A WINDOW-BASED CHARACTERIZATION METHOD FOR BIOPHYSICAL TIME SERIES

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF INFORMATICS
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

DENİZ KATIRCIOĞLU

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR
THE DEGREE OF DOCTOR OF PHILOSOPHY
IN
MEDICAL INFORMATICS

MAY 2017

A WINDOW-BASED CHARACTERIZATION METHOD FOR
BIOPHYSICAL TIME SERIES

Submitted by **DENİZ KATIRCIOĞLU** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy in Medical Informatics, Middle East Technical University** by,

Prof. Dr. Deniz Zeyrek Bozşahin
Director, **Graduate School of Informatics**   _____

Assoc. Prof. Dr. Yeşim Aydın Son
Head of Department, **Health Informatics**   _____

Prof. Dr. Nazife Baykal
Supervisor, **Information Systems**   _____

Examining Committee Members:

Assoc. Prof. Dr. Yeşim Aydın Son
Health Informatics, Middle East Technical University   _____

Prof. Dr. Nazife Baykal
Information Systems, Middle East Technical University   _____

Assist. Prof. Dr. Aybar C. Acar
Health Informatics, Middle East Technical University   _____

Prof. Dr. H. Altay Güvenir
Computer Engineering, Bilkent University   _____

Assist. Prof. Dr. Öznur Taştan Okan
Computer Engineering, Bilkent University   _____

Date: _____

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name: DENIZ KATIRCIOĞLU

Signature   :

ABSTRACT

A WINDOW-BASED CHARACTERIZATION METHOD FOR BIOPHYSICAL TIME SERIES

Katırcıoğlu, Deniz

Ph.D., Department of Medical Informatics

Supervisor   : Prof. Dr. Nazife Baykal

May 2017, 96 pages

In thesis, we propose a robust similarity score-based time series characterization method, termed as **W**indow-based **T**ime series **C**haracterization (**WTC**). Specifically, WTC generates domain-interpretable results and involves remarkably low computational complexity thereby rendering itself useful for densely sampled and populated time series datasets. In this study, we apply WTC to a proprietary action potential (AP) time series dataset on human cardiomyocytes and three precordial leads from a publicly available electrocardiogram (ECG) dataset. We, then, compare WTC with shapelet transform and fast shapelet transform (which constitutes an accelerated variant of the former), in terms of predictive accuracy and computational complexity. The results indicate that WTC achieves a slightly higher classification performance with significantly lower execution time when compared to its shapelet-based alternatives. With respect to its characterization capability, WTC has a potential to enable medical experts to explore definitive common trends in novel datasets.

Keywords: time series analysis, feature extraction, cardiac action potential, atrial fibrillation, electrocardiography

iv

ÖZ

BİYOFİZİKSEL ZAMAN SERİLERİ İÇİN PENCERE TABANLI NİTELEME YÖNTEMİ
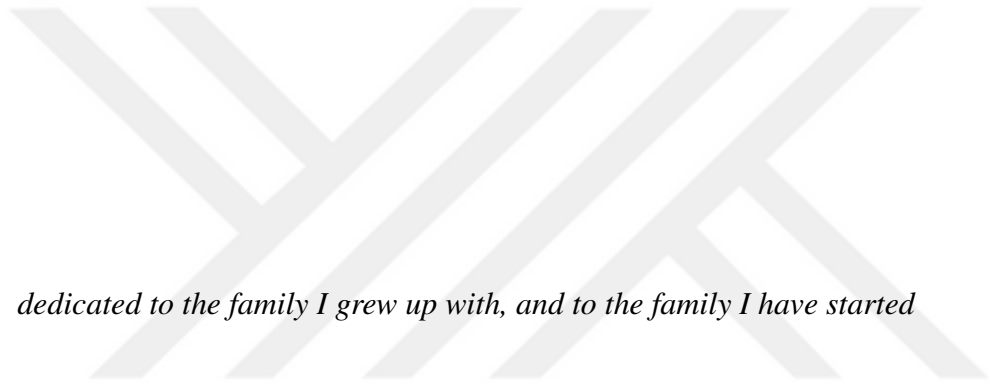
Katırcıoğlu, Deniz

Doktora, Tıp Bilişimi Bölümü

Tez Yöneticisi   : Prof. Dr. Nazife Baykal

Mayıs 2017, 96 sayfa

Bu tezde, WTC adında, benzerlik puanı tabanlı gürbüz bir zaman serisi niteleme yöntemi önerilmektedir. WTC, etkinlik alanı kullanıcıları tarafından yorumlanabilir sonuçlar üretir. Kayda değer düşüklükte hesaplama karmaşıklığına sahip olması sebebiyle de, yoğun örneklenmiş ve kalabalık zaman serisi veri kümeleri için uygundur. WTC, insan kardiyomiyositlerinden edinilen özel mülkiyetli bir aksiyon potansiyeli zaman serisi veri kümesine ve genel erişime açık, üç adet prekordiyal derivasyondan oluşan bir EKG veri kümesine uygulanmıştır. Sonrasında WTC, sınıflandırma doğruluğu ve hesaplama karmaşıklığı açısından, şekilcik dönüşümü ve bu dönüşümün hızlandırılmış bir türevi olan hızlı şekilcik dönüşümü metotlarıyla karşılaştırılmıştır. Sonuç olarak, WTC'nin şekilcik tabanlı alternatiflerine göre kayda değer düşüklükte hesaplama karmaşıklığıyla beraber nispeten yüksek sınıflandırma performansı elde ettiği gösterilmiştir. Niteleme yeteneği sayesinde, WTC, tıbbi uzmanların yeni zaman serisi veri kümelerinde tanımlayıcı ortak eğilimleri incelemesini kolaylaştırma potansiyeline sahiptir.

