SHAPE-BASED TIME SERIES MINING FOR PROCESS MONITORING AND ANOMALY DETECTION

by

Li Zhang

DISSERTATION

Submitted in partial fulfillment of the requirements

for the degree of Doctor of Philosophy at

The University of Texas at Arlington

August, 2021

Arlington, Texas

Copyright by

Li Zhang

2021

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

SHAPE-BASED TIME SERIES MINING FOR PROCESS MONITORING AND ANOMALY DETECTION

Li Zhang, Ph.D.

The University of Texas at Arlington, 2021

Supervising Professors: Chen Kan, Victoria C. P. Chen

Due to the rapid development of computing and sensing technology, Internet of Things (IoT)-enabled monitoring plays a crucial role for people suffering from cardiac problems. It is important to detect the abnormal ECG cycles during the cardiac monitoring for the early treatment. However, most existing methods focused on the full reading of time series, for the cycle-based time series, it is wasting time to read the whole time series while we can find the characteristic patterns instead. Characteristic patterns named shapelets are time series subsequences, which are explainable and discriminative features that can best classify time series. Shapelet-based classification that uses the similarity between a shapelet and a time series has been widely used recently in many applications. In this research, we extract the statistically significant shapelets from the cycle-based ECG data, and apply the support vector data description (SVDD) algorithm to statistical process control problem for the cardiac monitoring. The experimental results on the real-world MIT-BIH dataset demonstrate the effectiveness of proposed method.

Positive and unlabeled learning has attracted increasing interests in recent years. The setting of the positive and unlabeled learning is that we only

access the positive and unlabeled training data sets. Many methods have been proposed for the positive and unlabeled learning, however, only a few papers integrate the shapelet features into the positive and unlabeled learning. In this paper, we proposed the positive and unlabeled shapelet learning model for the time series classification, and the experiment results from the real-world data sets demonstrate the effectiveness of our proposed method.

TABLE OF CONTENTS

CHAPTER 1

Introduction

1.1 Shapelet-based ECG anomaly detection

According to the American Heart Association 2021 report, cardiovascular diseases (CVDs) remain the No. 1 cause of death in the US [1]. CVDs are a class of diseases involving the heart or blood vessels such as heart attack, stroke, congestive heart failure and other conditions. It is important to monitor the patients' heart conditions as early as possible so that the domain experts can take the timely interventions [2, 3]. An electrocardiogram (ECG) records the heart's electrical activity of the patient by electrodes placed on the chests and/or limbs, and it is the most commonly tool used to diagnose the cardiac related diseases by cardiologists [4]. Hence, it is important to detect the abnormal ECG cycles during the cardiac monitoring for the early treatment.

Despite the enormous ECG anomaly detection methods have been proposed, only a few papers use the shapelet as morphology features [5] and statistical control chart to monitor the ECG signal [6]. Shapelets as explainable and discriminative features were introduced in 2009 [7] for time series data mining, and can provide a model with better interpretability. Shapelets are time series sub-sequences, and represent the maximally discriminative segments of time series that split the time series into two classes. Discovering shapelets from time series has been increasing interests for researchers during the past decade. Control chart as the important tool for statistical process control is used to monitor the performance of the process over time and detect anomalies. In this paper, we extracted the statistically significant shapelet from ECG cycles, and then applied the shapelet transformation matrix with the support vector data description (SVDD) algorithm to construct the control chart to detect abnormal ECG cycles for IoT-enabled cardiac monitoring. The different methods with different features are experimentally studied on the MIT-BIH data sets to show the effectiveness of the proposed method.

Our contributions for this research topic are summarized as follows:

- We propose the shapelet-based method for the ECG anomaly detection, and the shapelets are the local features, which can provide better interpretability and robust result than others, which help the domain experts understand the model better behind results.
- Our method discovers the statistically significant shapelets based on the statistical tests that have the explanatory power supported by p-values, while the traditional shapelet discovery method does not have.
- We apply the shapelet-based features with the support vector data description to construct the control chart for the cardiac monitoring, and

the domain experts can easier see the anomaly ECG cycle and take the timely intervention.

1.2 Positive and unlabeled shapelet learning

Time series classification as a subset of the general classification problem, has attracted many interests in the research for both academic and industry people, as the data collected automatically by sensing and monitoring are time series. However, in many real world problems, collecting a large amount of the labeled data is costly, while the positive and unlabeled data are usually easily to be obtained. In such situation, only a small set of positive labeled data and a large amount of unlabeled data are available, which leads to the development of the positive and unlabeled learning [8]. Positive and unlabeled learning aims to learn a suitable binary classifier without the assistant of the negative data.

Despite a large amount of methods have been proposed for the positive and unlabeled learning, few efforts have been made to integrate the shapelets with the positive and unlabeled learning for time series classification [9–11]. Based on the previous algorithm large-margin label-calibrated support vector machines (LLSVM) for the positive and unlabeled learning [12], we integrated the shapelet features into the LLSVM to introduce the positive and unlabeled shapelet learning model. Benchmark time series data sets and MIT-BIH data sets are experimented to validate the proposed method.

Our contributions for this research topic are summarized as follows:

- This is the first effort of shapelet learning that focuses on the positive and unlabeled data setting.
- A new positive and unlabeled shapelet learning model is proposed that incorporates the derived distribution information from the unlabeled data.
- The proposed positive and shapelet learning method is applied for the ECG anomaly detection.

Bibliography

[1] Salim S Virani, Alvaro Alonso, Hugo J Aparicio, Emelia J Benjamin, Marcio S Bittencourt, Clifton W Callaway, April P Carson, Alanna M Chamberlain, Susan Cheng, Francesca N Delling, et al. Heart disease and stroke statistics—2021 update: a report from the american heart association. *Circulation*, 143(8):e254–e743, 2021.

[2] Jessica K Zègre-Hemsey, J Lee Garvey, and Mary G Carey. Cardiac monitoring in the emergency department. *Critical Care Nursing Clinics*, 28(3):331–345, 2016.

[3] Joy Liao, Zahira Khalid, Ciaran Scallan, Carlos Morillo, and Martin O'Donnell. Noninvasive cardiac monitoring for detecting paroxysmal atrial fibrillation or flutter after acute ischemic stroke: a systematic review. *Stroke*, 38(11):2935–2940, 2007.

[4] Robert O Bonow, Douglas L Mann, Douglas P Zipes, and Peter Libby.

Braunwald's heart disease e-book: A textbook of cardiovascular medicine. Elsevier Health Sciences, 2011.

[5] Mohammad Alshaer, Sandra Garcia-Rodriguez, and Cedric Gouy-Pailler. Detecting anomalies from streaming time series using matrix profile and shapelets learning. In *2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 376–383. IEEE, 2020.

[6] Yonghan Jung and Heeyoung Kim. Detection of pvc by using a wavelet-based statistical ecg monitoring procedure. *Biomedical Signal Processing and Control*, 36:176–182, 2017.

[7] Lexiang Ye and Eamonn Keogh. Time series shapelets: a new primitive for data mining. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 947–956, 2009.

[8] Fabien Letouzey, François Denis, and Rémi Gilleron. Learning from positive and unlabeled examples. In *International Conference on Algorithmic Learning Theory*, pages 71–85. Springer, 2000.

[9] Akihiro Yamaguchi and Takeichiro Nishikawa. One-class learning time-series shapelets. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 2365–2372. IEEE, 2018.

[10] Haishuai Wang, Qin Zhang, Jia Wu, Shirui Pan, and Yixin Chen. Time series feature learning with labeled and unlabeled data. *Pattern Recognition*, 89:55–66, 2019.

[11] Shen Liang, Yanchun Zhang, and Jiangang Ma. Pu-shapelets: Towards pattern-based positive unlabeled classification of time series. In *International Conference on Database Systems for Advanced Applications*, pages 87–103. Springer, 2019.

[12] Chen Gong, Tongliang Liu, Jian Yang, and Dacheng Tao. Large-margin label-calibrated support vector machines for positive and unlabeled learning. *IEEE transactions on neural networks and learning systems*, 30(11): 3471–3483, 2019.

CHAPTER 2

SHAPELET-BASED ECG ANOMALY DETECTION FOR IOT-ENABLED CARDIAC MONITORING

Li Zhang, Chen Kan, Victoria C. P. Chen

2.1 Abstract

Due to the rapid development of network and sensing technology, Internet of Things (IoT)-enabled monitoring plays a crucial role for people suffering from cardiac problems. It is essential to detect the abnormal ECG cycles during cardiac monitoring for early treatment. However, most existing methods focused on the entire reading of time series; for the cycle-based time series, it is wasting time to read the whole time series while we can find the characteristic patterns instead. Distinctive patterns named shapelets are time series subsequences, explainable and discriminative features that can best classify time series. The shapelet-based classification that uses the similarity between a shapelet and a time series has been widely used in many applications. In this research, we extract the statistically significant shapelets from the cycle-based ECG data and apply the support vector data description (SVDD) algorithm to the statistical process control problem for cardiac monitoring. The experimental results on the real-world MIT-BIH dataset demonstrate the effectiveness of the proposed method.

2.2 Introduction

According to the American Heart Association 2021 report, cardiovascular diseases (CVDs) remain the No. 1 cause of death in the US [1]. CVDs are diseases involving the heart or blood vessels such as heart attack, stroke, congestive heart failure, and other conditions. It is essential to monitor the patients' heart conditions as early as possible so that the domain experts can take the timely interventions [2, 3]. An electrocardiogram (ECG) records the heart's electrical activity by electrodes placed on the chests and/or limbs, and it is the most common tool used to diagnose the cardiac-related diseases by cardiologists [4]. Therefore, long-term ECG monitoring can track the patient's cardiac activities, and it is valuable and necessary for those who suffer CVDs. As shown in Figure 1, the typical ECG signal has three most significant components: P wave, QRS complex, and T wave. The P wave represents the atria depolarization; the QRS complex represents the ventricles' depolarization, and the T wave indicates the ventricles' repolarization.

Due to the rapid development of computing and sensing technology, the Internet of Things (IoT)-enabled cardiac monitoring devices make it available for people suffering the cardiac-related problems to regularly monitor their health no matter where they are and what they are doing, and for doctors remotely check and analyze the monitoring data. Many portable ECG monitoring systems are proposed as the healthcare-based cardiac monitoring for home monitoring, remote monitoring or diagnosis, and intensive care unit (ICU), for example, Singh et al. [5] presented the wearable smartphone based wireless cardiac activity monitoring sensor, Wan et al. [6] proposed wearable IoT cloud based health monitoring system, Verma and Sood [7] displayed the IoT-based real-time remote ECG monitoring system, and Basu et al. [8] proposed the fog assistant-IoT enabled patient health monitoring in smart homes. Hence, it is necessary to analyze the IoT-enabled cardiac monitoring.

As shown in Figure 2.1, in the normal ECG signal, the P wave, QRS complex, and T wave should be similar over time at a frequency ranging from 60

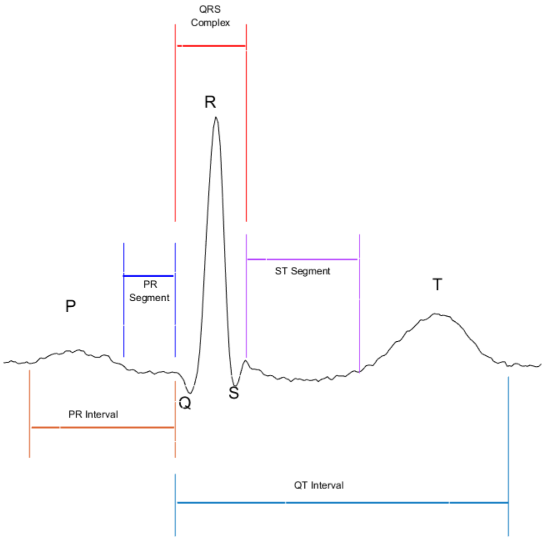


Figure 2.1: A typical ECG signal with the corresponding notations

to 100bpm [9]. Based on the paper [10], anomalies are defined as any deviation from normal behavior, in general. Hence, the ECG anomalies represent the irregular heart activity and anomaly detection for the ECG signals can act as an assistant for doctors to diagnose a cardiac condition [9]. During the past decades, ECG anomaly detection is one of the most popular research topics for these IoT-based healthcare monitoring systems, as it can get rapid response to the acute heart related illnesses and improve the diagnostic efficiency and accuracy [11–13]. Many methods have been proposed to analyze and classify the ECG signals: for example, Risk et al. [14] proposed the self-organizing maps for the beat detection and classification of ECG; Srinivasan et al. [15] used the autoregressive modeling to classify the cardiac arrhythmias; Özbay et al. [16] integrated the c-means clustering and wavelet transform for the ECG classification; and Rai et al. [17] used multi-resolution wavelet transform and artificial neural network for the ECG anomaly detection. Despite the fact that

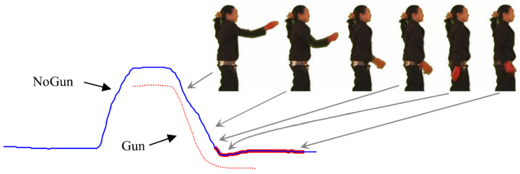


Figure 2.2: Top shapelet selected for GunPoint time series [20]

the enormous ECG anomaly detection methods have been proposed, only a few papers use the shapelet as morphology features [18] and statistical control chart to monitor the ECG signal [19].

Shapelets as explainable and discriminative features were introduced in 2009 [20] for time series data mining and can provide a model with better interpretability. Shapelets are time series sub-sequences and represent the maximally discriminative segments of time series that split the time series into two classes. For example, GunPoint time series (available in UCR time series repository [21]), which has been studied a lot in the time series classification [20, 22–24], as shown in Figure 2.2, the solid red line represents one of the best-selected shapelets. GunPoint time series has two classes: Gun-Draw and Point: for the Gun-Draw, the actor took the gun from the hip-mounted holster to point at a target; for the Point, the actor did the same action with index finger instead of a gun to point a target. The shapelet shown in Figure 2.2 displayed the "dip" without the gun, which means that the actor is trying to correct the action, while the actor is more careful to return the hand with the gun. The shapelet captures the inherent characteristic of the time series and provides an interpretation for the people who do not understand the inside processes or algorithms.

