Conditional Random Fields for Morphological Analysis of Wireless ECG Signals

Annamalai Natarajan
School of Computer Science
University of Massachusetts
Amherst, MA
anataraj@cs.umass.edu

Edward Gaiser
Department of Psychiatry
Yale University
New Haven, CT
edward.gaiser@yale.edu

Gustavo Angarita
Department of Psychiatry
Yale University
New Haven, CT
gustavo.angarita@yale.edu

Robert Malison
Department of Psychiatry
Yale University
New Haven, CT
robert.malison@yale.edu

Deepak Ganesan
School of Computer Science
University of Massachusetts
Amherst, MA
dganesan@cs.umass.edu

Benjamin Marlin
School of Computer Science
University of Massachusetts
Amherst, MA
marlin@cs.umass.edu

ABSTRACT

Thanks to advances in mobile sensing technologies, it has recently become practical to deploy wireless electrocardiograph sensors for continuous recording of ECG signals. This capability has diverse applications in the study of human health and behavior, but to realize its full potential, new computational tools are required to effectively deal with the uncertainty that results from the noisy and highly non-stationary signals collected using these devices. In this work, we present a novel approach to the problem of extracting the morphological structure of ECG signals based on the use of dynamically structured conditional random field (CRF) models. We apply this framework to the problem of extracting morphological structure from wireless ECG sensor data collected in a lab-based study of habituated cocaine users. Our results show that the proposed CRF-based approach significantly out-performs independent prediction models using the same features, as well as a widely cited open source toolkit.

Categories and Subject Descriptors

I.2 [**Computing Methodologies**]: Artificial Intelligence—*Machine Learning*; J.3 [**Computer Applications**]: Life and Medical Science—*health*

General Terms

Experimentation, Performance, Algorithms

Keywords

Electrocardiogram, Machine Learning, Mobile Health

BCB '14 , September 20 - 23 2014 Newport Beach, CA, USA
Copyright 2014 ACM 978-1-4503-2894-4/14/09 ...$15.00.
http://dx.doi.org/10.1145/2649387.2649414 .

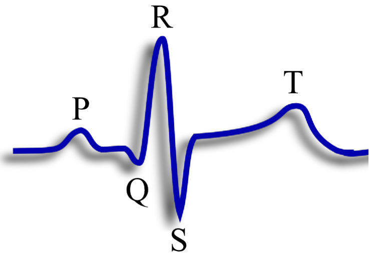


Figure 1: ECG trace of a single normal cardiac cycle with associated peak labels.

1. INTRODUCTION

An electrocardiograph is an instrument that measures changes in the electrical potential on the surface of the skin caused by the polarization and depolarization of the muscles of the heart [9, 11]. A typical normal heartbeat produces a sequence of five deflections away from the baseline potential that are referred to as the P wave, Q wave, R wave, S wave, and T wave, as illustrated in Figure 1 [9, 11]. The Q, R and S waves taken together are referred to as the *QRS complex*. The shapes of the individual waves, the locations of their peaks, and the intervals between pairs of waves carry detailed information about heart function. A classical problem in ECG analysis is the use of these morphological features to classify individual heartbeats as being normal or as expressing one of a set of arrhythmias [5].

Recent advances in mobile systems technology have made it possible to continuously record ECG data outside of laboratory and clinical settings using unobtrusive wearable sensors such as the AutoSense sensor suite [10] and the Zephyr BioHarness chest band [35]. This capability has diverse applications in the study of human health and behavior. However, to realize the full potential of these mobile ECG sensors, new computational tools are required to reliably extract detailed morphological information from ECG traces while effectively dealing with the uncertainty that results from the noisy and highly non-stationary signals.

In this work, we present a new approach to the problem of extracting the morphological structure of ECG signals based on the use of dynamically structured conditional random field (CRF) models [22]. We view the morphology extraction problem as the problem of identifying and labeling the peaks of each P, Q, R, S, and T wave in a given ECG trace. Our approach begins by over-generating candidate peak locations using a peak detection algorithm. Next, we use the output of the peak detector to dynamically instantiate a CRF graph where each label variable corresponds to a candidate peak location. Edges are created in the CRF between successive peak locations. Feature vectors are extracted from a local window around each candidate peak location and are associated with the corresponding label variable. Finally, exact probabilistic inference is used to jointly infer the most likely labeling of the complete sequence. We introduce an additional label N to represent candidate peak locations that do not correspond to valid waves. Since the model is chain-structured, exact inference is computationally efficient, scaling linearly with the number of candidate peaks in a sequence.

A drawback of a model-based framework is that before the model can be applied, it's parameters must be learned from data. Learning in chain-structured CRFs is also computationally efficient, but it requires labeled training sequences. To generate training sequences, we run the peak detection algorithm to extract candidate peaks, and then manually supply labels for those locations only. For the proposed approach to be useful in practice, it must generalize to new subjects given no or very limited training data. To this end, we evaluate our proposed framework in several learning settings including learning across-subject models, learning subject-specific models independently, and learning subject-specific models using transfer learning.

To evaluate our approach, we focus on the challenging domain of morphology extraction from wireless ECG data in the presence of cocaine use [28, 16]. The electrophysiology of the heart is directly affected by the presence of drugs like cocaine and atropine. These drugs have a well-understood large-scale impact on the cardiovascular system, causing an overall increase in heart rate [32]. They are also reported to induce a variety of specific morphological changes detectable in ECG traces including prolongation or shortening of the QT interval and flattening of the T wave [13, 24, 25, 34]. There is thus significant interest in the use of ECG morphological features to identify drug use events both for the purpose of monitoring individuals and for furthering the understanding of addiction [28, 16]. To support the evaluation of our proposed approach, we manually labeled over 20,000 candidate ECG peaks from six wireless ECG traces of habituated cocaine users who participated in a NIDA-approved clinical study of cocaine use. We use this data to assess the performance of our proposed approach compared to logistic regression and the well known ECGPUWave toolbox [30]. Our results show that our CRF framework out-performs both alternative approaches across a wide range of settings.

2. BACKGROUND AND RELATED WORK

In this section we briefly review ECG data analysis, the use of ECG data in mHealth, and the the CRF and sparse coding models that our proposed framework is based on.

2.1 ECG Data Analysis

While the computational analysis of ECG signals has been investigated since the 1960s [33], the vast majority of past work has focused on two specific data analysis problems: identification of QRS complexes and heartbeat classification. Pan and Tompkins developed a widely used and widely cited QRS complex detection algorithm based on simple features of the ECG trace. Their approach achieves a QRS detection accuracy rate of 99.325% on the well-known MIT-BIH data set [31]. However, systematic errors were noted in cases where the ECG signals contained stretches of noise, baseline shifts, unusual morphology and other artifacts. More recent work on QRS complex detection has focused on methods based on various transforms including the curve length transform [36] and the wavelet transform [27]. Both of these approaches give QRS complex identification precision and recall rates above 99.5% on standard databases.

