Evaluation of Cardiovascular Disease Risk in the China Kadoorie Biobank Using Novelty Detection

Yanting Shen\textsuperscript{1}, Zhengming Chen\textsuperscript{2}, Robert Clarke\textsuperscript{2}, David A. Clifton\textsuperscript{1}

\textbf{Abstract}—We evaluate the risks of cardiovascular disease to the Chinese population by i) detecting "abnormality" using 3 one-class classification methods (a discriminative one-class support vector machine (SVM), a generative kernel density estimate (KDE), and a discriminative KDE), and ii) predicting probabilities of "normality", arrhythmia, and ischemia using 3-class classification method (a discriminative SVM).

For one-class classification we constructed 5 possible criteria for "normality" and used 14 automatically extracted ECG features by the Mortara device and blood pressure data. We evaluated classification performance in balanced test sets and the discriminative SVM model achieved area under curve (AUC) 0.79-0.83 and 71.6\%-75.6\% accuracy, depending on the different criteria for defining "normality". The generative KDE achieved 0.72-0.80 AUC, and 74.8\% accuracy.

For three-class classification we propose a set of inclusion criteria for "normal", "arrhythmia" and "ischemia", and added 6 new features (amplitudes of P, Q, R, S, and T waves and ST-level) extracted from time-series ECG. With new features the accuracy improved from 64.6\% to 74.5\% for 3-class and from 74.5\% to 88.7\% for classification between normal and ischemia in balanced test sets.

\textbf{Index Terms}—ECG, Novelty Detection, SVM, KDE, Biobank.

\section{I. Introduction}

Cardiovascular disease (CVD), including coronary heart disease, cerebrovascular disease, peripheral arterial disease, rheumatic heart disease, congenital heart disease, deep vein thrombosis, and pulmonary embolism, is the leading cause of mortality worldwide \cite{1} and in China \cite{2}. An estimated 3.76 million people died from CVD in China in 2013 \cite{2}. There are large geographical and economic variations of CVD mortality in China \cite{2}, suggesting appropriate measures may be taken for prevention and effective treatment of the disease. For example, the World Health Organisation advises people with high CVD risk to access early detection and management facilities \cite{1}. Studies have shown effective reduction of CVD mortality rates by promoting awareness of risk factors and adopting lifestyle changes \cite{1}. Identifying risks to the Chinese population will help provide advice to people to improve their lifestyle and help clinicians to discover appropriate treatments for specific conditions; this promises to reduce mortality and healthcare expenditure.

Many CVD risk factors have been identified by long-term follow-up studies, such as obesity \cite{3}–\cite{5}, diabetes \cite{6}, metabolic syndrome \cite{7}, smoking \cite{1}, and hypertension \cite{8}, among many others. Recently, multiple biochemical and genetic risk factors have been identified via conventional follow-up approaches, and by machine learning in large datasets \cite{9}.

While much research has been focusing on the effects of certain risk factors, personalised medicine requires integrating all risk factors to which a person is exposed and then predicting risks for specific diseases, so that preventative measures may be taken. Examples of research for this purpose include \cite{10}, in which mathematical models were established by scoring systems using multiple risk factors (blood pressure, total cholesterol, high-density lipoprotein cholesterol, smoking, glucose intolerance, and left ventricular hypertrophy) as inputs to predict the risks of myocardial infarction, coronary heart disease, CVD, and death from these diseases. With an increasing number of risk factors being identified, and especially with abundant genetic and lifestyle data now available, it can be expected that such an approach will face difficulty as the "healthy" range of the newly-identified factors are hard to obtain or quantify.

Machine learning has the advantage of estimating the associations between risk factors and diseases without prior knowledge of accurate reference values of the risk factors. This approach has grown popular in risk evaluation and diagnostics for chronic diseases \cite{11}–\cite{14}. In CVD, Knuiman et al. have predicted coronary mortality in the Busselton cohort using a discriminative decision tree \cite{15}; Lapuerta et al. used a neural network in the prediction of coronary disease risk from serum and lipid profiles \cite{16}; Chattier et al. used logistic regression and multilayer perceptrons to predict CVD risk from the INDIANA (Individual Data Analysis of Antihypertensive Intervention Trials) cohort \cite{9}; and Das et al. performed heart disease diagnosis using ensembles of neural networks \cite{17}.