Discovering shapelets from time series has been increasing interest for researchers during the past decade [20, 25–29]. Xing et al. [26] extracted the local shapelets for the early time series classification. Lines et al. [30] proposed the shapelet transform for the time series classification so that the feature vectors can be used with any classifier. Ghalwash and Obradovic [31] extracted the interpretable shapelets for the multivariate early time series classification. Li et al. [32] proposed the shapelet-neural network approach for the multivariate time series classification.

As an essential tool for statistical process control, the control chart is used to monitor the performance of the process over time and detect anomalies [33]. For example, Sun and Tsung [34] proposed the kernel-distance-based charts (K charts) based on a support vector data description (SVDD) algorithm using the monitoring statistic derived from the distance between the new observation and the decision boundary and showed that the proposed method is better than Hotelling's T^2 control chart when the data is non-normal data. Later, Kumar et al. [35] proposed the robust K charts and revealed that the method efficiently handled the autocorrelated process data. Moreover, Zhang et al. [36] applied the one-class SVM-based control chart for anomaly detection in computer networking applications. Moreover, for healthcare, Jung and Kim [19] applied Hotelling's T^2 control chart for the PVC detection in the ECG signal.

However, previous paper [18] proposed the unsupervised approach to learn the shapelets from the same streaming ECG time series, and they did not consider ECG samples from more patients while we can get them easy right now, and also only a few research did the statistical tests to select the significant shapelets for the ECG time series [37]. In this paper, we extracted the statistically significant shapelet from each training model, including more different patients, and then applied the shapelet transformation matrix with the support vector data description (SVDD) algorithm to construct the control chart to detect abnormal ECG cycles for IoT-enabled cardiac monitoring.

The highlights of this paper are summarized as follows:

- We propose the shapelet-based method for the ECG anomaly detection, and the shapelets are the local features, which can provide better interpretability and robust result than others, which help the domain experts understand the model better behind results.
- Our method discovers the statistically significant shapelets based on the statistical tests that have the explanatory power supported by p-values, while the traditional shapelet discovery method does not have.
- We apply the shapelet-based features with the support vector data description to construct the control chart for the cardiac monitoring, and the domain experts can easier see the anomaly ECG cycle and take the timely intervention.

The remainder of the paper is organized as follows. Section 2.3 reviews the ECG anomaly detection methods. We propose the methods for extracting the statistically significant shapelets to construct the control chart in section 2.4. Section 2.5 shows the experimental study and results. Section 2.6 concludes the paper and points to future work.

2.3 Literature review

In this section, we review the existing research related to our work. ECG anomaly detection has two main parts: feature extraction and model training.

For the feature extraction, many morphological features or derived features have been studied in the past research. For example, the previous paper extracted the representation features mean, and trend [38], the paper [39] extracted the QRS complex as the morphology feature, and the paper [40] used the Dynamic Time Warping(DTW) distance between a cycle segmentation and the median cycle segmentation as the features. Most of the previous research combined the morphological and derived features together to get the higher classification accuracy, for example, Ye et al. [41] proposed the average R-R interval morphology and dynamic features for the anomaly detection. However, only a few papers used the shapelets as the feature for the ECG anomaly detection [18]. The previous paper [18] used the shapelet to detect the anomalies from the steaming ECG signal, but that method is the unsupervised learning without the prior labels for each ECG cycle, which is a suitable method for the unlabeled data. While for the ECG signal, there are many cardiologists that have annotated the ECG cycles, we can make full use of the annotations to get a better prediction. To solve this problem, we proposed the shapelet-based ECG anomaly detection features. The previous research that is most similar to our method is the motif discovery anomaly detection proposed by Sivaraks and Ratanamahatana [42]. While the motif method was focused on the clean normal ECG cycles without the abnormal cycles, the motif best represented the normal cycles. However, the difference between the normal and abnormal classes may lie in other places, and the two classes share the same motif. Shapelets that represents the maximally discriminative segments of time series that split the time series into two classes is the best way to make up this shortage for the motif discovery method.

Based on the models used for training, there are clustering, traditional machine learning classification, and deep learning classification for the ECG anomaly detection [9]. For example, Veeravalli et al. [43] proposed the K-means clustering-based algorithm for real-time and personalized anomaly detection from wearable health care; Li et al. [44] presented the weighted transductive one-class SVM, and Li et al. [45] proposed the convolutional neural network-based classification methods. Although much research has been proposed for all kinds of training methods, only a few papers integrated the classification with statistical process control (SPC). Jung and Kim [19] showed the PVC detection based on the SPC, while the monitoring statistics constructed are based on Hotelling's T^2 statistics, which is good at the normal distribution assumption [33], and difficult to be applied to non-normal data. Kernel-distance-based charts (K charts) based on the support vector data description (SVDD) algorithm proposed by Sun and Tsung [34] is a suitable substitution. After the feature extraction, we constructed the monitoring procedure using the K charted-based SVDD algorithm.

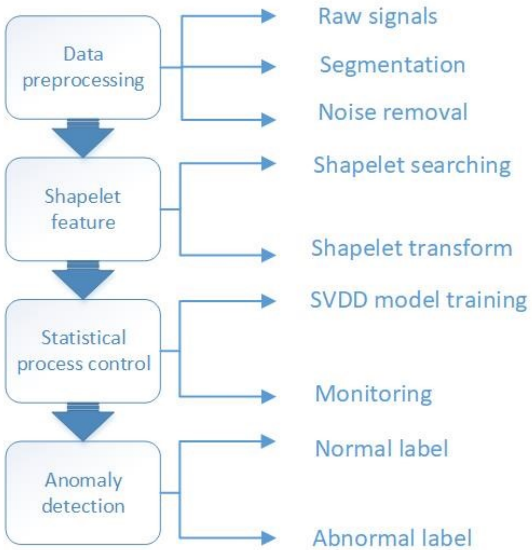


Figure 2.3: Flow chart of the proposed method: after the input of the ECG signal, the first step is to segment the signal into cycles, then remove noise with the Daubechies wavelet transform, and then extract features based on the wavelet coefficients, the last step is to apply the one class SVDD algorithm with the features to detect the abnormal ECG cycles

2.4 Research methodology

In this section, we propose the shapelet-based ECG anomaly detection method for cardiac monitoring. Figures 2.3 shows the overall view of the proposed method. Follow the flow chart of Figures 2.3, we will first present the data pre-processing, as it is an essential part of ECG anomaly detection.

2.4.1 Data pre-processing

The ECG signal monitored from the IoT-based devices is the raw ECG signal, and it may include the noise because of the non-stationary characteristic of the ECG recording. We apply the discrete wavelet transform [46] for every cycle segmentation. Since the Daubechies wavelet family is similar to the ECG recording [47], we selected it as the transformation. Low-order Daubechies wavelets have high time resolution but low-frequency resolution, while high-order ones have high-frequency resolution and low time resolution. As the previous paper [48] did, we employ order 2, but with 2 levels of decomposition instead of 4, which is more close to the original recording. After the discrete wavelet transform, i.e., db2(2), the final cycle segmentations are used to extract the shapelets.

2.4.2 Feature selection

The proposed method for feature selection has two steps. The first step is to search the shapelets from the final cycle segmentations and select statistically significant shapelets. The second step is to do the shapelet transform.

Shapelet searching

We now introduce the process of the shapelet searching. For each ECG cycle segmentation, it is a univariate time series T with a set of ordered numerical observations $t_1, t_2, \ldots, t_m$. A subsequence S of T is $s_i, s_{i+1}, \ldots, s_{i+l-1}$ with length $l \leq m$, $1 \leq l \leq m - l + 1$. The distance between subsequence S and time series T is calculated by Euclidean Norm, $dist(S, T) = \sqrt{\sum_{i=1}^{m}(s_i - t_i)^2}$ if subsequence S and T have the same length. If they have different lengths, and $|S| < |T|$, let S_i be the subsequence of T with $|S_i| = |S|$, $dist(S, T) = min(dist(S, S_i))$. The shapelet is defined as the most distinctive time series subsequence. Therefore, time series subsequence S is the shapelet candidate.

First, we will extract all of the subsequences with the length between the minimum length and maximum length as the shapelet candidates. For each shapelet candidate S, we calculate the minimum Euclidean distance between this candidate and each time series in the training dataset as the similarity $similarity(S) = min(dist(S, T))$, and store this similarity with the accordance time series label. To find the best threshold for the given shapelet, we compare the threshold candidates with the similarity to divide the training dataset into two groups for which the information gain is calculated. The threshold candidate with the maximum information gain is the best distance threshold for this given shapelet candidate.

Next, we need to assess if this shapelet candidate is significant by the statistical indicator p-value which can be obtained by the statistical association tests. For each shapelet candidate, there is the best distance threshold θ which is used to construct the contingency table between two groups as shown in Table 2.1, where $dist(S, T)$ is the minimum distance between the shapelet S and the time series T. Both of the Fisher's exact test [49] and Pearson's χ^2 test [50] can be used for the test for the contingency table. Since χ^2 test is more appropriate for the larger sample sizes, we will use the χ^2 test here to calculate the p-value based on the contingency table. The χ^2 statistic is defined as $\chi_c^2 = \frac{N(n_1n_4-n_2n_3)^2}{(n_1+n_2)(n_3+n_4)(n_1+n_3)(n_2+n_4)}$.

This shapelet candidate is significant if the accordance p-value from the statistical test is less than a certain significance level α. Since the number of the shapelet candidates is very large, it is likely to have a large number of false positives which means that the shapelet candidate will be mistak-

Table 2.1: Contingency table between two classes

Class label	$dist(S,T) \leq \theta$	$dist(S,T) > \theta$	Total
normal	n_1	n_2	n_{r1}
abnormal	n_3	n_4	n_{r2}
Total	n_{c1}	n_{c2}	N

enly considered statistically significant. This is also known as the multiple hypothesis testing problem. To solve this problem, the Bonferroni correction factor is introduced, which is the number of the statistical tests we have for all of the shapelet candidates. As stated in [37], the correction factor is the product of the training cycle segments number, the number of the distance threshold candidates, and the number of the shapelet candidates we extract from each cycle segment. Hence, the Bonferroni correction significance level is: $\alpha_{corrected} = \frac{\alpha}{Correctionfactor}$.

Once all of the significant shapelet candidates have been selected, they may have the similar pattern or morphology from the same cycle segment. We store the sorted shapelet candidates by the ascending p value in the same cycle segment, and remove the shapelet candidates if they have any overlap with the previous one. We keep all of the left shapelets candidates together in the ascending p value, and retain the top 50 so that we have the shapelet candidates from different cycle segments. Finally, we apply the hierarchical clustering with the top 50 shapelet candidates to have five clusters, and select the best one from each cluster.

In the shapelet searching, it is time consuming to calculate the similarity between each shapelet candidate and cycle segment. Here in this research, we follow the Mueen's ultra-fast algorithm for similarity search (MASS) [51].

The most important part for this algorithm is the convolution calculation. As shown in paper [52], the Euclidean distance can be rewritten in the following:

$$dist(S,T) = \sqrt{\sum_{i=1}^{m}(s_i - t_i)^2} \approx \sqrt{2m(1 - corr(S,T))} \tag{2.1}$$

where $corr(S,T) = \frac{\sum st - m\mu_s\mu_t}{m\sigma_s\sigma_t}$ is the correlation coefficient between shapelet candidate S and cycle segment T. $\mu_s = \frac{1}{m}\sum s$, $\sigma_s^2 = \frac{1}{m}\sum s^2 - \mu_s^2$. The sliding dot product $\sum st$ can be calculated by convolution. Since $\sum st$, $\sum s$, $\sum t$, $\sum s^2$ and $\sum t^2$ are the sufficient statistics, then the correlation coefficient is a constant operation. Hence, this algorithm can speed up the calculation between shapelet candidate and the cycle segments. The whole shapelet searching method is shown in Algorithm 1.

Algorithm 1: Shapelet extraction

```
Input: the training data $D$, minimum length $minL$ and maximum length $maxL$
Result: shapelets
finalS = $\varnothing$
for each $T$ in $D$ do
    candidates = ExactCandiates($T$, $minL$, $maxL$)
    for each $S$ in candidates do
        distance = MASS($T$, $S$)
        (threshold, informationGain) = inforCalculation($S$)
        $p$ value = ChiSquaredTest($S$)
        if $p$ value < Bonferroni correction $\alpha$ then
            Spool add $S$
        end
    end
    removeSimilar(Spool)
end
sort(Spool)
finalS = top50(Spool)
```

Shapelet transform

Shapelet transform was proposed by Lines et al. [30] to downsize the long time series into the feature vector, and at the same time, it preserves the shape information for the dataset. After the shapelet transform, we can disassociate the shapelet extracting from building the classifier. Most importantly, after the shapelet transform, the feature vectors can be used with any classifier. For the final selected shapelets, the shapelet transformed matrix is formed by calculating the minimum Euclidean distance between this shapelet and testing cycle segmentations to represent the features to do the classification.

2.4.3 Statistical process control

Control chart as the important tool for statistical process control is used to monitor the performance of the process over time to keep the process in an in-control state [33]. Here, it will be used as the ECG anomaly detection for the cardiac monitoring.

The statistical process control with control chart is implemented in two phases. Phase I is also called retrospective phase, that analyzes the in-control data from the training examples to estimate the control limit for the next step. In our method, phase I is the model training procedure, and the input will be the shapelet transformed matrix, then using the support vector data description (SVDD) algorithm [53] to get the control limit as the monitoring statistic. Phase II is called the prospective or monitoring phase, that based on the control limit from phase I to construct the control chart for the testing cycle segmentation to monitor the new cycle, and then to finish the ECG anomaly detection.

