and H. Yin, "Wavelet analysis in novelty detection for combustion image data," *Proc. The 10th CACSC*, Liverpool, UK, 2004, pp. 79–82.

[39] L. Clifton, H. Yin, and Y. Zhang, "Support vector machine in novelty detection for multi-channel combustion data," *Proc. ISNN (3)*, Chengdu, China, 2006, pp. 836–843.

[40] W. B. Ng, E. Clough, K. J. Syed, and Y. Zhang, "The combined investigation of the flame dynamics of an industrial gas turbine combustor using high-speed imaging and an optically integrated data collection method," *Measurement Science and Tachnology*, vol. 15, pp. 2303–2309, 2004.

[41] I. Daubechies, "Orthonormal bases of compactly supported wavelet," *Communications on Pure and Applied Mathematicse*, vol. 41, pp. 909–996, 1988.

[42] S. G. Mallat, "A theory for multiresolution signal decomposition: the wavelet representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, no. 7, pp. 674–693, 1989.

[43] L. Clifton, H. Yin, D. A. Clifton, and Y. Zhang, "Combined support vector novelty detection for multi-channel combustion data," *Proc. of IEEE ICNSC*, London, UK, 2007, pp. 495–500.

[44] F. Letouzey, F. Denis, and R. Gilleron, "Learning from positive and unlabeled examples," *Procs. of the 11th International Conference on Algorithmic Learning Theory*, 2000, pp. 71–85.

[45] D. Zhang, "A simple probabilistic approach to learning from positive and unlabeled examples," *In Proc. of the 5th Annual UK Workshop on Computational Intelligence*, 2005.

[46] C. Elkan and K. Noto, "Learning classifiers from only positive and unlabeled data," *Proc. 14th International Conference on Knowledge Discovery and Data Mining*, 2008.

[47] C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, pp. 121–167, 1998.

[48] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning, second ed.* Berlin: Springer-Verlag, 2009.

[49] C. M. Bishop, "Novelty detection and neural network validation," *Proceedings of IEE Conference on Vision and Image Signal Processing*, vol. 141, no. 4, pp. 217–222, 1994.

[50] N. P. S. Association, "Safer care for acutely ill patients: Learning from serious accidents," NPSA, Tech. Rep., 2007.

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

Lei Clifton received a BSc and MSc in electrical engineering from Beijing Institute of Technology, China, and a PhD degree in electrical engineering from Manchester University, U.K. After six years of post-doctoral research at the University of Oxford, UK, she was appointed as a Medical Statistician at the Centre for Statistics in Medicine, University of Oxford. Her research interests include statistical signal processing, and machine learning for intelligent health monitoring systems.

David A. Clifton is a member of faculty in the Department of Engineering Science at the University of Oxford, from which he previously graduated in 2009. He is the group leader of the Computational Health Informatics (CHI) laboratory in that Department, and the Associate Director of the Oxford Centre for Affordable Healthcare Technology. He is a Research Fellow of the Royal Academy of Engineering; a Fellow of Kellogg College, Oxford; a Fellow of Mansfield College, Oxford; and a College Lecturer in Engineering at Balliol College, Oxford.

Yang Zhang received a BEng from Zhejiang University, China, and a PhD in the Engineering Department of Cambridge University. After his post-doctoral research in Cambridge University, he moved to UMIST, and then the University of Manchester before taking the Chair of Combustion and Energy in Sheffield.

Peter Watkinson is an Intensive Care Physician at Oxford University Hospitals NHS Trust, and is one of two clinical leads for the Critical Care research group at the Kadoorie Centre in Oxford. His research focus combines the fields of bioengineering and acute medicine to generate innovative methods for identification of at-risk patients, both in and out of hospital.

Lionel Tarassenko received a BA in engineering science, and a DPhil in medical engineering, both from the University of Oxford, UK, in 1978, and 1985, respectively. He then held a number of positions in academia and industry, before taking up an Assistant Professorship in Oxford in 1988. He has been the holder of the Chair in Electrical Engineering at Oxford University since October 1997. He is the author of 150 journal papers, 160 conference papers, and 3 books; and holds 24 granted patents. He was the founding Director of the Oxford Institute of Biomedical Engineering in 2008, and has been the Director of the Centre of Excellence in Medical Engineering funded by the Wellcome Trust and EPSRC since October 2009. Prof. Tarassenko was awarded the 2006 Silver Medal of the Royal Academy of Engineering for his contribution to British engineering.

Hujun Yin has been with The University of Manchester, School of Electrical and Electronic Engineering since 1996. He received BEng, and MSc degrees from Southeast University, China; and a PhD degree from University of York, UK, in 1983, 1986, and 1996 respectively. His main research interests are neural networks, self-organising systems in particular, pattern recognition, bio- & neuro-informatics, and face recognition. He has published over 150 peer-reviewed articles in a range of topics. He is a senior member of the IEEE. He had been an Associate Editor of the IEEE Transactions on Neural Networks from 2006 to 2010, and has been a member of the Editorial Board of the International Journal of Neural Systems since 2005.