The problem of interest in this work is morphological labeling of the ECG trace including the identification of each P, Q, R, S and T wave, when present. The most common approach to this problem is to first identify QRS complexes using one of the methods described above. A set of rules and a local search procedure are then used to identify the individual waves [19, 27]. A downside of these approaches is that a large number of threshold parameters are involved in the local search procedure. The method of Martinez et al. [27], for instance, depends on fifteen threshold parameters that are set by hand. More recent work has used supervised learning to select the set of scales used in the wavelet decomposition [6].

The work of Hughes et al [18] and de Lannoy et al [8] has addressed the ECG segmentation problem using hidden Markov models (HMMs). However, Hughes et al. specify the HMM directly over raw ECG samples and partially specify the transition structure by hand. De Lannoy et al. specify the HMM over coefficients of multiple mother wavelets and additionally make an assumption that all windows of ECG data start with a P wave. Both approaches are forced to introduce self transition constraints into the model to counter the natural geometric distribution of self transition times inherent in an HMM. Our approach deals with this issue more elegantly by defining the CRF graph over candidate peak locations instead on raw ECG sample values. Our approach also has the natural advantages inherent in the use of a discriminative model over a generative model when applied to a discriminative task [22]. Finally, we note that CRFs have been applied to ECG data previously, but for the problem of heartbeat classification, not peak labeling or segmentation [7]. In the work of de Lannoy et al, the CRF labels correspond to the beat type of each complete cardiac cycle. In fact, their work uses the method of Martinez et al. to extract morphological features [7].

2.2 ECG in Mobile Health

A substantial body of work has explored the use of mobile ECG sensors for applications in health and behavioral science, primarily in the context of understanding physiological stress [1, 14], assessing cognitive load [12], detection of arrhythmias caused specifically by atrial fibrillation [3, 4, 17], and detection of drug use including cocaine use [28, 16]. Nearly all of these studies have been based on features derived from R-R intervals only (heart rate, heart rate variability). In the cocaine use context, cocaine causes a gross increase in heart rate, however, similar increases can be caused by a variety of physical activities. Hossain et al. address this issue by combining ECG data with actigraphy data (accelerometer readings) [16]. Our focus is on making the extraction of nuanced morphological features maximally reliable in an offline data analysis context with the hope that

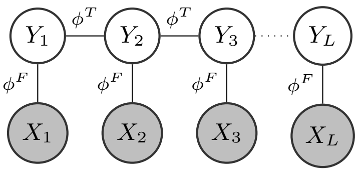


Figure 2: Linear Chain CRF

they will find wide application across a variety of domains, including the study of addiction to drugs like cocaine.

2.3 Conditional Random Fields

Our approach to ECG peak labeling is based on exact probabilistic inference in chain-structured conditional random fields [22]. CRFs are a sub-class of probabilistic graphical models [20] that generalize independent probabilistic classifiers like logistic regression [15] to the case of structured prediction. CRF models contain feature variables and label variables connected in a graph that captures problem-specific probabilistic dependencies between the variables.

Figure 2 shows a linear chain CRF. The shaded nodes $\mathbf{X}_1$ to $\mathbf{X}_L$ represent the feature variables, and the unshaded nodes Y_1 to Y_L are the corresponding label variables. We assume the label variables take values in the set $\mathcal{V}$. The feature variables $\mathbf{X}_i$ typically represents a D-dimensional vector of feature values X_{id}. Each feature variable X_{id} and label value v are associated with a feature potential ϕ^F_{dv} that captures the dependence between the features and the associated labels. Each pair of adjacent labels are associated with a transition potential $\phi^T_{vv'}$ to capture the structure of the transitions between adjacent label values.

In a CRF model, the probability of a sequence of labels $\mathbf{y} = [y_1, ..., y_L]$ conditioned on the observed feature variables $\mathbf{x} = [\mathbf{x}_1, ..., \mathbf{x}_L]$ is given by,

$$P_{\mathbf{W}}(\mathbf{y}|\mathbf{x}) = \frac{\exp(-E_{\mathbf{W}}(\mathbf{y}, \mathbf{x}))}{Z_{\mathbf{W}}(\mathbf{x})} \tag{1}$$

where E is the energy function of the model and Z is the partition function. The feature and transition potentials that define a CRF model are parametrized by a set of weights $\mathbf{W} = [\mathbf{W}^F, \mathbf{W}^T]$. The energy function is given by,

$$\begin{aligned} E_{\mathbf{W}}(\mathbf{y}, \mathbf{x}) = -\Big(& \sum_{i=1}^{L} \sum_{d=1}^{D} \sum_{v \in \mathcal{V}} W^F_{dv}[y_i = v] x_{id} \\ & + \sum_{i=1}^{L-1} \sum_{v \in \mathcal{V}} \sum_{v' \in \mathcal{V}} W^T_{vv'}[y_i = v][y_{i+1} = v'] \Big) \end{aligned} \tag{2}$$

The partition function is given by,

$$Z_{\mathbf{W}}(\mathbf{x}) = \sum_{\mathbf{y} \in \mathcal{V}^L} \exp(E_{\mathbf{W}}(\mathbf{y}, \mathbf{x})) \tag{3}$$

The unknown parameters $\mathbf{W} = [\mathbf{W}^F, \mathbf{W}^T]$ must be learned from training data before the model can be applied. They can be estimated by maximizing the ℓ_2 regularized conditional log likelihood

$$\mathcal{L}(\mathbf{W}|\mathcal{D}) = \sum_{n=1}^{N} \log P_{\mathbf{W}}(\mathbf{y}_n|\mathbf{x}_n) - \lambda ||\mathbf{W} - \mathbf{W}_R||_2^2 \tag{4}$$

given a data set $\mathcal{D} = \{(\mathbf{y}_n, \mathbf{x}_n)\}_{n=1:N}$ of fully labeled training sequences. In standard CRF learning, the regularization target parameters $\mathbf{W}_R = 0$, which yields a standard ridge regression estimator. However, the regularization target parameters $\mathbf{W}_R$ can also be set to model parameter estimates derived from an alternative data set, effecting a simple, but powerful form of transfer learning. In either the ridge or transfer case, this objective function is strongly convex so gradient-based methods are guaranteed to find the unique optimal solution. Computing the gradients requires all single label variable marginal probabilities as well as pairwise marginal probabilities for all pairs of adjacent label variables [20]. All of these marginal distributions can be found in time linear in the length of the chain (as can the partition function) using the well-known sum-product belief propagation algorithm [20]. Once a CRF model is learned, belief propagation (max-product or sum-product) can be used to infer labels for an entire sequence in time linear in the length of that sequence [20]. In this work, we employ a max marginal labeling approach.