Support vector data description (SVDD) is a method to solve the one-class classification problem by mixing SVM and the data description method together [53]. SVDD is to find the minimal enclosing sphere to provide a hypersphere boundary around the data. Let a be the center of the hypersphere, and R^2 be the radius of the hypersphere. $\{x_i\} \in X$ for $i = 1, 2, \ldots, N$ is the training observations. The objective for SVDD is to:

$$minimize R^2 + C \sum_{i=1}^{N} \xi_i \tag{2.2}$$

with constraint:

$$\parallel x_i - a \parallel^2 \leq R^2 + \xi_i \tag{2.3}$$

$$\xi_i \geq 0, \forall i \tag{2.4}$$

The parameter C is a constant, which controls the trade-off between the volume of the hypersphere and the errors. The previous paper [54] defined the user-specified parameter f to represent the fraction of the training data outside the decision boundary:

$$f = 1/NC$$

where N is the number of the training target observations. As the increasing of f, the volume of the hypersphere becomes smaller and the misclassification error in the training class becomes larger. This objective function with constrains can be solved by using the Lagrange multipliers and KKT complementarity conditions of Fletcher [55], and the optimization problem finally becomes:

$$L = \sum_{i} \alpha_i K(X_i, X_i) - \sum_{i,j} \alpha_i \alpha_j K(X_i, X_j) \tag{2.5}$$

The solution can be obtained by maximizing this equation subject to $0 \leq \alpha_i \leq C$.

As with conventional SVM, the SVDD algorithm can generate more flexible decision boundaries by replacing the inner product with kernel functions:

$$K(X_i, X_j) = exp(-\left|X_i - X_j\right|^2 / S^2) \tag{2.6}$$

where $S > 0$ is the width of the Gaussian kernel that controls the complexity of the SVDD boundary. The control boundaries of SVDD are decided by the input parameters f and S: for the same f value, as the increasing of S, the shape of the control boundary becomes smoother; for the same S value, as the increasing of f, the control boundary becomes smaller.

For the new testing data point z, D^2 that measures the distance between z and the center a can be calculated by the following equation:

$$D^2 = K(z, z) - 2\sum_{i} \alpha_i K(z, X_i) + \sum_{i,j} \alpha_i \alpha_j K(X_i, X_j) \tag{2.7}$$

For classification, the new observation z is classified as the target or in-control if D^2 is less than or equal to R^2.

2.5 Experimental study

Following the method proposed in section 2.4, the extensive experiments are carried out to validate the effectiveness of the method in this section. We introduce the MIT-BIH data used for the experiment in section 2.5.1. Section 2.5.2 demonstrates the data preprocessing used. Finally, in section 2.5.3, we show the experiment results of cardiac monitoring using the SVDD-based control charts, and compare the proposed method with the commonly used and the state-of-the-art methods.

2.5.1 Data set

We evaluate our method with the publicly available real-world data sets from the MIT-BIH ECG arrhythmia database from PhysioNet [56, 57]. There are 48 records from 47 subjects in the dataset. Each record has a 30-min two-channel ECG recording. One channel is modified limb lead II, and the other one is modified lead V1, V2, V4 or V5. Each record has the computer-readable reference annotations for each beat. For all records, the experiment only used the first channel modified limb lead II, as many previous papers [58] did. There are two groups of data in this database: the first group includes the representative common samples from the arrhythmias in routine clinical practice; the second group includes the less common but clinically significant arrhythmias, such as ventricular, junctional, and supraventricular arrhythmias. According to AAMI recommended practice [59], the paced records (#102, #104, #107 and #217) are excluded in the experiment. Then the first group has 20 records

Table 2.2: Cycle classes

Classes	Labels	Types
normal	N	Normal beat, atrial escape beat, and junctional escape beat
abnormal	L	Left bundle branch block beat
	R	Right bundle branch block beat
	S	Atrial premature beat, aberrated atrial premature beat, junctional premature beat, and supraventricular premature beat
	V	Premature ventricular contraction, and ventricular escape beat
	F	Fusion of ventricular and normal beat
	Q	Paced beat, fusion of paced and normal beat, and unclassifiable beat

with label # starting from 1, and the second group has 24 records with label # starting from 2.

As most previous papers [48, 58, 60–62] did, in our experiment, the first group of data is used as the training and the second group is used as the testing. The model is trained individually for each record in the testing group, as some papers [48, 62] did. The data used for training individual patient record includes two parts: the common representative cycles, which are the same for each patient record, and the patient-specific cycles from each patient record. The common representative cycles are selected from all training records, and the patient-specific cycles are from the first 5 minutes of each testing record, which is in compliance with AAMI recommended practice [59]. The remaining 25 minutes of the record was used for testing.

In our experiment, the cycles are classified into two classes: normal and abnormal shown in Table 2.2, as the normal is the target class, the model is to detect the abnormal cycles for individual patient record. Based on the AAMI recommended practice, most previous papers [58, 60–62] applied the five class labels: N, S, V, F and Q, some previous papers [48, 63] used seven class labels: N, L, R, S, V, F and Q. Here, we treat N as the normal class label, and the rest L, R, S, V, F and Q are the abnormal class labels.

2.5.2 Data pre-processing

The modified limb lead II ECG recording of each record is segmented into a sequence of cycles. Each cycle is segmented based on its R peak marked in the database. As the previous paper [48] did, each cycle segmentation has the fixed length 255, and includes the 0.25 seconds length of ECG recording before the detected R peak and the 0.45 seconds after the R peak. After the discrete wavelet transform, i.e. db2(2), the final cycle segmentation length is fixed to 66 which reduced the computational time for the experiment.

2.5.3 Results

In the experiments, we randomly selected 45 common representative normal cycles from each training record, as it would be time consuming if we use all of the cycles from the recording. The number chosen was decided carefully based on the number of abnormal cycles in the training records. However, we didn't select all of the abnormal cycles from the training records, we want to keep the data as balance as possible. If the number of specific abnormal label is less than or equal to 43 in the training record, we selected all of the abnormal labels, otherwise, we only randomly selected 43 cycles. For example, in Record 118, the number of abnormal label A is 1, we selected 1 cycle with label A; the number of abnormal label V is 109, we only selected 43 cycles with label V. Finally, in the common representative part for the training, there are 720 normal cycles, and 664 abnormal cycles (including 86 L, 86 R, 139 A, 333 V,

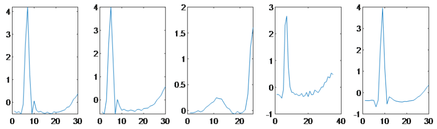


Figure 2.4: Five shapelets selected for Record 221

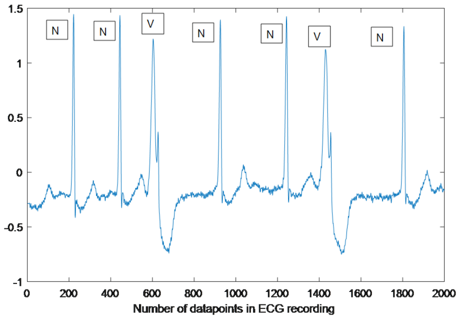


Figure 2.5: First 2000 datapoints of Record 221 ECG signal, N is the normal label, and V is the abnormal label

13 F and 7 Q).

We set the shapelet length to be 20, 25, 30, 40, and 45, and the sliding window to be 2. For the Bonferroni correction significance level, we set $\alpha = 0.05$, and then $\alpha_{corrected} = 0.05/(Correction factor)$. As shown in Figures 2.4, it represents the five best shapelets selected from the training model for record 221. Figures 2.5 displays the first 2000 data points from the ECG lead II in Record 221. As seen from Figures 2.5, the difference lies in the QRS between label N and V, it is easier for the shapelets to detect the abnormal label V.

For the evaluation of the performance, we used the accuracy (acc), sensitivity (sen), specificity (spe), and $F1$ score as the previous paper [45, 48, 60–62, 64, 65] did.

$$acc = \frac{TP + TN}{TP + TN + FP + FN}$$

$$sen = \frac{TP}{TP + FN}$$

$$spe = \frac{TN}{TN + FP}$$

$$F1 = \frac{2 * TP}{2 * TP + FN + FP}$$

TP, TN, FP and FN represent the true positive, true negative, false positive and false negative for the binary classification.

For the construction of the optimal shapelet-based SVDD classifier, the shapelet-based K-nearest neighbor (KNN) based control charts are constructed to compare with the shapelet-based SVDD control charts as shown in Table 2.3. Since some testing records only have one class, then the cell may be indicated with NaN. The highest accuracy is marked with bold. As can be seen from Table 2.3, in most testing records, the accuracy is higher in SVDD than KNN based control chart. And then it shows the ability of SVDD based control chart for classifying the ECG cycles.

In the following, six other different benchmark methods are also experimentally studied with the same training and testing cycles to demonstrate the effectiveness of proposed method, but with the principle-component-analysis (PCA) based features and Autoregressive model (AR) based features instead of shapelet-based features. Here, the first 12 principal components are kept in the PCA, and Burg's method is used to estimate the AR coefficients with order 4. For the PCA-based features, these four classifiers used are random forest [66], multilayer perceptron (MLP) of Artificial Neural Network (ANN) [67], multivariate Hoteling's T^2 control chart [68] and multivariate exponentially

Table 2.3: Comparison between shapelet-based SVDD and shapelet-based KNN control chart

	Proposed shapelet-based SVDD				Shapeled-based KNN control chart			
Record	acc	sen	spe	F1	acc	sen	spe	F1
200	**0.9788**	0.9411	0.9979	0.9676	0.9359	0.9658	0.9207	0.9103
201	**0.9658**	0.8571	0.9966	0.9172	0.6822	0.8750	0.6275	0.5490
202	**0.9834**	0.6338	0.9972	0.7438	0.9497	0.6901	0.9600	0.5104
203	**0.9597**	0.8836	0.9734	0.8698	0.9343	0.9418	0.9330	0.8137
205	**0.9945**	0.8462	1.0000	0.9167	0.9936	0.8462	0.9991	0.9041
207	**1.0000**	1.0000	NaN	1.0000	**1.0000**	1.0000	NaN	1.0000
208	0.9019	0.8043	0.9862	0.8837	**0.9089**	0.8707	0.9419	0.8985
209	**0.9079**	0.4826	0.9818	0.6081	0.8209	0.7212	0.8382	0.5440
210	**0.9873**	0.8763	0.9980	0.9239	0.9732	0.9330	0.9771	0.8599
212	0.9877	0.9886	0.9861	0.9906	**0.9947**	0.9946	0.9950	0.9960
213	**0.9303**	0.6314	0.9968	0.7673	0.7955	0.8310	0.7876	0.5965
214	**1.0000**	1.0000	NaN	1.0000	0.8504	0.8504	NaN	0.9191
215	0.9871	0.7537	0.9989	0.8487	**0.9964**	0.9552	0.9985	0.9624
219	**0.9927**	0.9424	0.9983	0.9626	0.8867	0.9634	0.8781	0.6301
220	**0.9545**	0.2258	0.9969	0.3529	0.9539	0.2688	0.9938	0.3906
221	**1.0000**	1.0000	1.0000	1.0000	**1.0000**	1.0000	1.0000	1.0000
222	**0.7905**	0.0000	0.9870	0.0000	0.5730	0.1473	0.6789	0.1207
223	**0.9700**	0.9300	0.9831	0.9387	0.9172	0.9558	0.9045	0.8508
228	**0.9812**	0.9508	0.9878	0.9477	0.9783	0.9869	0.9764	0.9421
230	**1.0000**	1.0000	1.0000	1.0000	0.9925	1.0000	0.9925	0.1250
231	**0.9898**	0.9949	0.9735	0.9933	0.7940	0.7590	0.9073	0.8491
232	**1.0000**	1.0000	NaN	1.0000	0.6175	0.6175	NaN	0.7635
233	**0.9965**	0.9915	0.9984	0.9936	0.9875	1.0000	0.9828	0.9777
234	**0.9782**	0.0566	1.0000	0.1071	0.9773	0.0566	0.9991	0.1034

weighted moving average (MEWMA) control chart [69]. For the AR-based features, the two classifiers are support vector machine (SVM) and k-nearest neighbors (KNN). Table 2.4 shows the comparison between the shapelet-based SVDD control chart and the PCA-based classifiers random forest and ANN. the proposed shapelet-based SVDD method has 17 highest accuracies out of total 24 testing records, while PCA-based random forest has 7 highest and PCA-based ANN has 2 highest. In most of the testing records, the accuracies from the shapelet-based SVDD control chart are higher than the PCA-based classifiers (random forest and ANN), which shows the effectiveness of the proposed shapelet-based features. Table 2.5 compares the accuracy between the shapelet-based SVDD control chart and AR-based classifiers SVM and KNN, and shapelet-based SVDD control charts has 21 highest accuracies, AR-based classifier SVM has 7 highest accuracies, and AR-based classifier KNN has 0, which again shows the effectiveness of the proposed shapelet-based features.

Table 2.6 displays the comparison results between the shapelet-based SVDD control chart and the PCA-based control charts: T^2 and MEWMA. The proposed shapelet-based SVDD control chart has 20 highest accuracies, while PCA-based multivariate T^2 control charts have 4 highest and MEWMA only have one. In most of the testing records, the accuracies from the shapelet-based SVDD control chart are higher than the PCA-based control chart methods (Hoteling's T^2 and MEWMA), which shows the effectiveness of proposed SVDD method.

