Risk Prediction for Cardiovascular Disease using ECG Data in the China Kadoorie Biobank

Yanting Shen1, Yang Yang1,2, Sarah Parish3, Zhengming Chen3, Robert Clarke3, David A. Clifton1

Abstract—We set out to use machine learning techniques to analyse ECG data to improve risk evaluation of cardiovascular disease in a very large cohort study of the Chinese population. We performed this investigation by (i) detecting “abnormality” using 3 one-class classification methods, and (ii) predicting probabilities of “normality”, arrhythmia, ischemia, and hypertrophy using a multiclass approach.

For one-class classification, we considered 5 possible definitions for “normality” and used 10 automatically-extracted ECG features along with 4 blood pressure features. The one-class approach was able to identify abnormality with area-under-curve (AUC) 0.83, and with 75.6% accuracy.

For four-class classification, we used 86 features in total, with 72 additional features extracted from the ECG. Accuracy for this four-class classifier reached 75.1%. The methods demonstrated proof-of-principle that cardiac abnormality can be detected using machine learning in a large cohort study.

I. INTRODUCTION

Cardiovascular diseases (CVD) are the leading causes of mortality worldwide and in China [1]. There are large geographical and economic variations of CVD mortality in China [2], suggesting appropriate measures are needed for prevention and effective treatment of the disease. The World Health Organisation advises people at high CVD risk to access early detection and treatment for prevention of CVD [1]. Identifying risks in individuals in the population could help to provide advice to people to improve their lifestyle and help clinicians to discover appropriate treatments for specific conditions to reduce mortality and healthcare expenditure.

Traditional CVD risk factors include smoking, hypercholesterolaemia, hypertension, diabetes, and obesity [2] among many others. Traditional risk factors do not fully explain the risk of CVD in populations. Personalised medicine requires integrating all risk factors to which a person is exposed and then predicting risks for specific diseases, so to optimise preventative measures for individuals. Examples of research for this purpose include [3], in which 5-year mortality rate was predicted on 20 risk factors from 498,103 participants in UK Biobank using proportional hazard models and Harrell’s C-index. 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 is difficult to 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 [4]. For example, in CVD, Knuiman et al. have predicted coronary mortality in the Busselton cohort using logistic regression[5]. The electrocardiogram (ECG), being an important measurement of cardiac function that is relatively easy to obtain, is surprisingly rarely used for risk prediction. The aim of this paper is 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 dataset. This paper describes two risk evaluation tasks: (i) “abnormality” detection and (ii) prediction of probabilities of “normal”, arrhythmia, ischemia, and hypertrophy. 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 “normality”, “arrhythmia”, “ischemia”, and “hypertrophy” using a multiclass approach.

II. DATASET DESCRIPTION

The China Kadoorie Biobank (CKB) is a prospective cohort study of over 520,000 adults from 10 areas in China during 2004–2008 [6]. Data were collected using questionnaires and anthropometric and physiological measurements were recorded at baseline and all participants provided a blood sample. Information on cause of death rates was collected from health insurance data and mortality and disease registries. After five years, approximately 25,000 surviving participants were resurveyed with further questionnaires, measurements, and blood collection. We have institutional ethics approval to use the data. Public access to the CKB data can be found at http://www.ckbiobank.org/site/Data+Access. The data available for our study include:

A. ECG time series

Standard 12-lead ECG (10-s duration, 500Hz) was recorded on 24,369 participants using a Mortara ELIxi50 device in 2013-2014. Also available is a “typical cycle” from each lead for each participant, which was generated by the device using a proprietary algorithm.

B. ECG Features

The Mortara device provides 10 main features (age, average RR interval, P wave duration, the time point of QRS offset, PR interval, QRS duration, QT duration, P axis, QRS axis, and T axis) which were automatically extracted from the “typical cycles” for each participant. A schematic
representation of a selection of these features is shown in Figure 1.

C. Blood pressure data

Systolic blood pressure and diastolic blood pressure were recorded twice on each individual using standard methods.

D. 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 were produced by application of the Minnesota coding system which is a heuristic scheme defined. Sinus rhythm and normal ECG are the most commonly-observed textual labels in our dataset, representing 83.2% and 44.1% of all records, respectively.

The outputs of the Mortara device included a label for abnormality for the waveform (abnormal ECG), which was coded as 1 for abnormality and 0 for normality.

