Motif Discovery in Long Time Series: Classifying Phonocardiograms

Hajar Alhijailan[1,2] [0000−0002−4169−7911] and Frans Coenen[1] [0000−0003−1026−6649]

[1] Department of Computer Science, University of Liverpool, Liverpool, United Kingdom
{h.alhijailan,coenen}@liverpool.ac.uk
[2] College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia
halhujailan@ksu.edu.sa

Abstract. A mechanism is presented for classifying phonocardiograms (PCGs) by interpreting PCGs as time series and using the concept of motifs, times series subsequences that are good discriminators of class, to support nearest neighbour classification. A particular challenge addressed by the work is that PCG time series are large which renders exact motif discovery to be computationally expensive; it is not realistic to compare every candidate time series subsequence with every other time series subsequence in order to discover exact motifs. Instead, a mechanism is proposed the firstly makes use of the cyclic nature of PCGs and secondly adopts a novel time series pruning mechanism. The evaluation, conducted using a canine PCG dataset, illustrated that the proposed approach produced the same classification accuracy but in a significantly more efficient manner.

Keywords: Phonocardiograms · Time Series Segmentation · Frequent Motif Discovery · Time Series Analysis · Classification.

1 Introduction

A phonocardiogram (PCG) is a recording of the sound of the heart; it is essentially an univariate time series [1,13,22]. The sound is cyclic and comprises two phases (S1 and S2), the Systole phase when the heart contracts, and the Diastole when the heart relaxes. The systole phase starts and ends with two sound components, when the *atrioventricular* valves close and when the aortic and pulmonary valves close respectively. The diastole phase is marked by the relative absence of sound. There will also be unwanted background noise and, in an unhealthy heart, what are known as murmurs, indicators of abnormal activity. A PCG is collected using a phonocardiograph, more commonly known as an electronic or digital stethoscope [8]. By analysing PCGs, it is possible to detect heart conditions of various sorts; this process can be automated using machine learning techniques, typically supervised learning (classification).

A common method of classifying time series where repeating patterns are known to exist is Motif Discovery [7,10,28,29]. The idea is to discover and store reoccurring patterns, known as *motifs* which are representative of class-labels [15,17]. A good motif, in the context of this paper, is one that appears frequently and at the same time is associated with only a single class. The discovered motifs can then be used to label (classify)

previously unseen time series [1,22,28]. However, finding motifs that are good representatives of class-labels is computationally challenging, especially given long time series (as in the case of PCG data). Exact motif discovery requires the comparison of every candidate time series subsequence with every other subsequence that exists in the dataset; a computationally expensive enterprise. One solution is to adopt an approximate approach [22]. Another is to reduce the complexity of the motif discovery process by preprocessing the time series so as to reduce the number of computations needed later in the process.

In this paper, an exact PCG motif discovery mechanism is presented which is effective in terms of classification accuracy and is efficient in terms of runtime. The idea is to limit the number of time series subsequences to be considered by first identifying cycles, using a PCG segmentation mechanism founded on the approach presented in [12], but with modifications. Next, to prune cycles that will not result in good motifs using a novel "zero motif" mechanism. Then, to process the retained candidate motifs further so as to extract good discriminators of class. The approach was evaluated using a canine PCGs dataset comprised of four classes. The first three classes described stages of Mitral Valve disease, as defined by the the European College of Veterinary Internal Medicine [2,23], and the fourth was a control class (no disease). The evaluation results obtained indicated that the proposed mechanisms was more efficient than alternative algorithms considered, whilst obtaining the same accuracy.

The rest of this paper is organised as follows. Section 2 gives a review of previous work regarding the research domain. The proposed PCG frequent motif selection and extraction mechanism, and its processes, are then presented in Section 3. Section 4 considers the evaluation strategy, followed by presentation and discussion of the results. The paper is completed with some concluding remarks in Section 5.

2 Previous Work

Time series analysis is concerned with the processing of time series data so as to extract knowledge. Typical applications include the discovery of distinguishing patterns, the clustering of time series collections and the modelling of the domain from which the time series are drawn. In the case of pattern identification, one type of pattern, and that of interest with respect to the work presented in this paper, is the *motif* [5,11,14]. A motif is a reoccurring subsequence in a time series that is a good indicator of class. A subsequence (candidate motif) is said to be reoccurring, and hence a motif, if there is at least one non-trivial match with another subsequence in a given time series according to some predefined similarity threshold [20,27]; a "trivial match" is where two subsequences overlap. To measure how well two subsequences match, a distance function is required. Euclidean Distance (ED) is widely used in the literature with some evidence suggesting its competitiveness with, or superiority to, other more complex measures [9]. The "brute force" approach to identifying motifs entails a significant computational overhead. A number of more efficient, but approximate, motif discovery algorithms have therefore been proposed [18,19,22], while a tractable exact algorithm remains a research challenge [3,22]. The later is, in part, the research focus of this paper. The exact algorithm presented in [22], the MK algorithm, is of particular relevance to this

paper because it is used as a comparator approach with which to compare the operation of the proposed approach.