Anahtar Kelimeler: zaman serisi analizi, öznitelik çıkarımı, kardiyak aksiyon potansiyeli, atriyal fibrilasyon, elektrokardiyogram

dedicated to the family I grew up with, and to the family I have started

ACKNOWLEDGMENTS

TABLE OF CONTENTS

LIST OF TABLES

xi

LIST OF FIGURES

LIST OF ABBREVIATIONS AND ACRONYMS

AdaBoost   Adaptive Boosting

ADTree   Alternating Decision Tree

AF   Atrial Fibrillation

AP   Action Potential

APD20   AP Duration at 20% of Repolarization

APD90   AP Duration at 90% of Repolarization

CART   Simple Classification and Regression Tree

CD   Critical Difference

DCT   Discrete Cosine Transform

ECG   Electrocardiogram

END   Ensemble of Nested Dichotomies

EUTRAF   The European Network for Translational Research in Atrial Fibrillation

ICD-10   International Statistical Classification of Diseases and Related Health Problems - 10th Revision

LWL   Locally Weighted Learning

MI   Myocardial Infarction

NR   Normal Rhythm

PAA   Piecewise Aggregate Approximation

PART	Partial Decision Tree Classifier
PIP	Perceptually Important Point
PLT	Plateau Potential
PMF	Probability Mass Function
PTBDB	Physikalisch-Technische Bundesanstalt DataBase
RIDOR	Rippledown Rule Learner
RIMARC	Ranking Instances by Maximizing the Area under the ROC Curve
RMP	Resting Membrane Potential
SAX	Symbolic Aggregate Approximation
SCB	Shape Confidence Band
SR	Sinus Rhythm
STEMI	ST Segment Elevation Myocardial Infarction
WEKA	Waikato Environment for Knowledge Analysis
WTC	Window-based Time Series Characterization

CHAPTER 1

INTRODUCTION

1.1 Motivation and Objectives of the Thesis

In the medical domain, biophysical signals in the form of time series are frequently used for diagnostic and prognostic purposes and are specifically relevant to documenting the history [1] and clinical course of a disease [2], [3]. Many decisive traits emerge from biophysical signals as rules of thumb suggested by health professionals based on a visual inspection [4]. Conversely, computer-aided methods extract common patterns among time series and establish a more objective data assessment framework. Thus, they are of immense practical value in interpreting data with respect to diagnostic, prognostic, and therapeutic perspectives [5].

Computerized analysis of medical data is an active and inter-disciplinary research area. Moreover, data mining studies involving time series data, whether it is medical or not, demand efficient extraction of powerful features. Medical experts, on the other hand, are expected to interact with and contribute to this feature selection process better once the selected features are "human readable". We define human readability as the convenience in the interpretation of the extracted features by the medical experts who are capable of and competent in inspecting the time series data of interest as suggested by their conventional medical training. On the other hand, computational complexity of any feature extraction method operating on time series data potentially suffers from the curse of dimensionality as the number of sample points and the number of time series instances increase.

The results of this thesis are expected to have beneficial impact on the research community interested in characterization of and feature extraction from medical time series. As such; we hope that addressing the aforementioned points about medical time series, this study will yield a method recognized as an essential analysis tool for the experts.

1.2 Contribution of the Thesis

In this section, after reviewing the literature about time series analysis methods from a medical domain perspective, we explain the contribution of this thesis.

Recent all-purpose time series analysis studies have revealed a wide variety of approaches borrowing techniques from computational geometry, Gaussian process modeling, morphology analysis, template matching, scale space representation, shapelet discovery, etc. The choice of approach for the analysis of a particular time series depends on the desired representation of the time series data [6] or the domain (spatial or spectral) in which the decomposition is to be performed.

In the computational geometry domain, it is assumed that "B-spline" constitutes one of the most efficient surface representation methods [7], [8]. Significant curve reconstruction efforts focused on using various spline paradigms [9], [10] especially in the cardiac domain [11]. The proposed methods are usually based on the continuous B-Spline representation for describing the curve instances in continuous-time rather than discrete-time for which (re-)sampled points of the curve instances are used. Although continuous B-spline representation or other types of splines mostly avoid the sampling errors and achieve a high level of accuracy, using them for describing biophysical signals to assess the differences that are indicative of a certain condition is not very suitable, as they do not produce human-readable and summarizing models.

Another concept from computational geometry is the "turning functions" examined in [12]. This approach is motivated by the fact that any digitized curve can be regarded as a polygon (possibly with a large number of vertices) without loss of information. The possible distortions can be obviated by approximating the original curve with the one having a similar perceptual appearance. Three types of free matching methods, namely, plain, polygonal and penalty, that use the turning functions are discussed for curve representations. These methods basically search for the best "angle of orientation" between the curves in order for their turning function distances to be the smallest. In the case of biophysical time series, however, finding the best angle is not a forefront necessity and the observation that the method may exhibit poor performance with "near-to-closed curves" and especially for curves with strong turns in different directions limits the use of turning functions in many real-life cases [12]. In addition to that, a scaling effort is deemed necessary with the cases of scale mismatch. To remedy such cases, a scaling step is performed on the basis of curve length. However, this step only achieves its goal when a priori knowledge of scale of the curve is available.

"Correspondence-based matching" is another approach for designing a time series descriptor [13]. This method incorporates each point in a time series as a candidate feature, potentially resulting in an extremely large feature set. Since all points are assigned as features, performance of this method is undesirably sensitive to noise in the data. For the matching part of this feature set, Hausdorff distance is used [14]. An advantage of using Hausdorff distance is that time series instances can be partially matched. Nevertheless, exact matching is computationally expensive due to the invariance in translation, scale and rotation.

Another widely used descriptor is the "convex hull" which is the minimal convex cover of a shape. In the extraction of the convex hull, both boundary tracing methods and morphological methods can be used [15]. Convex deficiency, a concept brought by the convex hull, is a useful approach for finding similar features among boundaries of the curves and shapes. It is defined as the set difference between the convex hull and the boundary it encompasses [16]. Simply by following the trajectory of a curve's

2

boundary and determining points falling into the area of convex deficiency, features for establishing similarity criteria among different curves, such as the points of deviation, can be extracted. Then, curves become eligible to be classified by the selected features in their "convex deficiency trees" [17]. Irregularities in signal digitization, possible artifacts and noise that are inherent in the majority of the biophysical signals make this approach and its utilities quite misleading in terms of curve classification owing to their sensitivity to the boundary features.

