Published in final edited form as:

J Biomed Inform. 2018 January ; 77: 97–110. doi:10.1016/j.jbi.2017.12.003.

Finding Representative Electrocardiogram Beat Morphologies with CUR

Emily P. Hendryx[a,1], **Béatrice M. Rivière**[a], **Danny C. Sorensen**[a], and **Craig G. Rusin**[b]

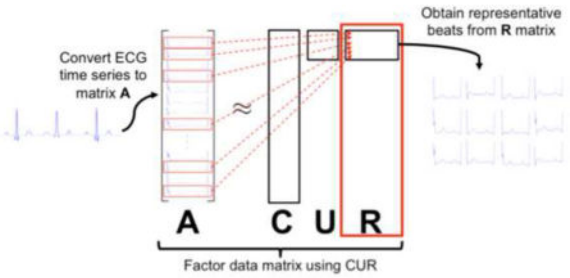
[a]Department of Computational and Applied Mathematics, Rice University, Houston, TX, United States

[b]Department of Pediatric Cardiology, Baylor College of Medicine, Houston, TX, United States

Abstract

In this paper, we use the CUR matrix factorization as a means of dimension reduction to identify important subsequences in electrocardiogram (ECG) time series. As opposed to other factorizations typically used in dimension reduction that characterize data in terms of abstract representatives (for example, an orthogonal basis), the CUR factorization describes the data in terms of actual instances within the original data set. Therefore, the CUR characterization can be directly related back to the clinical setting. We apply CUR to a synthetic ECG data set as well as to data from the MIT-BIH Arrhythmia, MGH-MF, and Incart databases using the discrete empirical interpolation method (DEIM) and an incremental QR factorization. In doing so, we demonstrate that CUR is able to identify beat morphologies that are representative of the data set, including rare-occurring beat events, providing a robust summarization of the ECG data. We also see that using CUR-selected beats to label the remaining unselected beats via 1-nearest neighbor classification produces results comparable to those presented in other works. While the electrocardiogram is of particular interest here, this work demonstrates the utility of CUR in detecting representative subsequences in quasiperiodic physiological time series.

Graphical abstract


Obtain representative beats from **R** matrix
Convert ECG time series to matrix **A**
≈
A
C
U
R
Factor data matrix using CUR

[1]Corresponding author at: 1600 Main Street - MS 134, Houston, TX 77005-1892. emily.hendryx@rice.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Keywords

Temporal data analysis; CUR matrix factorization; Dimension reduction; Electrocardiogram

1. Introduction

The identification of patterns in temporal biomedical data has been a popular topic for a number of years. Pattern identification plays a role in analyzing a wide variety of temporal data: from detecting patterns in hepatitis lab results regarding the effectiveness of treatment [1], to recognizing temporal patterns in electromyogram signals for controlling the movement of prostheses and orthoses [2], to identifying subsequences of event codes over time in electronic medical record data that may hold clinical relevance [3]. In this work we focus on the simultaneous recognition of both common and unusual subsequences in quasiperiodic physiological waveforms. While the methods presented here can be extended to other signals, we are particularly interested in the electrocardiogram, or ECG.

The ECG has provided clinical information about the heart since the beginning of the twentieth century [4]. Because the potential difference in each of the different ECG leads provides insight into cardiac function, physicians can use this data to monitor patient status over time and form diagnoses. However, performing a detailed analysis of all ECG signals continuously over longer periods of time for each patient is a nontrivial task. Therefore, there is a role for automated ECG analyses in supporting physicians in clinical decision-making. The goal of this work is to demonstrate an effective means of identifying a subset of heart beats that summarizes the different types of beats seen throughout a longer ECG recording.

In the clinical setting, the ECG is often described by looking at trends in smaller features contained within individual beats. Common features within each beat are the P, Q, R, S, and T waves; in the case that the left and right ventricles do not depolarize at the same time – as is observed in the presence of a bundle-branch block – an R' peak may also be present. Though other features may be present in the data, these six waves, depicted in Figure 1, are among those most commonly seen in the ECG and are the primary features of interest in this work. For a more extensive overview regarding the ECG, see, for example, Dale Dubin's *Rapid Interpretation of EKG's* [4].

Despite the fact that automated ECG analysis has been an area of interest for a while, a 2013 literature review by Velic, Padavic, and Car suggests that there is still much to be done in strengthening algorithms for clinical use [5]. This claim is supported, for example, by the need for monitoring systems that are designed for specific populations, such as pediatric patients with parallel circulation [6].

Our contribution in this work is to describe a framework for automatically identifying a representative subset of beats that appear in the ECG over extended periods of time. We provide a foundation for future algorithms by demonstrating the effectiveness of the CUR matrix factorization in identifying not only common, but also rare ECG beat morphologies within unlabeled data sets. These morphologies can then be used to train new classifiers for populations exhibiting a wide variety of beat shapes. This ability to identify a representative subset of beats is particularly relevant as it is now possible to store large amounts of physiological waveform data for retrospective studies, such as that carried out by Rusin et al. (2016) in the development of predictive models for specific patient classes [6]. For retrospective studies including millions, if not billions, of beats, this ability to reduce the data set to a representative subset can greatly decrease the computational cost of additional analyses and, if necessary, make the expert-labeling of different classes feasible. In addition to aiding in retrospective data analysis, the approach presented here can quickly provide physicians with a representative summary of beat morphologies seen over a given time frame to inform their clinical decision-making at the bedside.

As noted above, extensions of this work go beyond the ECG to other physiological waveforms that are periodic in nature to include data summarization as well as the development of template libraries to be used in waveform classification and predictive modeling incorporating other time series. In fact, though discussed only briefly herein (see the end of Section 4), the methods presented here are also applicable to other data types beyond time series. This paper is organized as follows. Section 2 provides further context regarding ECG data analysis and the CUR factorization. A more detailed description of the methods used in this work is provided in Section 3, followed by a discussion of our results in Section 4. In this latter section, we test the ability of our method to detect the presence of classes under a few different beat labeling systems, use selected beat morphologies to classify the remaining unlabeled beats, and evaluate the sensitivity of the algorithm's subset selection to automated beat delineation and noise. We conclude with a brief summary of our work and future directions of interest in Section 5.

2. Background and Related Work

2.1. Electrocardiogram Analysis

With the capability of collecting and/or storing ECG data digitally, the automated analysis of this data has been a topic of interest over the last several decades. In particular, there has been a growing interest in the classification of ECG beats for diagnostic and predictive purposes. Authors have classified ECG data using approaches involving different data representations such as wavelets [8], distance measures such as dynamic time warping [9], [10], [11], [12], and classifiers such as neural networks [13], [14], [15], among others.

To form a classifier in a supervised manner, one typically needs to know what classes to expect within the data. In the absence of a large annotated training set, this requires the identification of representative ECG signals within and across a variety of populations. Ceylan, Özbay, and Karlik (2009) use fuzzy clustering to identify class representatives for training a neural network classifier [15]. Similarly, Yeh, Wang, and Chiou (2010) use fuzzy clustering to identify representative ECG feature vectors for classifying unlabeled beats [16], and Annam, Mittapalli, and Bapi (2011) use dynamic time warping with k-medoid clustering to identify the classes of QRS complexes based on distances to cluster centers [17].

Cuesta-Frau, Pérez-Cortés, and Andreu-Garcí (2003) also use clustering to identify representative morphologies in ECG data; the authors use preclustering with dynamic time warping to reduce the data set, and then evaluate two different types of clustering (k-medians, or k-medoids, and Max-Min) with different types of data representation and temporal alignments. The authors point out the need to ensure that less common beats are detected separately and not clustered with more common beats in order for these less common beats to be of better diagnostic use [18].

Stemming from this work, Syed, Guttag, and Stultz (2007), also use dynamic time warping with Max-Min clustering to identify ECG beat classes and define a symbolic representation of longer time series [19]. While it is expected that clusters should contain beats from the same known ECG class, Syed et al. also allow for further divisions within a class and define their own class labels; where other works often confine the number of classes to those already known in the literature, this work seeks to refine class definitions to identify more subtle changes within the ECG [19].

Our work presented here has a similar goal in that we seek to implement an algorithm that will identify representative beat morphologies within large data sets with little to no prior knowledge about what classes should be expected in the data set. However, as opposed to clustering algorithms, or even motif and anomaly detection algorithms in the time series literature [20], we utilize the underlying structure of the data through the CUR matrix factorization to identify a broad representation of the beats present. Where other matrix factorization schemes (such as principal component analysis, or PCA) can also provide dimension reduction within large data sets, the CUR factorization provides a reduced data set consisting of original ECG beats. In this way, the reduced set maintains its clinical interpretability in contrast to the derived features from other dimension-reducing matrix factorizations.

2.2. CUR Factorization

The CUR factorization provides a means of identifying key rows and columns of the data matrix, **A**, approximating **A** as the product **CUR**, where **C** is a matrix consisting of a subset of k columns from **A** and the matrix **R** consists of a subset of k rows from **A**. Also sometimes called the matrix pseudoskeleton, the CUR factorization can be formed in a variety of ways. For instance, Goreinov, Tyrtyshnikov, and Zamarashkin (1997) propose defining the factorization by finding submatrices with maximal volume in the matrices of left and right singular vectors from the singular value decomposition [21]. Though not yet called the "CUR" factorization, the decomposition can also be derived using a quasi-Gram-Schmidt algorithm as was demonstrated in a 1999 work by Stewart [22]. A more commonly used method is to again consider the singular value decomposition, computing "leverage scores" to determine which rows/columns are used in forming **C** and **R** [23], [24]. For our implementation, we use the recently proposed discrete empirical interpolation method (DEIM) induced CUR with incremental QR because of its demonstrated improved performance over the use of leverage scores and its extendability to larger data sets [25]. As opposed to determining the set of reduced indices via leverage scores, which are computed using information from all singular vectors at once, DEIM-CUR determines the row and column indices of interest by considering each singular vector in turn, taking advantage of the fact that each individual singular vector holds different information about the space in which the data lives.

CUR acts simultaneously as a common motif detector and an anomaly detector, requiring previous knowledge about the data or further analysis in order to tell the two cases apart. While the detection of both standard and rare beat morphologies is an advantage of CUR, one limitation of this approach is that noisy perturbations of previously detected beats may be identified as independent events or anomalies. This is a consequence of using a framework that identifies the beats that are most different from those already considered. Making CUR more robust to noise is something we would like to consider in the future.

The CUR factorization has been applied to a number of topics, including traffic networks [26], music transcription [27], and toxicogenomics [28]. In addition, CUR has been applied to EEG (electroencephalogram) data for the purposes of compression. For instance, Dauwels, Srinivasan, Ramasubba, and Cichocki (2011) compare the approximation error of the CUR factorization via leverage scores to other EEG data compression schemes, concluding that CUR may not be optimal if the goal is solely to accurately approximate the original data set under a certain measure of error [29] (which is not the goal here). Lee and Choi (2008) demonstrate the use of CUR as a precursor to applying the nonnegative matrix factorization to EEG data [30]. To our knowledge, however, CUR–much less DEIM-CUR–has not been applied to the ECG or similar quasiperiodic physiological signals for simultaneous motif and anomaly identification.

3. Methods

To test our approach in identifying representative morphologies, we use several different data sets: a synthetic data set with different levels and types of added variability, the well-studied MIT-BIH Arrhythmia Database, the Massachusetts General Hospital-Marquette Foundation (MGH/MF) Waveform Database, and the St.Petersburg Institute of Cardiological Technics (Incart) 12-lead Arrhythmia Database. Our treatment of these data sets for subset selection, our choice in CUR implementation, and an overview of the additional tests discussed in Section 4 are described throughout the remainder of this section.