The efficiency of motif discovery algorithms, exact or approximate, can be enhanced by preprocessing the input data. This is typically conducted using knowledge of the application domain to reduce the number of calculations to be considered, usually by considering the characteristics of the subsequences to be considered. The simplest technique is to restrict the comparisons to potential non-trivial matches [5]. This technique is widely used in many proposed motif discovery algorithms [5,22] and is adopted with respect to the work presented in this paper. For some applications, it is possible to exclude some sequences because they are known in advance not to be relevant, but this requires very specific domain knowledge. Another technique for reducing the complexity of the motif discovery process is to adopt the concept of "early abandonment" whereby a similarity comparison is stopped when the dissimilarity between two potential motifs being compared reaches a pre-specified threshold at which point it can be safely assumed that the two subsequences cannot be motifs. The threshold can be user-defined; or, as in the case of [22], derived.

There has been considerable work directed at analysing PCG data, although not in terms of motif discovery, for the purpose of PCG classification. A segmentation process is typically applied first to identify cycles. This is usually achieved with respect to a reference signal, either an ECG signal recorded at the same time and/or a Carotid Pulse (CP) [16,24,25]. In the case of the PCG signals collected using electronic stethoscopes, the application focus of the work presented in this paper, no such reference signal is typically available. In such cases, the components of the PCG signal can still be extracted by processing the signal. This is usually achieved according to the "energy" of the signal and one or more *energy thresholds* [4,12]. The well-known Shannon Energy is frequently used [21] as it maintains time series features. However, there are situations where not all of the features are needed, as in the case of the work presented in this paper, in which case alternative energy methods can be used as long as the required salient features are preserved. Extracting the cardiac components from PCGs using empirically defined static thresholds is usually inappropriate because of the varying amplitudes recorded. This is due to difference between subjects in: the thickness of the chest wall [32], subject age [30], subject mood [30] and further subjective factors [30]. The alternative is to use dynamically computed thresholds [4,12]; this is the approach adopted with respect to the work presented in this paper.

3 PCG Frequent Motif Selection and Extraction

This section presents the proposed PCG frequent motif selection and extraction process. The input is a set of time series $T = \{\langle P_1, c_1\rangle, \langle P_2, c_2\rangle, \dots\}$. The output is a set of motifs, H''; a set of frequently occurring PCG cycles that are considered to be good discriminators of class. These motifs can then be used to classify previously unseen PCGs. The proposed mechanism is a three stage process:

1. Cycle segmentation.
2. Candidate motif selection.
3. Frequent motif extraction.

In the first stage, the set T is processed to produce the set H, a set of heartbeat cycle and class pairs $\langle h_i, c_i \rangle$. In the second stage, the set H is pruned by removing infrequent cycles so as to produce a set of cycles H'. This is then further processed in Stage 3 to identify the set of k motifs, the most frequent cycles within H' that are good discriminators of class; these are held in a set H'' which can be then used as a "data bank" in a Nearest Neighbour Classification (NNC) model which in turn can be used to label previously unseen time series. Each stage is discussed in further detail in the following three subsections, Subsections 3.1 to 3.3.

3.1 Cycle Segmentation

PCGs cycles comprises: (i) a heartbeat, (ii) some murmurs and clicks, if diseased, and (iii) noise. A cycle is measured from the start of the S1 component to the start of the following S1 component. The idea was to segment a training set of labelled PCGs into a collection of cycles and then to group the cycles according to class-labels. This idea is common in the Signal Processing field and has been applied to PCG signals to, for example, study the duration of S1 or to find "click positions" [21,30]. The proposed mechanism differs from this previous work; the focus is on "whole cycles" rather than their components. The mechanism, is founded on that presented in [12], but with modifications, and operates using a dynamic threshold, computed for each PCG signal, so as to detect the beginning of cycles.

The process is as follows. For each point series $P_i = \{p_{i_1}, p_{i_2}, \dots\} \in T$ the *standardised signal energy envelope* (V) is calculated according to the signal (time series) energy E which is standardised to give E'. The energy $E = \{e_1, e_2, \dots\}$ is computed by squaring the amplitude values (p_{i_j}) in P_i, for each $e_j \in E$, using Equation 1. There are many other ways to calculate the energy of a signal, such as using absolute value, Shannon entropy or Shannon energy [12]. As the aim in the context of the work presented in this paper is to detect the beginning of cycles, the start of the S1 component, usually the component with highest amplitude (the loudest) [8], the above method of calculating the energy was adopted because samples with high amplitude will be favoured over those with low amplitude. This will in turn facilitate S1 detection. The standardised energy e'_j, given a value e_j, is then calculated using Equation 2, where μ_e and σ_e are the mean and standard deviation of E respectively.

$$e_j = p_{i_j}^2 \qquad (1) \qquad\qquad e'_j = \frac{e_j - \mu_e}{\sigma_e} \qquad (2)$$