In addition to the convex hull, contour descriptors such as "circular" and "elliptical variances" are apt for comparing a time series with a predetermined template [18]. They are elements of directional statistics [19] where circular variance is defined as the variance of pointwise deflection of all points on a shape's contour from a circle with the same area as the shape of interest. In a similar fashion, elliptical variance is the variance of pointwise deflection from an ellipse having an equal covariance matrix with the same shape. These classic features are used for irregularity comparison and defect classification [20]. Such simple circular and elliptical templates, however, lack in generalizing biophysical time series with highly irregular contours. For this particular case, a suitable approach might be piecewise template generation with various features which in turn would increase the computational complexity [20].

Furthermore, the "scale space" method is proposed with the purpose of extracting hierarchical traits out of the curves [21]. The scale space representation of a curve is created by tracking the position of inflection points along the curve filtered by low-pass Gaussian filters of variable widths. The wider the Gaussian filter, the more the inflections eliminated from the boundary and smoother the shape becomes. The inflection points that remain present in the representation are expected to be "significant" object characteristics, defined as corners, smooth joints, ends, cranks, and bumps [22]. Even though this method requires a threshold setting to be given in order to produce an intended degree of characterization, it stands as a successful summarization tool.

Various "symbolic curve matching" paradigms listed in [6] deal the time series in spectral domain whereas some stay strictly in time domain. They are basically proposed to improve signal representation just as the local and piece-wise polynomial models constructed for an efficient representation of time series datasets [23], [24].

Recently, Gaussian Process modeling has gained attraction for analyzing time series, especially in the medical domain. Within a Bayesian framework, [25] makes an analysis by solving a regression problem assuming a dataset consisting of observations and outcome variables. The analysis considers first a function mapping the observed data to an outcome variable regardless of the ordinality of the observed data. Then, a curve fitting step is taken assuming the outcome variable is controlled by the observed data in order to extrapolate the hypothesized curve based on the observed data. This approach is suitable for predicting the missing outcome values of biological time series which often suffer from the irregularities in sampling and missing observations [25], [26].

Another characterization method has recently been developed to discover significant subsequences called "shapelets" from a time series dataset [27], [28]. A shapelet is a time series subsequence identified as the representative of a certain dataset. Discover-

ing shapelets requires a brute-force traversal of all overlapping subsequences whose length and number parameters are left to the users of the method. Shapelet transform produces human readable outputs and therefore we use them for benchmarking purposes in this study. We elaborate on the shapelet transform in Section 2.3.

In this thesis, we address the problem of extracting definitive time-domain features from a given time series dataset to be used in data mining applications with the ultimate goal of devising a computationally efficient algorithm. We propose a time series characterization method called Window-based Time Series Characterization (WTC), which is the main contribution of this thesis, summarizing the class-dependent behavior within consecutive time windows and deriving an overall similarity score for this behavior. In this context, characterization refers to the effort of representing information content of time series data in a compact and graphical way to reveal behaviors (i.e. patterns or trends) that pertain to class labels. In a manner similar to other characterization approaches in [29], [30], [31], [32] , [33], [34] and [35], the aim of the proposed method involves identifying temporal features that define the character of instances in a certain target class that are later used to discriminate between instances among other classes. In order to demonstrate the effectiveness of the proposed method, we employ a human cardiac action potential (AP) dataset and an ECG dataset consisting of recordings from three different leads.

Cardiac APs are bioelectrical signals that are recorded from cardiac tissue obtained during heart surgery from patients with defined heart rhythms. Based on a patient's heart rhythm at the time of surgery, APs may either exhibit a "spike-and-dome" shape for physiological sinus rhythm (SR) or a triangular shape for atrial fibrillation (AF) [36]. AF is known to be the most common arrhythmia in clinical practice with an approximate prevalence of $0.4 - 1\%$ in the general population and is usually associated with stroke, heart failure, and a significant increase in all-cause mortality [37]. Patients with AF are diagnosed based on thorough clinical examinations and ECG recordings. The resulting AP signals, however, may exhibit SR characteristics depending on the stage of the disease. As a highly information bearing signal, cardiac AP attracts significant attention from the researchers primarily motivated to reveal dynamics of the cardiac ion channels governing the electrical activity within the heart. Understanding the dynamics of the ion channels is essential for identifying the effects of certain agents in related drug studies [36]. AP signal acquisition is a labor-intensive task appearing as a limiting factor for widespread studies especially in the area of statistical time series analysis. There have been studies producing desired AP properties by fitting empirical models, [9], in order to simulate the real-life behavior of AP. In this thesis, apart from validating the proposed model, AP dataset serves for the purpose of stress testing the proposed method for a dataset consisting of crowded and densely sampled time series. To the best of our knowledge, this is the first study in which a data mining analysis is conducted on a cardiac AP time series dataset.

Conversely, the ECG dataset, consisting of patients with acute myocardial infarction and control subjects, is specifically selected to present a clinical application of the proposed method. As a fatal cardiac disease, myocardial infarction (MI) corresponds to irreversible loss of the heart muscle due to ischemia caused by blood flow interruption [38]. Prominent guidelines recommend the use of ECGs especially for emergency cases in addition to certain cardiac biomarkers with their superior sensitivity to MI [39]. ECG is an effective and non-invasive technique for the confirmation

4

of MI diagnosis during initial evaluation. Particularly for ST elevation myocardial infarction (STEMI), 12-lead ECGs exhibit certain morphological changes in the expected waveform such as ST-segment elevations or depressions and wave inversions or losses in different leads depending on the localization of MI [40]. There are public repositories publishing anonymized digital datasets in the form of time series. "PhysioNet" is one such repository consisting of a large and balanced amount of MI and control subjects [41]. In this thesis, we study a particular selection of PTB Diagnostic ECG Database available in the PhysioNet repository [42].

1.3 Organization of the Thesis

The rest of the thesis is organized as follows: In Chapter 2, we provide background information about the biophysical signals and the benchmark methods studied in this thesis. In Chapter 3, we explain the proposed Window-based Time Series Characterization (WTC) method. Throughout Chapter 4, we present the results for the classification experiments performed for the proposed WTC method along with two other shapelet-based benchmark methods using the cardiac AP and the three-lead ECG time series datasets. In Chapter 4, we also compare computational complexities of these methods. The medical and technical aspects of the results generated in Chapter 4 are discussed followed by the concluding remarks in Chapter 5.