The electrocardiogram (ECG), being an important measurement of cardiac function that is relatively easy to obtain, is surprisingly seldom used in prediction. This report attempts to address the need for risk metrics that include ECG-derived features by analysing CVD risks associated with abnormal ECG, using the China Kadoorie Biobank ECG dataset. This report include two risk evaluation tasks: i) "abnormality" detection and 2) prediction of probabilities of "normal", arrhythmia, and ischemia. Since abnormality is relatively rare in this database, novelty detection (the aim of which is to classify an under-sampled "abnormal" class) is an appropriate approach to address the first task. To address the second task we build models of "normal", "arrhythmia", and "ischemia" on which probabilistic prediction of the "borderline" data will be based.

This report includes a brief literature review of novelty

1 Department of Engineering Science, University of Oxford, UK
2 Nuffield Department of Population Health, University of Oxford, UK
Author to whom all correspondence should be directed is Yanting Shen <yanting.shen@eng.ox.ac.uk>.
Novelty detection is the task of classifying undersampled (typically "abnormal") classes whose distribution is difficult to model explicitly from the training set. It is especially useful in medical monitoring [18], where the abnormal possibilities are so varied as to be almost impossible to be modelled explicitly. Novelty detection is also termed one-class classification and outlier detection [19]. The principle of novelty detection is to estimate the distribution of the most populous class, or the "normal" class, \( H_0 \). We then generate a novelty score \( z(x) \) for each test sample \( x \), and then classify the test sample as abnormal if the \( z(x) \) exceeds a decision threshold, often termed the novelty threshold. Comprehensive reviews can be found in [20], [21].

Pimentel et al. [20] have categorised novelty detection methods into five families based on how \( H_0 \) is estimated: (i) probabilistic, (ii) distance-based, (iii) reconstruction-based, (iv) domain-based, and (v) information theory-based methods.

Probabilistic novelty detection estimates the probability density function (pdf) of the normal class and classifies the test sample by evaluating if it is from the same distribution as the normal data. The decision can be based on statistic hypothesis tests or thresholding of the novelty score. Probabilistic methods can be further divided into parametric and non-parametric methods.

Parametric methods such as Gaussian mixture models (GMMs) make assumptions of the normal class distribution, and typically need a large dataset to evaluate the model parameters accurately; for time-series novelty detection, hidden Markov models [19], [22], [23], and Kalman filters [24] are widely used, which assume data are generated by unobserved states and where transitions between these states are time-dependent and can be described by probabilistic distributions.

Non-parametric methods include kernel density estimates (KDEs) [25], also called the Parzen window method, in which multivariate Gaussian distributions are centred on each data point to learn the distribution of the normal class. The variance of the Gaussian distribution in the KDE controls the smoothness of the estimated distribution, and can be evaluated heuristically as being, for example, the average distance between every point and its 10 nearest neighbours [19]. Another non-parametric method is negative selection [26], [27], in which abnormal data are detected by their characteristic features, which are absent in the normal set.

Distance-based methods include nearest neighbour and clustering methods. Nearest neighbour-based approaches assume that normal data points are located in a close neighbourhood of other data, while abnormal data are located far from the majority [28].

Clustering-based approaches assume that the normal data are characterised by a small number of "prototypes", and use the distance to the nearest prototype to quantify abnormality [29].

Reconstruction-based novelty detection is represented by neural networks [30], [31] and subspace-based methods [32]. Data are reconstructed based on predefined rules and the abnormality of test samples is then quantified by the reconstruction error. Subspace methods assume data can be better represented in a lower dimension than the original feature dimension for classification. The original feature space is mapped to the lower-dimensional subspace, where test samples are compared with artificially-generated data. Both methods are useful when normal data is well-characterised.