The set of standardised energy values, $E' = \{e'_1, e'_2, \dots\}$ are then used to define the envelope V, which is then used to detect the beginning of cycles using an amplitude "cut-off" value t. The process is illustrated in Figure 1. The process starts by identifying the oscillation in V with the highest energy value, the magenta oscillation in Figure 1(a). Then, a predefined α threshold, a percentage of the width of the oscillation with the highest amplitude (the start and end of an oscillation can be identified from trend changes in V), is used to determine the value for t, the cyan line shown in Figure 1(b) (and Figure 1(c)). The value for t is used to find ascending intersection

points in V as shown in Figure 1(c). All oscillations in the energy envelope V whose amplitude falls bellow t are ignored, because they are deemed to be S2s, clicks and murmurs. Using this process, some ascending intersection points demarcating S2 components will still be retained, as illustrated in Figure 1(c). To remove these, the distance between intersection points is considered, if this falls below the average distance then we have an S2 intersection point which should be ignored. The retained points are then used to "track" back along the energy envelope V until a change in trend is discovered; this marks the start of an S1 component and thus the start of a cycle, the cycle ends with the start of the following S1 component. In this manner a set of cycles (heartbeats), $H = \{\langle h_1, c_1\rangle, \langle h_2, c_2\rangle, \dots\}$, for the given collection of time series (PCGs) T, is obtained. Note that each heartbeat h_j has a class-label c_j associated with it where c_j is taken from a set of class-labels C.

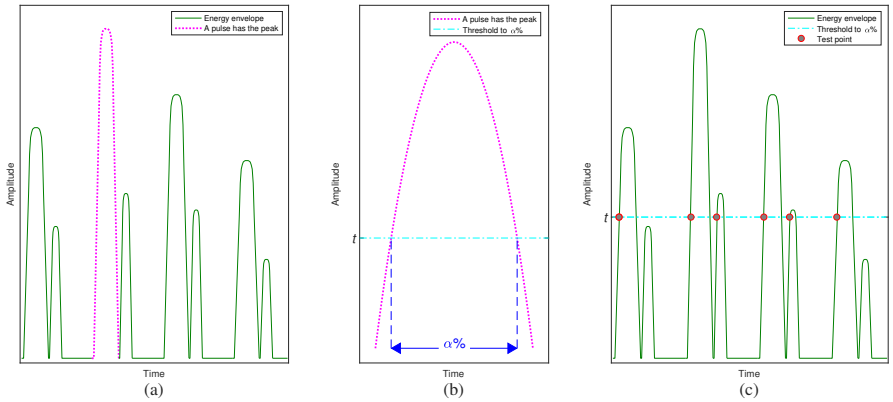


Fig. 1. Dynamic t value calculation using a PCG envelope signal: (a) example PCG envelope with the highest amplitude oscillation highlighted, (b) t value calculation and (c) intersect points.

3.2 Candidate Motif Selection

In Stage 2, the collection of heartbeats (cycles) H, generated during Stage 1, each with an associated class-label, is pruned by removing cycles that are infrequent so as to retain a set of candidate frequent cycles, H'. The assumption was that frequent cycles were likely to be better indicators of class than infrequent cycles. To find cycle frequency, a novel mechanism was adopted whereby a hypothetical cycle, r, referred to as the "zero motif", holding only zero values was used, $r = \{r_j : r_j = 0, j = 1\ to\ j = |h|\ \forall h \in H\}$. The similarity between each heartbeat $h_i \in H$, and r was calculated using a Euclidean Distance similarity function (Equation 3). However, given that r is a vector of zeros, the similarity function could be simplified to give Equation 4. Since the length ($|h_i|$) of each cycle h_i is not fixed, the similarity value was normalised by dividing it by cycle length (Equation 5).

$$d(r, h_i) = \sqrt{\sum_{j=1}^{j=\omega} (r_j - h_{i_j})^2} \quad (3) \qquad d_r(h_i) = \sqrt{\sum_{j=1}^{j=\omega} h_{i_j}^2} \quad (4)$$

$$d_r(h_i) = \frac{\sqrt{\sum_{j=1}^{j=\omega} h_{i_j}^2}}{|h_i|} \quad (5)$$

The obtained similarity values were used to define a Gaussian distribution, with bins holding similarity values arranged along the x-axis, for which the mean (μ_d) and sigma (σ_d) values were calculated. The cycles associated with bins falling within a given number of standard deviations, defined by a parameter ζ, were then retained to give a set H' ($H' \subset H$).

3.3 Frequent Motif Extraction

The third and final stage in the proposed process, given a set of candidate frequent cycles H', is to identify the most frequent cycles (motifs) that are deemed to be the best discriminators and store these in a set H''. Frequent motifs were defined using a threshold σ; if the frequency count of a cycle was greater than σ, the cycle was considered to be frequent and we have a candidate motif. Preliminary experiments, not reported here, indicated that the number of remaining cycles in H' could still be large and that not necessarily all of them would be good discriminators of class. An optional mechanism for limiting the number of candidate frequent cycles to be considered was thus introduced using a parameter max, whereby the max most frequent candidates from the set H' was chosen. If this option was not chosen, all the frequent candidates in the set H' would be considered. We distinguish the two approaches as the Max and All approaches respectively.