1) Labels for one-class novelty detection: Sinus rhythm, normal ECG, and abnormal ECG=0 were the only three labels available in our dataset that were used as potential criteria for defining normality. There were 19,925, 10,550, and 18,397 records labelled sinus rhythm, normal ECG, and abnormal ECG=0, respectively. Sinus rhythm and abnormal ECG=0 had 15,628 records in overlap. Normal ECG, which was a subset of the union of sinus rhythm and abnormal ECG=0, included 10,550 records. There are only 1,675 complete records without any of the three labels for defining normality.

Since the labels were produced according to the heuristic Minnesota criteria, their relationships to diseases require future study. 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 being 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.

2) Labels for four-class classification: For 4-class classification of “normal”, “arrhythmia”, “ischaemia” and “hypertrophy”, it is important to construct class models as precisely as possible. We also require criteria for the 4-class setting, and here we consider schemes according to [7] which are shown in Table I.

III. Feature Extraction

Six additional features were extracted from each of the 12-lead “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) constitute 72 additional features which were added to the 14 features in the one-class classification for use in the four-class classification. 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. We note that those 72 features derived from the ECG time-series were in addition to the 10 features automatically produced by the Mortara device, and the 4 blood pressure measurements, described earlier.

The positions of the onsets and offsets of these waves are provided by the Mortara device (part of the Mortara set of features described above).

IV. 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.

1) Cross-Validation and Partitioning of Training and Test Sets: The training and test sets were generated using a 5-fold cross-validation by permuting the entire dataset and assigning a different 20% for each fold of cross-validation as the test set. Results shown later are the mean values over this 5-fold cross-validation. All training and test sets were normalised and standardised according to the training sets.
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 and AUC in balanced 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 remained unbalanced. The balanced test sets under criteria C1-5 contain 1,824, 5,688, 2,452, 3,586, and 688 data points respectively.

3) Generative Kernel Density Estimator: We adapted the model described in [8]. In brief, the normal probability density function 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:

\[
p(x) = \frac{1}{N(2\pi)^{D/2} \sigma^D} \sum_{i=1}^{N} e^{-\frac{|x-x_i|^2}{2\sigma^2}}
\]

A novelty score, \( y \), is then calculated using equation 2.

\[
y(x) = -\log p(x)
\]

\[
P(C|y) = \frac{P(y|C)P(C)}{P(y)}
\]

We propose treating this novelty score as a univariate summary of the 14-dimensional data, which may then subsequently used by probabilistic models to predict the probability of test data belonging to the abnormal class. First, we estimated 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 \( P(C) \) was 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

\[
P_{test}(C|y) = \frac{P_{train}(y|C)P_{train}(y)P_{test}(C)}{P_{train}(C)P_{test}(y)}
\]

The threshold for posterior \( P_{test}(C|y) \) was set at 0.5 for classification.

4) Discriminative KDE: Alternatively, we added the novelty score into a discriminative framework by solving the inference problem \( P(C|y) \) directly. The posterior was estimated in the training set by binning \( y \) and calculating the frequency of observing a particular class in a bin. For example, for normal class \( C_0 \), the posterior in each bin was calculated as the proportion of data points belonging to \( C_0 \) in the bin. Ideally the bin size should be determined by cross-validation [9]. In this paper, the bin size was set to \( \delta y = 1 \). The posterior of a set with a different prior, in this paper the balanced test set, was calculated according to equation 4. Similarly the posterior was thresholded at 0.5 for classification.

5) Discriminative Support Vector Machine for one-class classification: To compare the results of KDE, we also used SVM. The coefficients \( C \) and \( \sigma \) (when using a Gaussian kernel) control the “flexibility” of the separation boundary, and were optimised by a grid search via 5 fold cross-validation on the training set. The classification score was mapped to probabilities and thus the training set posterior \( P(C|y) \) was learned.

B. Four-class classification

1) Constructing the training and test sets: We obtained balanced training and test sets by taking all data from the smallest class (Table I) and the same number of data points from each of the other classes were randomly selected to construct the training and test sets for 5-fold cross-validation. For example, the four-class balanced training-and-test set contains 1808 × 4 = 7472 datapoints. To illustrate the distinctiveness of each class, three-class and two-class classifications of any combinations of the “normal”, “ischemia”, “arrhythmia”, and “hypertrophy” were also performed for comparison using the same approach to balance classes.