CHAPTER 2

BACKGROUND

In this chapter, background information about the key elements covered in the course of this thesis is given. The first two sections introduce the domains to which the two time series datasets used in the evaluation of the proposed method belong. Section 2.1 starts with the electrophysiological concept of "action potential" (AP) of the heart muscle cells and their crucial role in the autonomous contraction that physically creates the "heart rhythm" phenomenon. Following that, the rhythm disorder called atrial fibrillation and its brief etiology that accompanies the impairment of this sensitive electrophysiological mechanism are explained. Section 2.2 focuses on preliminary definitions of acute "myocardial infarction" (MI) and how the ECG examinations assist in diagnosing early to advanced stage infarcts, while describing expected morphologies in disease-related ECG leads. Finally, Section 2.3 provides detailed insight for the shapelet-based time series feature extraction methods used for benchmarking in the classification experiments throughout Chapter 4.

2.1 Cardiac Action Potentials (AP) in Atrial Fibrillation (AF)

Almost all eukaryotic organisms have voltage differences (transmembrane potential) across their cell membranes [43]. While the resting conditions are in effect (i.e., no stimulation), the "resting membrane potential" (RMP) values generally vary between -10 and -100 mV [44], [45]. This state of electrical polarization is maintained and governed by ion transporters and voltage-sensitive ion channels, situated in the cell membrane, that open and close in accordance with the changes in the voltage across the membrane [46]. Unlike the others, the nerve and muscle cells in animals exhibit brief directional changes in the polarization of their membranes, leading to the fluctuations in the voltage levels. These fluctuations create the cyclic phenomenon of AP that begins with a rapid take-off of the voltage difference (depolarization) followed by a rapid decline called repolarization. The cardiac AP is a particular type that belongs to the cell membrane of heart cells called "myocytes". Unlike APs in other tissues, the cardiac AP is not initiated by nervous activity [47]. In order to pump blood, all parts of the heart must contract in concert, and electrical excitation of the cardiomyocytes is a prerequisite for coordinated contraction. The first impulse is produced by the sinoatrial node (SA node) in the heart. SA node is made up of specialized pacemaker cells located murally on the right atrium of the heart [48]. From these pacemaker cells initiating the electrical activity, APs propagate throughout the heart
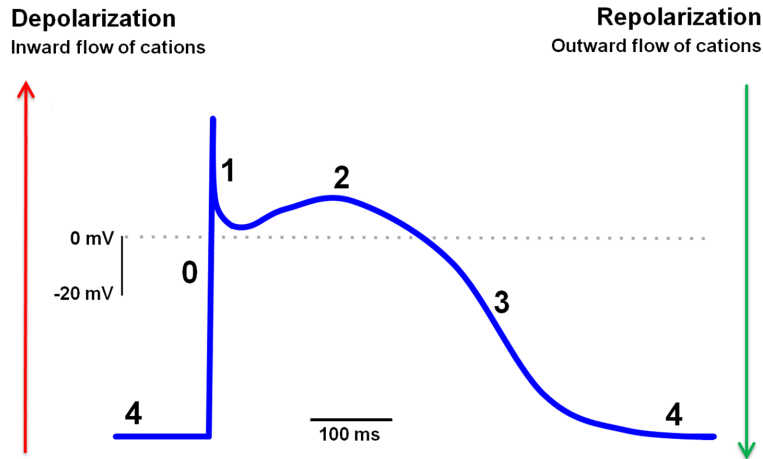
Figure 2.1 : Schematic illustration of a human cardiac AP in 5 phases, namely, phase 0: rapid depolarization; phase 1: early rapid repolarization; phase 2: "plateau"; phase 3: final repolarization; phase 4: resting membrane potential (adapted from [52]).

by gap junctions [49] which enable the passage from one cell to the next and stimulate the myocardium to trigger the contraction [50]. At rest, around 60 to 100 APs are generated per minute in humans [51].
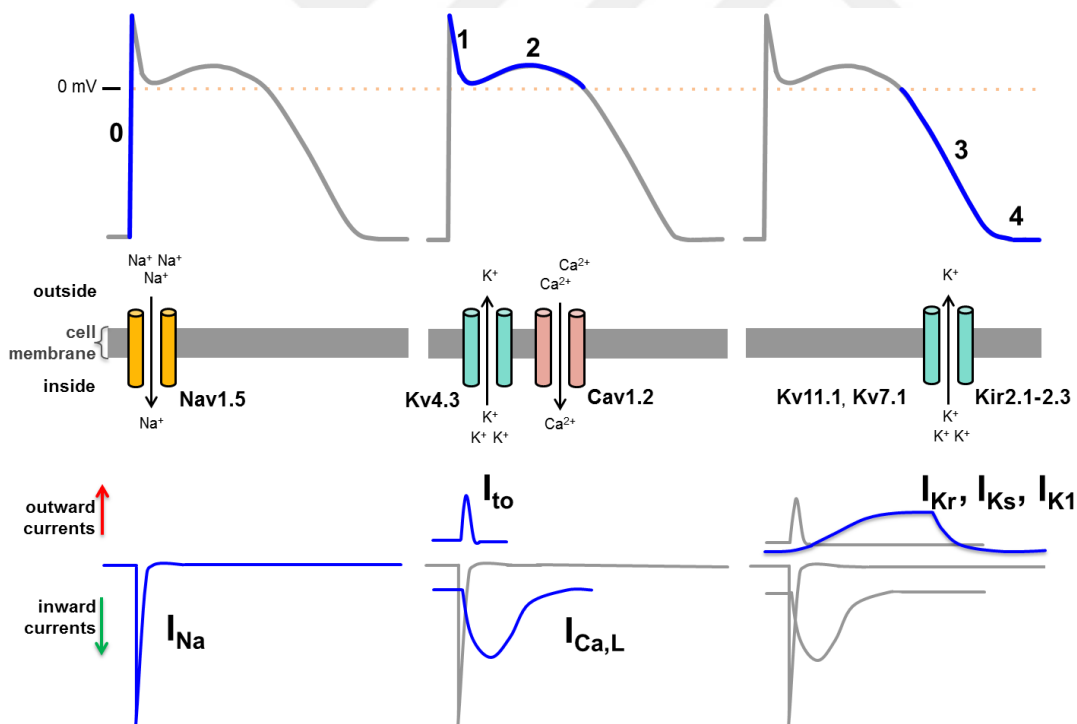
Figure 2.2 : Schematic of human cardiac APs (upper row), ion channels (middle row), and major ion current flows (lower row) (adapted from [52]). Nav1.5, cardiac Na^+ channel conducting Na^+ current (I_{Na}); Kv4.3, K^+ channel conducting transient outward current (I_{to}); Cav1.2, L-type Ca^{2+} channel conducting L-type Ca^{2+} current ($I_{Ca,L}$); Kv11.1, K^+ channel conducting rapidly activating, delayed outward rectifier K^+ current (I_{Kr}); Kv7.1, slowly activating outward rectifier K^+ current (I_{Ks}); Kir2.1, Kir2.3, inward rectifier K^+ current (I_{K1}).