The cycles associated with each class was processed in turn by creating a subset A from H' comprised of cycles that belong to c_i. Next, the frequency count f_i for each cycle $h_j \in A$ was determined using Euclidean Distance as the similarity measure and a threshold λ to define whether two cycles were similar or not; if the Euclidean distance between two cycles was less than λ, the two cycles were deemed to be similar. In this manner, a frequency count for each $h_i \in A$ was obtained. In each case, if the count was less than $\sigma\%$ of $|A|$, the cycle was removed from the subset A. Note that if a high σ threshold value is used, the set A may become empty, thus the value for σ must be selected appropriately. The set A is then ordered according to the frequency count. If the Max approach has been adopted only the max most frequent cycles in A are retained, otherwise all the cycles in A are retained. Whatever the case, the next step was to select the k most "discriminative" motifs from the set A. The most discriminative motifs are considered to be the most frequently occurring cycles which are associated with only one class (there are no similar cycles associated with other classes). It is then necessary to compare each cycle in A with all other cycles in H' (similarity is measured in the same way as before using Euclidean Distance and the λ threshold). The identified discriminative cycles are stored in the set H''. This set can then be used as the "data bank" in a Nearest Neighbour Classification (NNC) model.

4 Evaluation

This section presents the evaluation of the proposed mechanism. For the evaluation, a dataset of canine PCGs was used; this is described in Subsection 4.1. Subsection 4.2 presents the experimental set-up in terms of the parameters used. The following three subsections, Subsections 4.3 to 4.5, report on experiments designed to evaluate the operation of the cycle segmentation subprocess, the candidate motif selection subprocess and the frequent motif extraction subprocess respectively. Subsections 4.6 and 4.7 consider the runtime and accuracy of the proposed mechanism in comparison with two competitor approaches.

4.1 Evaluation Data

The dataset used for the evaluation was a set of 59 PCGs, encapsulated as WAVE files, collected using an electronic stethoscope, from animals with and without Mitral Valve disease. The average length of a single (PCG) point series was approximately $800K$ points. Each point series had a class-label associated with it selected from the class attribute set $\{B_1, B_2, C, Control\}$. The first three class attributes represented the three stages of Mitral Valve disease according to the European College of Veterinary Internal Medicine (ECVIM) classification [2,23]. The last class attribute was the control class, no disease.

4.2 Experimental Set-up

Recall that the proposed mechanism required six parameters: λ, σ, ζ, max, k and α. The selected values for these parameters all affect the number of frequent motifs identified and consequently the quality of any further utilisation of the motifs. Clearly, the higher the σ frequency threshold value, the fewer motifs that would be identified because the criteria for frequency would become stricter as σ increased. Inversely, the higher the similarity λ threshold value, the greater the number of motifs that would be identified because the criteria for similarity would become less strict as λ increased. As the value for ζ is increased, the number of selected motifs would also increase, but the average frequency of occurrence would decrease. The values for max and k would also affect the number of identified candidate frequent motifs and, it was conjectured, would thus also influence the number of frequent motifs eventually selected. For the experiments, ranges of values for λ and σ were used, $\{17e5, 91e5, 164e5, 238e5\}$ and $\{0.1, 1\}$ respectively. Similarly, a range of three values was used for both ζ and k, $\{1, 2, 3\}$ and $\{10, 20, 30\}$ respectively. The value for max was set to 60 although any value greater than k could have been used. The α parameter, the oscillation-width threshold to decide where the t cut-off was located, was fixed at 70%; this was the value was suggested in [12].

4.3 Cycle Segmentation Subprocess Evaluation

As noted earlier, the proposed cycle segmentation subprocess used a dynamic cut-off value t, computed using a user-specified α threshold that expressed a percentage-width

of the oscillation with the highest amplitude in a given time series. A method also adopted in [12] where $\alpha = 70$ was suggested to detect the S1 and S2 PCG components. The focus with respect to this paper was to detect the start of cycles, the start of the S1 component, therefore $\alpha = 70$ was used to detect all S1 components (and some S2 components which were discarded later). An example of the results obtained is given in Figure 2 using a fragment of one of the evaluation PCG time series. In Figure 2(a), the envelope signal is given for the raw signal given in Figure 2(b). From the figure, it can be seen that all S1s are identified (and in this case no S2s because these are all below the "cut-off" line) but no noise points. Using $\alpha = 70$, applied to the evaluation dataset, resulted in the identification of 2139 cardiac cycles, an average of 36.25 cycles per PCG (time series), stored in the set H.