Finally, we note that multinomial logistic regression (MLR) is a special case of a CRF were the edges between the label variables have been removed. The probability that a single label variable Y_i takes value v given feature vector $\mathbf{x}_i$ is given below. We note that this model can also be learned using regularized maximum likelihood, including the use of a transfer-based regularizer.

$$P_{\mathbf{W}}(Y_i = v|\mathbf{x}_i) = \frac{\exp(\sum_{d=1}^{D} \sum_{v \in \mathcal{V}} W^F_{dv}[y_i = v] x_{id})}{\sum_{v \in \mathcal{V}} \exp(\sum_{d=1}^{D} \sum_{v \in \mathcal{V}} W^F_{dv}[y_i = v] x_{id})} \tag{5}$$

2.4 Sparse Coding

Sparse coding is a method for re-representing a D-dimensional data vector $\mathbf{x}$ in terms of sparse linear combinations $\sum_{k=1}^{K} \alpha_k \mathbf{B}_k$ of a set of K basis vectors $\mathbf{B}_k$ [29]. Given a set of basis vectors $\mathbf{B}_k$, the sparse coefficient vector α is computed as the solution to the following ℓ_1 regularized optimization problem:

$$\arg\min_{\alpha} \left|\left| \mathbf{x}_n - \sum_{k=1}^{K} \alpha_k \mathbf{B}_k \right|\right|_2^2 + \lambda ||\alpha||_1 \tag{6}$$

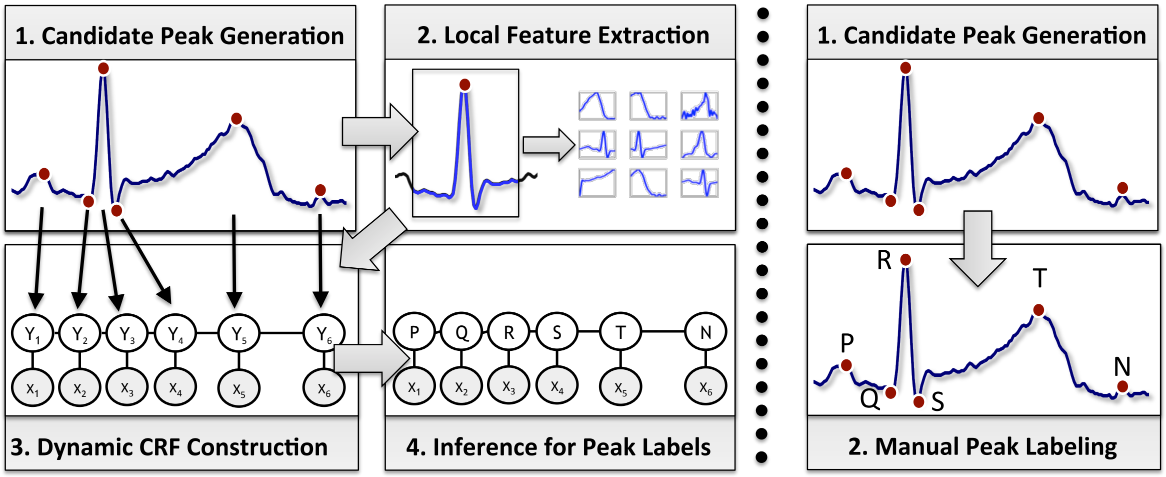
Given a data set $\mathcal{D} = \{\mathbf{x}_n\}_{n=1:N}$, the basis itself is learned to minimize the sum of the errors between each data case and it's reconstruction under the constraint of sparse coefficient vectors, as seen below. The typical approach to solving this problem is an alternating minimization strategy. We used the SPAMS toolbox [26] to perform sparse coding.

$$\min_{\alpha_{1:N}, \mathbf{B}} \sum_{n=1}^{N} \left(\left|\left| \mathbf{x}_n - \sum_{k=1}^{K} \alpha_{nk} \mathbf{B}_k \right|\right|_2^2 + \lambda ||\alpha_n||_1 \right) \tag{7}$$

The advantage of sparse coding over methods like principal components analysis (PCA) is that it produces sparse feature vectors, which can help to reduce over-fitting when these features are used for classification. Unlike PCA, sparse coding can also be used to learn an over-complete basis ($K > D$). This can help to make classification problems easier by making the feature vectors more linearly separable than the original data in the higher-dimensional feature space.

combination

3. DYNAMIC CRFS FOR ECG MORPHOLOGY EXTRACTION


1. Candidate Peak Generation
2. Local Feature Extraction
1. Candidate Peak Generation
Y1
Y2
Y3
Y4
Y5
Y6
X1
X2
X3
X4
X5
X6
P
Q
R
S
T
N
X1
X2
X3
X4
X5
X6
R
T
P
N
Q
S
3. Dynamic CRF Construction
4. Inference for Peak Labels
2. Manual Peak Labeling

(a) Morphology Extraction Pipeline

(b) Ground Truth Data Labeling Pipeline

Figure 3: (a) illustrates the ECG morphology extraction pipeline. (b) illustrates the ground truth data labeling pipeline.

In this section, we describe our CRF framework for ECG morphology extraction. We view the morphology extraction problem as the problem of identifying and labeling the peaks of each P, Q, R, S, and T wave in a given ECG trace. We first describe our morphology extraction framework. We then turn to the problems of data labeling and model learning.

3.1 Morphology Extraction Pipeline

Our approach to peak labeling consists of four primary steps: candidate peak generation, feature extraction, CRF graph generation and CRF inference. These steps are illustrated in Figure 3. Before applying these steps, we perform a small amount of pre-processing on the raw ECG data.

• **Pre-processing:** Raw ECG data is measured in millivolts and is typically recorded at hundreds of samples per second. Over the extended time periods, typically encountered in mHealth settings, ECG data from wireless on-body sensors exhibits significant baseline drift. We apply a standard low-pass Gaussian filter with a standard deviation of 600ms to estimate the baseline drift. We subtract the estimated drift from the raw data to yield baseline corrected data. All of our subsequent processing is based on baseline corrected ECG data.

• **Candidate Peak Generation:** The core of our approach is based on the idea of over-generating a set of candidate peak locations that will subsequently be labeled. Our aim is for this set to include the locations of all valid P, Q, R, S and T peaks, as well as a minimal number of additional peaks caused by noise and other artifacts in the ECG trace. Candidate peak generation is illustrated in Step 1 of Figure 3a. In this work, we apply Billauer's PeakDet method as we have found it be simple, fast and robust to noise [2]. However, we note that any peak detection algorithm that is robust to noise can be used for this component of the framework.

• **Feature Extraction:** Given a set of candidate peak locations, we next extract features from the ECG data in the local neighborhood of each candidate peak. Specifically, we define a window of width 204ms (±25 samples) centered at each peak location and extract features from the ECG data contained in that window. In this work, we use sparse coding [29] to learn an over-complete basis from ECG data in a fully unsupervised manner. Sparse coding is an attractive choice for this application as it aims to describe each 204ms waveform as resulting from a sparse linear combination of basis vectors. The sparse coefficient vectors of these linear combinations are the sparse coding feature vectors. Sparse coding feature extraction is illustrated in Step 2 of Figure 3a. We combine the sparse coding feature vectors with additional features representing the height of each candidate peak. However, we note that any set of feature extraction methods could be used for this component of the framework, including a fixed wavelet basis or other adaptive models such as PCA.