The following shows the comparison between the proposed method with the state-of-the-art convolutional neural network method proposed [45] for MIT-

Table 2.4: Comparison between shapelet-based control chart and PCA-based classifiers

	Shapelet-based		PCA-based classification			
	SVDD control chart		Random forest		ANN	
Record	acc	F1	acc	F1	acc	F1
200	**0.9788**	0.9676	0.9709	0.9569	0.9280	0.8997
201	**0.9658**	0.9172	0.9651	0.9163	0.9164	0.7678
202	**0.9834**	0.7438	0.9118	0.3478	0.8840	0.2791
203	**0.9597**	0.8698	0.9496	0.8474	0.8247	0.6253
205	**0.9945**	0.9167	0.9914	0.8774	0.9918	0.8696
207	**1.0000**	1.0000	0.9906	0.9953	0.9975	0.9987
208	0.9019	0.8837	**0.9807**	0.9792	0.9319	0.9303
209	**0.9079**	0.6081	0.8797	0.3399	0.8809	0.3304
210	**0.9873**	0.9239	0.9837	0.9077	0.9782	0.8776
212	0.9877	0.9906	**0.9961**	0.9970	0.6550	0.7910
213	0.9303	0.7673	**0.9548**	0.8778	0.9333	0.7810
214	**1.0000**	1.0000	0.9878	0.9938	0.9984	0.9992
215	0.9871	0.8487	**0.9989**	0.9887	0.9914	0.9070
219	**0.9927**	0.9626	0.9150	0.6908	0.9449	0.6263
220	0.9545	0.3529	**0.9776**	0.7683	0.9457	0.0213
221	**1.0000**	1.0000	0.9985	0.9952	0.9965	0.9890
222	0.7905	0.0000	0.7470	0.1301	**0.8005**	0.0094
223	**0.9700**	0.9387	0.9199	0.8491	0.8985	0.7722
228	0.9812	0.9477	**0.9965**	0.9901	0.9759	0.9311
230	**1.0000**	1.0000	0.9849	0.0667	0.9871	0.0000
231	**0.9898**	0.9933	0.8794	0.9268	0.7635	0.8659
232	**1.0000**	1.0000	0.9414	0.9698	**1.0000**	1.0000
233	**0.9965**	0.9936	0.9895	0.9808	0.9867	0.9758
234	**0.9782**	0.1071	**0.9782**	0.4565	0.9694	0.0541

Table 2.5: Comparison between shapelet-based SVDD and AR-based classifiers

	Shapelet-based		AR-based classification			
	SVDD control chart		SVM		KNN	
Record	acc	F1	acc	F1	acc	F1
200	**0.9788**	0.9676	0.9603	0.9392	0.8906	0.8346
201	**0.9658**	0.9172	0.9553	0.8882	0.9125	0.7726
202	**0.9834**	0.7438	0.9717	0.4536	0.8631	0.2147
203	**0.9597**	0.8698	0.7420	0.4255	0.6574	0.3541
205	0.9945	0.9167	**0.9955**	0.9333	0.9873	0.8228
207	**1.0000**	1.0000	0.9969	0.9984	0.9485	0.9736
208	0.9019	0.8837	**0.9688**	0.9655	0.9364	0.9305
209	**0.9079**	0.6081	0.8523	0.0053	0.8189	0.1231
210	**0.9873**	0.9239	0.9492	0.7037	0.8484	0.4698
212	**0.9877**	0.9906	0.9869	0.9900	0.9759	0.9816
213	**0.9303**	0.7673	**0.9303**	0.7632	0.9014	0.7412
214	**1.0000**	1.0000	**1.0000**	1.0000	0.9196	0.9581
215	**0.9871**	0.8487	0.9803	0.8197	0.9613	0.6197
219	**0.9927**	0.9626	0.9218	0.3605	0.9019	0.3529
220	**0.9545**	0.3529	0.9451	NaN	0.9510	0.2095
221	**1.0000**	1.0000	0.9946	0.9827	0.9624	0.8911
222	0.7905	0.0000	**0.7939**	0.0091	0.7882	0.1284
223	**0.9700**	0.9387	0.7675	0.1824	0.7921	0.5360
228	**0.9812**	0.9477	0.9806	0.9477	0.8984	0.7539
230	**1.0000**	1.0000	**1.0000**	1.0000	0.9935	0.1429
231	**0.9898**	0.9933	0.9366	0.9594	0.9193	0.9463
232	**1.0000**	1.0000	0.9475	0.9730	0.8586	0.9239
233	**0.9965**	0.9936	0.8488	0.6466	0.8148	0.6965
234	**0.9782**	0.1071	**0.9782**	0.1071	0.9773	0.1875

Table 2.6: Comparison between shapelet-based SVDD and PCA-based control charts

	Shapelet-based		PCA-based control chart			
	SVDD control chart		T^2		MEWMA	
Record	acc	F1	acc	F1	acc	F1
200	**0.9788**	0.9676	0.9645	0.9479	0.6617	0.0027
201	**0.9658**	0.9172	0.9382	0.8605	0.7533	0.0053
202	**0.9834**	0.7438	0.7840	0.2406	0.7422	0.0837
203	**0.9597**	0.8698	0.7622	0.5597	0.8484	0.0259
205	**0.9945**	0.9167	0.9859	0.8098	0.9386	0.4000
207	**1.0000**	1.0000	**1.0000**	1.0000	0.9994	0.9997
208	0.9019	0.8837	**0.9142**	0.9144	0.4975	0.2312
209	0.9079	0.6081	**0.9392**	0.7941	0.8872	0.5860
210	**0.9873**	0.9239	0.9505	0.7615	0.9110	0.0485
212	**0.9877**	0.9906	0.9873	0.9904	0.6524	0.7886
213	**0.9303**	0.7673	0.8577	0.6837	0.8181	NaN
214	**1.0000**	1.0000	0.9941	0.9971	0.0895	0.1642
215	**0.9871**	0.8487	0.9617	0.7147	0.9520	NaN
219	**0.9927**	0.9626	0.8778	0.6136	0.8987	0.0000
220	0.9545	0.3529	**0.9835**	0.8409	0.9415	0.0000
221	**1.0000**	1.0000	0.9401	0.8389	0.8411	0.0123
222	0.7905	0.0000	0.5258	0.3097	**0.7986**	0.0000
223	**0.9700**	0.9387	0.9568	0.9103	0.7525	0.0000
228	**0.9812**	0.9477	0.9777	0.9406	0.8180	0.0000
230	**1.0000**	1.0000	0.9451	0.0192	0.9962	0.0000
231	**0.9898**	0.9933	0.9303	0.9564	0.7635	0.8659
232	**1.0000**	1.0000	0.9542	0.9766	0.4451	0.6160
233	**0.9965**	0.9936	0.9168	0.8679	0.7250	0.0000
234	**0.9782**	0.1071	0.9764	0.4906	0.8189	0.2004

BIH dataset. As we used the different cycle classes with most of the previous papers [45, 60–62, 64, 65], we want to make the results comparable, then we removed the testing records with the abnormal label L and R (#207, #212, #214, #231 and #232) when doing the comparison. Even we have the same cycle classes as paper [48], since this paper didn't have results for every testing record, then we just compared the classification result of every testing record with the paper [45]. Table 2.7 shows the comparisons of accuracy for every testing record except the abnormal label L and label R (#207, #212, #214, #231 and #232) between proposed method and previous paper [45]. The previous paper [45] only has the classification results based on abnormal label S and V, and then we recalculated the paper's results based on all abnormal labels: all types [45] column from Table 2.7 is recalculated based on the data from the previous paper [45]. Although the proposed method only has few testing records outperform the previous paper, it still shows the comparable accuracy with the state-of-art research, and most of important, our proposed method with shapelets has the good interpretation ability and the shapeletes are statistically significant

Finally, we apply the testing records to evaluate the performance of the proposed shapelet-based SVDD control chart for the individual cardiac monitoring. The common ECG cycles combined with the first five minutes ECG cycles from the individual testing record are used in Phase I training. The optimal value of the Gaussian kernel width is found based on the highest value of the F measure in the training dataset. Figures 2.6 displays the Phase II from SVDD based control charts for testing record 221. The samples with red color

Table 2.7: Comparisons of accuracy for every testing record between proposed method and paper [45]

Record	Cycle labels					V[45]	S[45]	All types[45]	Proposed all types
	N	S	V	F	Q				
200	1436	28	700	2	0	0.9780	0.9810	0.9751	**0.9788**
201	1193	126	198	2	0	0.9930	0.9760	0.9776	0.9658
202	1798	55	15	1	0	0.9950	0.9580	0.9535	**0.9834**
203	2101	0	373	1	4	0.9760	0.9970	0.9873	0.9597
205	2122	2	64	11	0	0.9980	1.0000	1.0000	0.9945
208	1306	2	824	301	2	0.9790	0.9860	0.9869	0.9019
209	2144	372	1	0	0	0.9990	0.9620	0.9612	0.9079
210	2008	20	165	9	0	0.9900	0.9960	0.9917	0.9873
213	2208	27	195	268	0	0.9830	0.9940	0.9834	0.9303
215	2659	2	131	1	0	1.0000	1.0000	1.0000	0.9871
219	1713	7	51	0	0	0.9650	0.9390	0.9131	**0.9927**
220	1599	93	0	0	0	1.0000	0.9730	0.9730	0.9545
221	1702	0	316	0	0	0.9980	1.0000	1.0000	**1.0000**
222	1905	209	0	0	0	0.9830	0.9230	0.9074	0.7905
223	1668	66	455	8	0	0.9240	0.9710	0.9589	**0.9700**
228	1396	3	302	0	0	0.9990	0.9980	0.9975	0.9812
230	1856	0	1	0	0	0.9990	1.0000	0.9990	**1.0000**
233	1857	4	692	6	0	0.9910	0.9960	0.9949	**0.9965**
234	2236	50	3	0	0	0.9990	0.9870	0.9870	0.9782

represents the abnormal cycles, and the blue color represents the normal cycles. As we have 2020 testing monitoring statistics, it is difficult to see if using the line connection between points. Here, we omit the line connection. The control limit is determined by phase I training model. As shown in Figures 2.6, abnormal cycles are clearly detected, and the accuracy is 1. Figures 2.7 displays the Phase II for testing record 233, and the accuracy is 0.9965. Three false positive cycles are identified with square sign, and the arrows point to the six false negative cycles. The two subplots show the corresponding ECG cycles detected for the first false positive and false negative samples. Hence, it proves the effectiveness of the proposed shapelet-based SVDD control chart in monitoring the ECG cycles.

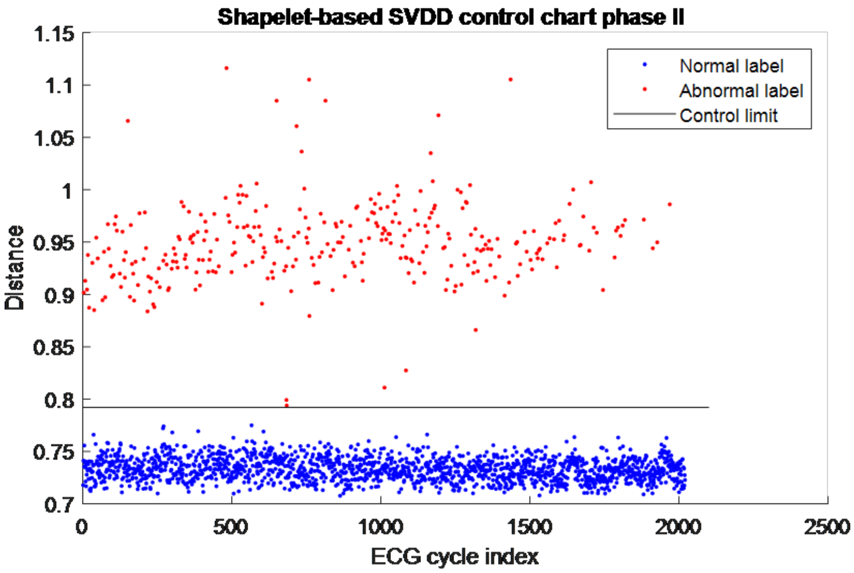


Figure 2.6: Phase II of the shapelet-based SVDD control chart for Record 221, the black line represents the control limit calculated from Phase I, the blue dot represents the normal labeled cycles, and the red dot represents the abnormal labeled cycles

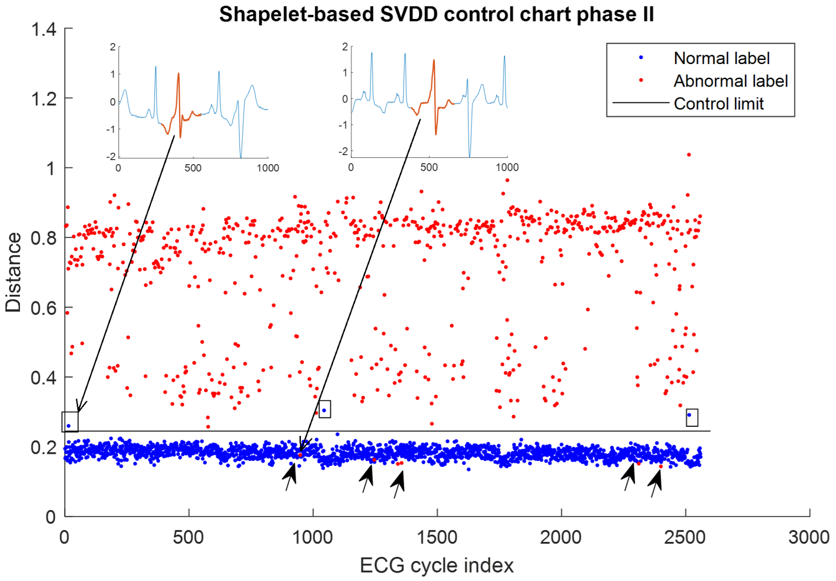


Figure 2.7: Phase II of the shapeled-based SVDD control chart for Record 233, the square shows the false positive, and the arrow points the false negative cycle

2.6 Conclusions and future work

As the increasing of the cardiac related disease, it is important to detect the abnormal ECG cycles during the cardiac monitoring for the early treatment. The IoT-enabled cardiac monitoring devices make it available for people suffering the cardiac related problems to regularly monitor their health, and for doctors remotely analyze the monitoring data to provide the timely service. We proposed a method to extract the statistically significant shapelets from the long time ECG monitoring, and applied the shapelet-based feature with the SVDD algorithm to construct the control chart in statistical process control for the ECG anomaly detection. The different methods are experimentally studied, and the comparison results demonstrate the effectiveness of the proposed shapelet-based SVDD control chart method. The experimental results on the real-world MIT-BIH dataset show that our proposed method is comparable with the state-of-art method. Since we only used one lead ECG signal in our experiments, as future work, we plan to extend the method to the multivariate time series so that we can make full use of the available ECG signals. Because the IoT-based devices require the algorithm to be more efficient during the calculation, it is also a good direction to propose a speed-up algorithm for shapelet-based feature in the future.