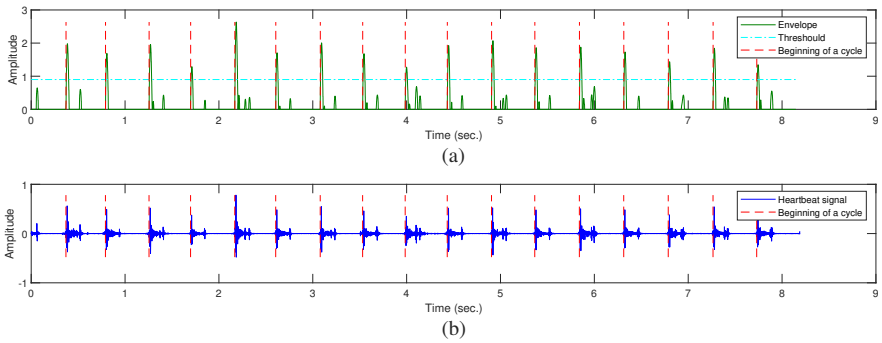


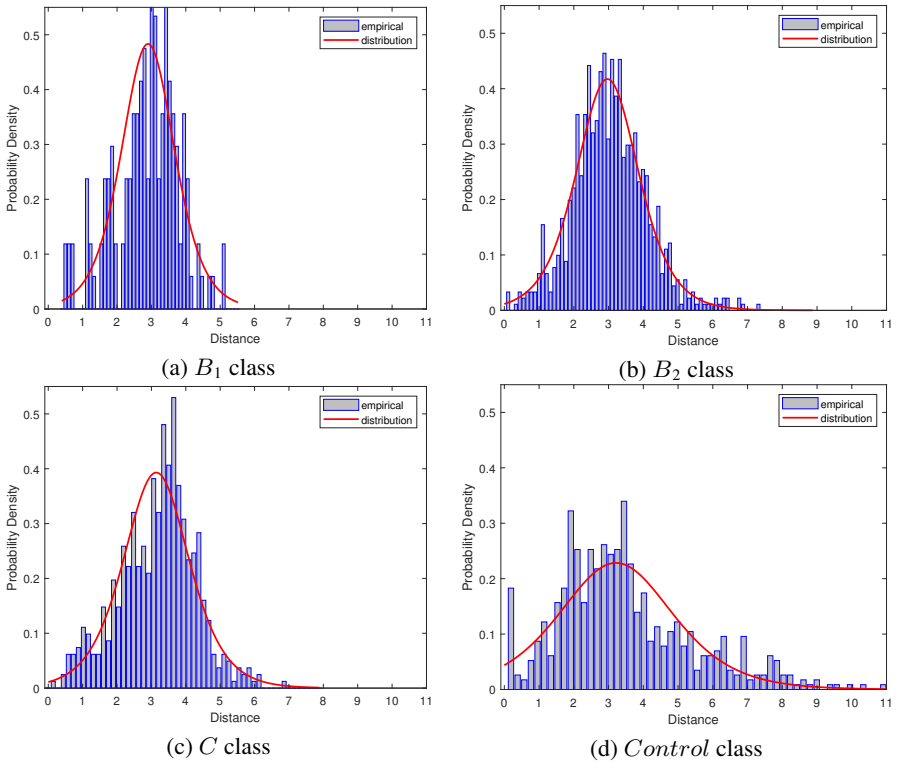
Fig. 2. Extraction of cardiac cycles from the signal energy envelope; (a) envelope PCG signal with "cut-off" line, and (b) raw PCG signal with "start points".

4.4 Candidate Motif Selection Subprocess Evaluation

The Candidate Frequent Cycle Selection subprocess was used to identify and select the most frequent cycles in H, and prune the remainder. The proposed method involved determining the Gaussian distribution of the similarity values (distances), calculated using a novel zero motif approach, and then selecting those that were within a number of standard deviations as prescribed by the user-supplied ζ threshold. The effectiveness is illustrated in Figure 3 which shows the distribution of distances for each class in the evaluation dataset. A normal distribution (bell-shape) curve can be fitted to these distributions (the red line in the figure). The "68.3-95.5-99.7 empirical rule" was adopted, which assumes that 68.27%, 95.45% and 99.73% of the distances fall with 1, 2 and 3 standard deviations respectively from the mean (μ_d), to select frequent cycles and store them in a set H'.

4.5 Frequent Motif Extraction Subprocess Evaluation

During the Frequent Motif Extraction subprocess, cycles were ordered according to their frequency within the set H' and then either: (i) a given number (max) of the most



(a) B_1 class (b) B_2 class

(c) C class (d) $Control$ class

Fig. 3. The best-fit distribution curve for distance similarity values for each class.

frequent cycles were selected or (ii) all cycles were considered. The retained cycles were then processed to determine discriminating cycles, the motifs. The k most frequent best discriminating cycles were then retained, these were then the set of motifs to be used for classification purposes. Evaluation of the process indicated that using the Max approach, some classes had no motifs associated with them at all. A modification was therefore made to the Max approach to ensure that each class had at least one motif associated with it. This change solved the problem.