2) Training the 4-class model with Support Vector Regression: A \( P(C|x) \) was estimated for each of the classes using support vector regression in a one-vs-all approach; i.e. the regressor \( i \) was learned in a training set with only the class \( i \) labelled 1 and other classes were labelled 0. The class probability \( P(C|x) \) was calculated from the predicted value of the regressor \( i \) according to Equation 5. Finally the data point was classified to the class with the highest probability:

\[
P_i(C_j|x) = \frac{e^{-|y-x_i|}}{2\pi \sum_{j=1}^{4} P_j(C_j|x)}, i = 1, 2, 3, 4
\]

C. Results and Discussion

1) One-class classification: In the balanced hold-out test set, the discriminative SVM achieved high accuracy, 71.6% to 75.6%, for all 5 candidate criteria for defining “normality”, while the generative KDE had comparable result (74.8%) under the criteria C5 (Figure II).

The discriminative KDE had similar AUC values as the generative KDE, but lower accuracy, which implies a better optimisation of this method may be needed.

For both KDE and SVM C5 was the best-performing criterion, suggesting C5 may be the most appropriate criterion among the 5 studied to be a “gold standard” for training algorithms for one-class classification.

2) Four-class classification: The four-class classification results using support vector regression are shown in Figure 2. The 72 new features improved the results in all cases, most markedly in classifications involving ischemia and hypertrophy. This agrees with our expectation since the 10 Mortara features do not contain information concerning the amplitudes of the peaks, while ischemia and hypertrophy
TABLE II
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</th>
<th>Discriminative</th>
<th>Generative KDE</th>
<th>Discriminative KDE</th>
</tr>
</thead>
<tbody>
<tr>
<td>AUC</td>
<td>Accuracy %</td>
<td>AUC</td>
<td>Accuracy %</td>
</tr>
<tr>
<td>C1</td>
<td>0.79</td>
<td>71.7</td>
<td>0.73</td>
</tr>
<tr>
<td>C2</td>
<td>0.79</td>
<td>71.6</td>
<td>0.73</td>
</tr>
<tr>
<td>C3</td>
<td>0.82</td>
<td>77.1</td>
<td>0.72</td>
</tr>
<tr>
<td>C4</td>
<td>0.80</td>
<td>73.7</td>
<td>0.75</td>
</tr>
<tr>
<td>C5</td>
<td>0.83</td>
<td>75.6</td>
<td>0.81</td>
</tr>
</tbody>
</table>

Fig. 2. Classification accuracies. Numbers on the edges are classification accuracies between the two classes on the nodes; numbers in the centres of the triangles are classification accuracies of the 3 classes on the nodes of the triangle. The red and black numbers are results with and without the 72 features derived from the ECG, over and above the basic set of the 10 features provided by the Mortara device and the 4 blood pressure features, respectively.

were highly correlated with amplitude abnormalities, especially ST-levels, and R-S amplitudes.

It is encouraging that the classification accuracies with the new features are 30 to 50 percentage points higher than those that would be obtained by chance, suggesting machine learning methods can achieve high agreement with clinical knowledge, without resorting to a complex rule-based system. However it is worth noting that all labels here were provided by the Mortara device, and our models need further validation by comparing with labels from human experts in ECG resurvey.

V. 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 was the least stringent of all possibilities, and is thus appropriate for detection of outliers. In view of the relatively good performance of the discriminative SVM, in the second task of this study we modelled the normal, arrhythmia, ischemia, and hypertrophy using multiclass support vector regression, aiming to produce accurate models for prediction of the unclassified data points. The encouraging results suggest the multiclass 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. 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. Future work will link our analysis of ECG data to electronic health records for each participant, obtained from the Chinese medical insurance system, disease and mortality registries.

ACKNOWLEDGMENT
The authors gratefully acknowledge the support of the RCUK Digital Economy Programme grant EP/G036861/1 (Oxford Centre for Doctoral Training in Healthcare Innovation), China Scholarship Council (CSC), the K.C. Wong Fellowship, the EPSRC, and the Royal Academy of Engineering. The China Kadoorie Biobank study was supported by grants from the MRC, Wellcome Trust, and the British Heart Foundation.

REFERENCES