• **Dynamic CRF Construction:** Given a set of candidate peak locations and their corresponding features, we construct a dynamic CRF model. We instantiate one label variable Y_i and one feature variable $\mathbf{X}_i$ for each candidate peak location i. Importantly, we augment the label set with an additional label N to indicate candidate peaks that do not correspond to the extrema of valid waves. We set the feature vector $\mathbf{x}_i$ to the feature vector extracted for candidate peak i in the previous step. Finally, we connect adjacent label variables to form chain-structured graph. This process is illustrated in Step 3 of Figure 3a. Compared to the default approach of associating one label variable with each ECG sample, our approach constructs a CRF with nearly 50 times fewer label variables on average.

• **Inference for Peak Labels:** Once a CRF has been dynamically instantiated given the candidate peak locations, standard probabilistic inference methods can use used to infer the most likely values for the labels of the candidate peaks taking into account all of the information contained in the CRF model. The restriction to a chain-structured graph permits the application of linear-time exact inference methods. Compared to an independent classification model like logistic regression, the CRF model is able to leverage the high degree of regularity in the ECG peak label transitions to aid in determining labels in regions of high noise. Compared to a method for morphological extraction based on QRS complex detection followed by local search for other peaks, the CRF model has the advantage that it determines all peak labels jointly. This makes it more robust in cases where the local evidence for identifying QRS complexes is weak due to transient noise, but other peaks like P or T are clearly discernible. Inference for an ECG trace with six peaks is illustrated in Step 4 of Figure 3a.

3.2 Data Labeling and Model Learning

The primary disadvantage of a model-based approach to ECG morphology extraction is that it requires labeled data to estimate the model parameters. An advantage of our approach is that it is not necessary to fully label the raw ECG data to indicate which wave each individual sample belongs to. Instead, we first run the peak detection method to generate a set of candidate peak locations and then manually specify labels for the candidate peak locations only. This makes the entry of label information much faster. This approach is illustrated in Steps 1 and 2 in Figure 3b.

We also note that it is not necessary to fully label each sequence of candidate peak locations. For a chain-structured CRF, the learning algorithm only needs access to labels for pairs of adjacent label variables to estimate the transition parameters $\mathbf{W}^T$. For each available ECG trace, we fully labeled all candidate peaks in multiple short segments consisting of one to three cardiac cycles. We refer to these segments as *clusters*. We designed a simple GUI to implement this labeling approach. Once a set of labeled clusters is available, standard CRF learning can be applied to estimate both the feature parameters $\mathbf{W}^F$ and the transition parameters $\mathbf{W}^T$.

4. EMPIRICAL PROTOCOLS

In this section we describe the details of our data set, training protocols, feature extraction pipelines, morphology extraction methods, and evaluation metrics.

4.1 Data Set

Wireless ECG data was collected from six habituated cocaine users in a NIDA-approved clinical study. The subjects wore a wireless single-channel ECG chest band . The wireless sensor on these chest bands samples ECG data at 250Hz and transmits the data to a smartphone via bluetooth. Data was collected from subjects in both in the presence and absence of cocaine use. We manually labeled over 20,000 candidate peak locations in nearly 1,500 clusters across the six subjects. The details of the data set are listed in Table 1. Importantly, the use of the candidate peak generation step reduces the number of locations considered by the CRF during inference by more than 27 times over all the subjects.

4.2 Train, Validation and Test Splits

We randomly partition the available data for each subject into a training set consisting of 10% of labeled clusters, a validation set consisting of 45% of labeled clusters and a test set consisting of 45% of labeled clusters, up to a total of 135 clusters, which is the minimum number across all subjects. These splits remain fixed for each subject throughout all experiments. The training sets are used to train the CRF model. The validation sets are used to select the CRF regularization parameter as well as to select between different feature sets. The test sets are used to evaluate model performance.

4.3 Learning and Evaluation Protocols

Our evaluation uses three different learning protocols: within-subjects, across-subjects, and transfer learning. In the within-subjects protocol, we use the training and validation set for each subject s to learn a subject-specific model and evaluate the model on the test data for subject s. In the across-subjects evaluation, for a given subject s, we pool the training set and the validation set for the subjects other than s and use this pooled data to learn a model. We evaluate this model on the data for subject s. In the transfer learning evaluation, for a given subject s, we begin by learning the across subjects model. We then use the learned weights from the across subjects model to define a data-dependent regularizer when learning the within-subjects model for subject s.

4.4 Feature Extraction and Normalization

We set the size of the sparse coding basis to $K = 100$ and the sparsity parameter to $\lambda = 0.01$. The basis vectors $\mathbf{B}_k$ were learned on ECG data extracted from a window of size 51 samples (204ms) centered at each candidate peak location. These values were found to yield good performance in preliminary testing. For within-subjects training, we learn a separate set of sparse coding basis vectors from all of the data windows available for each subject s. In across subjects training and transfer learning, we learn the sparse coding basis for subject s using all of the available data windows for each subject other than subject s. Thus, in across subjects training we are assessing both the generlizability of the sparse coding basis and the CRF model to a new subject. We also make the height and the height squared of each candidate peak location available as additional features. We consider three different feature sets when learning a model: sparse coding only (SC), sparse coding with peak height (SCH), and sparse coding with height and height squared ($SCHH^2$).

We also consider several different ways of normalizing the data within each window prior to extracting the features. We consider subtractive normalization (SN) where we shift the data to have zero mean within each window; subtractive and divisive normalization where we shift the data to have zero mean within each window and re-scale it to have unit standard deviation within each local window (SDN_L); and subtractive and divisive normalization where we shift the data to have zero mean within each window and and jointly re-scale all of the windows to have unit standard deviation globally (SDN_G).

In each of our experiments, we consider nine possible feature extraction pipelines given by the cross product of a choice of feature set from $\{SC, SCH, SCHH^2\}$ and a choice of data normalization framework from $\{SN, SDN_L, SDN_G\}$. For each model, we select one of the nine possible feature extraction pipelines using the validation set in each experiment.

4.5 Morphology Extraction Methods

In each of our experiments, we consider three different methods for extracting ECG peak locations and labels including our dynamic CRF approach, an independent multinomial logistic regression model (MLR) and the open-source ECGPUWave toolbox (PUW) [30, 23]. The only difference between our CRF framework and the MLR model is that the CRF model includes edges between adjacent candidate peak locations while the MLR model makes independent predic-

Subject	Session Length	# Samples	# Candidate Peaks	# Labeled Peaks	# Clusters
1	6h36m	5,624,954	217,941	3145	175
2	7h01m	5,649,203	214,563	4558	462
3	7h42m	6,537,902	301,317	3231	141
4	11h01m	9,492,152	333,165	4104	219
5	11h55m	6,736,003	245,995	2341	135
6	15h45m	13,565,502	450,256	3966	332
Total	60h	47,605,716	1,763,237	21,345	1464

Table 1: Data set details including the total data set sizes and the number of labeled peaks per subject.

tions. The MLR and CRF models otherwise have access to identical candidate peak locations, feature sets, and labels during training, validation and testing.