4.6 Runtime Evaluation

To determine the runtime complexity, 18 sets of experiments were conducted using $\zeta = \{1, 2, 3\}$, $k = \{10, 20, 30\}$ and either the All or Max approach. Experiments, not reported here, were also conducted using a range of λ and σ values, but it was found that this did not affect the runtime, so the results are presented here in the context of the ζ and k parameters, and the approach used. The results are presented in Table 1, these are average runtimes obtained using five evaluation runs. In the table, runtimes are presented for: (i) Cycle Segmentation, (ii) Candidate Motif Selection and (iii) Frequent Motif Extraction. The last column in the table presents the average runtime to process

a single PCG time series (the sum of the values in the previous three columns divided by 59, the number of records in the test dataset).

Table 1. Runtime for PCG Frequent Motif Selection and Extraction (seconds).

ζ	k	Cycle Extraction	Candidate Motif Selection	Frequent Motif Extraction		Average	
				All	Max	All	Max
1	10	15.97	184.03	10.55	15.10	3.57	3.57
	20			11.67	11.40	3.59	3.58
	30			12.43	11.12	3.60	3.58
2	10		184.05	12.10	29.94	3.60	3.90
	20			13.94	24.58	3.63	3.81
	30			15.29	23.78	3.65	3.65
3	10		184.06	11.87	35.77	3.59	4.00
	20			14.02	31.34	3.63	3.92
	30			15.67	30.34	3.66	3.90

From the table, it can be seen that the difference between the runtimes, using different ζ values, when selecting candidate frequent cycles, was negligible. As anticipated, the larger the k value, the more runtime that was required to discover the frequent motifs using the All approach and the less runtime that was required to discover the frequent motifs using the Max approach. The reason for the difference in runtime between the All and Max approaches was unclear: most of the Max experiments required a longer runtime, however two of them ($\zeta = 1$ with $k = 20$ and $\zeta = 1$ with $k = 30$) featured a runtime that was less than the All approach. As also anticipated, when using the Max approach the runtime increased as ζ increased because a larger number of cycles required processing. However, the total runtime required for a single time series to be processed, on average, was similar in all cases.

The average runtime for the combination of parameters that gave the best accuracy (accuracy is discussed in further detail in the next subsection) was 3.65 sec/record which is much faster compared with the motif discovery mechanisms and algorithms reported in [1] and [22] where best accuracy runtimes of 700.20 sec/record and 4500.00 sec/record were recorded respectively. Note that the proposed algorithm, and the two comparator algorithms, were implemented using the Java programming language and run on an iMac Pro (2017) computer with 8-Cores, 3.2GHz Intel Xeon W CPU and 19MB RAM.

4.7 Classification Accuracy

The experimental results presented in the previous subsection demonstrated that the proposed process speeded up the runtime compared with the comparator mechanisms considered. However, for this speed up to be of value, the accuracy should not be adversely affected. The experiments reported on in this subsection sought to investigate this. The parameter settings used were as follows:

- $\zeta = \{1, 2, 3\}$.
- $k = \{10, 20, 30\}$.
- $max = 60$.
- $\langle \lambda, \sigma \rangle = \{\langle 17e5, 0.1\rangle, \langle 91e5, 0.1\rangle, \langle 164e5, 0.1\rangle, \langle 238e5, 0.1\rangle, \langle 238e5, 1\rangle\}$.
- Discrimination approach = { All , Max }.

The adopted process for classifying previously unseen cycles (motifs) was the well-known Nearest Neighbour Classification (NNC) method [6], because this was frequently used in the context of time/point series analysis [26,31]. For the experiments, $k_{NNC} = 1$ and $k_{NNC} = 3$ were used. The dataset was divided into training and testing subsets and five cross validation was applied. The idea was that the accuracy of the classification would provide an indicator of the quality of the proposed approach; the metric used were accuracy (Acc.).

Given that each query PCG to be labelled comprised a number of cycles, each of which would be labelled separately, there was a chance that more than one class-label would be associated with the query PCG. To select the most "appropriate" class-label, three different methods were considered: (i) Shortest Distance (SD), (ii) Shortest Total Distances (STD) and (iii) Highest Votes (HV). The SD method simply chooses the class-label associated with the most similar motif. The STD method chooses the class-label associated with the lowest accumulated distance. The HV method chooses the most frequently occurring class-label. In each case, if more than one class-label was nominated, one of the other class selection methods was applied.

Analysis of the results indicated some interesting patterns. It was found that $\zeta = 2$ usually produces the best accuracy regardless of the k_{NNC} value, approach or classification method used. The best results, with regard to the $\langle \lambda, \sigma \rangle$ combinations considered, are presented in Table 2. The table includes the average runtimes recorded (secs). The best obtained accuracy was 72.0% when $\langle \lambda, \sigma \rangle = \langle 164e5, 0.1\rangle$, the Max approach, the HV classification method, $\zeta = 2$, $k = 30$ and $k_{NNC} = 3$; a runtime of only 3.65 sec/record was recorded. Comparing this best recorded accuracy with that obtained using the comparator mechanism described in [1], and the motif discovery approach presented in [22], it was found that the same level of accuracy was obtained but much more efficiently.