Bibliography

[1] Salim S Virani, Alvaro Alonso, Hugo J Aparicio, Emelia J Benjamin, Marcio S Bittencourt, Clifton W Callaway, April P Carson, Alanna M

Chamberlain, Susan Cheng, Francesca N Delling, et al. Heart disease and stroke statistics—2021 update: a report from the american heart association. *Circulation*, 143(8):e254–e743, 2021.

[2] Jessica K Zègre-Hemsey, J Lee Garvey, and Mary G Carey. Cardiac monitoring in the emergency department. *Critical Care Nursing Clinics*, 28(3):331–345, 2016.

[3] Joy Liao, Zahira Khalid, Ciaran Scallan, Carlos Morillo, and Martin O'Donnell. Noninvasive cardiac monitoring for detecting paroxysmal atrial fibrillation or flutter after acute ischemic stroke: a systematic review. *Stroke*, 38(11):2935–2940, 2007.

[4] Robert O Bonow, Douglas L Mann, Douglas P Zipes, and Peter Libby. *Braunwald's heart disease e-book: A textbook of cardiovascular medicine*. Elsevier Health Sciences, 2011.

[5] Mandeep Singh, Gurmohan Singh, Jaspal Singh, and Yadwinder Kumar. Design and validation of wearable smartphone based wireless cardiac activity monitoring sensor. *Wireless Personal Communications*, pages 1–17, 2021.

[6] Jie Wan, Munassar AAH Al-awlaqi, MingSong Li, Michael O'Grady, Xiang Gu, Jin Wang, and Ning Cao. Wearable iot enabled real-time health monitoring system. *EURASIP Journal on Wireless Communications and Networking*, 2018(1):1–10, 2018.

[7] Prabal Verma and Sandeep K Sood. Fog assisted-iot enabled patient

health monitoring in smart homes. *IEEE Internet of Things Journal*, 5 (3):1789–1796, 2018.

[8] Samik Basu, Anwesha Sengupta, Anindita Das, Mahasweta Ghosh, and Soma Barman. Iot-based real-time remote ecg monitoring system. In *International Conference on Computers and Devices for Communication*, pages 23–28. Springer, 2019.

[9] Hongzu Li and Pierre Boulanger. A survey of heart anomaly detection using ambulatory electrocardiogram (ecg). *Sensors*, 20(5):1461, 2020.

[10] Varun Chandola. *Anomaly detection for symbolic sequences and time series data*. University of Minnesota, 2009.

[11] Mehmet Akif Ozdemir, Onan Guren, Ozlem Karabiber Cura, Aydin Akan, and Aytug Onan. Abnormal ecg beat detection based on convolutional neural networks. In *2020 Medical Technologies Congress (TIPTEKNO)*, pages 1–4. IEEE, 2020.

[12] Fenghuan Li, Kehai Chen, Jie Ling, Yinwei Zhan, and Gunasekaran Manogaran. Automatic diagnosis of cardiac arrhythmia in electrocardiograms via multigranulation computing. *Applied Soft Computing*, 80: 400–413, 2019.

[13] Ziqian Wu, Tianjie Lan, Cuiwei Yang, and Zhenning Nie. A novel method to detect multiple arrhythmias based on time-frequency analysis and convolutional neural networks. *IEEE Access*, 7:170820–170830, 2019.

[14] Marcel0 R Risk, Jamil F Sobh, and J Philip Saul. Beat detection and classification of ecg using self organizing maps. In *Proceedings of the 19th Annual International Conference of the IEEE Engineering in Medicine and Biology Society.'Magnificent Milestones and Emerging Opportunities in Medical Engineering'(Cat. No. 97CH36136)*, volume 1, pages 89–91. IEEE, 1997.

[15] N Srinivasan, DF Ge, and SM Krishnan. Autoregressive modeling and classification of cardiac arrhythmias. In *Proceedings of the Second Joint 24th Annual Conference and the Annual Fall Meeting of the Biomedical Engineering Society][Engineering in Medicine and Biology*, volume 2, pages 1405–1406. IEEE, 2002.

[16] Yüksel Özbay, Rahime Ceylan, and Bekir Karlik. Integration of type-2 fuzzy clustering and wavelet transform in a neural network based ecg classifier. *Expert Systems with Applications*, 38(1):1004–1010, 2011.

[17] Hari Mohan Rai, Anurag Trivedi, and Shailja Shukla. Ecg signal processing for abnormalities detection using multi-resolution wavelet transform and artificial neural network classifier. *Measurement*, 46(9):3238–3246, 2013.

[18] Mohammad Alshaer, Sandra Garcia-Rodriguez, and Cedric Gouy-Pailler. Detecting anomalies from streaming time series using matrix profile and shapelets learning. In *2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 376–383. IEEE, 2020.

[19] Yonghan Jung and Heeyoung Kim. Detection of pvc by using a wavelet-based statistical ecg monitoring procedure. *Biomedical Signal Processing and Control*, 36:176–182, 2017.

[20] Lexiang Ye and Eamonn Keogh. Time series shapelets: a new primitive for data mining. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 947–956, 2009.

[21] Yanping Chen, Eamonn Keogh, Bing Hu, Nurjahan Begum, Anthony Bagnall, Abdullah Mueen, and Gustavo Batista. The ucr time series classification archive, July 2015. `www.cs.ucr.edu/~eamonn/time_series_data/`.

[22] Chotirat Ann Ratanamahatana and Eamonn Keogh. Three myths about dynamic time warping data mining. In *Proceedings of the 2005 SIAM international conference on data mining*, pages 506–510. SIAM, 2005.

[23] Hui Ding, Goce Trajcevski, Peter Scheuermann, Xiaoyue Wang, and Eamonn Keogh. Querying and mining of time series data: experimental comparison of representations and distance measures. *Proceedings of the VLDB Endowment*, 1(2):1542–1552, 2008.

[24] Xiaopeng Xi, Eamonn Keogh, Christian Shelton, Li Wei, and Chotirat Ann Ratanamahatana. Fast time series classification using numerosity reduction. In *Proceedings of the 23rd international conference on Machine learning*, pages 1033–1040, 2006.

[25] Zhengzheng Xing, Jian Pei, and S Yu Philip. Early classification on time series. *Knowledge and information systems*, 31(1):105–127, 2012.

[26] Zhengzheng Xing, Jian Pei, Philip S Yu, and Ke Wang. Extracting interpretable features for early classification on time series. In *Proceedings of the 2011 SIAM international conference on data mining*, pages 247–258. SIAM, 2011.

[27] Bastian Hartmann and Norbert Link. Gesture recognition with inertial sensors and optimized dtw prototypes. In *2010 IEEE International Conference on Systems, Man and Cybernetics*, pages 2102–2109. IEEE, 2010.

[28] Bastian Hartmann, Ingo Schwab, and Norbert Link. Prototype optimization for temporarily and spatially distorted time series. In *2010 AAAI Spring Symposium Series*, 2010.

[29] T Shajina and P Bagavathi Sivakumar. Human gait recognition and classification using time series shapelets. In *2012 international conference on advances in computing and communications*, pages 31–34. IEEE, 2012.

[30] Jason Lines, Luke M Davis, Jon Hills, and Anthony Bagnall. A shapelet transform for time series classification. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 289–297, 2012.

[31] Mohamed F Ghalwash and Zoran Obradovic. Early classification of multivariate temporal observations by extraction of interpretable shapelets. *BMC bioinformatics*, 13(1):1–12, 2012.

[32] Guozhong Li, Byron Choi, Jianliang Xu, Sourav S Bhowmick, Kwok-Pan Chun, and Grace Lai-Hung Wong. Shapenet: A shapelet-neural network approach for multivariate time series classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 8375–8383, 2021.

[33] Thuntee Sukchotrat, Seoung Bum Kim, and Fugee Tsung. One-class classification-based control charts for multivariate process monitoring. *IIE transactions*, 42(2):107–120, 2009.

[34] Ruixiang Sun and Fugee Tsung. A kernel-distance-based multivariate control chart using support vector methods. *International Journal of Production Research*, 41(13):2975–2989, 2003.

[35] S Kumar, AK Choudhary, Maneesh Kumar, R Shankar, and MK Tiwari. Kernel distance-based robust support vector methods and its application in developing a robust k-chart. *International Journal of Production Research*, 44(1):77–96, 2006.

[36] Zhisheng Zhang, Xuejun Zhu, and Jionghua Jin. Svc-based multivariate control charts for automatic anomaly detection in computer networks. In *Third International Conference on Autonomic and Autonomous Systems (ICAS'07)*, pages 56–56. IEEE, 2007.

[37] Christian Bock, Thomas Gumbsch, Michael Moor, Bastian Rieck, Damian Roqueiro, and Karsten Borgwardt. Association mapping in biomedical

time series via statistically significant shapelet mining. *Bioinformatics*, 34(13):i438–i446, 2018.

[38] Rob J Hyndman, Earo Wang, and Nikolay Laptev. Large-scale unusual time series detection. In *2015 IEEE international conference on data mining workshop (ICDMW)*, pages 1616–1619. IEEE, 2015.

[39] Philip De Chazal and Richard B Reilly. A patient-adapting heartbeat classifier using ecg morphology and heartbeat interval features. *IEEE transactions on biomedical engineering*, 53(12):2535–2543, 2006.

[40] Zhancheng Zhang, Jun Dong, Xiaoqing Luo, Kup-Sze Choi, and Xiaojun Wu. Heartbeat classification using disease-specific feature selection. *Computers in biology and medicine*, 46:79–89, 2014.

[41] Can Ye, BVK Vijaya Kumar, and Miguel Tavares Coimbra. Heartbeat classification using morphological and dynamic features of ecg signals. *IEEE Transactions on Biomedical Engineering*, 59(10):2930–2941, 2012.

[42] Haemwaan Sivaraks and Chotirat Ann Ratanamahatana. Robust and accurate anomaly detection in ecg artifacts using time series motif discovery. *Computational and mathematical methods in medicine*, 2015, 2015.

[43] Bharadwaj Veeravalli, Chacko John Deepu, and DuyHoa Ngo. Real-time, personalized anomaly detection in streaming data for wearable healthcare devices. In *Handbook of large-scale distributed computing in smart healthcare*, pages 403–426. Springer, 2017.

[44] Kang Li, Nan Du, and Aidong Zhang. Detecting ecg abnormalities via transductive transfer learning. In *Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Biomedicine*, pages 210–217, 2012.

[45] Feiteng Li, Jiaquan Wu, Menghan Jia, Zhijian Chen, and Yu Pu. Automated heartbeat classification exploiting convolutional neural network with channel-wise attention. *IEEE Access*, 7:122955–122963, 2019.

[46] Stephane G Mallat. A theory for multiresolution signal decomposition: the wavelet representation. 1987.

[47] P De Chazal, BG Celler, and RB Reilly. Using wavelet coefficients for the classification of the electrocardiogram. In *Proceedings of the 22nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (Cat. No. 00CH37143)*, volume 1, pages 64–67. IEEE, 2000.

[48] Saeed Saadatnejad, Mohammadhosein Oveisi, and Matin Hashemi. Lstm-based ecg classification for continuous monitoring on personal wearable devices. *IEEE journal of biomedical and health informatics*, 24(2):515–523, 2019.

[49] Ronald A Fisher. On the interpretation of χ 2 from contingency tables, and the calculation of p. *Journal of the Royal Statistical Society*, 85(1): 87–94, 1922.

[50] Karl Pearson. X. on the criterion that a given system of deviations from

the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):157–175, 1900.

[51] Chin-Chia Michael Yeh, Yan Zhu, Liudmila Ulanova, Nurjahan Begum, Yifei Ding, Hoang Anh Dau, Diego Furtado Silva, Abdullah Mueen, and Eamonn Keogh. Matrix profile i: all pairs similarity joins for time series: a unifying view that includes motifs, discords and shapelets. In *2016 IEEE 16th international conference on data mining (ICDM)*, pages 1317–1322. Ieee, 2016.

[52] Davood Rafiei and Alberto Mendelzon. Similarity-based queries for time series data. In *Proceedings of the 1997 ACM SIGMOD international conference on Management of data*, pages 13–25, 1997.

[53] Xien Liu, Mengjun Li, Yuanlun Sun, Xiaoyan Deng, et al. Support vector data description for weed/corn image recognition. *Journal of Food, Agriculture and Environment*, 8(1):214–219, 2010.

[54] David MJ Tax and Robert PW Duin. Support vector data description. *Machine learning*, 54(1):45–66, 2004.

[55] Roger Fletcher. *Practical methods of optimization.* John Wiley & Sons, 2013.

[56] George B Moody and Roger G Mark. The impact of the mit-bih arrhyth-

mia database. *IEEE Engineering in Medicine and Biology Magazine*, 20 (3):45–50, 2001.

[57] Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, 101(23):e215–e220, 2000.

[58] Xiaolong Zhai, Zhanhong Zhou, and Chung Tin. Semi-supervised learning for ecg classification without patient-specific labeled data. *Expert Systems with Applications*, 158:113411, 2020.

[59] R Mark. Aami-recommended practice: Testing and reporting performance results of ventricular arrhythmia detection algorithms. *Association for the Advancement of Medical Instrumentation, Arrhythmia Monitoring Subcommittee, AAMI ECAR, 1987*, 1987.

[60] Wei Jiang and Seong G Kong. Block-based neural networks for personalized ecg signal classification. *IEEE Transactions on Neural Networks*, 18 (6):1750–1761, 2007.

[61] Turker Ince, Serkan Kiranyaz, and Moncef Gabbouj. A generic and robust system for automated patient-specific classification of ecg signals. *IEEE Transactions on Biomedical Engineering*, 56(5):1415–1426, 2009.