The ECGPUWave toolbox follows a traditional two-stage approach based on first identifying QRS complex locations and then performing a local search to identify the the peak locations within each cardiac cycle. The ECGPUWave toolbox can operate in conjunction with a number of different QRS complex detectors. The classical detector used with ECGPUWave is the Pan-Tompkins detector [31]. We found that the more recent open-source WQRS detector of Zong et al. performed significantly better on our data. The WQRS detector is based on the curve length transform and has been shown to be very robust, achieving a QRS sensitivity of 99.65% and a gross QRS positive predictive accuracy of 99.77% on the MIT-BIH Arrhythmia Database [36].

Since our data is labeled in terms of candidate peak locations and the CRF and MLR models are restricted to making predictions only at these locations, it is straightforward to assess their prediction performance. ECGPUWave can predict peaks at arbitrary locations so evaluating its accuracy requires some care. We apply a minimum weighted bipartite matching algorithm to the ground truth and ECGPUWave label locations to establish a correspondence between the true and predicted labels based on the distance between their time points [21]. We allow the ECGPUWave predictions to match ground truth labels within a window of plus or minus four samples (16ms). We define an ECGPUWave prediction as being correct if it is matched to a ground-truth label of the correct type. As a result of the matching window constraint, all correct peak labels must be within plus or minus four samples of a ground truth label of the correct type. Also due to the matching window constraint, some ECGPUWave predictions may not match any ground truth label locations. These predictions are considered as matching a ground truth label of N (not a valid peak location), which counts as a labeling error. We performed a preliminary analyses of the effect of window size on the number of matched ECGPUWave predictions and determined that the number of matches remains nearly constant as the window size is increased to nearly the average width of a full cycle. This indicates that the lack of a match for ECGPUWave typically means it did not identify a given wave type within a complex at all. Failure to identify a given ground truth wave is assessed as a prediction of N (not a valid peak) for that ground truth label. By contrast, the CRF and MLR methods are required to match the ground truth label locations exactly for their predictions to be considered correct.

4.6 Evaluation Metrics

We evaluate the three morphology extraction methods described above using several different metrics. All of the results that we report are averaged over the test set performance of our six subjects and the standard error of the mean is also reported. The first metric we employ is average labeling accuracy over all six label types (P,Q,R,S,T,N). We also report confusion matrices where we list the fraction of each ground truth label that is predicted to be of each label type. This allows for a detailed analysis of the types of prediction errors that each method tends to make.

We are also interested in assessing the impact of morphology extraction accuracy on the computation of ECG morphological feature values. We use the distance between the Q and T peaks as an example feature related to cocaine use. We assess the recall and precision of QT distances as well as the error in the distance for recalled QT pairs. The recall is the number of complexes where the ground truth contained a QT pair and both Q and T peaks were predicted to be present, divided by the number of complexes where the ground truth contained a QT pair. The precision is the number of complexes where the ground truth contained a QT pair and both Q and T peaks were predicted to be present, divided by the number of complexes that were predicted to contain a QT pair. The error in the QT distance is defined to be the absolute difference between the predicted QT distance (the distance between the predicted peaks) and the ground truth QT distance.[1]

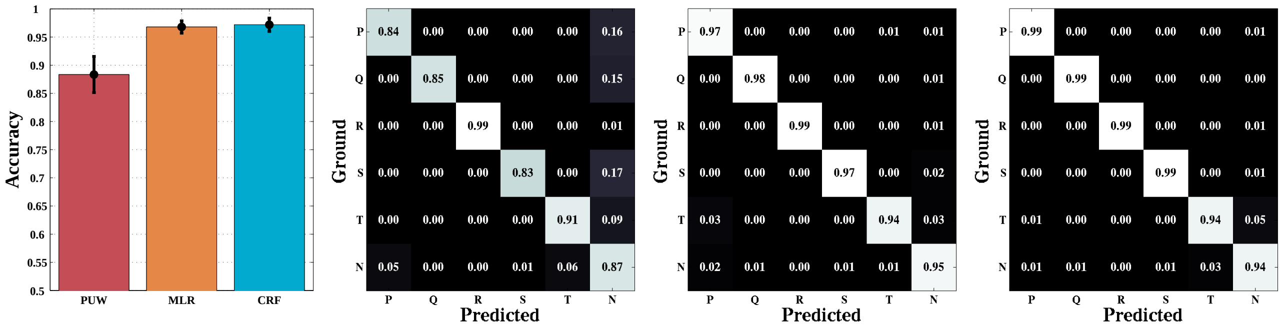
5. RESULTS

In this section we describe the results of our empirical evaluation including within-subjects evaluation, across subjects evaluation, transfer learning evaluation and QT feature extraction evaluation. Throughout this section, PUW refers to ECGPUWave using the WQRS detector, MLR refers to multinomial logistic regression, and CRF refers to our dynamic CRF framework.

5.1 Within-Subjects Evaluation

The results of the within-subjects evaluation as shown in Figure 4. Figure 4a shows the average prediction accuracy results for each of the three methods. We can see that the CRF and MLR methods both achieve the same average accuracy above 0.95, while PUW performs substantially worse with an average accuracy of about 0.87. The confusion matrices shown in Figures 4b-4d provide a more detailed look

[1] Note that the clinical definition of the QT interval is the difference between the onset of the Q wave and the end of the T wave. We use the QT peak-to-peak distance as a more convenient surrogate in this work. The recall and precision numbers would be identical for the standard QT interval as opposed to the peak-to-peak distance. The QT error of the proper interval would depend on the accuracy of an additional wave delineation step, but note that the same delineation method can be used with any set of peak labels.



(a) Labeling Accuracy (b) PUW (c) MLR (d) CRF

Figure 4: (a) Show the average labeling accuracy for within-subject training. (b)-(d) show the corresponding confusion matrices for PUW, MLR and CRF.

at the performance of the methods on a per-peak type basis. We can see that the prediction profiles for both the CRF and MLR models are nearly identical. We can also see that the distribution of errors for PUW is highly non-uniform. Consistent with past results for the WQRS detector, the PUW's identification of R peaks is highly accurate (99%). However, performance for all of the other peak types is much worse. In essentially all cases, this poor performance is caused by PUW failing to identify valid peaks, resulting in a prediction of N (not a valid peak).