Domain-based novelty detection defines a boundary between abnormal and normal space. The support vector machine (SVM) is representative of this family, which uses a hyperplane to separate normal and abnormal classes in a high-dimensional space [20]. The SVM is natively suitable for two-class classification [21]. The one-class SVM for novelty detection was proposed by Schölkopf et al. [33]. There are two strategies to find the optimal boundary: maximal margin and minimal radius. The former finds the boundary which separates the two classes with the maximum margin; the latter finds the minimal radius to contain the normal class in the high-dimensional space defined by the SVM.

Information theory-based methods calculate the information content (e.g., entropy) of the dataset, and quantify abnormality by the change of information content associated with the elimination or addition of a test point or subset [20].

The thresholding of a novelty score, used in the above methods, is another factor for novelty detection methods. In probabilistic novelty detection the novelty score is itself a probability density value with no probabilistic interpretation [20]. Tarassenko et al. [34] have proposed using a cumulative density function in place of the pdf and thresholding it with empirical values; however this approach is difficult to adapt to a more general case, especially to systems where data dimensionality may change. Extreme value theory (EVT) was proposed to model the distribution of the decision boundary, and is useful in time-series novelty detection [20]. Clifton et al. demonstrated that extensions of EVT are required when the data are generated by multimodal or multivariate sources [35].

### III. Dataset Description

The China Kadoorie Biobank (CKB) is a collection of questionnaire responses, physical measurements, and blood samples acquired from over 520,000 adults from 10 areas in China during 2004-2008 [36]. Information concerning death rates and other healthcare outcomes, such as hospitalisation and the diagnosis of chronic diseases is also available. After five years, approximately 25,000 surviving participants were
resurveyed with further questionnaires, measurements, and blood acquisition. The data of this study include:

1) ECG time series: 10s duration, 500Hz, standard 12-lead ECG from 25019 participants was obtained by the Mortara device in 2008 and 2013-2014. Also available is a "typical cycle" from each lead for each participant; however, it is not clear how the typical cycles were generated. It is likely that the latter are the average of aligned and length-adjusted cycles in the 10s ECG signal, because the former are smoother than the latter.

2) features: 19 features were automatically extracted from the "typical cycles" for each participant by the VERITAS algorithm of the Mortara device. The meaning of the features were described in [37] and are illustrated and summarised in Figure 1 and Table I.

The R-peak of each cycle is aligned at \( t = 500 \text{ms} \) and the relative positions of the start and the end of the P, QRS, and T waves are extracted as features. Durations of the P, QRS, P-R, and Q-T intervals are subtracted from the position features. Corrected QT intervals \( QT_c \), \( QT_B \), and \( QT_cF \) were available, as were ECG axes (P axis, QRS axis, and T axis).

3) Textual labels: The Mortara device automatically provides up to 10 textual labels from 236 possibilities for each participant, such as sinus rhythm, normal ECG, and atrial fibrillation. These labels are produced by application of the Minnesota coding system [38], which is a heuristic scheme defined over 50 years ago. Sinus rhythm and normal ECG are the most commonly-observed in our dataset, representing 83.2% and 44.1% of all records, respectively.

The outputs of the Mortara device also include a label concerning abnormality for the waveform (abnormal ECG). The latter takes two values: 1 for abnormality and 0 for normality.

Labels for one-class novelty detection:
Sinus rhythm, normal ECG, and abnormal ECG=0 are the only three labels available in our dataset that are potential criteria for defining normality, and their relationship is shown in Figure 2. There are 19925, 10550, and 18397 records labelled sinus rhythm, normal ECG, and abnormal ECG=0, respectively. Sinus rhythm and abnormal ECG=0 have 15628 records in overlap. Normal ECG is a subset of the union of sinus rhythm AND abnormal ECG=0, which contains 10550 records. There are only 1675 complete records without any of the three labels for defining normality.