3.1. Synthetic Data Construction

Though there are synthetic waveform generators like ECGSYN [31] that produce more accurate synthetic ECG waveforms, for proof of concept and for more control over specific waveform characteristics, we take a simplified approach to synthetic signal generation. Where ECGSYN uses a dynamical model for signal formation that takes advantage of the Gaussian appearance of the individual beat features [31], here, each synthetic beat is constructed by simply summing the six different Gaussian curves in Equations (1) through (6) representing the P, Q, R, S, T, and R' waves in the ECG. Since the primary purpose of the synthetic data set is to test the sensitivity of our approach to very specific types of variability, the simple model described here allows for direct manipulation of each feature according to the variations of interest in a manner suitable to our purposes.

$$R\left(t\right)=R_{amp}\times exp\left(-\left(\frac{t-R_{peak}}{R_{width}}\right)^{2}\right), \quad (1)$$

$$Q\left(t\right)=-Q_{amp}\times exp\left(-\left(\frac{t-\left(\min\{R_{peak},R_{peak}+R'_{shift}\}-Q_{shift}\right)}{Q_{width}}\right)^{2}\right), \quad (2)$$

$$S\left(t\right)=-S_{amp}\times exp\left(-\left(\frac{t-\left(\max\{R_{peak},R_{peak}+R'_{shift}\}+S_{shift}\right)}{S_{width}}\right)^{2}\right), \quad (3)$$

$$P\left(t\right)=P_{amp}\times exp\left(-\left(\frac{t-\left(\min\{R_{peak},R_{peak}+R'_{shift}\}-P_{shift}\right)}{P_{width}}\right)^{2}\right), \quad (4)$$

$$T\left(t\right)=T_{amp}\times exp\left(-\left(\frac{t-\left(\max\{R_{peak},R_{peak}+R'_{shift}\}+T_{shift}\right)}{T_{width}}\right)^{2}\right), \quad (5)$$

and

$$R'\left(t\right)=R'_{amp}\times exp\left(-\left(\frac{t-\left(R_{peak}+R'_{shift}\right)}{R'_{width}}\right)^{2}\right). \quad (6)$$

The P, Q, S, T, and R′ waves are all defined relative to the location of the R peak as indicated by parameters with a subscript of “shift.” Note that the presence of the R′ peak can also affect the location of the other features in time. Parameters with a subscript of “amp” are indicative of the corresponding wave amplitude, and a subscript of “width” corresponds to a feature’s width, defining the standard deviation of the affiliated normal curve. The inclusion of these parameters not only allows for the construction of different beat morphology classes, but each class can also be constructed to have a certain level of within-class variability with respect to individual feature location, feature magnitude, feature width, as well as heart rate variability given by R-peak placement. The control cases for the twelve classes constructed for this work are shown in Figure 2. The parameters to construct these control morphologies using Equations 1 through 6 are presented in Table A.14 in the

Appendix.[2] The control heart rate for each class is 120 beats per minute to simulate an ECG that might be found in pediatrics.

Feature variability is included by allowing each of the above parameters to vary within a certain percentage of the given "control" parameter used in class definition. For instance, to add 10% variability from the T wave control width, T_{width_c}, a uniformly distributed random number on [−0.1, 0.1], r, is generated, and the perturbed width parameter is given as $(1-r)\,T_{width_c}$. Heart rate variability is added such that the heart rate is increased from the control rate within a desired range.

For each of the twelve constructed classes, heart rate, feature magnitude, and feature width variability are added separately for levels of 1%, 2%, 5%, 10%, 20%, 30% and 50%. These percentages are halved for adding variability in feature placement. Separate synthetic sets are constructed for each of the seven levels of variability for a given variability type. A total of 6,000 beats are included in each synthetic test set, with 500 beats coming from each of the twelve morphology classes.

To construct the data matrices for analysis, each synthetic time series is divided into individual beats through R-peak detection. A basic peak detector based on finding local maxima of the signal magnitude was used for automated beat delineation in this paper. Similar to the control beats shown in Figure 2, the beats in the matrix are defined from R peak to R peak since the R peak is a more clearly identifiable feature within the ECG (as opposed to trying to delineate beats by defining a cut-off point between T and P waves). If the R peak is less prevalent in some ECGs, then the larger downward peaks can be used to delineate beats; see for example classes 6, 7, 10, and 11 in Figure 2. Of note is that in the case of 50% amplitude variability, the coded R-peak finder for beat separation is unable to accurately detect all 6,000 beats due to the large amplitude fluctuations, and this case is removed from analysis. In the remaining sets, the individual beats are interpolated to have a length of 125 samples and then concatenated to form a matrix in which each column contains an individual beat.

In addition to variability in the typical ECG waves/peaks, we also test the performance of DIEM-CUR with incremental QR in the presence of random noise. While there are several sources of noise in ECG data, for initial experiments incorporating noise in the synthetic data, we add normally distributed random noise to the control data set. (More realistic types of noise are studied on the other data sets as described more below.) After forming the data matrix **A** to contain 500 identical control beats from each of the 12 classes, we add a randomly generated matrix, **E**, to **A** where the entries of **E** are normally distributed and $\|\mathbf{E}\|_2$ is a fraction of $\|\mathbf{A}\|_2$. As in the variability experiments constructed above, we generate noise matrices such that **E** has a norm that is 1%, 2%, 5%, 10%, 20%, 30% or 50% of the spectral norm of **A**. Examples of the Class 1 morphology under the varying levels of corruption are shown in Figure 3.

[2] The synthetic data sets, waveform generation codes, and other codes relating to the results presented throughout this paper can be found at https://github.com/ehendryx/deim-cur-ecg.

3.2. Real Patient Data

To test our methods on real data, we first use data downloaded from the MIT-BIH Arrhythmia Database [32] available on PhysioNet [33]. This data contains 48 files from 47 adult patients. These files each contain approximately 30 minutes of data recorded at 360 Hz for two ECG leads. For the purpose of this work, however, we analyze only the data from one of the leads. When provided, we used the MLII lead; otherwise, for the two files without MLII information, we use the V5 lead. Of note is that this use of different leads does not inhibit the use of CUR in our application; our proposed approach should identify representative ECG morphologies regardless of the lead from which individual beats come. The records are labeled with the following whole-beat annotations: normal beat (*N*), left bundle branch block beat (*L*), right bundle branch block beat (*R*), atrial premature beat (*A*), aberrated atrial premature beat (*a*), nodal (junctional) premature beat (*J*), supraventricular premature or ectopic beat (*S*), premature ventricular contraction (*V*), ventricular flutter wave (*!*)[3], fusion of ventricular and normal (*F*), atrial escape beat (*e*), nodal (junctional) escape beat (*j*), ventricular escape beat (*E*), paced beat (/), fusion of paced and normal (*f*), and unclassifiable beat (*Q*) [33].

With the MIT-BIH Arrhythmia Database used to identify an appropriate tolerance to be used in CUR, we test the performance of the algorithm with the selected parameter on the MGH-MF and Incart Databases, which are also both available on PhysioNet [33]. The MGH-MF Database [34] contains 250 records, though we disregard files mgh061, mgh127, mgh230, and mgh235 due to lack of annotations or unavailability of .mat files for ready analysis in MATLAB [35]. For comparison with results generated by Syed et al. [19], we focus primarily on the first 40 files. The data in this set also has sampling frequency of 360 Hz but has a wider range of record lengths with the typical record being about an hour long [33]. With some of the same PhysioNet labels that are present in the MIT-BIH Arrhythmia data set, the MGH-MF data set also contains labels for (atrial or nodal) supraventricular escape beats (*n*), R-on-T premature ventricular contractions (*r*), and beats that remain unclassified (?) [33]. Due to the larger size of this data, we used the WFDB Toolbox for Matlab and Octave [36] available on PhysioNet [33] to download the waveforms.

The Incart Database consists of 75 files each containing 30 minutes of 12-lead Holtermonitor data sampled at 257 Hz. In addition to some of the before mentioned PhysioNet labels, this data set also has beats with one additional annotation: (unspecified) bundle branch block beat (*B*) [33]. For both MGH-MF and Incart data sets, Lead II data was preferred for analysis via CUR when available.

Prior to constructing the corresponding data matrices for the different data sets, each lead is filtered using a zero-phase first order high pass Butterworth filter with a normalized cutoff frequency of $5 \times 10^{-3}\pi$ radians per second to reduce baseline wandering. To eliminate edge effects, 5% of the full signal is trimmed off of each end of the record prior to dividing the signal into individual beats from R peak to R peak using the annotation data provided in the corresponding database. A median of 2010, 5974.5, and 2088 beats per patient remain after

[3] This is referred to as a “non-beat annotation” on PhysioNet.

filtering the MIT-BIH, MGH-MF, and Incart data sets, respectively. Each beat is interpolated to contain 150 samples when constructing the data matrix.

As pointed out by Keogh and Kasetty [37] and Rakthanmanon et al. [38], the time series data should be normalized prior to further analysis. With the synthetic and real patient data, each beat is Z-normalized to have zero mean and a standard deviation of one before the CUR factorization is formed.

Despite the fact that we only study one lead from each of the records in these databases, we note that extension of our algorithm to multiple leads is possible. How the method is extended will depend on the application and purpose of the identified beat subset; in some cases it may be desirable to identify more beats from each lead, suggesting the need to form separate matrices for each lead. However, it is also possible to apply CUR to matrices consisting of data from multiple leads, tracking which matrix column corresponds to which lead. Another option is to construct **A** such that each column contains beats simultaneously recorded from multiple leads stacked on one another. For example, to identify representative R-R intervals based on both leads I and II, we can construct $\mathbf{A}_I \in \mathbb{R}^{m_I \times n}$ and $\mathbf{A}_{II} \in \mathbb{R}^{m_{II} \times n}$ to contain beats from leads I and II, respectively, and then define $\mathbf{A} \in \mathbb{R}^{(m_I + m_{II}) \times n}$ such that

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_I \\ \mathbf{A}_{II} \end{bmatrix}.$$

(Notice that each lead can have a different number of samples per beat if so desired.) With this construction, the column-selection algorithm will then have to consider both leads simultaneously to form **C** in the CUR factorization.

3.2.1. Classification and Sensitivity Tests—As discussed in more detail in Section 4, we also evaluate the performance of CUR in detecting morphologies under the AAMI and AAMI2 labeling schemes. This allows us to then compare our subset-selection method paired with the well-known one-nearest-neighbor classification algorithm to existing results in the literature.

After testing the performance of our approach on relatively clean data sets, we also test our method on a subset of the MIT-BIH Arrhythmia data set with added physiological noise. Where the synthetic data is evaluated under the addition of random noise, for a more realistic scenario, we add the noise signals available in the MIT-BIH Noise Stress Test Database (NSTDB) [39] available on PhysioNet [33]. This data set contains three separate records with baseline wandering (bw), muscle artifact (ma), and electrode motion artifact (em); each record contains two noisy signals which we add to the two leads in the MIT-BIH Arrhythmia database in the order presented. As described by Clifford, Behar, Li, and Rezek [40], we add the noise, $\mathscr{N}$, to a clean MIT-BIH signal $S_{\mathscr{C}}$ such that

$$S_{\mathscr{N}} = S_{\mathscr{C}} + \alpha \times \mathscr{N},$$

where $S_{\mathscr{N}}$ is the resultant noisy signal and α is given by

$$\alpha = \sqrt{exp\left(\frac{-ln\,(10)\,s}{10}\frac{P_{s_{\mathscr{C}}}}{P_{s_{\mathscr{N}}}}\right)}$$

for signal-to-noise ratio (SNR) s and respective signal powers given by P_x. For our experiments, we add noise with SNR equal to −6, 0, 6, 12, 18, and 24 decibels. Because the original noise waveforms in the NSTDB are typically shorter than those in the MIT-BIH Arrhythmia Database, $\mathscr{N}$ is formed by padding the end the original noisy signal with a partial mirror image of itself such that the noisy and clean signals are the same length.

In addition to the prior assumption of having relatively clean data, for almost all of the analyses, we assume that beat delineation in matrix construction is largely accurate for the sake of demonstrating the effectiveness of CUR. However, in practice, R-R intervals may have slightly different delineations depending on the peak-detection algorithm used and the quality of the data. Hence, in addition to including different types of realistic noise, we also compare CUR results generated using the beat delineations provided in PhysioNet with results generated using the simple automated peak detector originally implemented for use on the synthetic data. Under automatic delineation, each beat is labeled according to the nearest known annotation given in PhysioNet [33]. The automatic beat separation is carried out both on the "clean" and noisy versions of MIT-BIH Arrhythmia data.