[62] Serkan Kiranyaz, Turker Ince, and Moncef Gabbouj. Real-time patient-

specific ecg classification by 1-d convolutional neural networks. *IEEE Transactions on Biomedical Engineering*, 63(3):664–675, 2015.

[63] Tae Joon Jun, Hoang Minh Nguyen, Daeyoun Kang, Dohyeun Kim, Daeyoung Kim, and Young-Hak Kim. Ecg arrhythmia classification using a 2-d convolutional neural network. *arXiv preprint arXiv:1804.06812*, 2018.

[64] Yazhao Li, Yanwei Pang, Jian Wang, and Xuelong Li. Patient-specific ecg classification by deeper cnn from generic to dedicated. *Neurocomputing*, 314:336–346, 2018.

[65] Xiaolong Zhai and Chung Tin. Automated ecg classification using dual heartbeat coupling based on convolutional neural network. *IEEE Access*, 6:27465–27472, 2018.

[66] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[67] Mostafa Kahani, Mohammad Hossein Ahmadi, Afshin Tatar, and Milad Sadeghzadeh. Development of multilayer perceptron artificial neural network (mlp-ann) and least square support vector machine (lssvm) models to predict nusselt number and pressure drop of tio2/water nanofluid flows through non-straight pathways. *Numerical Heat Transfer, Part A: Applications*, 74(4):1190–1206, 2018.

[68] J Edward Jackson. Multivariate quality control. *Communications in Statistics-Theory and Methods*, 14(11):2657–2688, 1985.

[69] Sharad S Prabhu and George C Runger. Designing a multivariate ewma control chart. *Journal of Quality Technology*, 29(1):8–15, 1997.

CHAPTER 3

POSITIVE AND UNLABELED SHAPELET LEARNING FOR ECG ANOMALY DETECTION

Li Zhang, Chen Kan, Victoria C. P. Chen

3.1 Abstract

Positive and unlabeled learning has attracted increasing interest in recent years. The setting of the positive and unlabeled learning is that we only access the positive and unlabeled training data sets. Many methods have been proposed for the positive and unlabeled learning. However, only a few paper integrate the shapelet features into the positive and unlabeled learning. In this paper, we propose the positive and unlabeled shapelet learning model for the ECG anomaly detection, and the experiment results from the real-world data sets demonstrate the effectiveness of our proposed method.

3.2 Introduction

Time series classification as a subset of the general classification problem has attracted many interests in the research for both academic and industry people, as the data collected automatically by sensing and monitoring are time series. However, in many real-world problems, collecting a large amount of the labeled data is costly, while the positive and unlabeled data are usually easily to be obtained. In such a situation, only a small set of positive labeled data and a large amount of unlabeled data is available, which leads to the development of positive and unlabeled learning [1]. Positive and unlabeled learning aims to learn a suitable binary classifier without the assistant of the negative data. Here, the positive data can be exchanged to the target class.

Positive and unlabeled learning has been applied in many applications in recent years. For example, Wei and Li [2] proposed the software clone detection for software maintenance and evolution in computer science, Youngs et al. [3] introduced the protein function prediction in biology, Li et al. [4] applied the remote-sensing data with positive and unlabeled learning for the image classification, and Li et al. [5] identified the fake reviews for the online shopping website, etc.

Given the broad applications of the positive and unlabeled learning, many different algorithms have been proposed, such as two-step techniques, biased learning and class prior incorporation, etc. For the two-step techniques, the first step is to identify the reliable negative and positive from the unlabeled data, for example, Liu et al. [6] proposed the S-EM algorithm using the spy techniques, Yu et al. [7] proposed the PEBL algorithm using the 1-DNF mapping, Li and Liu [8] introduced the Roc-SVM method using the Rocchio algorithm; the second step is to build a set of classifiers by applying the classification algorithm and then select the best classifier like the traditional supervised learning, for example, S-EM algorithm used the expectation maximization algorithm, PEBL and Roc-SVM used the support vector machines (SVM). For the biased learning, the methods treat the unlabeled data as the negative data with class label noise, and then a penalty is considered on the misclassified positive data, for example, Liu et al. [9] used the biased SVM that penalize the misclassified positive and negative data differently with the standard SVM, Mordelet and Vert [10] used the bagging SVM that train on the positive data and a subset of the negative data, Lee and Liu [11] applied the weighted logistic regression to favor the correct positive classification over the correct negative classification by giving the larger weights to the positive data. For the class prior incorporation, it adapts the algorithm to incorporate the class prior information during the learning or preprocess the training data to assign weights to the unlabeled data, for example, Elkan and Noto [12] trained the classifier on the part of the data while keeping a separate validation set and then estimate the label frequency as the average predicted probability of a labeled validation set, Bekker and Davis [13] estimated the class prior using the decision tree induction.

Despite the fact that a large amount of methods have been proposed for the positive and unlabeled learning, few efforts have been made to integrate the shapelets with the positive and unlabeled learning for time series classification [14–16]. Shapelets as explainable and discriminative features were introduced in 2009 [17] for time series data mining, and can provide a model with better interpretability. Shapelets are time series sub-sequences, and represent the maximally discriminative segments of time series that split the time series into two classes. As shown in Figure 3.1, leftmost plots show the two shapelets learned from the coffee dataset [18]), which captures the inherent characteristic of the time series, and provides an interpretation for the people who do not understand the inside processes or algorithms. Hence, discovering shapelets from time series has been increasing interest for researchers during the past decade.

For the shapelet learning, Wang et al. [15] proposed the semi-supervised shapelet learning (SSSL) model, which treats the unlabeled data in a supervised way by using the pseudo-labels and then uses the regularized least-square technique to learn both shapelets and classification boundaries. Yamaguchi

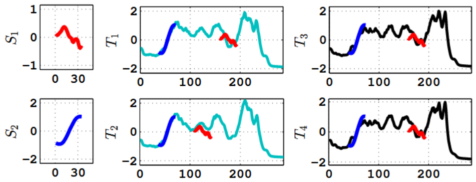


Figure 3.1: Two shapelets S1 and S2 learned from Coffee dataset [17]

and Nishikawa [14] introduced the one-class learning time series shapelet (OCLTS), which aims to learn the shapelets and classifiers using only training data for the majority class without the minority class. Although both of SSSL and OCLTS can be applied with the positive and unlabeled data, SSSL mainly focused on the labeled data with positive and negative, and then treated the labeled and unlabeled data in the same way, which may underestimate the importance of the positive labeled data and induce inaccurate classification, and OCLTS's target class is the majority, while in the positive and unlabeled setting, the target may be the minority class like the heart disease. Liang et al. [16] proposed the extracting method for positive and unlabeled shapelets, which is the only paper that combine shapelets and positive and unlabeled data together. However, they used the traditional way to label the unlabeled data first, and then extract the shapelets as the supervised classification, it is still time-consuming to find shapelets. And then in this paper, we integrate the shapelet with the exactly positive and unlabeled learning setting for the time series classification. We include the data distribution information that derived from unlabeled data as the paper [19] did to amend the label bias caused by the missing negative data and incorporate the hat loss to SVM for shapelet learning.

One of the important topics for time series classification is ECG anomaly detection. Anomalies are defined as any deviation from the normal behavior, and hence, the ECG anomalies represent the irregular heart activity. ECG anomaly detection can assist the domain experts to diagnose the cardiac condition, and provide the timely treatment for patients. Many methods have been proposed for the ECG anomaly detection in the past research including the supervised and unsupervised learning. However, since the supervised learning requires the annotation for each ECG cycle from domain experts prior to the training model and there are no labels for ECG cycles with unsupervised learning, semi-supervised learning would be a better choice. Because in the real situation, there are a lot of normal ECG cycles available, then in this paper, we apply the proposed positive and unlabeled shapelet learning method for the ECG anomaly detection.

The highlights of this paper are summarized as follows:

- This is the first effort of shapelet learning that focuses on the positive and unlabeled data setting.
- A new positive and unlabeled shapelet learning model is proposed that incorporates the derived distribution information from the unlabeled data.
- The proposed positive and shapelet learning method is applied for the ECG anomaly detection.

The remainder of the paper is organized as follows. Section 3.3 reviews the existing research related to our work. We propose the positive and unlabeled shapelet learning model (PUSL) in section 3.4. Section 3.5 shows the experimental study and results. Section 3.6 concludes the paper, and points the future work.

3.3 Literature review

Shapelet was first introduced by Ye and Keogh in 2009 for time series classification [17], and then it becomes as an important research area in the time series classification. The basic idea to find the shapelet is to consider all of the sub-sequences with the length less than the shortest time series, and assess the quality of all of the shapelet candidates based on the accuracy for the prediction of the training data. Because of the high number of candidates, the exhaustive search of shapelets is time-consuming, and then a series of the shapelets research has been focused on the speed-up technique. For example: sub-sequence distance early abandon and entropy pruning of the information gain [17]; computation intelligent cashing and reuse, and admissible pruning of the search space [20]; a random projection technique on the Symbolic Aggregate approximation (SAX) [21]; extracting the infrequent shapelet candidates to find shapelets [22]; a dynamic programming algorithm implemented in highly parallel graphic process units (GPUs) [23]; etc.

Although the speed-up technique has improved a lot on time using, they are still based on the exhaustive search, and it is time-consuming, and then learning-based shapelets were proposed by the researchers. Learning-based shapelet was first proposed by Grabocka et al. [24] to use the mathematical formulation to optimize the classification objective function, which can learn the near-to-optimal shapelets instead of searching from the candidates. This improves the time efficiency compared with the searching-based shapelet methods. However, this paper used the fully labeled training data, in the real world problems, it is costly to label all the time series data. Based on this work, Zhang et al. [25] proposed the unsupervised shapelet learning algorithm with the unlabeled data, Wang et al. [15] proposed the semi-supervised shapelet learning algorithm with the mixed labeled and unlabeled data, and Yamaguchi and Nishikawa [14] introduced the one class shapelet learning algorithm with the majority class labeled data. However, all the previous methods are not specified for the positive and unlabeled data setting, based on the large-margin label-calibrated support vector machines (LLSVM) algorithm for the positive and unlabeled learning proposed by Gong et al. [19], we integrate the shapelet features into the LLSVM model to introduce the positive and unlabeled shapelet learning model.

3.4 Research methodology

In this section, we first propose the positive and unlabeled shapelet learning model in section 3.4.1, and then explain the optimization process in section 3.4.2.

3.4.1 Positive and unlabeled shapelet learning model

The data set that contains n training examples is denoted as $T = (T_1, T_2, \ldots, T_n)$. Each time series has ordered real-valued observations and a class value -1 or

1. In the positive and unlabeled training problem setting, there are two types of time series: positive labeled P and unlabeled time series U. The classification is performed according to its class value, and to find the real-valued decision function f based on the training set $T = P \cap U$. For a sliding window of length L, a set of ordered segments can be obtained when the window slides along the time series. The segment starting at time j is defined as $(T_{(i,j)}, T_{(i,j+1)}, \ldots, T_{(i,j+L-1)})$ for T_i. As the shapelet definition said, time series segments are shapelet candidates and denoted by S. Shapelet transform was proposed by Lines et al. [26], and a new transformed shapelet matrix is formed where each column represents a shapelet, and each value is the distance between this shapelet and the corresponding time series. For a set of shapelets $S = (S_1, S_2, \ldots, S_m)$ and a set of time series $T = (T_1, T_2, \ldots, T_n)$, the distance between the $i - th$ time series T_i and $k - th$ shapelet S_k is defined as the minimum distance $X_i(S_k)$ among the distances between the shapelet S_k and each segment of time series T_i, where each segment has the same length as the shapelet S_k. Hence, the shapelet transformed matrix is represented as $X(S)$, and each element $X_i(S_k)$ is shown in the following equation:

$$X_i(S_k) = \min_{j=1,\ldots,Q-L+1} \frac{1}{L} \sum_{l=1}^{L} (T_{(i,j+l-1)} - S_{(k,l)})^2 \tag{3.1}$$

where L is the length of the shapelet and Q is the length of time series, $Q - L + 1$ is the total number of segments with length L in time series T_i. Since Equation 3.1 is not differential, we approximate it with the soft minimum function which is the differentiable approximation of the minimum function introduced by [24] in the following:

$$X_i(S_k) \approx \frac{\sum_{j=1}^{Q-L+1} d_{(i,k,j)} e^{\theta d_{(i,k,j)}}}{\sum_{j'=1}^{Q-L+1} e^{\theta d_{(i,k,j')}}} \tag{3.2}$$

where $d_{(i,k,j)} = \frac{1}{L}\sum_{l=1}^{L}(T_{(i,j+l-1)} - S_{(k,l)})^2$ is the distance between the $j-th$ segment of time series T_i and the $k-th$ shapelet S_k, and θ is the parameter that controls the precision of the Equation 3.2. As shown in [24], $\theta = -100$ is small enough to make the soft minimum yield exactly the same result as the true minimum. Hence, we set $\theta = -100$ in the following experiments.

Let n, p, and u as the sizes of T, P, and U, respectively, the model is formulated in the following as the LLSVM model proposed [19]:

$$\begin{aligned} \min \quad & \tfrac{1}{2}\|w\|^2 + \tfrac{\alpha}{p}\textstyle\sum_{i=1}^{p} \max(1 - w^T X_i(S), 0) \\ & + \tfrac{\beta}{u}\textstyle\sum_{i=p+1}^{p+u} \max(1 - |w^T X_i(S)|, 0) \\ & + \gamma \max(\tfrac{1}{u}\textstyle\sum_{i=p+1}^{p+u} \Phi(w^T X_i(S)) - t, 0) \end{aligned} \tag{3.3}$$

where $\alpha \geq 0$, $\beta \geq 0$, and $\gamma \geq 0$ are trade-off parameters, and $\Phi(y) = \frac{2}{\pi}\tan^{-1}(y)$ made the value of y between -1 and 1. Most importantly, we integrate the shapelet S into the features, and for $i-th$ example, the feature is denoted as $X_i(S)$. The first term $\frac{1}{2}\|w\|^2$ in the objective function is to prevent over-fitting. The second term in the objective function requires the label to be positive, which means that $w^T X_i(S) \geq 1$ in the training model if the feature is in the positive training examples. The third term makes the unlabeled training time series to have the clear label which means that $w^T X_i(S) \geq 1$ or $w^T X_i(S) \leq -1$, otherwise it will get penalized. The fourth term sets the upper bound of the mean value of unlabeled training time series' real labels to be parameter t which will be estimated using the class prior in the unlabeled examples.