Since the labels are produced according to the heuristic Minnesota criteria, it is likely that they do not correspond closely to clinical reality. In addition, these criteria have a large number of rules in disagreement due to contradictions and inconsistencies in the coding system. As a result, it may not be convincing to treat any of these labels alone as the "gold standard" for use in defining normality when training a classification algorithm. This report compares the performance of three algorithms against five reasonable combinations of the three labels. The five combinations considered by us are (C1) sinus rhythm, (C2) normal ECG, (C3) abnormal ECG=0, (C4) sinus rhythm AND abnormal ECG=0, and (C5) sinus rhythm OR abnormal ECG=0. These will be referred to as the "normal criteria" C1-5.

Labels for three-class classification:
For three-class classification of "normal", "arrhythmia", and "ischemia", it is important to construct class models as precisely as possible. We also require criteria for the 3-class setting, and the schemes considered here are shown in Table II.

Under our three-class criteria, there are only 157 overlaps between arrhythmia and ischemia, and no overlap between "normal" and either of the other two classes, which is considered acceptable as arrhythmia and ischemia are not mutually exclusive. However it is worth noticing that almost half of the dataset fall into none of the three classes. These may be the "borderline data" which does not strictly belong to any of the classes.

Fig. 1: Mortara features shown for lead I (reproduced from [37])

Fig. 2: Relationship of Sinus Rhythm, Normal ECG, and Abnormal ECG=0. The numbers indicate how many records in the continuous area are. For example, the number 4374 means there are 4374 records labelled sinus rhythm but not abnormal ECG=0.

Incomplete records: 650
TABLE I: 19 Mortara features and 4 blood pressure features. Features are excluded either because they can be expressed as functions of other features or because they are constant.

<table>
<thead>
<tr>
<th>Included Features</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td></td>
</tr>
<tr>
<td>Average RR interval</td>
<td>Average distance between two R peaks</td>
</tr>
<tr>
<td>QRS offset</td>
<td>End of the QRS complex</td>
</tr>
<tr>
<td>P wave duration</td>
<td></td>
</tr>
<tr>
<td>PR interval</td>
<td>Interval between the onset of P wave and the onset of R peak</td>
</tr>
<tr>
<td>QRS duration</td>
<td>Duration of the QRS complex</td>
</tr>
<tr>
<td>QT duration</td>
<td>Interval between Q wave onset and T wave offset</td>
</tr>
<tr>
<td>P axis</td>
<td>Determined by deflects of P wave in different leads</td>
</tr>
<tr>
<td>QRS axis</td>
<td>Determined by deflects of QRS complex in different leads</td>
</tr>
<tr>
<td>T axis</td>
<td>Determined by deflects of T wave in different leads</td>
</tr>
<tr>
<td>SBP first</td>
<td>Systolic blood pressure in the first resurvey</td>
</tr>
<tr>
<td>SBP second</td>
<td>Systolic blood pressure in the second resurvey</td>
</tr>
<tr>
<td>DBP first</td>
<td>Diastolic blood pressure in the first resurvey</td>
</tr>
<tr>
<td>DBP second</td>
<td>Diastolic blood pressure in the second resurvey</td>
</tr>
</tbody>
</table>

Excluded Features

- Ventricular rate = 60000/Average RR interval
- R peak = 500ms
- P wave onset = Q offset - PR duration - QRS duration
- P wave offset = P onset + P duration
- QRS onset = PR duration + P onset
- T wave offset = Q onset + QT duration
- $QT_c$ duration = QT duration + 154(1 - 60/ventricular rate)
- $QT_{cB}$ duration = QT duration/(average RR interval)$^{1/2}$
- $QT_{cF}$ duration = QT duration/(average RR interval)$^{1/3}$

TABLE II: 3 Class inclusion criteria and ratios

<table>
<thead>
<tr>
<th>Class</th>
<th>Number in the class</th>
<th>Inclusion Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>10739</td>
<td>normal ECG</td>
</tr>
<tr>
<td>Arrhythmia</td>
<td>2323</td>
<td>abnormal rhythm, atrial fibrillation, early repolarisation, preexcitation, premature ectopias, blocks, uncertain rhythm</td>
</tr>
<tr>
<td>Ischemia</td>
<td>774</td>
<td>explicitly stated ‘ischemia’</td>
</tr>
<tr>
<td>Unclassified</td>
<td>10533</td>
<td>None of the above</td>
</tr>
</tbody>
</table>

IV. PREPROCESSING AND VISUALISATION

A. Feature Selection

Of the 19 features available from the Mortara device, 9 are excluded because they can be expressed as functions of other features or because they have constant values. These are shown in Table I.