Each cardiac AP exhibits distinct phases (see Fig. 2.1): the resting membrane potential (phase 4), the rapid depolarization ("upstroke", phase 0), early rapid repolarization (phase 1), the "plateau" (phase 2); final repolarization (phase 3) and return to the resting membrane potential (phase 4). Excitable cells possess an inside-negative resting membrane potential due to electrochemical gradients of the cations Na^+ and K^+ across the cell membrane maintained by ion pumps, and voltage-dependent ion channels which open and close in a voltage- and time-dependent manner (see Fig. 2.2). We note that, the plateau phase of the cardiac AP is more extended than that of the typical neuronal AP [53].

When the threshold for activation is reached by a small depolarization of the cell membrane, Na^+ channels open rapidly and allow Na^+ to enter the cell. The influx of positive charge further depolarizes the cell membrane which in turn depolarizes neighboring cardiomyocytes triggering a domino effect. Following this activation, repolarization is caused by efflux of K^+ through various K^+ channels of different kinetics. Influx of Ca^{2+} via L-type Ca^{2+} channels triggers the release of further Ca^{2+} from intracellular stores, which is required for activation of the contractile machinery. Thus, the shape of an AP is governed by the superposition of all membrane currents flowing through various ion channels depicted along the lower row of Fig. 2.2.

Depolarization and repolarization phases of cardiac myocytes are quite different from those of other cells. In subjects with normal sinus rhythm (SR), the fact that the repolarization phase lasts relatively longer inhibits rapid stimulation of cardiac cells, preventing rhythmic disorders of the heart [54]. One such disorder, atrial fibrillation (AF or sometimes AFib) is known to be one the most common arrhythmias of tachycardic nature with supraventricular (i.e., above the ventricles of the heart) origin. Frequently diagnosed with three types, namely, paroxysmal, persistent and permanent, AF is more prevalent among higher age groups [55] and is closely associated to mortality [56] and comorbidities [57] such as hypertension, diabetes, heart failure, chronic obstructive pulmonary disease (COPD), chronic renal failure, ischemic stroke and major haemorrhage [58]. During AF, the SA node malfunctions yielding irregularly fast atrial rates (at the upper chambers of the heart) around 400-600 bpm. Conduction of these AP impulses to the ventricles (lower chambers of the heart) eventually causes an irregular ventricular rate around 100 to 200 bpm [59].

During assessment, if any irregularity in the heart rhythm is felt by the patient him/herself or noticed by a medical professional upon checking the pulse, patient is led to have an ECG to confirm the primary diagnosis of AF [60], [61]. Specifically, typical ECG (see Fig. 2.3 for an example 12-lead ECG of an AF patient) of a patient having an AF episode shows variable heart rate, no obvious P waves but rather smaller fibrillatory F waves [62] and varying R-R intervals[63]. The increase in the atrial rate causes the generated APs to fail in reaching their expected amplitude, which in turn results in the absence of the P waves (see Fig. 2.6 for an annotated ECG waveform).

Atrial APs from SR patients have a characteristic "spike-and-dome" shape. During AF, electrical activity becomes very fast and uncoordinated, leading to remodeling processes that change - amongst others - the expression of ion channels in the atrial cardiomyocytes and give rise to the typical triangular-shaped AP reflecting the shortened AP duration and effective refractory period (see Fig. 2.4), [36], [65]).