3.3. CUR Factorization

To identify representative beat morphologies within the data, we use the CUR factorization of the data matrices. In the CUR factorization, the data matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ is approximated as

$$\mathbf{A} \approx \mathbf{C}UR,$$

where $\mathbf{C} = \mathbf{A}(:, \mathbf{q}) \in \mathbb{R}^{m \times k}$ and $\mathbf{R} = \mathbf{A}(\mathbf{p},:) \in \mathbb{R}^{k \times n}$ for column index vector **q** and row index vector **p** selected via a variety of methods. The matrix $\mathbf{U} \in \mathbb{R}^{k \times k}$ is constructed such that the approximation of **A** via this matrix product holds; note that the CUR factorization is non-unique and **U** can be constructed to meet a number of approximation requirements. For example, Sorensen and Embree demonstrate that setting $\mathbf{U} = \mathbf{A}(\mathbf{p}, \mathbf{q})^{-1}$ yields a CUR factorization that exactly matches the **p** rows and **q** columns of **A** [25]. Like these authors (and others–see [22] and [24]), we ultimately choose to construct **U** such that $\mathbf{U} = \mathbf{C}^{\dagger}\mathbf{A}\mathbf{R}^{\dagger}$, where $\mathbf{C}^{\dagger} = (\mathbf{C}^{T}\mathbf{C})^{-1}\mathbf{C}^{T}$ and $\mathbf{R}^{\dagger} = \mathbf{R}^{T}(\mathbf{R}\mathbf{R}^{T})^{-1}$ are the left and right inverses of **C** and **R**, respectively. This particular construction of **U** is chosen because, as demonstrated by Stewart [22], it minimizes $\|\mathbf{A} - \mathbf{CUR}\|_2$ for a given **p** and **q**.

As discussed in Section 2, there are also a number of ways to select the indices held in **p** and **q** in forming the CUR decomposition. To construct **p** and **q**, we use the DEIM and incremental QR implementation of the CUR factorization. A brief description of this approach is presented below. Because the DEIM approach used here relies on the singular value decomposition (SVD) of **A** and can be used independently to determine the CUR factorization, we first present the DEIM algorithm and then describe how incremental QR can be used to approximate the SVD for DEIM in larger data sets. For a more detailed description of DEIM-CUR with incremental QR, the interested reader is referred to the 2016 work of Sorensen and Embree [25].

3.3.1. DEIM—Originally described by Chaturantabut and Sorensen in 2010 [41], DEIM provides a deterministic means of identifying the most important rows and columns of $\mathbf{A} \in \mathbb{R}^{m \times n}$. First, the singular value decomposition of $\mathbf{A}$ is formed such that $\mathbf{A} = \mathbf{VSW}^T$ where $\mathbf{V} \in \mathbb{R}^{m \times k}$ and $\mathbf{W} \in \mathbb{R}^{n \times k}$ are unitary matrices and $\mathbf{S}$ is a diagonal matrix containing the k nonzero singular values of $\mathbf{A}$. DEIM is then applied to the left and right singular vectors of $\mathbf{A}$ contained in $\mathbf{V}$ and $\mathbf{W}$, respectively, to construct the index vectors $\mathbf{p}$ and $\mathbf{q}$.

In forming $\mathbf{p}$, let $\mathbf{v}_j$ denote the j^{th} column of $\mathbf{V}$ and $\mathbf{V}_j$ denote the matrix containing the first j columns of $\mathbf{V}$. Similarly, let $\mathbf{p}_j$ contain the first j elements of $\mathbf{p}$ and $\mathbf{P}_j = \mathbf{I}(:, \mathbf{p}_j)$, where $\mathbf{I}$ is the identity matrix in $\mathbb{R}^{m \times m}$. Then, with $\mathbf{p_1}$ defined such that $|\mathbf{v}_1(\mathbf{p_1})| = \max(|\mathbf{v}_1|)$, we take the j^{th} interpolatory projector $\mathscr{P}_j$ as

$$\mathscr{P}_j = \mathbf{V}_j \left(\mathbf{P}_j^T \mathbf{V}_j\right)^{-1} \mathbf{P}_j^T.$$

For

$$\mathbf{r} = \mathbf{v}_j - \mathscr{P}_{j-1} \mathbf{v}_j,$$

the j^{th} element of $\mathbf{p}$ is defined to be the positive integer, p_j, such that

$$|\mathbf{r}(p_j)| = \max\left(|\mathbf{r}|\right).$$

Notice that the interpolatory nature of $\mathscr{P}_j$ comes from the fact that for any vector $\mathbf{x} \in \mathbf{R}^m$,

$$\mathscr{P}_j \mathbf{x}(\mathbf{p_j}) = \mathbf{P}_j^T \mathscr{P}_j \mathbf{x} = \mathbf{x}(\mathbf{p_j}).$$

Adapted from [25], Algorithm 1 uses MATLAB notation to demonstrate how one can implement DEIM. This same approach is used to construct $\mathbf{q}$ using the columns of $\mathbf{W}$.

Algorithm 1

DEIM Point Selection (Adapted from [25])

```
Input: $\mathbf{V}$, a matrix in $\mathbb{R}^{m \times n}$ with $m \geq n$
Output: $\mathbf{p}$, a vector in $\mathbb{R}^n$ containing distinct integral values from {1, ..,$m$}
$\mathbf{v} = \mathbf{V}(:, 1)$
$[, p_1] = \max(|\mathbf{v}|)$
$\mathbf{p} = p_1$
```

```
for $j = 2 : n$ do
    $\mathbf{v} = \mathbf{V}(:, j)$
    $\mathbf{c} = \mathbf{V}(\mathbf{p}, 1{:}\ j-1)^{-1}\mathbf{v}(\mathbf{p})$
    $\mathbf{r} = \mathbf{v} - \mathbf{V}(:, 1{:}\ j-1)\mathbf{c}$
    $[, p_j]=\max(|\mathbf{r}|)$
    $\mathbf{p} = [\mathbf{p}; p_j]$
end for
```

3.3.2. Incremental QR

Because there are cases in which the entirety of the matrix **A** cannot fit into memory at the same time, Sorensen and Embree have introduced the incremental QR algorithm [25]. While memory is not an issue for the data studied here, we apply this approach in the next section to test its general effectiveness.

The incremental formation of the QR factorization can be implemented prior to computing the SVD of **A**. Typically, the QR factorization of **A** is formed via a method such as Gram-Schmidt orthogonalization so that $\mathbf{A} = \mathbf{QT}$, where $\mathbf{Q} \in \mathbb{R}^{m\times n}$ is a unitary matrix and $\mathbf{T} \in \mathbb{R}^{n\times n}$ is an upper triangular matrix. (Here **T** is used instead of **R** to avoid confusion with the use of **R** in the CUR factorization.) In the incremental case, however, the factorization is approximated by eliminating any rows of **T** and corresponding columns of **Q** that do not contribute significantly to the cumulatively computed Frobenius norm of **T** after each orthogonalization step. In this way, if a column of **A** does not contribute much to the factorization formed from the columns considered up to that point, that column and its corresponding row of **T** and column of **Q** do not need to be stored in memory; this results in the formation of a rank-k approximation to **A** with an $m \times k$ orthogonal matrix **Q** and a $k \times n$ matrix **T** for $k \leq \min(m, n)$.

Depending on the size of k relative to m and n, the SVD of $\mathbf{T}=\hat{\mathbf{V}}\mathbf{S}\mathbf{W}^T$ may be more easily computed to approximate the SVD of **A** such that

$$\mathbf{A} \approx \mathbf{QT}=\mathbf{Q}\hat{\mathbf{V}}\mathbf{S}\mathbf{W}^T=\mathbf{V}\mathbf{S}\mathbf{W}^T,$$

where $\mathbf{V}=\mathbf{Q}\hat{\mathbf{V}}$. With that, DEIM is then used to select the rows and columns of **A** for the CUR factorization. (Note that when k is close in size to $\min(m, n)$, as is sometimes the case in this paper, forming the SVD of **T** to approximate **A** may not actually be much cheaper than computing the full SVD of **A** itself.)

A general outline of our implementation of the incremental QR algorithm is shown in Algorithm 2 for brevity. Note, however, that in line 6 of the algorithm below, we follow the suggestion of Sorensen and Embree and include the reorthogonalization step proposed by Daniel, Gragg, Kaufman, and Stewart [42] as presented in [25]. For more details on the implementation of incremental QR, see [25].

Algorithm 2

Incremental QR (Adapted from [25])

```
1:  $k = 1$
2:  $\mathbf{Q} = \mathbf{A}(:,1)/\|\mathbf{A}(:,1)\|$; $\mathbf{T} = \|\mathbf{A}(:,1)\|$; rownorms(1) = $\|\mathbf{T}(1,1)\|^2$
3:  $j = k + 1$
4:  while $j \le n$ do
5:    $\mathbf{a} = \mathbf{A}(:,j)$; $\mathbf{r} = \mathbf{Q}^T\mathbf{a}$; $\mathbf{f} = \mathbf{a} - \mathbf{Qr}$
6:    $\mathbf{c} = \mathbf{Q}^T\mathbf{f}$; $\mathbf{f} = \mathbf{f} - \mathbf{Qc}$; $\mathbf{r} = \mathbf{r} + \mathbf{c}$ (reorthogonalization step)
7:    $\rho = \|f\|$; $\mathbf{q} = \mathbf{f}/\rho$
8:    $\mathbf{Q} = [\mathbf{Q}, \mathbf{q}]$; $\mathbf{T} = \begin{bmatrix} \mathbf{T} & \mathbf{r} \\ 0 & \rho \end{bmatrix}$
9:    rownorms($i$) = rownorms($i$) + $\mathbf{r}(i)^2$ for $i = 1, \ldots, k$
10:   rownorms($k + 1$) = $\rho^2$
11:   FnormT = sum(rownorms)
12:   $[\sigma, i_{\min}]$ = min(rownorms($1 : k+1$))
13:   if $\sigma <= (tol)^2$(FnormT – rownorms($i_{\min}$)) then
14:     Eliminate least-contributing row of $\mathbf{T}$ and corresponding column of $\mathbf{Q}$ (corresponding to $i_{\min}$).
15:     Update rownorms appropriately.
16:   else
17:     $k = k + 1$
18:   end if
19:   $j = j + 1$
20: end while
```

Once the vectors **p** and **q** have been determined through DEIM, **R** is simply defined to be those rows of **A** given by the indices in **p**, and **C** is defined to be the columns of **A** given by the indices in **q**. In our application then, **R** contains key sampling points across all interpolated beats and **C** contains a subset of beats, with each column of **C** containing a single beat. While the rows of **R** may prove useful in the future, here we focus on the beat subset selection given through the columns of **C**. The performance of this approach in the ECG setting is described in the next section, with some results compared to those presented by others in both unsupervised and supervised (or semisupervised) settings.

One positive aspect of the presented method is that the only parameter to be selected in the formation the CUR factorization is the tolerance used in determining which rows of **T** and columns of **Q** are kept within the incremental QR algorithm (line 13 of Algorithm 2). This threshold then affects the rank of the approximate QR factorization of **A**, which in turn determines the rank of the approximate SVD of **A**. As implemented here, this rank-*k* SVD then determines the number of indices selected by DEIM, that is, the number of rows and columns selected from **A**. Hence, with the incremental QR threshold the only parameter in this approach, we limit the amount of user input and prior knowledge needed for selecting representative waveforms.

In terms of computational cost for row and column index selection, the cost is dominated by the formation of the approximate rank-k SVD of **A** [25]. In approximating the SVD of **A**, incremental QR has complexity $O(mnk)$ and forming the SVD of the $k \times n$ matrix **T** has complexity $O(nk^2)$. As noted by Sorensen and Embree, once the approximate SVD is computed, the selection of indices for both **p** and **q** is then $O(mk + nk)$ [25].

The computational cost for the Max-Min clustering algorithm with dynamic time warping used by Syed et al. is reported to be $O(mnc) + O(m^2ck)$ with a worst-case scenario-cost of $O(m^2nk)$, where c is the number of clusters found in a pre-clustering step [19]. We see, then, that the $O(mnk)$ cost of index selection in the algorithm presented here is perhaps comparable with the cost of the approach presented in [19] for cases in which $m << n$ and c is relatively small. In general, however, our approach is the less expensive approach.

4. Results and Discussion

To test our approach, the DEIM-CUR factorization with incremental QR is applied to both the synthetic and the real patient data. Since the selection of the representative beats is of primary interest in this paper, only those CUR results contained in the matrix **C** are discussed here. The factorization is considered to be effective in the unsupervised sense if at least one representative beat from each class is contained within the columns of **C**.

While we are able to compare CUR beat selection to another existing work ([19]), we have had difficulty finding other results in the literature based solely on the performance of similar unsupervised, class-detection algorithms applied to the patient data studied here. To further demonstrate the value of data summarization via CUR, then, we also compare the results of a 1-nearest neighbor (1NN) classifier based on the labels of the CUR-selected beats with other state-of-the-art classification methods. The MIT-BIH Arrhythmia and Incart data sets are used in generating these results.