We removed records with missing values for any of the features, and data for 24,369 participants were thus retained for use in our study.

B. Visualisation

We performed a 2-D visualisation of the 14-dimensional feature space according to the 5 normal criteria, C1-C5, using a Gaussian process latent variable model (GPLVM), which projects high-dimensional data into a lower-dimensional subspace. It may be seen from figure 3 that substantial overlap exists between the “normal” and “abnormal” classes, as is expected for a complex, realistic medical application.

C. Signal Quality Evaluation

A Signal Quality Index (SQI) was evaluated for all 12-lead ECG signals independently using in-house software [39]. The SQI, ranging from 0 to 1, depends on the agreement of 2 peak-detectors on the positions of the R-peaks. High agreement yields a high SQI, which corresponds to high signal quality. 97.11% of all 12-lead waveforms have high quality (SQI ≥ 0.9). Each lead has at least 92% signals with SQI ≥ 0.9. Lead V4 has the highest proportion (99.2%) of good-quality signals (SQI ≥ 0.9). The signal quality of our dataset is therefore sufficient to classify at least 92% of the 24,369 participants with complete ECG records.
D. Feature Extraction

6 additional features are extracted from lead I "typical cycles": amplitudes of P, Q, R, S, and T waves relative to the baseline, and ST level (which was approximated as the level of QRS offset). The baseline was approximated as the average level of the segment between P offset and Q onset, because this segment is used to define ST deviation in clinic. The positions of the onsets and offsets of these waves are provided by the Mortara device (part of the Mortara features).

V. ANALYSIS

A. One-class classification

We present three methods to predict the posterior probability of a feature vector belonging to the abnormal class, and hence predict their cardiovascular disease risk. The entire analysis process is illustrated in Figure 4, and is described in detail as follows:

1) Cross-Validation and Partitioning of Training and Test Sets: we performed 5-fold cross-validation by permuting the entire dataset and assigning a different 20% for each fold of cross-validation. Results shown later are the mean values over this 5-fold cross-validation.

   All sets were normalised column-wise according to the mean and standard deviation of the training set.

2) Balancing of the Test Sets: To make a fair comparison between the normal criteria C1-C5 which have different class ratios (i.e. balance between normal and abnormal data), we use the accuracy in balanced sets and AUC in both balanced and unbalanced sets for model evaluation.

   We therefore created a balanced test set (a subset of the unbalanced test set), containing all abnormal test data and the same number of normal data. The training set remains unbalanced.

3) Statistics: The definitions we use for accuracy, true positive rate (TPR), and false positive rate (FPR) are:

\[
\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)
\]

\[
\text{TPR} = \frac{TP}{TP + FN} \quad (2)
\]

\[
\text{FPR} = \frac{FP}{TN + FP} \quad (3)
\]

where TP: true positive, TN: true negative, FP: false positive, FN: false negative.

4) Generative KDE: We adapted the model described in [19]. In brief, the normal pdf was learned from the training set by placing a multivariate Gaussian distribution on each 14-dimensional data point. For ease of computation, we performed k-means clustering to summarise the normal data with 500 cluster centres in the 14-dimensional space. Only the most "normal" (i.e., those labelled Normal ECG) were used in clustering. The data likelihood is calculated via:

\[
p(x) = \frac{1}{N(2\pi)^{D/2} \sigma^D} \sum_{i=1}^{N} e^{-\frac{|x - x_i|^2}{2\sigma^2}} \quad (4)
\]
A novelty score, $y$, is then calculated using equation 5.

$$ y(x) = -\log p(x) $$

(5)

$$ P(C|y) = \frac{P(y|C)P(C)}{P(y)} $$

(6)