9
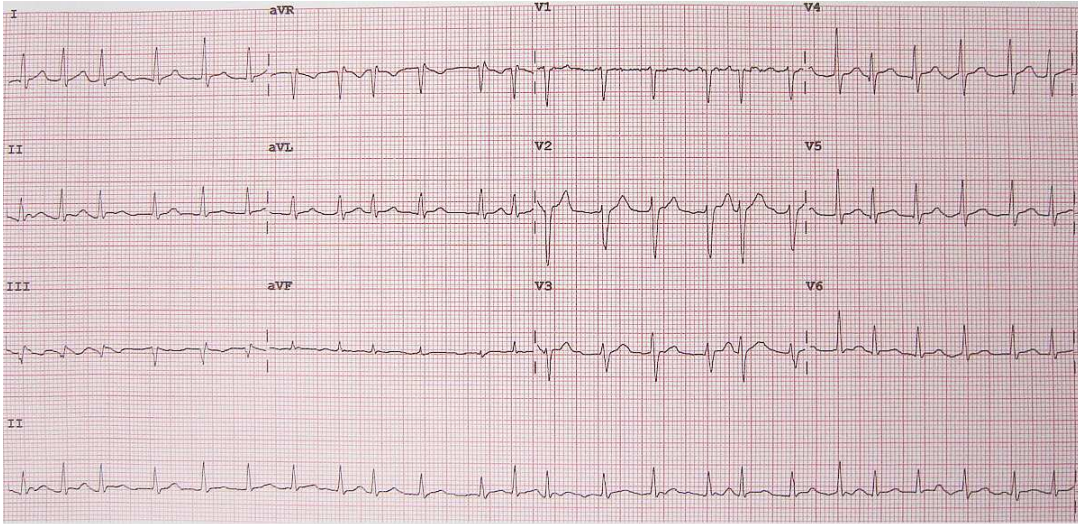
Figure 2.3 : An example ECG (by James Heilman, MD - own work [64], CC BY-SA 3.0) of a patient in AF.
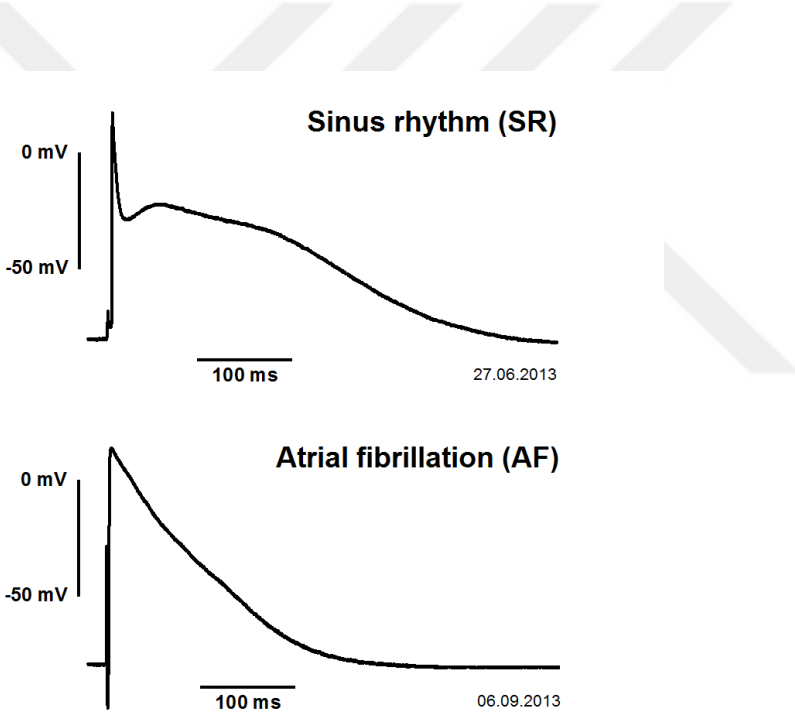
Figure 2.4 : Typical "spike-and-dome" shaped AP from right atrial tissue of an SR patient (top) and triangular shaped AP from an AF patient (bottom) (adapted from [66]).

2.2 Electrocardiograms (ECG) in Acute Myocardial Infarction (MI)

Acute MI (also AMI) corresponds to irreversible loss of the heart muscle due to ischemia caused by blood flow interruption [38]. This interruption is usually caused by a coronary blockage associated with intravenous plaque accumulation or embolism

[67]. Sometimes undiagnosed, MI can occur as a chronic lifelong condition, or as a severe event that causes sudden death [68]. One of the underlying causes of MI is known to be the coronary artery disease [69].

ECG is the most commonly used diagnostic tool to assess the presence and extent of infarction and its anatomical location [70]. It is a non-invasive and non-risk bearing cardiac procedure which basically records the electrical activity arising from the de-polarization and repolarization of the cardiac muscle. ECG is recorded from the body surface via probes fixed at several specific locations [71]. The ECG analysis provides a deeper understanding of the underlying dynamics of the cardiac pathophysiologies. As the magnitude and direction of the electrical field evolves in time, ECG captures the resulting signal and produces its tracing output. MI causes alterations in the generated electrical field that are observable in ECG traces. As depicted in Fig. 2.7 ECG is the superposition of the action potentials occurring during the contractions of cardiac myocytes.

Table 2.1 : Lead configurations in conventional 12-lead ECG

Lead Types	Lead Names
Bipolar limb leads (Einthoven)	I
	II
	III
Augmented leads (Goldberg)	AvL
	AvR
	AvF
Unipolar chest leads (Wilson)	v1
	v2
	v3
	v4
	v5
	v6

(a) Precordial (chest) leads   (b) Extremity leads
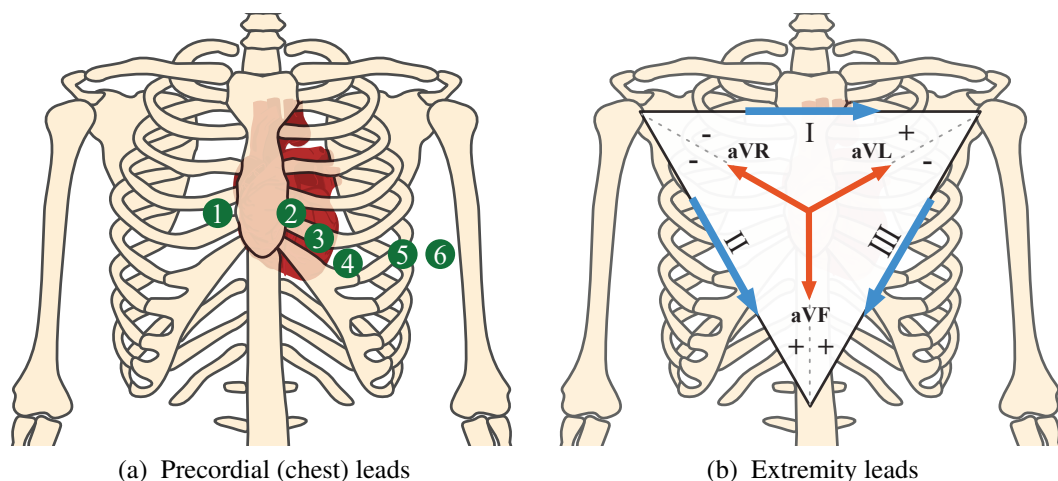
Figure 2.5 : Precordial (a) and extremity leads (b) in conventional 12-lead ECG (adapted from [72])

During a conventional 12-lead ECG, the electrical potential is measured from 12 dif-

ferent angles with 10 distinct probes mounted on the body surface (i.e., skin) as shown in Figs. 2.5a and 2.5b with the lead configuration in Table 2.1.