4.1. CUR with Synthetic Data

In each of the synthetic experiments, **A** is a 125 by 6,000 data matrix containing 12 classes of beats with a given type and level of variability. The results for each of the four variability types-heart rate, feature placement in time, feature magnitude, and feature width–as well as added random noise are shown in Table 1 for the seven variability levels tested. Recall that placement variability is actually added at half of the percent change shown in the first column. The stopping tolerance within the QR factorization is taken to be 1×10^{-5}. For each experiment, the percent dimension reduction and the total number of missed classes are presented.

These results in Table 1 demonstrate that, for this particular stopping tolerance, we are able to achieve greater than 97% dimension reduction with CUR for all of the tested variation types and levels. There are a few cases with higher feature variability levels in which classes are missed, but in most cases, each of the twelve classes are represented in the reduced data set. Hence, we see that we are able to maintain synthetic class detection while greatly reducing the size of the data set under consideration in the cases of heart rate and feature variability.

The added noise results are somewhat different in that the amount of dimension reduction is constant for all tested noise levels and there are fewer missed classes in the presence of more noise. The underlying reason for these results is not exactly clear. However, the constant level of dimension reduction is likely due to the fact that the addition of normally-distributed random noise results in a full rank (or nearly full rank) matrix with singular values that exhibit little decay. Also, the detection of more classes in the cases with higher noise levels may be related to the fact that, as seen in Figure 3, with enough noise, the original beat morphologies are lost; since the classes are evenly distributed, it is perhaps not surprising that CUR would select one beat from each class under such corruption of the data. While the noise present in real ECG data is not expected to be random, we gain a sense of the behavior of the CUR algorithm in the presence of such noise in varying amounts. There are, however, a variety of other noise sources in real patient data, and the effects of the different types of noise on CUR class identification are studied further on the real patient data below.

The results for the magnitude case, on the other hand, seem particularly good, with no classes missed and greater than 99% dimension reduction for all levels of added variability. On top of the effects of normalization, this is likely due to the fact that, by constructing the amplitude-varying beats as a sum of scaled Gaussians, all discretized beats in a class will lie in the same vector subspace spanned by the set of discretized Gaussians (which have fixed width and placement in time). Because each class is constructed to lie in a low-dimensional space, it is no surprise that we see greater dimension reduction for this variability type.

As noted in Section 3, we have traded realism in our synthetic data for more direct control over parameters in order to understand the sensitivity of our chosen dimension reduction technique to very specific perturbations in the data. However, to better understand the performance of DEIM CUR with incremental QR in practice, it is critical that the method be applied to more realistic data such as data simulated using ECGSYN, which still affords some control over generated features [31], or data recorded from real patients. Hence, we turn now to the application of our method to real patient data.

4.2. CUR with Patient Data

4.2.1. Unsupervised Subset Selection—To evaluate the performance of our proposed approach on patient data, DEIM-CUR with incremental QR is first applied to the data matrices constructed from each patient file in the MIT-BIH Arrhythmia Database. We again consider our implementation to be successful if the resulting subset selection from CUR contains representatives from each annotation class. Table 2 provides a summary of these results for each of the 16 annotations delineated in Section 3. For the eight different tested incremental QR stopping tolerances, each column of the table presents the percentage of patients in which an annotation was detected among all patients with that particular beat type.

Table 3 compares our results for annotations *N*, *A*, *V*, /, *f*, *F*, *j*, *L*, *R*, and *a* with stopping tolerance 5×10^{-5} to those presented in the 2007 work by Syed, Guttag, and Stultz [19]. (Note that / is referred to as *P* in the paper of interest.) Syed et al. also report that, though annotations *E* and *e* are reported in only one patient each, their algorithm was indeed able to detect these classes; this is also reflected in Table 3. As is done in [19], the CUR results presented in Table 3 consider only those patients with a given annotation present in three or more beats. We see that our results for annotations *N*, *A*, *V*, /, *f*, *F*, *L*, *R*, *a*, *E*, and *e* are all the same as those presented by Syed et al., with the exception being that our percent detection for annotation *j* (junctional escape beats) is 66.67% where the percent detection by Syed et al. is 100% [19]. In looking closer at our results for annotation *j* (detailed more below), we see that only three patient files contained greater than two beats with said annotation, and CUR failed to detect a *j* beat in one of those files.

Again considering even those annotations with fewer than three beats present in a file, Table B.15 in Appendix B provides a closer look at some of the results from Table 2. This more detailed table shows the results of analyzing each individual patient file via CUR with a stopping tolerance of 5×10^{-5} within incremental QR. Each row corresponds to a different file number and each column corresponds to one of the 16 annotations included in the data set. In each table entry, the number of CUR-selected beats for a given annotation is presented with respect to the total number of beats with that annotation in the filtered data from that file. An entry containing “ – ” indicates that there were no beats with that particular annotation for that patient file. Notice that some annotations (including *E* and *e*) are only present in one file. The bottom row of the table contains the CUR annotation representation with respect to the total number of files containing that annotation. (These fractions in the bottom row are used to generate the entries in Table 2 for the given stopping tolerance.)

In looking at the entries in Table B.15, we see a similar level of dimension reduction as that seen in the synthetic data. Because our implementation of the DEIM-CUR factorization selects the same number of rows and columns from the data matrix, **A**, the dimensionality of the reduced set is limited by the minimum dimension of **A**; hence, it is not surprising that the original beat set is reduced to no more than 150 beats in each experiment. With this reduction, however, we see that the different beat types are still typically retained in the selected subset. In particular, there are several cases in which annotations are detected even though there are only a few beats in the file with that annotation. For example, though there are few beats with annotation *Q*, CUR still selects at least one *Q* beat per patient file. This indicates that the algorithm is indeed able to identify both common and rare morphologies in selecting representatives from within the data set.

Based on the tolerance experiments on the MIT-BIH Arrhythmia results, we then apply CUR with an incremental QR tolerance of 5×10^{-5} to the other two real-patient data sets.

For the MGH-MF set, we first present the results on the first 40 files of the data set in Table 4. If we look at all 40 of these files and include results for annotations in files with only one or two beats present, we get the results presented on the first row of Table 4. For the interest of comparison with previously presented results by Syed et al. [19], we also exclude cases with fewer than three representative beats and exclude files mgh002 and mgh026. Note that Syed et al. state their use of three beats as the representation cut-off threshold because it is 1% of the length the record times in the MIT-BIH database; it is unclear what the cut-off threshold was in the presentation of the MGH-MF database results, especially given the larger variability in file size for this database. Also note that Syed et al. do not report results for supraventricular premature or ectopic beat (S) detection on the MGH-MF database, so this table entry is left blank.

With a representation cut-off of three or more beats, we see that we are again able to obtain similar results as those reported in a previous work using a clustering approach, the difference being that CUR misses premature ventricular contractions in two files: mgh025 and mgh028.

For completeness, we also test our approach on the full MGH-MF database (excluding files mgh061, mgh127, mgh230, and mgh235). The results on this full data set are shown in Table 5. Even with the addition of over 200 more files, we see results that are similar to those generated on only a fraction of the data set for the annotations present in the first 40 files.

Though we have not found in the literature any comparable results with this type of unsupervised class detection (relying on no a priori knowledge) for the Incart database, we present the Incart CUR performance results in Table 6. We again show the patient annotation representation for all files as well as for those cases in which only three or more beats are present with a given annotation. Note that annotations n and j are only present in two files, and in both cases, these annotations were detected in only one of the two files; similarly, S is only present for ≥ 3 beats in three records and was detected in only one such record, resulting in the poorer class detection. In general, the class-detection results for the Incart data set are comparable to the class-detection results generated on the MIT-BIH database.

While it is unclear why some annotations are missed in CUR dimension reduction, one possibility is that timing information used in annotating is lost when converting the data to matrix form. Relative feature timing should be maintained within individual beats, but if overall beat length is a major factor in defining a beat class, then it seems natural that CUR should miss this class.

4.2.2. AAMI and AAMI2 Class Representation—In testing on both synthetic and real data, we have also tested CUR on data sets that are different in regard to the distribution of classes within the data set. Where the synthetic data is designed such that each class is equally represented, the patient databases clearly do not have this structure among classes. Though we see that CUR is able to detect some of the more poorly represented classes in the patient data (a desirable quality for some tasks), such granularity may not always be necessary. In some works, such as those by de Chazal, O’Dwyer, and Reilly [43] and Llamedo and Martínez [44], the MIT-BIH class labels presented on PhysioNet [33] are summarized by a smaller number of classes as recommended by the Association for the Advancement of Medical Instrumentation (AAMI) [45]. With this labeling, beats with annotation N, L, R, e, and j are all considered members of the “normal” class $\left(\hat{N}\right)$, beats with annotations A, a, J, and S are considered members of the “supraventricular ectopic beat” class $\left(\hat{S}\right)$, beats with annotations V and E are placed in the “ventricular ectopic beat” class $\left(\hat{V}\right)$, the fusion of ventricular and normal beats (F) still retain their own “fusion beat” class, $\left(\hat{F}\right)$, and beats with annotations /, *f*, and Q are all included in the "Unknown beat" class $\left(\hat{Q}\right)$ [43].

Ignoring MIT-BIH Arrhythmia Database patient files with paced beats in accordance with the AAMI recommendations [45], Llamedo and Martínez discard the $\hat{Q}$ class due to poor representation. Similarly, because the $\hat{F}$ class has a somewhat low representation relative to the other data sets and consists only of beats that are a fusion of ventricular and normal beats, the authors combine $\hat{F}$ and $\hat{V}$ into one class, which we will call $\hat{V}'$. Llamedo and Martínez refer to this even further summarization of the MIT-BIH Arrhythmia annotations as the AAMI2 labeling [44].

Though neither the AAMI and AAMI2 labeling systems result in a truly even distribution of classes, the presence of "rare" classes is greatly decreased for the MIT-BIH Arrhythmia data set. The class detection results for the AAMI and AAMI2 labeling of this database are shown in Table 7 for an incremental QR tolerance of 5×10^{-5}. Note that for these labeling systems, "≥ 3 beats" corresponds to at least three beats present with a given annotation under the original PhysioNet labeling system; these cases were removed from consideration prior to translating the PhysioNet [33] labels to AAMI/AAMI2 classes. As in [43] and [44], these results are generated without the inclusion of ventricular flutter waves in file 207m and without records containing paced beats (though the inclusion of these records has no effect on the percentages shown). We also leave out the representation results for the unknown beats $\hat{Q}$ in the AAMI2 labeling results as in [44], though the AAMI results show that unclassified beats (Q) are still detected despite their small class representation. It should be noted that in all of the reported AAMI2 results, beats belonging to the $\hat{Q}$ class remain in the data set and can be detected as representative beats by CUR. Because all records with fusion of ventricular and normal beats also have CUR-detected premature ventricular contractions (as can be noted in Table B.15), the AAMI2 class of $\hat{V}'$ sees 100% detection despite the fact that the fusion beats are not detected in some cases. From these results we see that CUR performs well in detecting the more broadly defined and generally well-represented classes given by the AAMI and AAMI2 labeling systems.

Overall, our implementation of DEIM-CUR with incremental QR seems to successfully summarize the types of morphologies present in the ECG time series data. The fact that we are able to achieve similar results to those presented by Syed et al. [19] using a matrix factorization technique for dimension reduction as opposed to clustering on the entire data set is encouraging. Not only do we see that applying CUR reduces the problem size, but in doing so, rare beat events are preserved alongside more commonly seen patterns. In this way, CUR provides a broad representation of the ECG data, a desirable result for understanding the types of morphologies that occur within the data over longer periods of time. The value of this broad representation is demonstrated through the use of the CUR-selected subset in classifying unlabeled data, as discussed further below.

4.2.3. Using CUR-Selected Beats in Classification—

For comparison with state-of-the-art methods in the literature, we perform 1NN classification with the Euclidean distance using the annotations from the CUR-selected beats to label the remaining beats in each file of the MIT-BIH Arrhythmia and Incart data sets. We compare this semi-supervised classification approach with the results produced by methods presented in four other works. de Chazal, O'Dwyer, and Reilly (2004) present a method that uses linear discriminants to classify beats based on features derived from the data [43], and Llamedo and Martínez (2011) present a compensated linear discriminant classifier (with unequal weights), also based on features derived from the ECG [44]. In a separate work, Llamedo and Martínez (2012) also present a method that pairs a linear discriminant classifier with an expectation-maximization clustering approach that can be implemented in one of three modalities–automatic, slightly assisted, or assisted–depending on the amount of expert input in assigning labels to clusters [46]. In the most recent work considered here, Oster, Behar, Sayadi, Nemati, Johnson, and Clifford (2015) use a switching Kalman filter approach to classify a beat as either "normal," "ventricular," or possibly a member of the "X-factor" class (consisting of beats that look different from the normal and ventricular beats, potentially due to noise); when the X-factor is included as a class option, beats identified as belonging to the X-factor class are removed from further analysis [47].