Many methods have been proposed to estimate the class prior such as [13, 27–31]. Here, we follow the decision tree based method introduced in [13], which estimates the class prior by estimating the probability that a positive example is selected to be labeled. Specifically, let label frequency $c = P(s = 1|y = 1)$ be the constant probability to be selected to be labeled, and here, y is the true class, s is the positive labeled class. Following the "selected completely at random" assumption, the label frequencies are equal in any subdomain A, which means the label frequency is the ratio of the probabilities to be labeled positive and to be positive in any subdomain A: $c = \frac{P(s=1|x\in A)}{P(y=1|x\in A)}$. Since the probability $P(y = 1|x \in A)$ is at most to be 1, then in general, $c \geq P(s = 1|x \in A)$.

Hence, the lower bound on c can be estimated in any subset of data with L labeled positive and N total examples, i.e. $c \geq L/N$. Because of the stochastic nature of the labeling positive, the error term derived from the one-sided Chebyshev inequality is included, and the probabilistic lower bound for c is shown in the following:

$$P(c \leq \frac{L}{N} - \sqrt{\frac{(1-\delta)c(1-c)}{\delta N}}) \leq \delta \tag{3.4}$$

where the number of the labeled examples L exceeding the expected number by at least λ, the labeling variance is σ^2, $\delta = \sigma^2/(\sigma^2 + \lambda^2)$, and the error term is shrinking with the total sample size N.

The next step is to use the decision tree induction to split the dataset into two separate sets in each node to look for the interesting highly labeled subdomains using one set and estimate label frequency c using the other set by taking the maximum lower bound. There is one parameter $\delta = \frac{1}{1+4\epsilon^2 N_{min}}$ that needs to be tuned, here, we follow the rule in the paper [13] that bigger dataset with more than 10000 examples needs 1000 samples to update with an error $\epsilon = 0.1$, and smaller dataset needs one tenth of the sample size. Hence, $\delta = \max(0.025, \frac{1}{1+0.004N})$.

Hence, class prior $\tau_T = P(s = 1)/c$ can be estimated from the label frequency c, but the estimated class prior is in the total sample size, we need to change it back to the class prior in the unlabeled examples: $\tau_U = (\tau_T n - p)/u$, and the t is estimated in the following equation:

$$t = \frac{u\tau_U - u(1-\tau_U)}{u} = 2\tau_U - 1 \tag{3.5}$$

To make the objective function differentiable everywhere, as shown in [19] applying the squared hinge loss, squared label calibration term and the smooth Gaussian-like function [32] approximation of the hat loss, the final objective function is shown in the following:

$$\begin{aligned} \min \quad & \tfrac{1}{2}\|w\|^2 + \tfrac{\alpha}{p}\textstyle\sum_{i=1}^{p} \max(1 - w^T X_i(S), 0)^2 \\ & + \tfrac{\beta}{u}\textstyle\sum_{i=p+1}^{p+u} e^{-3(w^T X_i(S))^2} \\ & + \tfrac{\gamma}{u}\textstyle\sum_{i=p+1}^{p+u} (\max(\Phi(w^T X_i(S)) - t, 0))^2 \end{aligned} \tag{3.6}$$

3.4.2 Optimization

Because of the non-convex term of objective function 3.6, we use the minibatch SGD as [19] for optimization. Divide the training dataset T into N nonoverlapped minibatches which means that the minibatch size $m = n/N$, and then update w and S successively by batches. Rewrite the objective function 3.6 into positive and unlabeled, respectively as the followings:

$$F_i(w,S) = \frac{1}{2}\|w\|^2 + \frac{\alpha}{p}\sum_{i=1}^{p} \max(1 - w^T X_i(S), 0)^2, X_i(S) \in P \tag{3.7}$$

$$\begin{aligned} &\tfrac{1}{2}\|w\|^2 + \tfrac{\beta}{u}\textstyle\sum_{i=p+1}^{p+u} e^{-3(w^T X_i(S))^2} \\ &+\tfrac{\gamma}{u}\textstyle\sum_{i=p+1}^{p+u}(\max(\Phi(w^T X_i(S)) - t, 0))^2, X_i(S) \in U \end{aligned} \tag{3.8}$$

The gradient on w are:

$$\frac{\partial F_i(w,S)}{\partial w} = w + \frac{2\alpha}{p}\min(w^T X_i(S) - 1, 0)X_i(S), X_i(S) \in P \tag{3.9}$$

$$\begin{aligned} &w - \tfrac{6\beta}{u} w^T X_i(S) e^{-3(w^T X_i(S))^2} X_i(S) \\ &+\tfrac{4\gamma}{\pi u(1+(w^T X_i(S))^2)} \max(\Phi(w^T X_i(S)) - t, 0)X_i(S), X_i(S) \in U \end{aligned} \tag{3.10}$$

The gradient on S are:

$$\frac{\partial F_i(w,S)}{\partial S} = \frac{2\alpha}{p}\min(w^T X_i(S) - 1, 0)w^T \frac{\partial X_i(S)}{\partial S}, X_i(S) \in P \tag{3.11}$$

$$\begin{aligned} &-\tfrac{6\beta}{u} w^T X_i(S) e^{-3(w^T X_i(S))^2} w^T \tfrac{\partial X_i(S)}{\partial S} \\ &+\tfrac{4\gamma}{\pi u(1+(w^T X_i(S))^2)} \max(\Phi(w^T X_i(S)) - t, 0)w^T \tfrac{\partial X_i(S)}{\partial S}, X_i(S) \in U \end{aligned} \tag{3.12}$$

where for each shapelet S_k, $\frac{\partial X_i(S_k)}{\partial S_k} = \frac{\partial X_i(S_k)}{\partial d_{(i,k,j)}} \frac{\partial d_{(i,k,j)}}{\partial S_k}$, and $\frac{\partial d_{(i,k,j)}}{\partial S_k} = \frac{2}{L}(S_{(k,l)} - T_{(i,j+l-1)})$, $\frac{\partial X_i(S_k)}{\partial d_{(i,k,j)}} = \frac{e^{-100 d_{(i,k,j)}(1-100(d_{(i,k,j)}-X_i(S_k)))}}{\sum_{j'=1}^{J} e^{-100 d_{(i,k,j')}}}$. Hence, the updating w and S for each iteration will be:

$$\begin{aligned} w &:= w - \tfrac{\eta}{m} \textstyle\sum_{i=1}^{m} \frac{\partial F_i(w,S)}{\partial w} \\ S &:= S - \tfrac{\eta}{m} \textstyle\sum_{i=1}^{m} \frac{\partial F_i(w,S)}{\partial S} \end{aligned} \tag{3.13}$$

where η is the step size. The algorithm of the PU shapelet learning model is shown in Algorithm 2:

Algorithm 2: PU shapelet learning

```
Input: the non-negative parameters $\alpha$, $\beta$, and $\gamma$
Number of minibatches N
Maximum number of epochs epoch
Initialization of $w$ and $S$
Step size $\eta$
Result: Optimal $w$ and $S$
for $ite = 1 : epoch$ do
    for $batch = 1 : N$ do
        update $w$
        update $S$
    end
end
```

3.5 Experimental study

In this section, we demonstrate the experiment results for the different data sets to validate the effectiveness of the proposed PUSL model, and compare PUSL with the state-of-art SSSL method. Experiments of three data sets from UCR time series repository are shown in section 3.5.1, and section

Table 3.1: Data set description

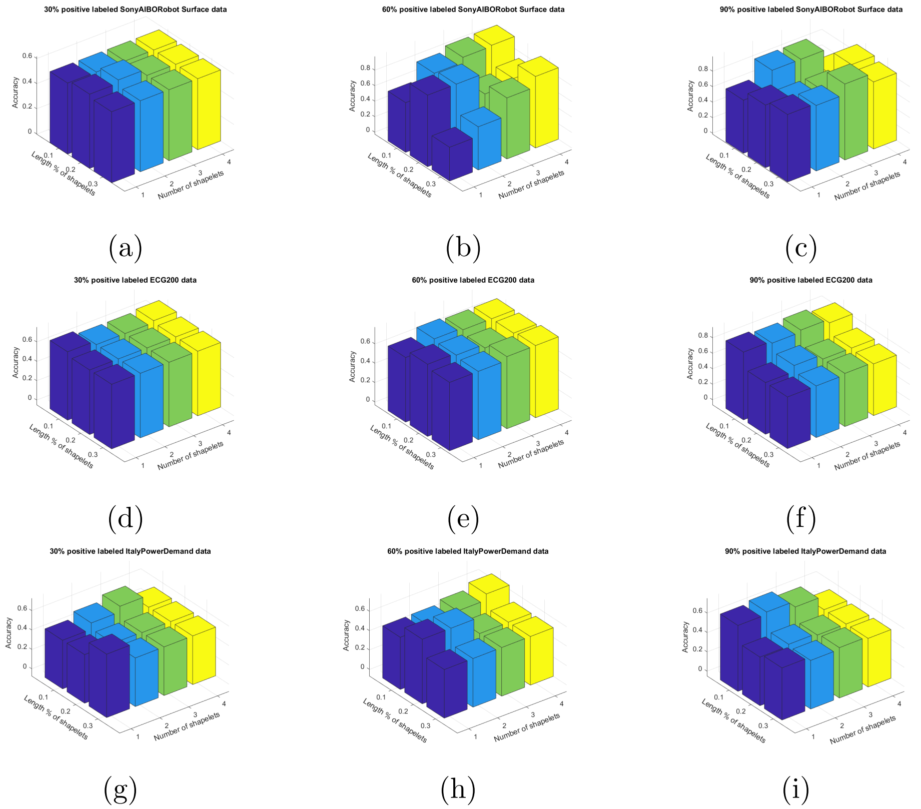
Data set	Size of data set	Length
SonyAIBORobot Surface	621	70
ECG200	200	96
ItalyPowerDemand	1096	24

3.5.2 displays the experiments for the ECG anomaly detection from MIT-BIH database.

3.5.1 Real-world data sets from the UCR time series repository

We evaluate our proposed PUSL algorithm on the three publicly available real-world data sets from the UCR time series repository [18], and the details of the size of data and length of each time series are shown in Table 3.1. In the experiments, we use the 75% for the training and 25% left for the testing. For each training data set, we randomly treat 30%, 60% and 90% positive data as labeled and leave the remaining 70%, 40% and 10% positive data as well as all the negative data as unlabeled. The performance of the proposed algorithm and state-of-art algorithm is measured by the classification accuracy: $(tp + tn)/(tp + fp + tn + fn)$, where tp is true positive, fp is false positive, tn is true negative, fn is false negative. For each data set, we repeated all experiments 10 times, and the best result is reported.

Figure 3.2 shows the classification accuracy of each data set based on the different parameters: the number of the shapelets and the length percentage of the shapelet. For each data set, we did the experiment with the 0.1, 0.2 and 0.3 of the time series length on the four different shapelets. The learning



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Figure 3.2: Parameter study based on the classification accuracy of PUSL with respect to the number of the shapelets and length percentage of the shapelets

rate is set to 0.01 and the number of the non-overlapped minibatches is set to 40. The parameters in the objective function are searched from 10^{-4} to 10^{4}. As shown in Figure 3.2, the algorithm can perform well for each dataset with the small length percentage of the shapelets and less number of shapelets.

Table 3.2 shows the comparisons of the classification accuracy for each data set on the different label ratios between proposed PUSL and SSSL. For the comparison SSSL method, the parameters of the objective function are set as the paper had, and the length percentage of the shapelet and the number of the shapelet are set as our experiment: 0.1, 0.2, 0.3 for the length percentage

Table 3.2: Comparison of the classification accuracy between PUSL and SSSL

Data set	Ratio of positive labeled	Best accuracy achieved	
		SSSL	PUSL (proposed)
SonyAIBORobot Surface	30%	**0.7051**	0.6154
	60%	0.8269	**0.9551**
	90%	0.8846	**0.9808**
ECG200	30%	0.6863	**0.7059**
	60%	0.7059	**0.7451**
	90%	0.7255	**0.8750**
ItalyPowerDemand	30%	**0.6982**	0.6545
	60%	**0.6945**	0.6667
	90%	**0.7127**	0.7055

of shapelet, and four different shapelets. As shown Table 3.2, the proposed algorithm performs better for the ECG200 data set with the three different ratios of the positive labeled; for the SonyAIBORobot Surface data set, the proposed method performs better for the 60% and 90% positive labeled. For the ItalyPowerDemand dataset, the length of each time series is 24, and the SSSL performs better., which indicate that our proposed method may not be good for the short time series.

Figure 3.3 shows the classification accuracy of the proposed PUSL algorithm for each data set with respect to the different ratios of the positive labeled data, and the accuracy is increasing as the ratio of the positive labeled data increases, which demonstrate the effectiveness of our proposed PUSL algorithm.