Table 2. The best classification accuracy results.

$\langle \lambda, \sigma \rangle$		Parameters					Results	
λ	σ	Discrim. Approach	Class. Method	ζ	k	k_{NNC}	Acc.	Runtime (Sec.)
17e5	0.1	All	HV	3	10	1	0.667	3.59
91e5	0.1	Max	HV	3	20	3	0.704	3.92
164e5	0.1	Max	HV	2	30	3	0.720	3.65
238e5	0.1	Max	HV	2	30	3	0.695	3.65
238e5	1	Max	HV	2	30	3	0.695	3.65

5 Conclusions

An approach to PCG classification, using the concept of Motifs has been described. The proposed process addresses the challenge of finding discriminative motifs in long time series using three pipelined mechanisms: (i) cycle segmentation and (ii) candidate motif selection and (iii) frequent motif extraction. The first mechanism, has been relatively well studied, but as a means of analysing PCG cycles and not as precursor to motif discovery. The second mechanism featured a novel approach, that did not require every candidate frequent subsequence to be compared to every other subsequence, to prune time series subsequences (cycles) that could not be considered to be frequent. The third involved the extraction of motifs that were good discriminators of class from the retained candidate frequent motifs. The performance of the proposed approach was analysed in terms of runtime and the quality of the identified motifs in the context of a classification scenario, with respect to two comparator algorithms. The results obtained demonstrated a similar classification accuracy, but a significant runtime improvement.

References

1. Alhijailan, H., Coenen, F., Dukes-McEwan, J., Thiyagalingam, J.: Segmenting sound waves to support phonocardiogram analysis: The pcgseg approach. In: Geng, X., Kang, B.H. (eds.) PRICAI 2018: Trends in Artificial Intelligence. pp. 100–112. Springer International Publishing, Cham (2018)
2. Atkins, C., Bonagura, J., Ettinger, S., Fox, P., Gordon, S., Haggstrom, J., Hamlin, R., Keene, B., Luis-Fuentes, V., Stepien, R.: Guidelines for the diagnosis and treatment of canine chronic valvular heart disease. Journal of Veterinary Internal Medicine **23**(6), 1142–1150 (2009). https://doi.org/10.1111/j.1939-1676.2009.0392.x
3. Bagnall, A., Hills, J., Lines, J.: Finding motif sets in time series. CoRR (07 2014)
4. Cherif, L.H., Debba, S.: Variability of pulmonary blood pressure, splitting of the second heart sound and heart rate. Journal of Clinical & Experimental Cardiology **8**(10), 1–3 (2017). https://doi.org/10.4172/2155-9880.1000550
5. Chiu, B., Keogh, E., Lonardi, S.: Probabilistic discovery of time series motifs. In: Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 493–498. KDD '03, ACM, New York, NY, USA (2003). https://doi.org/10.1145/956750.956808
6. Dasarathy, B.V.: Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques. IEEE Computer Society Press tutorial, IEEE Computer Society Press (1991), the University of Michigan
7. Dau, H.A., Keogh, E.: Matrix profile v: A generic technique to incorporate domain knowledge into motif discovery. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 125–134. KDD '17, ACM, New York, NY, USA (2017). https://doi.org/10.1145/3097983.3097993
8. Delgado-Trejos, E., Quiceno-Manrique, A., Godino-Llorente, J., Blanco-Velasco, M., Castellanos-Dominguez, G.: Digital auscultation analysis for heart murmur detection. Annals of Biomedical Engineering **37**(2), 337–353 (Feb 2009). https://doi.org/10.1007/s10439-008-9611-z
9. Ding, H., Trajcevski, G., Scheuermann, P., Wang, X., Keogh, E.: Querying and mining of time series data: Experimental comparison of representations and distance measures. Proc. VLDB Endow. **1**(2), 1542–1552 (Aug 2008). https://doi.org/10.14778/1454159.1454226