As is done in these papers, we classify unlabeled beats according to the AAMI and AAMI2 classes. We also remove from analysis ventricular flutter waves (beats associated with the '!' annotation), and we again exclude records with paced beats. For comparison purposes, we first select the incremental QR threshold on a subset of the MIT-BIH Arrhythmia Database (referred to as DS1) and evaluate the performance of the method with the chosen threshold on a separate subset of the MIT-BIH Arrhythmia (DS2) data and Incart data sets. The partition used for splitting the MIT-BIH data into subsets DS1 and DS2 is given by de Chazal et al. [43].

Like Oster et al., we focus our attention on the classification of beats in the AAMI and AAMI2 ventricular class and perform parameter selection based on the results for this class in DS1 [47]. In selecting the incremental QR threshold, we look at the sensitivity (Se) and the positive predictive value ($+P$) as described in [44], and use the harmonic mean of these two values, the F_1 score [47], to select the appropriate threshold. As shown in Table 8, the highest F_1 values are achieved for a tolerance of 5×10^{-3} for both the $\hat{V}'$ and $\hat{V}$ class labelings. We then use this threshold to classify the beats in the DS2 and Incart data sets. The result comparison for the DS2 and Incart sets as reported by Oster et al. [47] is shown in Tables 9 and 10, respectively, with the addition of the results from the CUR-1NN classification method presented in this work. The corresponding confusion matrices for the AAMI classification are shown in Tables 11 and 12; the AAMI2 confusion matrices can be derived from these tables given the direct relationship of the two labeling systems.

From these results, we see that $+P$ tends to be higher than Se for our simple classification method. Despite this, we see that CUR-1NN classification has the highest F_1 score among those presented for $\hat{V}'$ classification in the DS2 data set, and comes in second only to the classifier with X-factor by Oster et al. for DS2 $\hat{V}$ classification. While the CUR-1NN F_1 scores are less impressive compared to some of the other reported results on the Incart data due to the method's lower Se values, we see that the $\hat{V}'$ results using our method are still comparable to some of the results from the previously presented state-of-the-art methods, even with using 1NN and the Euclidean distance.

Of note is that in our preprocessing steps, several beats are lost prior to constructing the data matrix; future analysis will be needed to determine the impact of the preprocessing on the CUR-1NN results shown here. Also, where methods presented in the comparison papers can make use of multiple leads, our method currently makes use of only one lead at a time. While this can be beneficial depending on the amount of ECG data available for a given analysis, extensions of this approach to handle multiple leads may also be of interest in the future. Additionally, we have selected the incremental QR threshold solely based on the $\hat{V}'$ and $\hat{V}$ classes; other means of parameter selection should be considered in the future, taking into account performance on multiple classes at a time.

Given the approach presented in this work, however, we are generally able to achieve comparable performance to existing classification algorithms using an unsupervised learning method (with only one parameter) followed by the 1NN classifier with Euclidean distance. Our results could possibly be improved with the use of another similarity measure and/or classifier, but we have chosen a simple classification method to highlight the utility of CUR in the unsupervised step. From the results presented in this section, we see that CUR is indeed a viable unsupervised means of representative subset selection for use in annotating the larger data set when analyzing clean data with accurate peak delineation. In the following subsection, we test ability of CUR to detect class representatives under noisier conditions.

4.2.4. Sensitivity Analysis—Having demonstrated the utility of CUR subset selection on relatively clean data with known beat delineations, we now turn to testing the sensitivity of our approach to different beat definitions and added noise. We perform these tests on the DS2 subset of the MIT-BIH Arrhythmia Database under the AAMI labeling scheme. Results are first generated using both the PhysioNet and automatically generated beat delineations on the “clean” DS2 data. Both beat detectors are then applied to data with one of the three types of noise (em, ma, or bw) added at varying levels.

Table 13 shows results from DS2 class detection with an incremental QR threshold of 5×10^{-3} for the two different beat delineation approaches, both including and excluding cases in which the original PhysioNet labels were present in fewer than 3 files. The median beat count per file is 1978 for the original known delineation and 1916 for the automatic delineation. With the exception of perhaps annotation $\hat{F}$, the results are within 10% of one another for the two methods. It is worth noting, however, that the basic automated peak detector is applied to each file without careful tuning and labels are assigned based solely on proximity to the known annotations, regardless of whether or not a given “beat” is actually a true R-R interval. Hence, while this experiment appears to generally support the robustness of CUR under beat delineation, it may be of value to perform a more thorough study of the effects of beat separation on CUR performance in the future.

The class detection results given the above mentioned noise types, signal-to-noise ratios, and beat delineation methods are presented in Figure 4. Figure 4a shows the results for the original beat labeling, and Figure 4b shows the results for the basic automatic peak detector in the presence of noise. Note that where the original beat labeling is fixed regardless of the amount of noise, a variable number of beats can be detected per file depending on the type and level of added noise for the automatic beat separator.

From these figures, we see that $\hat{Q}$ detection is typically more sensitive to noise and the delineation method of choice. Considering that this class is not as well represented as the other four, this is perhaps not very surprising. As seen in the noiseless beat detection results in Table 13, $\hat{F}$ detection under noisy conditions is typically poorer with the basic peak detector. Also poorly represented relative to the other classes, $\hat{F}$ detection is consistently worse as the SNR increases. This behavior is also seen in the results for the $\hat{Q}$ and $\hat{S}$ classes and is not well understood. However, as was a possibility in the presence of larger random noise in the synthetic data, as noise increases (SNR decreases), it may be the case that the beat morphologies lose their distinguishing characteristics and the beat selection is determined more by noise than the original defining features; that is, the addition of greater amounts of noise affects the likelihood that a member of a particular class is selected as it is no longer the base morphology dictating beat selection but instead the added noise.

With $\hat{N}$ detected in all cases, $\hat{V}$ is better detected in settings with less noise. Detection of $\hat{S}$ varies some but does not demonstrate a consistent trend as the SNR varies across the different noise types for either beat detector. While CUR class detection seems to vary more under em noise with basic peak detection, the larger classes are somewhat robust to the presence of noise. As noted above, we do see that the detection of more poorly represented classes can be influenced by noise and peak detection; with this observation, we must keep in mind that larger fluctuations in the detection of classes such as $\hat{F}$ and $\hat{Q}$ may be apparent due the smaller number of files containing beats with these annotations. In general, however, CUR appears to be reasonably robust to noise and beat delineation. Of course clean data with appropriate beat detection is always preferable, but CUR is able to identify different beat classes under noise in a number of cases. With some rare beat detection still possible in these cases, better performance is expected among classes that are well-represented in the data.

Though the results presented in this paper are generally encouraging regarding the utility of DEIM CUR with incremental QR in selecting representative ECG morphologies, there are a couple of points that should be noted regarding this approach. While we are able to detect a broad representation of classes in the data, some classes may be represented multiple times within the selected subset. If granularity in morphology detection is desired, this may be helpful; otherwise, a post-processing step to further refine the number of selected representatives may be desired. In addition, because each beat is interpolated to have the same length, it is important to recognize that the CUR-detected beats have lost their temporal context and are representative only with respect to morphology. This limitation should be considered in using the selected subsets for further analysis; while the morphologies may indeed be representative, they may lose some interpretability without information regarding their original timing. Hence, in selecting representative beat morphologies, it may be of interest to also maintain a record of the original timing of each beat.

It is also important to note that, where we have focused on the application of DEIM CUR to electrocardiogram data, this approach can be used for subset selection in other data types. In general, the method presented here can be used in any setting in which dimension reduction is desired without loss of the underlying structure of the data, from physiological time series to electronic health records to gene expression data. For an example beyond the temporal setting, as mentioned in Section 2, the CUR factorization based on leverage scores has already been applied to select subsets of genes from microarray data in toxicogenomics [28]. Sorensen and Embree also apply both DEIM CUR and leverage-score-based CUR to genetic data for the identification of features (or probes) that can be used for binary classification of patient data, demonstrating that the superiority of one method over the other depends on the problem and the measure of success [25].

While we have only highlighted in this work the beats selected via DEIM for the **C** matrix in the CUR decomposition, there are other data sets in which both **C** and **R** may hold valuable information–data sets in which not only representative observations are desired, but also representative features within these observations. Reaching much further than the quasiperiodic time-series analysis presented here, the underlying DEIM CUR framework is applicable to a number of different problems across the entire biomedical informatics spectrum.

5. Conclusion

In this paper, we have demonstrated the utility of DEIM-CUR with incremental QR in identifying representative beat morphologies in ECG waveform data. Testing on both synthetic and real patient data, we see that our approach is able to select both common and rare morphologies occurring within the data, offering an unsupervised means of reducing the data set to a robust representation of the whole. Through comparison with classification methods presented in the existing literature, we have also demonstrated that the CUR-selected subset can be effectively used classifying beats with AAMI and AAMI2 labels. We have also demonstrated that CUR is reasonably robust in the presence of realistic noise and automated beat detection, particularly for classes that are well represented in the data.

Future work for this approach includes investigating ways of making the CUR factorization less sensitive to noise within the data without losing rare beat detection, as well as the extension of this approach to larger data sets. In addition, the impact of other factors in the treatment of the ECG prior to analysis should be more closely considered. For example, a more detailed analysis given a wider variety of filtering and beat delineation techniques should be analyzed in the future. Further study regarding the role of CUR-selected beats in classification algorithms is also a topic of interest for future work.

With extendability to larger data sets and other waveform types (not to mention a variety of other data types), this method can serve as a springboard for identifying important subsequences in physiological data. Whether the selected subsequences are themselves used, or the subsequences are further analyzed for predictive model construction, the application of CUR to physiological time series can lead to the development of improved clinical decision support tools.

Acknowledgments

The authors would like to thank Sebastian Acosta for his helpful input throughout the completion of this research. In addition, the authors would like to thank Chris Jermaine for his comments on this manuscript, as well as the anonymous reviewers for their helpful suggestions in making improvements to this paper. This research was primarily supported by a training fellowship from the Keck Center of the Gulf Coast Consortia, on the NLM Training Program in Biomedical Informatics (NLM Grant No. T15LM007093). Partial funding has been received from the following: NSF DMS-0739420, NSF DMS-1312391, NSF DMS-1318348, AFOSR FA9550-12-1-0155, NSF CCF-1320866, NIH 1R56HL131574, and Award 16BGIA27490024 from the American Heart Association and The Children’s Heart Foundation.

Appendix A. Synthetic Data Parameters

Table A.14 presents the parameters for constructing the control set of the synthetic beats shown in Figure 2. The P_{shift}^{*} and T_{shift}^{*} parameters given in the table are first multiplied by the time allotted for the individual beat of interest in order to define the P_{shift} and T_{shift} used in Equations 4 and 5, respectively. In addition, $R_{peak}^{*}=0$ in Table A.14 is defined relative to the middle of the beat. That is, in constructing each beat individually, R_{peak} is taken to be in the middle of the beat.

Table A.14

Parameters for constructing the control set of synthetic beats. Note that P^{*}_{shift} and T^{*}_{shift} are multiplied by the total beat time to arrive at the true parameters used in defining the corresponding feature Gaussian curves. Also, R^{*}_{peak} is added to the midpoint of the allotted beat time interval such that the R peak occurs in the middle of the beat.