The fact that MLR and CRF have similar performance in the within-subjects case indicates that the feature representation provided by sparse coding, as described in Section 2.4, is rich enough and the amount of data is large enough that there is no marginal benefit to structured prediction. However, the full within-subjects training protocol is based on hundreds of peak labels per subject. The need to label this much data for each individual subject is highly prohibitive. To assess the performance of the MLR and CRF methods given less data, we repeated the within-subjects evaluation while varying the number of labeled clusters available during training between 1 and 14 (each cluster contains 15 labeled peaks on average). The results of this assessment are given in Figure 5a. We can see that performance of MLR and CRF are strongly differentiated in the more realistic low data limit. With only one cluster of labels, the CRF still out-performs PUW on average, while MLR does not. We can also see that as more data becomes available, the CRF is able to improve it's performance significantly faster than MLR.

5.2 Across-Subjects Evaluation

A natural alternative to learning models for each individual subject is to learn models from an existing database of and apply that model to new subjects. The across-subjects evaluation assesses the performance of this approach when a model is learned using data from 5 subjects and then evaluated on the 6th held-out subject. We report results averaged over holding out each subject. Figure 6 gives the results of this assessment. We can see that both MLR and CRF suffer a decrease in performance relative to the full-data within subjects case. However, the CRF still out-performs PUW in the across subjects setting while MLR performs worse on average. The confusion matrices show that MLR confuses a variety of similar wave types in this setting (P for T, R for P and T, T for P). The CRF makes similar types of errors, but to a much lesser extent. This discrepancy can be explained by the fact that the CRF's transition parameters are able to exploit the regularity in the ordering of the waves within a complex to compensate for feature parameters tuned for other subjects. By contrast, MLR only has access to features values. When there is a poor match between the shapes of the waves across subjects, it's performance thus degrades much more quickly.

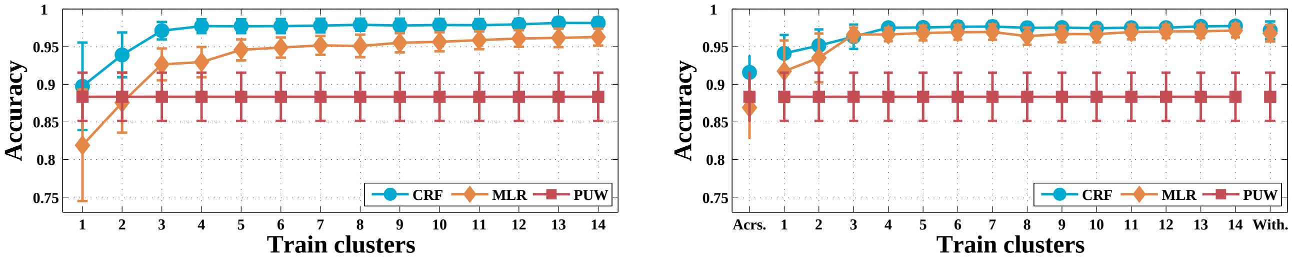
5.3 Transfer Learning Evaluation

The drop in performance of MLR and CRF in the across subjects setting motivates the evaluation of a third training protocol: transfer learning. Under the transfer learning approach we employ, as outlined in Section 4.3, data from other subjects is used to create a prior distribution over the model parameters. In the absence of any data for a given subject, the learned model falls back to the across-subjects model. As more data becomes available for an individual subject, transfer learning can smoothly interpolate between the across subjects model and the within-subjects model. Figure 5b shows the results of this analysis. We can see that transfer learning is able to dramatically improve the performance of both MLR and the CRF in the low data limit. With just one cluster of labels observed (approximately 16 labels), both MLR and CRF out-perform PUW and their corresponding across-subjects results.

5.4 QT Feature Extraction Evaluation

From the perspective of mHealth research, an important question is how differential accuracy in ECG peak labeling relates to the accuracy of ECG feature extraction. As a case study, we consider the problem of extracting QT distances from ECG data. The standard approach to this problem is to first identify the individual peak locations, and then compute QT distances using the identified waves. The potential problem with this approach is that failure to predict either the Q or T peak results in the absence of a QT feature. Complexes for which feature values could not be extracted are typically discarded from subsequent analysis. However, this can lead to a systematic bias in the subsequent analysis if there is a relationship between the true value of a feature and the ability of a feature extraction method to extract it reliably.

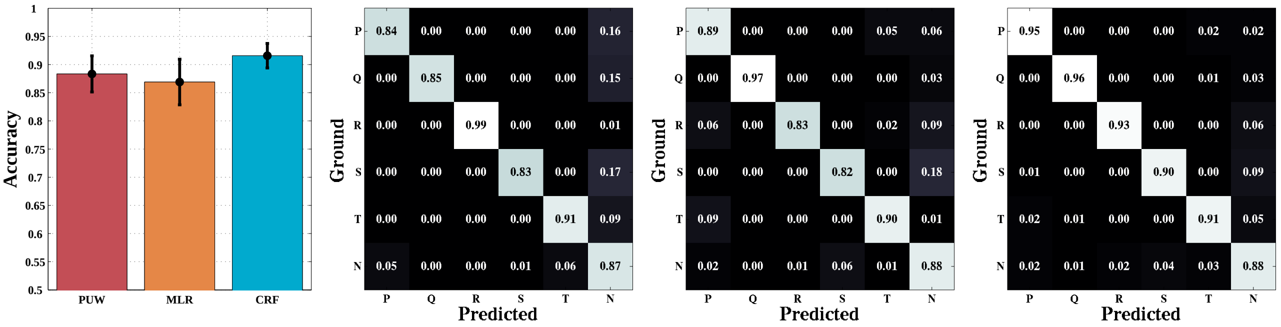
To assess the extent of this issue in our data, we used the MLR and CRF models trained using transfer learning with four clusters of labeled data to give a more realistic scenario for comparing subject-specific models to ECGPUWave. The results are summarized in Table 2. We can see that the



(a) Within-Subject Training: Labeling Accuracy

(b) Transfer Learning: Labeling Accuracy

Figure 5: (a) shows average labeling accuracy as a function of number of training label clusters for within-subjects training. (b) shows the same results using transfer learning-based training.



(a) Labeling Accuracy

(b) PUW

(c) MLR

(d) CRF

Figure 6: (a) Show the average labeling accuracy for across-subject training. (b)-(d) show the corresponding confusion matrices for PUW, MLR and CRF.

Model	Error	Recall	Precision
PUW	8.5914±12.8231	0.8733	0.9689
MLR	0.8469±13.5030	0.9549	0.9912
CRF	1.9085±17.4729	0.9854	0.9830

Table 2: QT interval evaluation for PUW, MLR and CRF.

lower accuracy of PUW results in significantly lower recall and precision of QT distances, as expected. We can also see that PUW has much higher mean error for the QT intervals that are retrieved than either MLR or PUW. Details of how we compute QT errors, precision and recall are explained in Section 4.6.