A single ECG cycle representing a single heart beat is made up of a typical P wave (reflects atrial depolarization), QRS complex (atrial repolarization and ventricular depolarization happening simultaneously) and a T wave (reflects ventricular repolarization) as depicted in Fig 2.6.
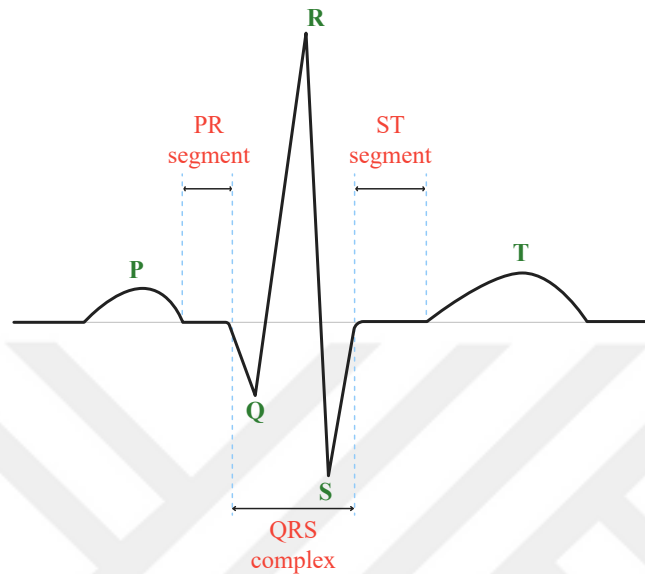
Figure 2.6 : An ECG depiction with annotated elementary waveforms and intervals

ECGs of emergency patients with signs indicating ST-segment elevation acute MI (STEMI) provide prognosis, therefore, help confirming the eventual diagnosis. Obviously, ECG eases the detection of life-threatening arrhythmias by enabling effective triage [74], [75]. As shown in Fig. 2.6, the ST segment reflecting the ventricular repolarization marks the specific interval starting from the completion of the S wave and ending before the onset of the T wave [76]. The ECGs performed on the patients with STEMI exhibit a progressively distinct pattern compared to those of the healthy patients. This progressively changing pattern usually begins with the appearance of hyper-acute (taller) T waves, soon to be followed by ST segment elevation and negative T waves. At later stages, pathological Q waves may be observed [77], even though such Q waves may also appear at the very beginning [40]. Pathological Q waves are not one of the earliest manifestations of MI. They rather indicate a possible prior history of MI for that patient [68]. In cases where ST elevation is not observed (as in NSTEMI), certain cardiac biomarkers such as troponin are examined to diagnose MI [78]. STEMI accounts for about 25 to 40 percent of acute coronary syndrome cases [39].

The abnormal patterns expected during STEMI in a 12-lead ECG are diversified depending on the localization of the myocardium [79]. For STEMI, with anterior or anteroseptal localizations, hyper-acute T waves either exist solely in leads V2-V4 or they are accompanied by the loss of R wave height (see Fig. 2.8b) in leads V2-V6 (predominantly in V2 and V3). For anterior localization, subtle ST segment elevation may be observed in leads I, aVL and V5, with reciprocal ST depression in lead III as