Class	P^{*}_{shift}	P_{amp}	P_{width}	Q_{shift}	Q_{amp}	Q_{width}	R^{*}_{peak}	R_{amp}	R_{width}	S_{shift}	S_{amp}	S_{width}	T^{*}_{shift}	T_{amp}	T_{width}	R'_{shift}	R'_{amp}	R'_{width}
1	0.25	0.1	0.02	0.01	0.2	0.01	0	1	0.01	0.01	0.25	0.01	0.3	0.15	0.03	0	0	0.008
2	0.25	0.1	0.02	0.01	0.25	0.01	0	1	0.01	0.01	0	0.01	0.3	0.15	0.03	0.0225	0.5	0.006
3	0.25	0	0.02	0.01	0.2	0.01	0	1	0.01	0.01	0.25	0.01	0.3	0.15	0.03	0	0	0.008
4	0.25	0.1	0.02	0.01	0.25	0.01	0	1	0.01	0.01	0.2	0.01	0.3	0.15	0.03	0.0225	0.7	0.006
5	0.25	0.1	0.02	0.01	0.25	0.01	0	1	0.01	0.01	0.4	0.01	0.15	0.15	0.03	0	0	0.008
6	0.25	0.1	0.02	0.01	0.2	0.01	0	0	0.01	0.01	0.5	0.01	0.15	0.15	0.03	0	0	0.008
7	0.25	0.04	0.02	0.011	0.35	0.01	0	0	0.01	0.013	−0.02	0.07	0.25	−0.05	0.05	0	0	0.008
8	0.25	0.04	0.03	0.011	0.05	0.01	0	0.4	0.01	0.03	0.02	0.08	0.25	0.005	0.05	0	0	0.008
9	0.25	0.02	0.03	0.011	0	0.01	0	0.4	0.01	0.04	0.02	0.06	0.28	−0.06	0.04	0	0	0.008
10	0.25	0.02	0.02	0.01	0	0.01	0	0.2	0.01	0.01	0.3	0.01	0.3	0.03	0.035	0	0	0.008
11	0.45	0.02	0.02	0.01	0	0.01	0	0.2	0.01	0.01	0.3	0.01	0.4	0.03	0.035	0	0	0.008
12	0.4	0.1	0.02	0.01	0.25	0.01	0	1	0.01	0.01	0.2	0.01	0.4	0.15	0.03	0.0225	0.7	0.006

Appendix B. Full MIT-BIH Database Class Detection Results

Table B.15

The number of CUR-selected beats out of the total number of beats with a given annotation for each patient file in the MIT-BIH Arrhythmia Database. The last row shows the fraction of files in which an annotation was detected via CUR given that the annotation was present. (CUR stopping tolerance 5 × 10^{-5}.)

File ID	N	A	V	Q	/	f	F	j	L	a	J	R	!	E	S	e
	CUR/Tot.	CUR/Tot.	CUR/Tot.	CUR/Tot.	CUR/Tot.	CUR/Tot.	CUR/Tot.	CUR/Tot.	CUR/Tot.	CUR/Tot.	CUR/Tot.	CUR/Tot.	CUR/Tot.	CUR/Tot.	CUR/Tot.	CUR/Tot.
100	127/2011	21 / 31	1/1	–	–	–	–	–	–	–	–	–	–	–	–	–
101	145/1664	2/3	–	2/2	–	–	–	–	–	–	–	–	–	–	–	–
102	13/87	–	2/4	–	120/1841	14/35	–	–	–	–	–	–	–	–	–	–
103	147/1874	2/2	–	–	–	–	–	–	–	–	–	–	–	–	–	–
104	11/158	–	1/2	15/15	70/1226	52/604	–	–	–	–	–	–	–	–	–	–
105	133/2277	–	14/34	2/5	–	–	–	–	–	–	–	–	–	–	–	–
106	90/1347	–	59/471	–	–	–	–	–	–	–	–	–	–	–	–	–
107	–	–	25/50	–	124/1871	–	–	–	–	–	–	–	–	–	–	–
108	146/1550	0/3	3/14	–	–	–	0/2	0/1	–	–	–	–	–	–	–	–
109	–	–	22/34	–	–	–	–	–	127/2239	–	–	–	–	–	–	–
111	–	–	1/1	–	–	–	–	–	148/1905	–	–	–	–	–	–	–
112	147/2283	2/2	–	–	–	–	–	–	–	–	–	–	–	–	–	–
113	145/1611	–	–	–	–	–	–	–	–	4/5	–	–	–	–	–	–
114	133/1628	3/9	11/42	–	–	–	1/4	–	–	–	1/2	–	–	–	–	–
115	149/1755	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
116	137/2058	1/1	11/108	–	–	–	–	–	–	–	–	–	–	–	–	–
117	148/1380	1/1	–	–	–	–	–	–	–	–	–	–	–	–	–	–
118	–	20/86	9/15	–	–	–	–	–	–	–	–	120/1940	–	–	–	–
119	102/1378	–	47/413	–	–	–	–	–	–	–	–	–	–	–	–	–
121	147/1662	1/1	1/1	–	–	–	–	–	–	–	–	–	–	–	–	–
122	149/2217	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
123	146/1360	–	3/3	–	–	–	–	–	–	–	–	–	–	–	–	–
124	–	0/2	25/47	–	–	–	3/5	1/5	–	–	6/29	114/1369	–	–	–	–
200	88/1570	4/27	57/750	–	–	–	0/2	–	–	–	–	–	–	–	–	–
201	97/1372	5/30	20/198	–	–	–	1/2	0/8	–	26/94	0/1	–	–	–	–	–
202	117/1849	9/36	13/18	–	–	–	0/1	–	–	10/19	–	–	–	–	–	–
203	130/2266	–	16/412	3/4	–	–	–	–	–	0/2	–	–	–	–	–	–
205	133/2313	2/3	14/71	–	–	–	0/11	–	–	–	–	–	–	–	–	–
207	–	–	24/69	–	–	–	–	–	81/1438	–	–	3/58	35/428	6/70	–	–
208	91/1427	–	49/880	1/2	–	–	8/350	–	–	–	–	–	–	–	0/2	–
209	127/2365	21/350	1/1	–	–	–	–	–	–	–	–	–	–	–	–	–
210	116/2187	–	23/167	–	–	–	3/8	–	–	7/20	–	–	–	–	–	–
212	56/811	–	–	–	–	–	–	–	–	–	–	93/1668	–	–	–	–
213	113/2339	10/22	13/211	–	–	–	11/350	–	–	2/3	–	–	–	–	–	–
214	–	–	57/233	1/2	–	–	1/1	–	90/1794	–	–	–	–	–	–	–
215	118/2880	1/3	30/142	–	–	–	0/1	–	–	–	–	–	–	–	–	–
217	50/244	–	14/156	–	25/1337	60/254	–	–	–	–	–	–	–	–	–	–
219	135/1863	1/7	12/55	–	–	–	1/1	–	–	–	–	–	–	–	–	–
220	112/1746	37/90	–	–	–	–	–	–	–	–	–	–	–	–	–	–
221	113/1819	–	36/375	–	–	–	–	–	–	–	–	–	–	–	–	–
222	104/1800	14/208	–	–	–	–	–	31/212	–	–	0/1	–	–	–	–	–
223	85/1803	13/66	48/456	–	–	–	1/14	–	–	0/1	–	–	–	–	–	2/15
228	133/1511	0/3	16/326	–	–	–	–	–	–	–	–	–	–	–	–	–
230	149/2015	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
231	61/312	1/1	2/2	–	–	–	–	–	–	–	–	85/1068	–	–	–	–
232	–	126/1246	–	–	–	–	–	0/1	–	–	–	23/353	–	–	–	–
233	63/2015	0/5	86/737	–	–	–	0/11	–	–	–	–	–	–	–	–	–
234	139/2426	–	3/3	–	–	–	–	–	–	–	7/50	–	–	–	–	–
Representation	40/40	22/26	36/36	6/6	4/4	3/3	9/15	2/5	4/4	5/7	3/5	6/6	1/1	1/1	0/1	1/1

References

1. Hirano S, Tsumoto S. Mining Similar Temporal Patterns In Long Time-Series Data And Its Application To Medicine. Data Mining, 2002. 2002:219–226. ICDM 2003. Proceedings. 2002 IEEE International Conference on, IEEE.
2. Graupe D, Salahi J, Kohn KH. Multifunctional prosthesis and orthosis control via microcomputer identification of temporal pattern differences in single-site myoelectric signals. Journal of Biomedical Engineering. 1982; 4(1):17–22. [PubMed: 7078136]
3. Patnaik, D., Butler, P., Ramakrishnan, N., Parida, L., Keller, BJ., Hanauer, DA. Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM; 2011. Experiences with mining temporal event sequences from electronic medical records: initial successes and some challenges; p. 360-368.
4. Dubin, D. Rapid interpretation of EKG's. COVER Publishing Company; 2000.
5. Velic, M., Padavic, I., Car, S. EUROCON, 2013. IEEE, IEEE; 2013. Computer aided ECG analysisstate of the art and upcoming challenges; p. 1778-1784.
6. Rusin CG, Acosta SI, Shekerdemian LS, Vu EL, Bavare AC, Myers RB, Patterson LW, Brady KM, Penny DJ. Prediction of imminent, severe deterioration of children with parallel circulations using real-time processing of physiologic data. The Journal of Thoracic and Cardiovascular Surgery.
7. Hendryx, E. Master's thesis. Rice University; 2015. Identifying ECG clusters in congenital heart disease. http://hdl.handle.net/1911/88126
8. Khadra L, Al-Fahoum A, Al-Nashash H. Detection of life-threatening cardiac arrhythmias using the wavelet transformation. Medical and Biological Engineering and Computing. 1997; 35(6):626–632. [PubMed: 9538538]
9. Huang B, Kinsner W. ECG frame classification using dynamic time warping. Electrical and Computer Engineering, 2002. 2002:1105–1110. IEEE CCECE 2002. Canadian Conference on, Vol. 2 IEEE.
10. Tuzcu, V., Nas, S. Systems, Man and Cybernetics, 2005 IEEE International Conference on. Vol. 1. IEEE; 2005. Dynamic time warping as a novel tool in pattern recognition of ECG changes in heart rhythm disturbances; p. 182-186.
11. Shorten, G., Burke, M. 2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE; 2011. A time domain based classifier for ECG pattern recognition; p. 4980-4983.
12. Raghavendra, B., Bera, D., Bopardikar, AS., Narayanan, R. World of Wireless, Mobile and Multimedia Networks (WoWMoM), 2011 IEEE International Symposium on a. IEEE; 2011. Cardiac arrhythmia detection using dynamic time warping of ECG beats in e-healthcare systems; p. 1-6.
13. Stamkopoulos T, Diamantaras K, Maglaveras N, Strintzis M. ECG analysis using nonlinear PCA neural networks for ischemia detection. IEEE Transactions on Signal Processing. 1998; 46(11): 3058–3067.
14. Jokić, S., Krčo, S., Delić, V., Sakać, D., Jokić, I., Lukić, Z. Neural Network Applications in Electrical Engineering (NEUREL), 2010 10th Symposium on. IEEE; 2010. An efficient ECG modeling for heartbeat classification; p. 73-76.
15. Ceylan R, Özbay Y, Karlik B. A novel approach for classification of ECG arrhythmias: Type-2 fuzzy clustering neural network. Expert Systems with Applications. 2009; 36(3):6721–6726.
16. Yeh YC, Wang WJ, Chiou CW. A novel fuzzy c-means method for classifying heartbeat cases from ECG signals. Measurement. 2010; 43(10):1542–1555.
17. Annam, JR., Mittapalli, SS., Bapi, R. India Conference (INDICON), 2011 Annual IEEE. IEEE; 2011. Time series clustering and analysis of ECG heart-beats using dynamic time warping; p. 1-3.
18. Cuesta-Frau D, Pérez-Cortés JC, Andreu-García G. Clustering of electrocardiograph signals in computer-aided Holter analysis. Computer methods and programs in Biomedicine. 2003; 72(3): 179–196. [PubMed: 14554133]
19. Syed Z, Guttag J, Stultz C. Clustering and symbolic analysis of cardiovascular signals: discovery and visualization of medically relevant patterns in long-term data using limited prior knowledge. EURASIP Journal on Applied Signal Processing. 2007; 2007(1):97–97.

20. Ratanamahatana, CA., Lin, J., Gunopulos, D., Keogh, EJ., Vlachos, M., Das, G. Data Mining and Knowledge Discovery Handbook. Springer; 2010. Mining time series data; p. 1049-1077.

21. Goreinov SA, Tyrtyshnikov EE, Zamarashkin NL. A theory of pseudoskeleton approximations. Linear Algebra and its Applications. 1997; 261(1):1–21.

22. Stewart G. Four algorithms for the the efficient computation of truncated pivoted QR approximations to a sparse matrix. Numerische Mathematik. 1999; 83(2):313–323.

23. Drineas P, Mahoney MW, Muthukrishnan S. Relative-error CUR matrix decompositions. SIAM Journal on Matrix Analysis and Applications. 2008; 30(2):844–881.