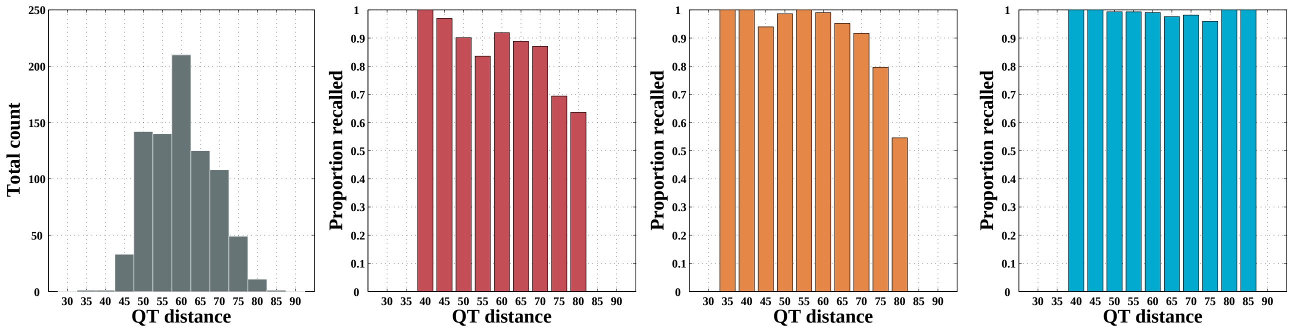
However, the interesting question is whether the recall rate for QT distances is uniform across all ground-truth QT distance values. Figure 7a shows the ground truth distribution of QT distances for our test data, pooled over all subjects. Figures 7b to 7d show the recall rate as a function of the ground truth QT distance (in bins of 5 samples). We can see that both PUW and MLR exhibit a strong differential recall rate as the ground truth QT distance increases. Only the CRF method achieves a nearly flat recall rate as a function of ground truth QT.

The final component of this case study looks at the distribution of QT values as a function of the study condition (cocaine vs no cocaine). Figure 8a presents the distribution of ground truth QT distances for both conditions pooled over all subjects. Figures 8b to 8d show the distribution of predicted QT distances for the complexes where both Q and T waves were identified. We can see that the CRF matches the ground truth distribution of QT distances quite closely for both the cocaine and no cocaine conditions as a result of its flat recall profile. On the other hand, PUW fails to identify any of the QT distances in bins $65, 80, 85$ under no cocaine and significantly skews the QT distribution in the presence of cocaine. MLR also misses a large number of cases in bins $75, 80, 85$ under no cocaine, but performs well in the cocaine setting.

6. CONCLUSIONS AND FUTURE WORK

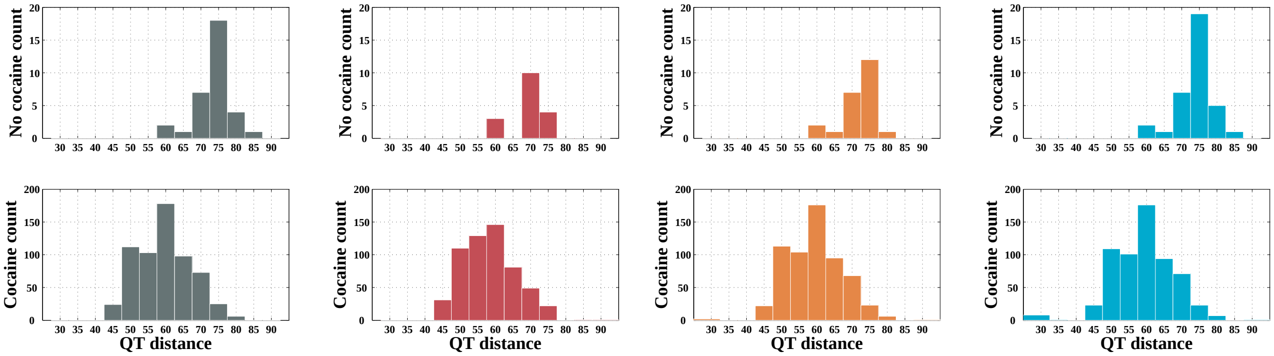
In this paper, we have presented a novel dynamic CRF framework for the extraction of morphological structure from ECG data that is designed to be robust to the type of noise and artifacts encountered using wireless sensors in the mHealth setting. We have evaluated our model relative to multinomial logistic regression and the open-source ECG-PUWave toolbox in three distinct learning settings: within subjects, across subjects and transfer learning. Our results show that the CRF out-performs ECGPUWave in all settings. Further, the CRF substantially out-performs multinomial logistic regression in the more realistic low data limit. Finally, our case study of ECG feature extraction in the cocaine use setting highlights the propagation and compounding of errors across the morphological structure identification and feature extraction pipelines, which can significantly alter feature distributions, misleading downstream analysis and prediction tasks. Given these results, we believe that in specialized mHealth studies where a large amount of data is collected from a relatively small number of subjects at a substantial cost (as in the cocaine study setting), the effort required to build highly accurate subject-specific models is justified. Our proposed combination of CRFs and transfer learning can substantially reduce this effort by minimizing the amount of labeled data required to learn robust subject-specific models.

There are many opportunities for future work in this area.



(a) GT (b) PUW (c) MLR (d) CRF

Figure 7: (a) shows the ground truth distribution of QT distances over all data. (b)-(d) show recall rates as a function of ground truth QT distance for each method. These results show that PUW exhibits a strong differential recall rate as a function of the ground truth QT interval, while the CRF does not.



(a) GT (b) PUW (c) MLR (d) CRF

Figure 8: Distribution of QT distances for cocaine vs no cocaine. (a) shows ground truth QT distance distribution. (b)-(d) shows distributions of predicted QT intervals for PUW, MLR, and CRF.

Both the candidate peak generation and feature extraction pipelines are completely modular. It may be possible to reduce the number of candidate peaks arising from noise and artifacts using more sophisticated approaches. Investigating the use of other feature sets in the CRF is also of interest. In this work, we have used unsupervised representation learning methods, but the investigation of fixed wavelet bases is also of interest. There are also a number of possible extensions to the CRF model itself including the incorporation of edge features.

7. ACKNOWLEDGMENTS

We would like to thank the staff of the Clinical Neuroscience Research Unit at the Connecticut Mental Health Center and the Hospital Research Unit at Yale New-Haven Hospital. This work was supported by the National Institute on Drug Abuse (K24 DA017899; R01 DA033733; P20 DA027844; RTM), the National Science Foundation (CNS-0910900; CNS-0855128), the Department of Mental Health and Addiction Services of the State of Connecticut, CTSA (UL1 RR024139) from the National Center for Research Resources and the National Center for Advancing Translational Science, components of the National Institutes of Health and the President's Science and Technology Fund, University of Massachusetts, Amherst. We would like to thank George Moody for his help with ECGPUWave toolbox and Gregory Shute for helping us label ECG peaks.

8. REFERENCES

[1] F. Alamudun, J. Choi, R. Gutierrez-Osuna, H. Khan, and B. Ahmed. Removal of subject-dependent and activity-dependent variation in physiological measures of stress. In *Pervasive Computing Technologies for Healthcare, Proceedings of the 6th International Conference on*, pages 115–122, 2012.

[2] E. Billauer. Peak detection. `http://billauer.co.il/peakdet.html`.