We propose treating this novelty score as a univariate summary of the 14-dimensional data, which may then subsequently be used by probabilistic models to predict the probability of test data belonging to the abnormal class. First we learn the likelihood $P(y|C)$ in the training set, by performing a kernel density estimation of the class-specific pdf. For the unbalanced test set, the prior $P(C)$ is set to the class ratio of the training set; for the balanced set, the class prior equals 0.5. The posterior $P(C|y)$ is

$$ P_{\text{test}}(C|y) = \frac{P_{\text{train}}(C|y)P_{\text{train}}(y)P_{\text{test}}(C)}{P_{\text{train}}(C)P_{\text{test}}(y)} $$

(7)

The posterior $P_{\text{test}}(C|y)$ is thresholded at 0.5 for classification.

5) **Discriminative KDE**: Alternatively we can feed the novelty score into a discriminative framework by solving the inference problem $P(C|y)$ directly [42]. The posterior is learned in the training set by binning $y$ and calculating the frequency of observing a particular class in a bin (figure 5). For example, for normal class $C_0$, the posterior in each bin is calculated as the proportion of data points belonging to $C_0$ in the bin. Ideally the bin size should be determined by cross-validation [40]. In this report, the bin size was set to $\delta y = 1$. The posterior of a set with a different prior, in this report the balanced test set, is calculated according to equation 7. Similarly the posterior is thresholded at 0.5 for classification.

6) **Discriminative SVM for one-class classification**: To compare the results of KDE we also used SVM. In brief, the original data space $I$ is projected into a higher dimensional feature space $F$ and a hyperplane in $F$ is used to separate the normal and the abnormal class with a maximal margin. The hyperplane is described by

$$ < w; x > + b = 0 $$

where the hyperplane coefficient $w$ belongs to $F$ and $b$ is real. The problem is equivalent to minimising

$$ \frac{1}{2} < w, w > + C \sum_i s_i $$

The coefficients $C$ and $\sigma$ (when using a Gaussian kernel) control the ”softness” of the separation boundary, and are optimised by grid search via 5 fold cross-validation on the training set. The classification score is mapped to probabilities and thus the training set posterior $P(C|y)$ is learned. For an unbalanced test set, the posterior is estimated to be equal to the training set, while the balanced test set posterior is reweighted using equation 9.

B. **Three class classification**

1) **Constructing the training and test sets**: We obtained balanced training and test sets by taking all data from the smallest class and the same number of data points from each of the other two classes were randomly selected to construct the training and test set for 5-fold cross-validation. To illustrate the distinctiveness of each class, two-class classifications between any two of the ”normal”, ”ischemia”, ”arrhythmia” are also performed for comparison. To evaluate how the ”normal” class differs from the other two classes, a ”normal vs abnormal” set was constructed with 50% data from ”normal” class and 25% from each of the other two classes.

2) **Training the 3-class model**: We used discriminative SVM to train three class models with or without the 6 additional features. Similar optimisation and posterior evaluation approaches as described in section V-A6 were used. For the remainder of the dataset which was not selected for training or testing, probabilities of class membership were predicted using trained posteriors.

C. **Results and Discussion**

1) **One-class classification**: Under all 5 criteria C1-C5, data from the normal class take overall lower novelty scores than data from the abnormal class, as shown in figure 6. However, the two classes are largely overlapped in the peak area of the abnormal class, suggesting that the generative KDE may not...
be able to separate them easily. Figure 6 shows the likelihood under the criteria C5, and the results for other criteria were similar.

In order to rule out the influence of different test sets to classification accuracy, we first evaluated the performance of discriminative SVM and generative KDE under the same (unbalanced) test set. Considering the substantiate overlap shown by the visualisation (Figure 3) and the likelihood (Figure 6), the generative KDE and discriminative SVM have relatively high AUC. Under C5 the AUC is high with both methods. The SVM model achieved higher AUC for all criteria, suggesting it may be a more robust method than KDE (Figure 7 and Table III).