24. Mahoney MW, Drineas P. CUR matrix decompositions for improved data analysis. Proceedings of the National Academy of Sciences. 2009; 106(3):697–702.

25. Sorensen DC, Embree M. A DEIM induced CUR factorization. SIAM Journal on Scientific Computing. 2016; 38(3):A1454–A1482.

26. Mitrovic, N., Asif, MT., Rasheed, U., Dauwels, J., Jaillet, P. 16th International IEEE Conference on Intelligent Transportation Systems (ITSC 2013). IEEE; 2013. CUR decomposition for compression and compressed sensing of large-scale traffic data; p. 1475-1480.

27. An, İ., Şimşekli, U., Cemgil, AT., Akarun, L. Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European. IEEE; 2012. Large scale polyphonic music transcription using randomized matrix decompositions; p. 2020-2024.

28. Žitnik M, Zupan B. Matrix factorization-based data fusion for drug-induced liver injury prediction. Systems Biomedicine. 2014; 2(1):16–22.

29. Dauwels, J., Srinivasan, K., Ramasubba, RM., Cichocki, A. Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on. IEEE; 2011. Multi-channel EEG compression based on matrix and tensor decompositions; p. 629-632.

30. Lee, H., Choi, S. Neural Networks, 2008. IEEE; 2008. CUR+ NMF for learning spectral features from large data matrix; p. 1592-1597.IJCNN 2008.(IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on

31. McSharry PE, Clifford GD, Tarassenko L, Smith LA. A dynamical model for generating synthetic electrocardiogram signals. IEEE Transactions on Biomedical Engineering. 2003; 50(3):289–294. [PubMed: 12669985]

32. Moody, GB., Mark, RG. Engineering in Medicine and Biology Magazine. Vol. 20. IEEE; 2001. The impact of the MIT-BIH arrhythmia database; p. 45-50.

33. Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PC, Mark RG, Mietus JE, Moody GB, Peng CK, Stanley HE. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation. 2000 Jun 13; 101(23):e215–e220. circulation Electronic Pages: http://circ.ahajournals.org/cgi/content/full/101/23/e215. DOI: 10.1161/01.CIR.101.23.e215 [PubMed: 10851218]

34. Welch J, Ford P, Teplick R, Rubsamen R. The massachusetts general hospital-marquette foundation hemodynamic and electrocardiographic database–comprehensive collection of critical care waveforms. Clinical Monitoring. 1991; 7(1):96–97.

35. MATLAB, version 8.2.0.701 (R2013b). The MathWorks Inc.; Natick, Massachusetts: 2013.

36. Silva I, Moody G. An Open-source Toolbox for Analysing and Processing PhysioNet Databases in MATLAB and Octave. Journal of Open Research Software. 2014; 2(1):e27. http://dx.doi.org/10.5334/jors.bi. [PubMed: 26525081]

37. Keogh EJ, Kasetty S. On the need for time series data mining benchmarks: A survey and empirical demonstration. Data Mining and knowledge discovery. 2003; 7(4):349–371.

38. Rakthanmanon, T., Campana, B., Mueen, A., Batista, G., Westover, B., Zhu, Q., Zakaria, J., Keogh, EJ. Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM; 2012. Searching and mining trillions of time series subsequences under dynamic time warping; p. 262-270.

39. Moody GB, Muldrow W, Mark RG. A noise stress test for arrhythmia detectors. Computers in cardiology. 1984; 11(3):381–384.

40. Clifford G, Behar J, Li Q, Rezek I. Signal quality indices and data fusion for determining clinical acceptability of electrocardiograms. Physiological measurement. 2012; 33(9):1419. [PubMed: 22902749]

41. Chaturantabut S, Sorensen DC. Nonlinear model reduction via discrete empirical interpolation. SIAM Journal on Scientific Computing. 2010; 32(5):2737–2764.

42. Daniel JW, Gragg WB, Kaufman L, Stewart G. Reorthogonalization and stable algorithms for updating the Gram-Schmidt QR factorization. Mathematics of Computation. 1976; 30(136):772–795.

43. De Chazal P, ’Dwyer M, Reilly RB. Automatic classification of heartbeats using ECG morphology and heartbeat interval features. IEEE Transactions on Biomedical Engineering. 2004; 51(7):1196–1206. [PubMed: 15248536]

44. Llamedo M, Martínez JP. Heartbeat classification using feature selection driven by database generalization criteria. IEEE Transactions on Biomedical Engineering. 2011; 58(3):616–625. [PubMed: 20729162]

45. Testing and reporting performance results of cardiac rhythm and ST segment measurement algorithms, ANSI/AAMI EC57.

46. Llamedo M, Martínez JP. An automatic patient-adapted ECG heartbeat classifier allowing expert assistance. IEEE Transactions on Biomedical Engineering. 2012; 59(8):2312–2320. [PubMed: 22692868]

47. Oster J, Behar J, Sayadi O, Nemati S, Johnson AE, Clifford GD. Semisupervised ECG ventricular beat classification with novelty detection based on switching Kalman filters. IEEE Transactions on Biomedical Engineering. 2015; 62(9):2125–2134. [PubMed: 25680203]

Highlights

- We propose CUR matrix factorization for representative beat selection in ECG data
- CUR reduces dimension and retains morphologies from each class in the data
- Both common and rare beat events are selected, providing broad data representation

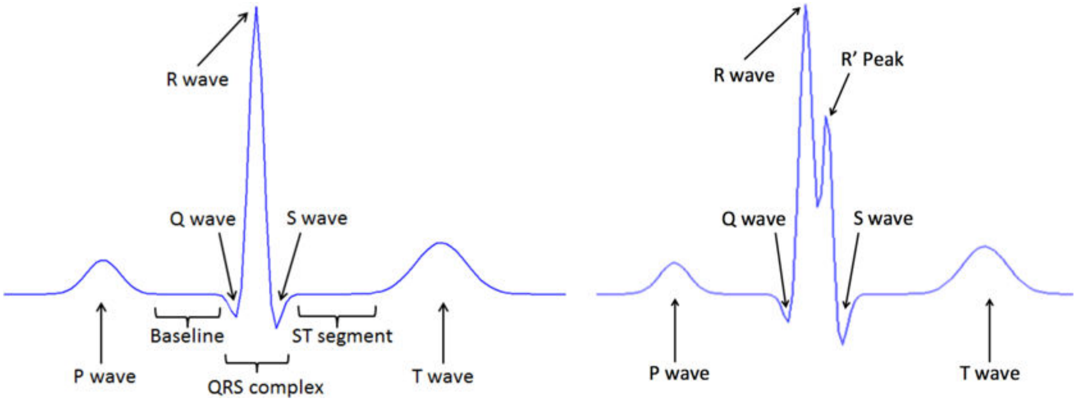


Figure 1.

Labeled features within synthetic ECG waveforms. Left: Some standard ECG features; this figure is taken from [7]. Right: Some standard ECG features with the addition of the R′ peak.

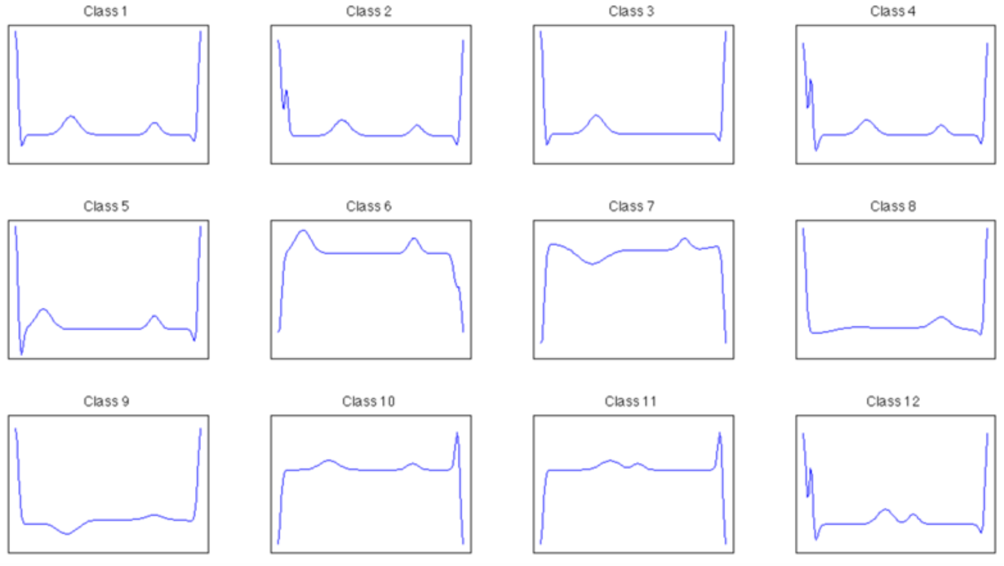


Figure 2.

Control morphologies from each of the twelve synthetic beat classes constructed.

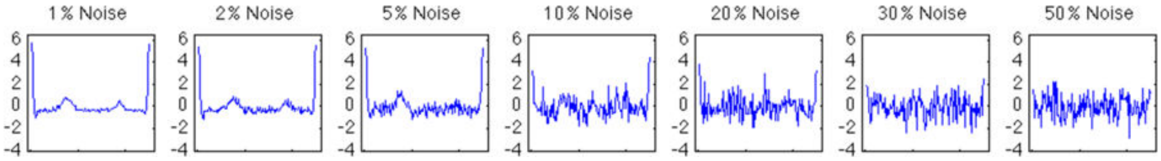


Figure 3.

Examples of the Class 1 synthetic beat morphology with the addition of different levels of random noise.

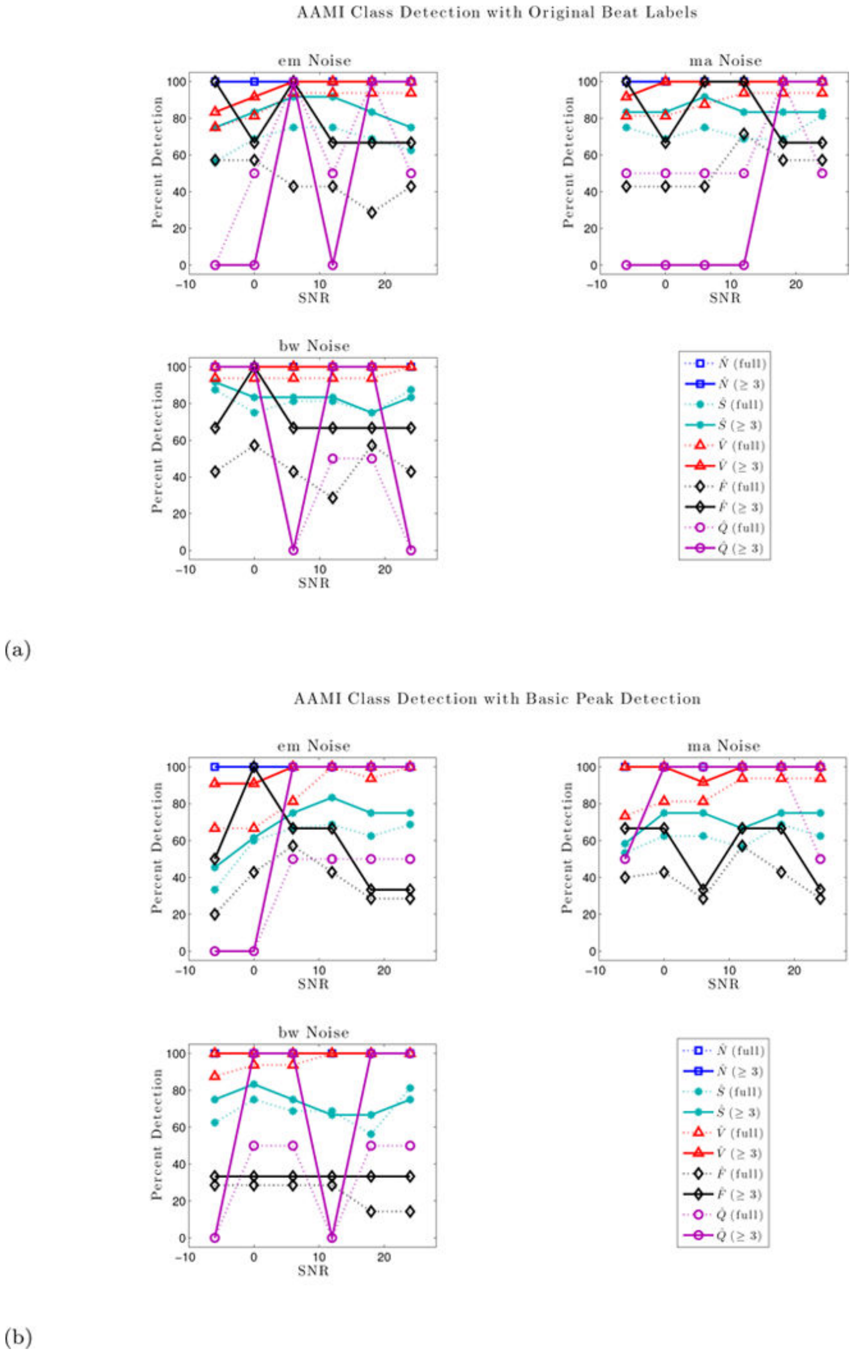


Figure 4.