3.5.2 MIT-BIH database

We consider another publicly available real-world data sets from the MIT-BIH ECG arrhythmia database from PhysioNet [33, 34]. We evaluate the proposed PUSL algorithm with two records 200 and 221. Each record has a 30-min

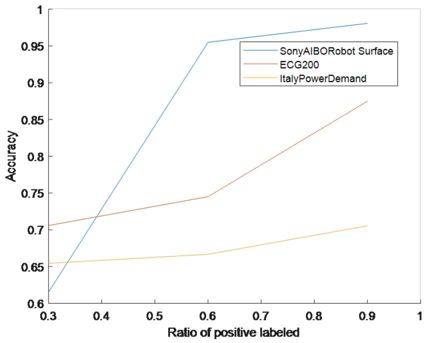


Figure 3.3: Classification accuracy of the proposed PUSL method with respect to the three different ratios of positive labeled data

two-channel ECG recording. We apply one channel modified limb lead II for these two records. As AAMI recommended practice [35] said, we use the first five minutes of the signal for training, and the rest 25 minutes are for testing. Each ECG cycle is segmented based on its R peak marked in the database. Each ECG cycle has the fixed length 255, and includes the 0.25 seconds length of ECG recording before the detected R peak and the 0.45 seconds after the R peak. Discrete wavelet transform db2(2) is applied for each ECG cycle segmentation to remove the noise. The length of the final cycle segmentation is fixed to 66.

The first 10000 datapoints of Record 221 are shown in Figure 3.4. There are two beat types: normal beat and abnormal beat (premature ventricular contraction). As shown in Figure 3.5, left plot shows one shapelet learned for Record 221, and last two plots show the pattern of two classes from the training data. We can see the shapelet captures the difference between normal and abnormal cycles.

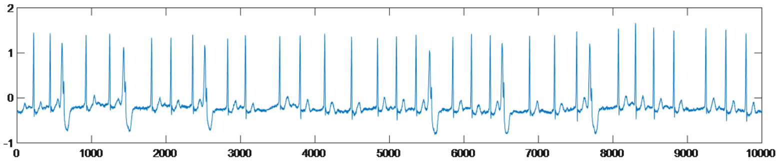
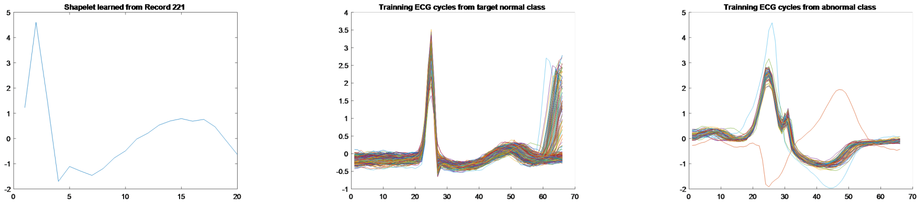


Figure 3.4: Part of ECG signal (first 10000 points) for Record 221



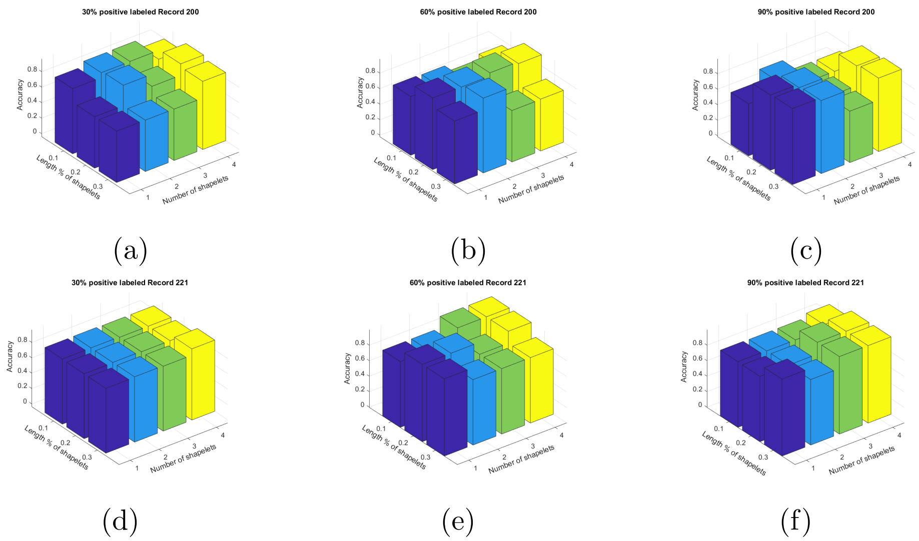
(a) (b) (c)

Figure 3.5: Training data pattern and one shapelet learned for Record 221

Figure 3.6 shows the parameter study of the ECG signal for Record 200 and Record 221 with three different lengths on four different shapelets, and the PUSL method can achieve high accuracy even with the 60% positive labeled data. Table 3.3 compares the classification accuracy between the proposed PUSL and state-of-art method SSSL, and the proposed PUSL performs better for the three different positive labeled ratios.

Table 3.3: Comparison of the classification accuracy between PUSL and SSSL for Record 200 and 221

Data set	Ratio of positive labeled	Best accuracy achieved	
		SSSL	PUSL (proposed)
Record 200	30%	**0.9492**	0.9275
	60%	0.9548	**0.9654**
	90%	0.9719	**0.9806**
Record 221	30%	0.8673	**0.9262**
	60%	0.9282	**0.9995**
	90%	0.9678	**0.9995**



(a) (b) (c)

(d) (e) (f)

Figure 3.6: Parameter study based on the classification accuracy of PUSL with respect to the number of the shapelets and length percentage of the shapelets for Record 200 and 221

3.6 Conclusions and future work

In this paper, we integrated the shapelet features into the positive and unlabeled learning setting, and based on the previous research model LLSVM, we proposed the PUSL algorithm. The experiment results on the real data sets demonstrate the effectiveness of our proposed algorithm. We compared our algorithm with the state-of-art method SSSL on the public available data sets, and it shows that our PUSL can achieve better in most of the data sets. Especially for the ECG data set, the proposed PUSL model performs well even with the lower positive labeled ratio, and then it is effective to apply our algorithm for the ECG anomaly detection. Since we only verify our method with ECG anomaly detection and benchmark time series data sets, then in the future, we will do more experiments with the different applications. Our algorithm focuses on the univariate time series data, and in the future, we may apply this with the multivariate time series data.

Bibliography

[1] Fabien Letouzey, François Denis, and Rémi Gilleron. Learning from positive and unlabeled examples. In *International Conference on Algorithmic Learning Theory*, pages 71–85. Springer, 2000.

[2] Huihui Wei and Ming Li. Positive and unlabeled learning for detecting software functional clones with adversarial training. In *IJCAI*, pages 2840–2846, 2018.

[3] Noah Youngs, Duncan Penfold-Brown, Richard Bonneau, and Dennis Shasha. Negative example selection for protein function prediction: the nogo database. *PLoS computational biology*, 10(6):e1003644, 2014.

[4] Wenkai Li, Qinghua Guo, and Charles Elkan. A positive and unlabeled learning algorithm for one-class classification of remote-sensing data. *IEEE transactions on geoscience and remote sensing*, 49(2):717–725, 2010.

[5] Huayi Li, Zhiyuan Chen, Bing Liu, Xiaokai Wei, and Jidong Shao. Spotting fake reviews via collective positive-unlabeled learning. In *2014 IEEE international conference on data mining*, pages 899–904. IEEE, 2014.

[6] Bing Liu, Wee Sun Lee, Philip S Yu, and Xiaoli Li. Partially supervised

classification of text documents. In *ICML*, volume 2, pages 387–394. Sydney, NSW, 2002.

[7] Hwanjo Yu, Jiawei Han, and Kevin Chen-Chuan Chang. Pebl: positive example based learning for web page classification using svm. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 239–248, 2002.

[8] Xiaoli Li and Bing Liu. Learning to classify texts using positive and unlabeled data. In *IJCAI*, volume 3, pages 587–592. Citeseer, 2003.

[9] Bing Liu, Yang Dai, Xiaoli Li, Wee Sun Lee, and Philip S Yu. Building text classifiers using positive and unlabeled examples. In *Third IEEE International Conference on Data Mining*, pages 179–186. IEEE, 2003.

[10] Fantine Mordelet and J-P Vert. A bagging svm to learn from positive and unlabeled examples. *Pattern Recognition Letters*, 37:201–209, 2014.

[11] Wee Sun Lee and Bing Liu. Learning with positive and unlabeled examples using weighted logistic regression. In *ICML*, volume 3, pages 448–455, 2003.

[12] Charles Elkan and Keith Noto. Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 213–220, 2008.

[13] Jessa Bekker and Jesse Davis. Estimating the class prior in positive and unlabeled data through decision tree induction. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[14] Akihiro Yamaguchi and Takeichiro Nishikawa. One-class learning time-series shapelets. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 2365–2372. IEEE, 2018.

[15] Haishuai Wang, Qin Zhang, Jia Wu, Shirui Pan, and Yixin Chen. Time series feature learning with labeled and unlabeled data. *Pattern Recognition*, 89:55–66, 2019.

[16] Shen Liang, Yanchun Zhang, and Jiangang Ma. Pu-shapelets: Towards pattern-based positive unlabeled classification of time series. In *International Conference on Database Systems for Advanced Applications*, pages 87–103. Springer, 2019.

[17] Lexiang Ye and Eamonn Keogh. Time series shapelets: a new primitive for data mining. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 947–956, 2009.

[18] Yanping Chen, Eamonn Keogh, Bing Hu, Nurjahan Begum, Anthony Bagnall, Abdullah Mueen, and Gustavo Batista. The ucr time series classification archive, July 2015. `www.cs.ucr.edu/~eamonn/time_series_data/`.

[19] Chen Gong, Tongliang Liu, Jian Yang, and Dacheng Tao. Large-margin label-calibrated support vector machines for positive and unlabeled learning. *IEEE transactions on neural networks and learning systems*, 30(11): 3471–3483, 2019.

[20] Abdullah Mueen, Eamonn Keogh, and Neal Young. Logical-shapelets:

an expressive primitive for time series classification. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1154–1162, 2011.

[21] Thanawin Rakthanmanon and Eamonn Keogh. Fast shapelets: A scalable algorithm for discovering time series shapelets. In *proceedings of the 2013 SIAM International Conference on Data Mining*, pages 668–676. SIAM, 2013.

[22] Qing He, Fuzhen Zhuang, Tianfeng Shang, Zhongzhi Shi, et al. Fast time series classification based on infrequent shapelets. In *2012 11th international conference on machine learning and applications*, volume 1, pages 215–219. IEEE, 2012.

[23] Kai-Wei Chang, Biplab Deka, Wen-Mei W Hwu, and Dan Roth. Efficient pattern-based time series classification on gpu. In *2012 IEEE 12th International Conference on Data Mining*, pages 131–140. IEEE, 2012.

[24] Josif Grabocka, Nicolas Schilling, Martin Wistuba, and Lars Schmidt-Thieme. Learning time-series shapelets. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 392–401, 2014.

[25] Qin Zhang, Jia Wu, Hong Yang, Yingjie Tian, and Chengqi Zhang. Unsupervised feature learning from time series. In *IJCAI*, pages 2322–2328. New York, USA, 2016.

[26] Jason Lines, Luke M Davis, Jon Hills, and Anthony Bagnall. A shapelet

transform for time series classification. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 289–297, 2012.

[27] Shantanu Jain, Martha White, and Predrag Radivojac. Estimating the class prior and posterior from noisy positives and unlabeled data. *Advances in neural information processing systems*, 29:2693–2701, 2016.

[28] Arun Iyer, Saketha Nath, and Sunita Sarawagi. Maximum mean discrepancy for class ratio estimation: Convergence bounds and kernel selection. In *International Conference on Machine Learning*, pages 530–538. PMLR, 2014.

[29] Mingming Gong, Kun Zhang, Tongliang Liu, Dacheng Tao, Clark Glymour, and Bernhard Schölkopf. Domain adaptation with conditional transferable components. In *International conference on machine learning*, pages 2839–2848. PMLR, 2016.

[30] Harish Ramaswamy, Clayton Scott, and Ambuj Tewari. Mixture proportion estimation via kernel embeddings of distributions. In *International conference on machine learning*, pages 2052–2060. PMLR, 2016.

[31] Marthinus Christoffel, Gang Niu, and Masashi Sugiyama. Class-prior estimation for learning from positive and unlabeled data. In *Asian Conference on Machine Learning*, pages 221–236. PMLR, 2016.

[32] Olivier Chapelle and Alexander Zien. Semi-supervised classification by

low density separation. In *International workshop on artificial intelligence and statistics*, pages 57–64. PMLR, 2005.

[33] George B Moody and Roger G Mark. The impact of the mit-bih arrhythmia database. *IEEE Engineering in Medicine and Biology Magazine*, 20 (3):45–50, 2001.

[34] Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, 101(23):e215–e220, 2000.

[35] R Mark. Aami-recommended practice: Testing and reporting performance results of ventricular arrhythmia detection algorithms. *Association for the Advancement of Medical Instrumentation, Arrhythmia Monitoring Subcommittee, AAMI ECAR, 1987*, 1987.

CHAPTER 4

Conclusion

In this research, we focused on the shapelet-based features for time series classification, and proposed two different methods to deal with the healthcare data. Because of the rapid development of computing and sensing technology, the Internet of Things (IoT)–enabled monitoring plays a crucial role for people suffering from the cardiac problems, and it is important to detect the abnormal ECG cycles during the cardiac monitoring. We extracted the statistically significant shapelets from the cycle-based ECG data, and applied the shapelet-based features with the support vector data description algorithm to construct the control charts in statistical process control for the IoT-enabled cardiac monitoring. The different methods are experimentally studied, and the comparison results demonstrate the effectiveness of the proposed shapelet-based SVDD control chart method. The experimental results on the real-world MIT-BIH dataset show that our proposed method is comparable with the state-of-art method.

Since it is costly to collect a large amount of the labeled data in the real-world applications, then we face the problem to deal with the situation that only a small set of positive and unlabeled data are available, which is called the positive and unlabeled (PU) learning. Based on the previous large-margin label-calibrated support vector machines (LLSVM) model, we integrated the shapelet-based feature into it, and learned the optimal shapelets to achieve high accuracy in time series classification. We compared with the state-of-art semi-supervised shapelet learning (SSSL) method on the benchmark time series data sets, and the experiment results demonstrate the effectiveness of our proposed positive and unlabeled shapelet learning (PUSL) method. We also did the experiment with the real-world MIT-BIH dataset, and our PUSL method can performs well even with the lower labeled data.