In the balanced set the discriminative SVM achieved high accuracy, 71.1% to 75.6%, for all 5 criteria, while the generative KDE has comparable result (74.8%) under the criteria C5 (Figure IV).

The discriminative KDE has similar AUC values as the generative KDE, but lower accuracy, which implies a better optimisation of this method may be needed. Figure 7 shows that for different algorithms, the most distinctive criteria is different. For the generative KDE and the discriminative SVM the best-performing criterion is C5, and C1, respectively, suggesting different algorithms may favour different criteria.

It is unexpected that the most stringent criteria, C3, being the subset of all other 4 criteria, has not yielded best results with any of the algorithms considered. The criteria C5 is predicted accurately by both the generative KDE and discriminative SVM, suggesting it may be more appropriate as the “gold standard” for training algorithms for one-class classification.

2) Three-class classification: The three-class classification results with or without the 6 new features using the discriminative SVM are shown in figure 8. The numbers on the edge are classification accuracies between the two classes on the node, while the numbers in the middle of the triangle is three-class classification accuracy. The red numbers are these results after adding 6 new features, while the black numbers are results of using only the 14 original features without the 6 new features. The new features improved the results in all cases, most markedly in classification of ischemia. This agrees with our expectation since the Mortara features do not contain information concerning the amplitudes of the peaks, while ischemia is highly correlated with amplitude abnormalities, especially ST-levels.

The three-class classification in the balanced test set has achieved 74.5% with new features under the new criteria, suggesting that the three classes are generally distinctive under
our criteria. However, it is worth noting that the new criteria cannot classify nearly half of the dataset in the sense that the criteria only label half of our data.

VI. CONCLUSIONS AND FUTURE WORK

The first task of this research was to detect “abnormality”, by exploring different one-class novelty detection algorithms under various criteria of “normality”. The algorithms favoured criteria C5, which is the least stringent of all possibilities, and is thus appropriate for detection of the “real” outliers. In view of the relatively good performance of the discriminative SVM, in the second task of this study we modelled the normal, arrhythmia, and ischemia using a three-class SVM, aiming to produce accurate models for prediction of the unclassified datapoints. The encouraging results under the new criteria suggest the three class models may be appropriate to predict the probability of class membership of the “borderline” data that are otherwise difficult to classify.

We can further improve the classification accuracy by extracting more features, such as heart rate variability and T wave alternans, from more leads. Also the original 10s signal may lend more information than the ‘typical cycle’ as the former contains more time-dependent information than the latter. The length of the signal is a major limitation to our feature extraction, because many informative features such as ST-level need longer (>60s) signals to be evaluated accurately [41].

ACKNOWLEDGMENTS

The author gratefully acknowledges the support of the RCUK Digital Economy Programme grant EP/G036861/1 (Oxford Centre for Doctoral Training in Healthcare Innovation), the CKB, and the China Scholarship Council (CSC).

REFERENCES


TABLE IV: AUC and accuracy of predicting the 5 normal criteria by generative KDE, discriminative KDE, and discriminative SVM in the balanced sets. Results are presented as the mean±standard deviation in 5-fold cross-validation.

<table>
<thead>
<tr>
<th>SVM Discriminative</th>
<th>Generative KDE</th>
<th>Discriminative KDE</th>
</tr>
</thead>
<tbody>
<tr>
<td>AUC</td>
<td>Accuracy %</td>
<td>AUC</td>
</tr>
<tr>
<td>C1</td>
<td>0.79 71.7</td>
<td>0.73 67.2</td>
</tr>
<tr>
<td>C2</td>
<td>0.79 71.6</td>
<td>0.73 63.5</td>
</tr>
<tr>
<td>C3</td>
<td>0.82 77.1</td>
<td>0.72 65.1</td>
</tr>
<tr>
<td>C4</td>
<td>0.80 73.7</td>
<td>0.75 64.5</td>
</tr>
<tr>
<td>C5</td>
<td>0.83 75.6</td>
<td>0.81 74.8</td>
</tr>
</tbody>
</table>