Percent detection of AAMI classes for original beat delineation and the basic automated peak finder with the three different kinds of added noise. Results are shown for the full files and for the removal of PhysioNet annotations with fewer than three beats represented.

Table 1

Results from the CUR factorization applied to the synthetic data sets with different types and levels of feature variability. (Recall that placement variability is added at one half of the percentages in the first column.) Dimension reduction is given as a percentage of trial size. This table shows that most classes are detected even with > 97% dimension reduction.

Change	Heart Rate		Placement		Magnitude		Width		Noise	
	Dimension Reduction	Missed Classes	Dimension Reduction	Missed Classes	Dimension Reduction	Missed Classes	Dimension Reduction	Missed Classes	Dimension Reduction	Missed Classes
1%	99.25	0	99.42	0	99.58	0	99.40	0	97.93	1
2%	99.05	0	99.35	0	99.55	0	99.33	0	97.93	2
5%	98.73	0	99.25	0	99.53	0	99.27	0	97.93	3
10%	98.47	0	99.15	0	99.50	0	99.20	0	97.93	3
20%	98.13	0	99.07	1	99.40	0	99.07	0	97.93	0
30%	98.02	0	99.07	0	99.22	0	99.02	0	97.93	0
50%	97.93	0	99.03	2	–	–	98.77	1	97.93	0

Table 2

The original DEIM-CUR representation for each annotation present in the MIT-BIH Arrhythmia Database and for the eight different CUR stopping tolerances tested. The values in each column represent the percentage of patients correctly identified as having a given annotation through CUR beat selection.

tol	N	A	V	Q	/	f	F	j	L	a	J	R	!	E	S	e
0.05	100	34.615	80.556	33.333	100	100	13.333	0	100	28.571	20	100	100	100	0	0
0.01	100	88.462	97.222	66.667	100	100	46.667	40	100	71.429	60	83.333	100	100	0	0
5×10^{-3}	100	69.231	100	83.333	100	100	60	40	100	85.714	60	100	100	100	0	0
1×10^{-3}	100	84.615	100	83.333	100	100	73.333	40	100	71.429	60	100	100	100	0	100
5×10^{-4}	100	84.615	100	83.333	100	100	53.333	40	100	71.429	60	100	100	100	0	100
1×10^{-4}	100	88.462	100	100	100	100	60	20	100	71.429	60	100	100	100	0	100
5×10^{-5}	100	84.615	100	100	100	100	60	40	100	71.429	60	100	100	100	0	100
1×10^{-5}	100	84.615	100	100	100	100	60	40	100	71.429	60	100	100	100	0	100

Table 3

Comparison of the annotation representation results from Syed et al. [19] and DEIM-CUR with incremental QR. The CUR results presented here are for stopping tolerance 5×10^{-5} and consider only patients with three or more beats present for a particular annotation.

Method	N	A	V	/	f	F	j	L	a	R	E	e
Syed, et al. (2007)	100	84.21	100	100	100	75	100	100	100	100	100	100
Inc. QR, DEIM-CUR	100	84.21	100	100	100	75	66.67	100	100	100	100	100

Table 4

The DEIM-CUR representation for each annotation present in the MGH-MF Database and for an incremental QR stopping tolerance of 5×10^{-5}. The values in each column represent the percentage of patients correctly identified as having a given annotation through CUR beat selection. The first row considers all patient annotations. The second row shows results only for those cases in which ≥ 3 beats with a given annotation are present and excluding files mgh002 and mgh026 for comparison with the results presented by Syed et al. [19] (shown in the last row).

Experiment	N	S	V	F	J	/	e
MGH-MF (first 40 files)	100	91.304	87.097	20	100	100	100
MGH-MF (≥ 3 beats, subset)	100	100	90	100	100	100	100
MGH-MF (Syed et al.)	100	–	100	100	100	100	100

Table 5

The DEIM-CUR representation for each annotation present in the full MGH-MF Database (excluding files mgh061, mgh127, mgh230, and mgh235) and for incremental QR stopping tolerance 5×10^{-5}. Each column shows the percentage of patients correctly identified as having a given annotation through CUR beat selection. The first row considers all patient files and annotations. The second row shows results only for those cases in which ≥ 3 beats with a given annotation are present.

Experiment	N	S	V	F	J	/	e	Q	n	f	a	r	?
MGH-MF (all files)	100	83.553	86.857	43.548	100	100	100	66.667	0	0	100	14.286	50
MGH-MF (≥ 3 beats, all)	100	91.379	91.912	54.545	100	100	100	62.5	–	–	100	20	–

Table 6

The DEIM-CUR representation for each annotation present in the Incart Database and for an incremental QR tolerance of 5×10^{-5}. The values in each column represent the percentage of patients correctly identified as having a given annotation through CUR beat selection. The first row considers all patient annotations, and the second row shows results only for those cases in which ≥ 3 beats with a given annotation are present.

Experiment	N	V	A	F	Q	n	R	B	j	S
Incart (all files)	100	100	71.875	50	20	50	100	0	50	25
Incart (≥ 3 beats)	100	100	89.474	64.286	–	50	100	–	50	33.333

Table 7

The DEIM-CUR representation for the AAMI and AAMI2 class labelings of the MIT-BIH Arrhythmia data for an incremental QR tolerance of 5 × 10^{-5}. The values in each column represent the percentage of patients correctly identified as having a given annotation through CUR beat selection. The first row considers all patient annotations, and the second row shows results only for those cases in which ≥ 3 beats with a given annotation are present. These results ignore records with paced beats.

Experiment	$\hat{N}$	$\hat{S}$	$\hat{V}$	$\hat{F}$	$\hat{Q}$	Experiment	$\hat{N}$	$\hat{S}$	$\hat{V}'$
AAMI (all files)	100	83.871	100	60	100	AAMI2 (all files)	100	83.871	100
AAMI (≥ 3 beats)	100	86.957	100	75	100	AAMI2 (≥ 3 beats)	100	86.957	100

Table 8

DS1 CUR classification results for $\hat{V}'$ and $\hat{V}$.

DS1	$\hat{\mathbf{V}}'$			$\hat{\mathbf{V}}$		
Tol.	***Se***	***+P***	$\boldsymbol{F_1}$	***Se***	***+P***	$\boldsymbol{F_1}$
0.05	79.91	91.51	85.32	79.70	82.25	80.96
0.01	89.51	97.17	93.18	90.36	89.17	89.76
5×10^{-3}	94.26	99.12	**96.63**	94.34	91.42	**92.85**
1×10^{-3}	95.59	96.24	95.92	96.17	89.58	92.76
5×10^{-4}	95.12	96.00	95.56	95.76	89.25	92.39
1×10^{-4}	94.81	96.31	95.55	95.70	89.56	92.53
5×10^{-5}	94.81	96.31	95.55	95.70	89.56	92.53
1×10^{-5}	94.81	96.31	95.55	95.70	89.56	92.53

Table 9

DS2 CUR classification results for $\hat{V}'$ and $\hat{V}$ with incremental QR tolerance 5 × 10^{-3}. This table is adapted from the results presented in Table IV of [47].

DS2	$\hat{\mathbf{V}}'$			$\hat{\mathbf{V}}$		
Method	***Se***	***+P***	$\boldsymbol{F_1}$	***Se***	***+P***	$\boldsymbol{F_1}$
de Chazal et al. [43]	86.5	47.2	57.1	85.1	81.9	83.5
Llamedo and Martínez [44]	95.3	28.6	44.0	94.6	88.1	91.2
Llamedo and Martínez (automatic) [46]	82.1	77.9	79.9	90.1	86.0	88.0
Llamedo and Martínez (assisted) [46]	91.4	96.9	94.1	93.8	96.7	95.2
Oster et al. (no X-factor) [47]	87.6	96.4	91.8	92.7	96.2	94.5
Oster et al. (with X-factor) [47]	90.5	99.96	95.2	97.3	99.96	98.6
This work (CUR-1NN)	94.4	97.9	96.1	92.5	98.3	95.3

Table 10

Incart CUR classification results for $\hat{V}'$ and $\hat{V}$ with incremental QR tolerance 5 × 10^{-3}. This table is adapted from the results presented in Table VI of [47].

Incart	$\hat{\mathbf{V}}'$			$\hat{\mathbf{V}}$		
Method	***Se***	***+P***	$\boldsymbol{F_1}$	***Se***	***+P***	$\boldsymbol{F_1}$
Llamedo and Martínez [44]	82	88	84.9	N/A	N/A	N/A
Llamedo and Martínez (automatic) [46]	88.0	96.0	91.8	N/A	N/A	N/A
Llamedo and Martínez (assisted) [46]	98.0	98.0	98.0	N/A	N/A	N/A
Oster et al. (no X-factor) [47]	94.7	99.3	97.0	95.4	99.3	97.3
Oster et al. (with X-factor)[47]	98.6	99.9	99.2	99.1	99.96	99.4
This work (CUR-1NN)	93.3	98.8	96.0	93.3	98.5	95.8

Table 11

DS2 AAMI CUR classification confusion matrix with incremental QR tolerance 5 × 10^{-3}.

DS2	$\hat{\mathbf{N}}_{pred}$	$\hat{\mathbf{S}}_{pred}$	$\hat{\mathbf{V}}_{pred}$	$\hat{\mathbf{F}}_{pred}$	$\hat{\mathbf{Q}}_{pred}$	
$\hat{\mathbf{N}}_{true}$	39267	350	25	35	1	39678
$\hat{\mathbf{S}}_{true}$	161	1521	6	0	0	1688
$\hat{\mathbf{V}}_{true}$	111	6	2698	102	0	2917
$\hat{\mathbf{F}}_{true}$	66	1	15	292	0	374
$\hat{\mathbf{Q}}_{true}$	6	0	0	0	1	7
	39611	1878	2744	429	2	44664

Table 12

Incart AAMI CUR classification confusion matrix with incremental QR tolerance 5 × 10^{-3}.

Incart	$\hat{\mathbf{N}}_{pred}$	$\hat{\mathbf{S}}_{pred}$	$\hat{\mathbf{V}}_{pred}$	$\hat{\mathbf{F}}_{pred}$	$\hat{\mathbf{Q}}_{pred}$	
$\hat{\mathbf{N}}_{true}$	138035	389	205	5	0	138634
$\hat{\mathbf{S}}_{true}$	630	1131	1	0	0	1762
$\hat{\mathbf{V}}_{true}$	1129	7	16927	83	0	18146
$\hat{\mathbf{F}}_{true}$	85	0	55	72	0	212
$\hat{\mathbf{Q}}_{true}$	3	0	2	0	1	6
	139882	1527	17190	160	1	158760

Table 13

The DEIM-CUR representation of the AAMI classes from the DS2 subset of the MIT-BIH Arrhythmia data for an incremental QR tolerance of 5 × 10^{-3}. The first row of the table contains results for the original beat delineation available on PhysioNet [33], and the second row contains results for beat delineation determined by the basic peak detector used on the synthetic data.

Experiment	$\hat{\mathbf{N}}$	$\hat{\mathbf{S}}$	$\hat{\mathbf{V}}$	$\hat{\mathbf{F}}$	$\hat{\mathbf{Q}}$	Experiment	$\hat{\mathbf{N}}$	$\hat{\mathbf{S}}$	$\hat{\mathbf{V}}$	$\hat{\mathbf{F}}$	$\hat{\mathbf{Q}}$
Original (all files)	100	75	100	57.143	50	Original (≥ 3 beats)	100	75	100	66.667	100
Automated (all files)	100	81.25	93.75	14.286	50	Automated (≥ 3 beats)	100	75	100	33.333	100