10. Gao, Y., Lin, J., Rangwala, H.: Iterative grammar-based framework for discovering variable-length time series motifs. In: IEEE International Conference on Data Mining. pp. 111–116. IEEE (11 2017). https://doi.org/10.1109/ICDM.2017.20
11. Guhneuc, Y.G., Antoniol, G.: Demima: A multilayered approach for design pattern identification. IEEE Transactions on Software Engineering **34**(5), 667–684 (Sept 2008). https://doi.org/10.1109/TSE.2008.48
12. Hamza Cherif, L., Debbal, S.M., Bereksi-Reguig, F.: Segmentation of heart sounds and heart murmurs. Journal of Mechanics in Medicine and Biology **8**(4), 549–559 (2008). https://doi.org/10.1142/S0219519408002759
13. Hannan, E.J.: Time series analysis. Chapman and Hall (1960)
14. Hutchins, L.N., Murphy, S.M., Singh, P., Graber, J.H.: Position-dependent motif characterization using non-negative matrix factorization. Bioinformatics **24**(23), 2684–2690 (2008). https://doi.org/10.1093/bioinformatics/btn526
15. Krejci, A., Hupp, T.R., Lexa, M., Vojtesek, B., Muller, P.: Hammock: a hidden markov model-based peptide clustering algorithm to identify protein-interaction consensus motifs in large datasets. Bioinformatics **32**(1), 9–16 (Jan 2016). https://doi.org/10.1093/bioinformatics/btv522
16. Lehner, R.J., Rangayyan, R.M.: A three-channel microcomputer system for segmentation and characterization of the phonocardiogram. IEEE Transactions on Biomedical Engineering **34**(6), 485–489 (June 1987). https://doi.org/10.1109/TBME.1987.326060
17. Li, N., Crane, M., Gurrin, C., Ruskin, H.J.: Finding motifs in large personal lifelogs. In: Proceedings of the 7th Augmented Human International Conference 2016. pp. 1–8. ACM, New York, NY, USA (2016). https://doi.org/10.1145/2875194.2875214
18. Lin, J., Keogh, E., Lonardi, S., Patel, P.: Finding motifs in time series. In: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 53–68 (2002)
19. Lin, J., Keogh, E., Wei, L., Lonardi, S.: Experiencing sax: a novel symbolic representation of time series. Data Mining and Knowledge Discovery **15**(2), 107–144 (Oct 2007). https://doi.org/10.1007/s10618-007-0064-z
20. Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., Alon, U.: Network motifs: Simple building blocks of complex networks. Science **298**(5594), 824–827 (2002). https://doi.org/10.1126/science.298.5594.824
21. Mubarak, Q.u.a., Akram, M.U., Shaukat, A., Ramazan, A.: Quality Assessment and Classification of Heart Sounds Using PCG Signals, pp. 1–11. Springer International Publishing, Cham (2019). https://doi.org/10.1007/978-3-319-96139-2_1
22. Mueen, A., Keogh, E., Zhu, Q., Cash, S., Westover, B.: Exact discovery of time series motifs. In: Proceedings of the 2009 SIAM International Conference on Data Mining. pp. 473–484 (2009). https://doi.org/10.1137/1.9781611972795.41
23. Nakamura, K., Kawamoto, S., Osuga, T., Morita, T., Sasaki, N., Morishita, K., Ohta, H., Takiguchi, M.: Left atrial strain at different stages of myxomatous mitral valve disease in dogs. Journal of veterinary internal medicine **31**(2), 316–325 (2017). https://doi.org/10.1111/jvim.14660
24. Oliveira, J., Sousa, C., Coimbra, M.: Coupled hidden markov model for automatic ecg and pcg segmentation. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1023–1027 (March 2017). https://doi.org/10.1109/ICASSP.2017.7952311
25. Ramli, D., Hooi, M., Chee, K.: Development of heartbeat detection kit for biometric authentication system. Procedia Computer Science **96**, 305 – 314 (2016). https://doi.org/10.1016/j.procs.2016.08.143

26. Stojanovi, M.B., Boi, M.M., Stankovi, M.M., Staji, Z.P.: A methodology for training set instance selection using mutual information in time series prediction. Neurocomputing **141**(Supplement C), 236 – 245 (2014). https://doi.org/10.1016/j.neucom.2014.03.006
27. Thijs, G., Marchal, K., Lescot, M., Rombauts, S., De Moor, B., Rouz, P., Moreau, Y.: A gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes. Journal of Computational Biology **9**(2), 447–464 (04 2004). https://doi.org/10.1089/10665270252935566
28. Torkamani, S., Lohweg, V.: Survey on time series motif discovery. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **7**(2), 1–8 (2017). https://doi.org/10.1002/widm.1199
29. Vahdatpour, A., Amini, N., Sarrafzadeh, M.: Toward unsupervised activity discovery using multi-dimensional motif detection in time series. In: Proceedings of the 21st International Jont Conference on Artifical Intelligence. pp. 1261–1266. IJCAI'09, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (2009)
30. Vaswani, A., Khaw, H.J., Dougherty, S., Zamvar, V., Lang, C.: Cardiology in a Heartbeat. Scion Publishing Limited (2015)
31. Wang, X., Fang, Z., Wang, P., Zhu, R., Wang, W.: A Distributed Multi-level Composite Index for KNN Processing on Long Time Series, pp. 215–230. Springer International Publishing, Cham (2017). https://doi.org/10.1007/978-3-319-55753-3_14
32. Zhao, Y., Xu1, D., Xiao, S., Yan, X., Liu, J., Liu, Y., Luo, L., Xia, G.: Measurement of two new indicators of cardiac reserve in humans, rats, rabbits, and dogs. Journal of Biomedical Science and Engineering **6**(10), 960–963 (10 2013). https://doi.org/10.4236/jbise.2013.610118