[3] R. Bouhenguel and I. Mahgoub. A risk and incidence based atrial fibrillation detection scheme for wearable healthcare computing devices. In *Pervasive Computing Technologies for Healthcare, Proceedings of the 6th International Conference on*, pages 97–104, 2012.

[4] R. Bouhenguel, I. Mahgoub, and M. llyas. An energy efficient model for monitoring and detecting atrial fibrillation in wearable computing. In *Body Area Networks, Proceedings of the 7th International Conference on*, pages 59–65, 2012.

[5] P. De Chazal, M. O'Dwyer, and R. B. Reilly. Automatic classification of heartbeats using ecg morphology and heartbeat interval features. *Biomedical Engineering, IEEE Transactions on*, 51(7):1196–1206, 2004.

[6] G. de Lannoy, A. De Decker, M. Verleysen, et al. A supervised learning approach based on the continuous wavelet transform for r spike detection in ecg. In

BIOSIGNALS (1), pages 140–145, 2008.

[7] G. de Lannoy, D. François, J. Delbeke, and M. Verleysen. Weighted conditional random fields for supervised interpatient heartbeat classification. *Biomedical Engineering, IEEE Transactions on*, 59(1):241–247, 2012.

[8] G. de Lannoy, B. Frénay, M. Verleysen, and J. Delbeke. Supervised ecg delineation using the wavelet transform and hidden markov models. In *4th European Conference of the International Federation for Medical and Biological Engineering*, pages 22–25, 2009.

[9] W. Einthoven. Ueber die form des menschlichen electrocardiogramms. *Archiv fÃijr die gesamte Physiologie des Menschen und der Tiere*, 60(3-4):101–123, 1895.

[10] E. Ertin, N. Stohs, S. Kumar, A. Raij, M. al'Absi, and S. Shah. Autosense: unobtrusively wearable sensor suite for inferring the onset, causality, and consequences of stress in the field. In *Proceedings of the 9th ACM Conference on Embedded Networked Sensor Systems*, pages 274–287, 2011.

[11] A. L. Goldberger. *Clinical electrocardiography: a simplified approach.* Elsevier Health Sciences, 2012.

[12] E. Haapalainen, S. Kim, J. F. Forlizzi, and A. K. Dey. Psycho-physiological measures for assessing cognitive load. In *Ubiquitous computing, Proceedings of the 12th ACM international conference on*, pages 301–310, 2010.

[13] M. C. Haigney, S. Alam, S. Tebo, G. Marhefka, A. Elkashef, R. Kahn, C. Chiang, F. Vocci, and L. Cantilena. Intravenous cocaine and qt variability. *Journal of cardiovascular electrophysiology*, 17(6):610–616, 2006.

[14] J.-H. Hong, J. Ramos, and A. K. Dey. Understanding physiological responses to stressors during physical activity. In *Ubiquitous Computing, Proceedings of the 2012 ACM Conference on*, pages 270–279, 2012.

[15] D. W. Hosmer Jr and S. Lemeshow. *Applied logistic regression.* John Wiley & Sons, 2004.

[16] S. M. Hossain, A. A. Ali, M. M. Rahman, E. Ertin, D. Epstein, A. Kennedy, K. Preston, A. Umbricht, Y. Chen, and S. Kumar. Identifying drug (cocaine) intake events from acute physiological response in the presence of free-living physical activity. In *Proceedings of the 13th international symposium on Information processing in sensor networks*, pages 71–82, 2014.

[17] S. Hu, Z. Shao, and J. Tan. A real-time cardiac arrhythmia classification system with wearable electrocardiogram. In *Body Sensor Networks, Proceedings of the 2011 International Conference on*, pages 119–124, 2011.

[18] N. Hughes and L. Tarassenko. Automated qt interval analysis with confidence measures. In *Computers in Cardiology, 2004*, pages 765–768, 2004.

[19] R. Jané, A. Blasi, J. García, and P. Laguna. Evaluation of an automatic threshold based detector of waveform limits in holter ecg with the qt database. In *Computers in Cardiology 1997*, pages 295–298, 1997.

[20] D. Koller and N. Friedman. *Probabilistic graphical models: principles and techniques.* MIT press, 2009.

[21] H. W. Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

[22] J. D. Lafferty, A. McCallum, and F. C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 282–289, 2001.

[23] P. Laguna, R. G. Mark, A. Goldberg, and G. B. Moody. A database for evaluation of algorithms for measurement of qt and other waveform intervals in the ecg. In *Computers in Cardiology 1997*, pages 673–676, 1997.

[24] K. Levin, M. Copersino, D. Epstein, S. Boyd, and D. Gorelick. Longitudinal ECG changes in cocaine users during extended abstinence. *Drug Alcohol Depend*, 95(1-2):160–163, 2008.

[25] A. Magnano, N. Talathoti, R. Hallur, D. Jurus, J. Dizon, S. Holleran, B. D. M., E. Collins, and H. Garan. Effect of acute cocaine administration on the QTc interval of habitual users. *The American journal of cardiology*, 97(8):1244–1246, 2006.

[26] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online learning for matrix factorization and sparse coding. *J. Mach. Learn. Res.*, 11:19–60, 2010.

[27] J. P. Martínez, R. Almeida, S. Olmos, A. P. Rocha, and P. Laguna. A wavelet-based ecg delineator: evaluation on standard databases. *Biomedical Engineering, IEEE Transactions on*, 51(4):570–581, 2004.

[28] A. Natarajan, A. Parate, E. Gaiser, G. Angarita, R. Malison, B. Marlin, and D. Ganesan. Detecting cocaine use with wearable electrocardiogram sensors. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, pages 123–132, 2013.

[29] B. A. Olshausen and D. J. Field. Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision research*, 37(23):3311–3325, 1997.

[30] L. Pablo, J. Raimon, B. Eudald, and V. A. David. Qrs detection and waveform boundary recognition using ecgpuwave. `http://physionet.org/physiotools/ecgpuwave/`.

[31] J. Pan and W. J. Tompkins. A real-time qrs detection algorithm. *Biomedical Engineering, IEEE Transactions on*, 32(3):230–236, 1985.

[32] B. G. Schwartz, S. Rezkalla, and R. A. Kloner. Cardiovascular effects of cocaine. *Circulation*, 122(24):2558–2569, 2010.

[33] F. W. Stallmann and H. V. Pipberger. Automatic recognition of electrocardiographic waves by digital computer. *Circulation research*, 9(6):1138–1143, 1961.

[34] W. Vongpatanasin, A. J. Taylor, and R. G. Victor. Effects of cocaine on heart rate variability in healthy subjects. *The American journal of cardiology*, 93(3):385–388, 2004.

[35] Zephyr. Bioharness 3. `http://www.zephyr-technology.com/products/bioharness-3/`.

[36] W. Zong, G. Moody, and D. Jiang. A robust open-source algorithm to detect onset and duration of qrs complexes. In *Computers in Cardiology, 2003*, pages 737–740, 2003.