Sinus Node

Atrial Muscle

A-V Node

Common Bundle

Bundle Branches

Purkinje Fibers

Ventricular Muscle

Ventricular Muscle

R

P

T

Q

S
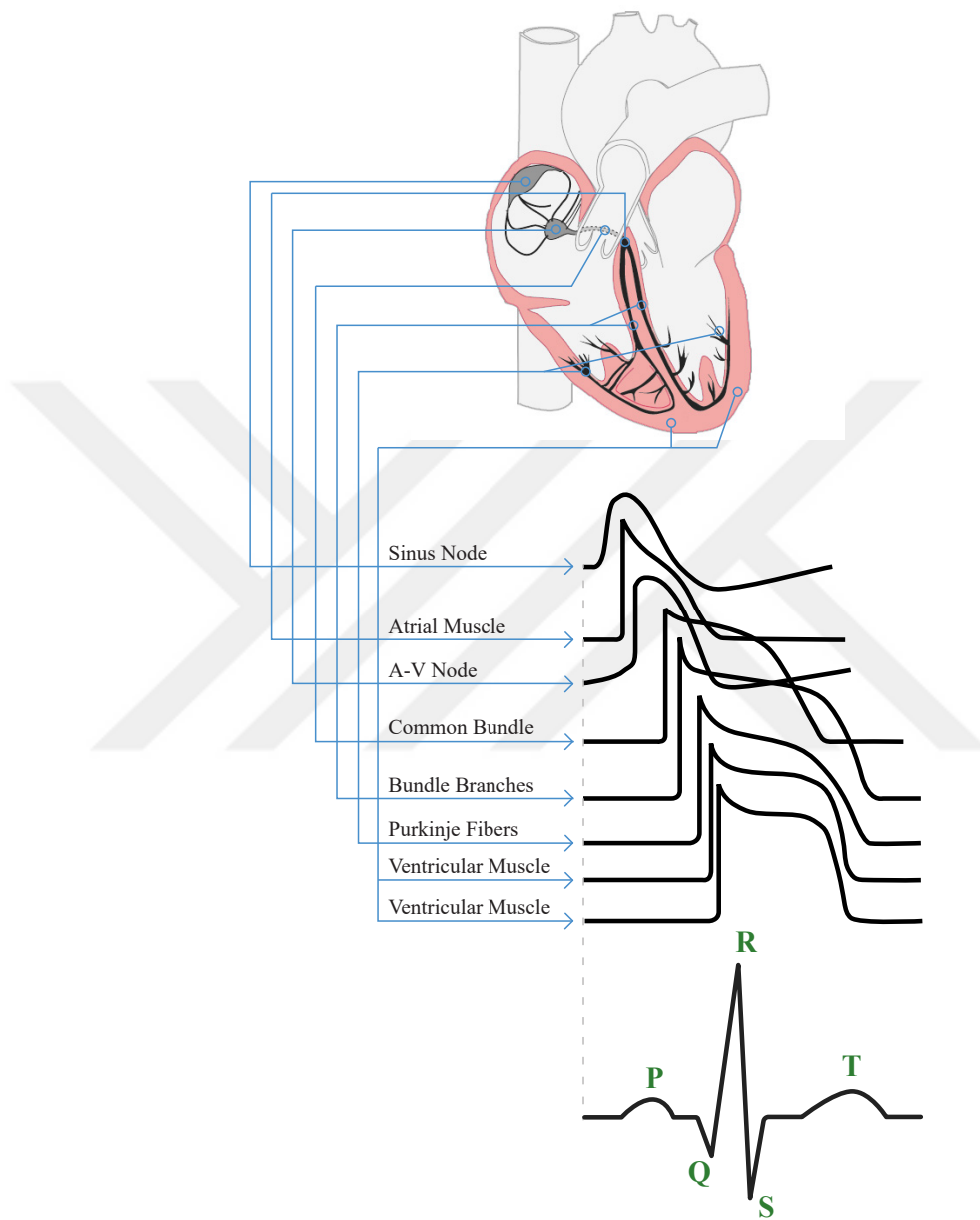
Figure 2.7 : Combined sum of APs from various parts of the heart forming the ECG signal (adapted from [73])
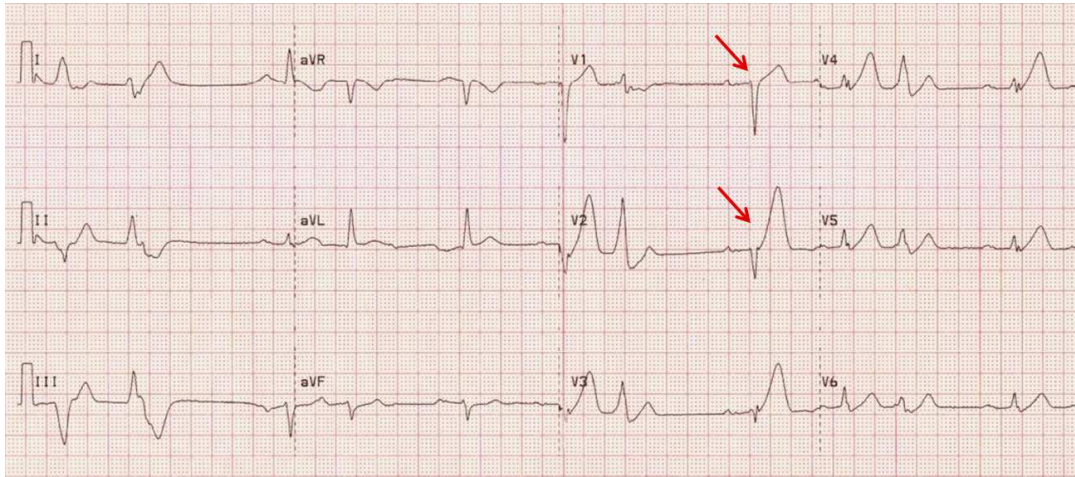
(a) Hyperacute Anterior STEMI
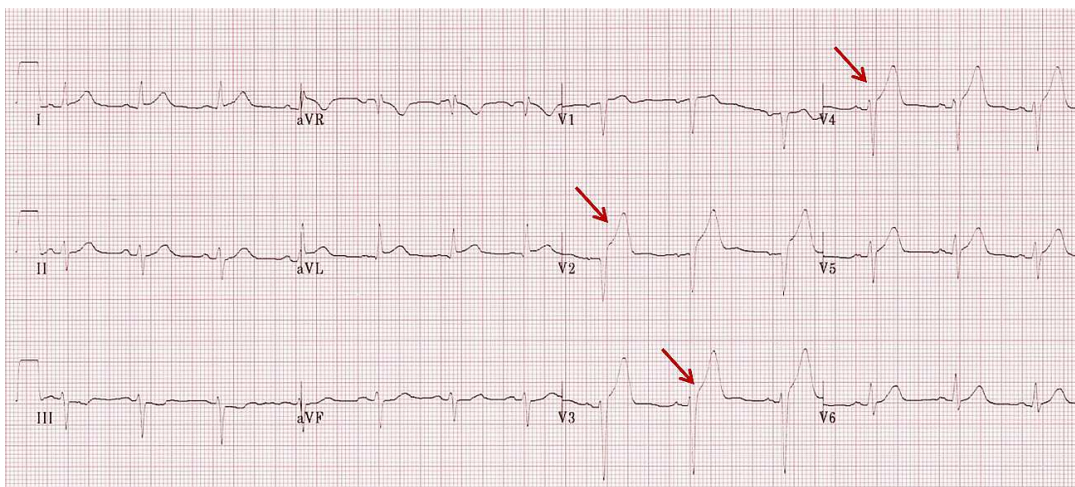
(b) Hyperacute Anteroseptal STEMI

Figure 2.8 : ECG abnormalities (indicated with arrows) for (a) anterior and (b) anteroseptal STEMI captured with 12-lead ECG

depicted in Fig. 2.8a. T waves from the normal ECGs are expected to be asymmetrical [80]. As such cardiac abnormalities appear, they tend to be more symmetrical, taller, biphasic or inverted [81]. Yet, these alone do not reveal much about the severity of the cardiac condition.

2.3   Shapelet-based Feature Extraction Methods

Shapelet transform is a feature extraction method to discover the so-called "shapelets" [28] from a given time series dataset. A shapelet is a defined as a subsequence identified as the representative of a certain time series dataset. Shapelets are essentially subsequences extracted from the time series instances themselves through an exhaustive sliding window-based search operation. Let L denote the desired shapelet length (specified by the user) of the shapelet transform. For K number of time series instances of length N, there exist $(N - L + 1)$ subsequences, i.e., shapelet candidates,

from each time series making a total of $(N-L+1)K$ candidates. The distance between each subsequence - time series pair is computed. A single distance computation requires sliding the subsequence across the time series of interest in all possible ways to find the closest possible match (see Figs 2.9a, 2.9b) for a generic illustration). Thus, a distance vector to all time series instances belonging to all class labels is obtained for each subsequence. A distance threshold is then determined for each subsequence yielding the most homogenous separation of instances, i.e., the highest information gain, regarding the class labels. Finally, user specified number of subsequences (with the highest information gain) are designated as shapelets (see Fig 2.9c).

The corresponding distance vectors of the chosen shapelets, then, become the feature vectors generated by the shapelet transform. These feature vectors, i.e., the transformed dataset, are used to train the classifier of interest. A novel time series instance to be classified is transformed into the feature space by taking its distance to each chosen shapelet to obtain its feature vector which is then feed into the trained classifier.

(a)   (b)

(c)
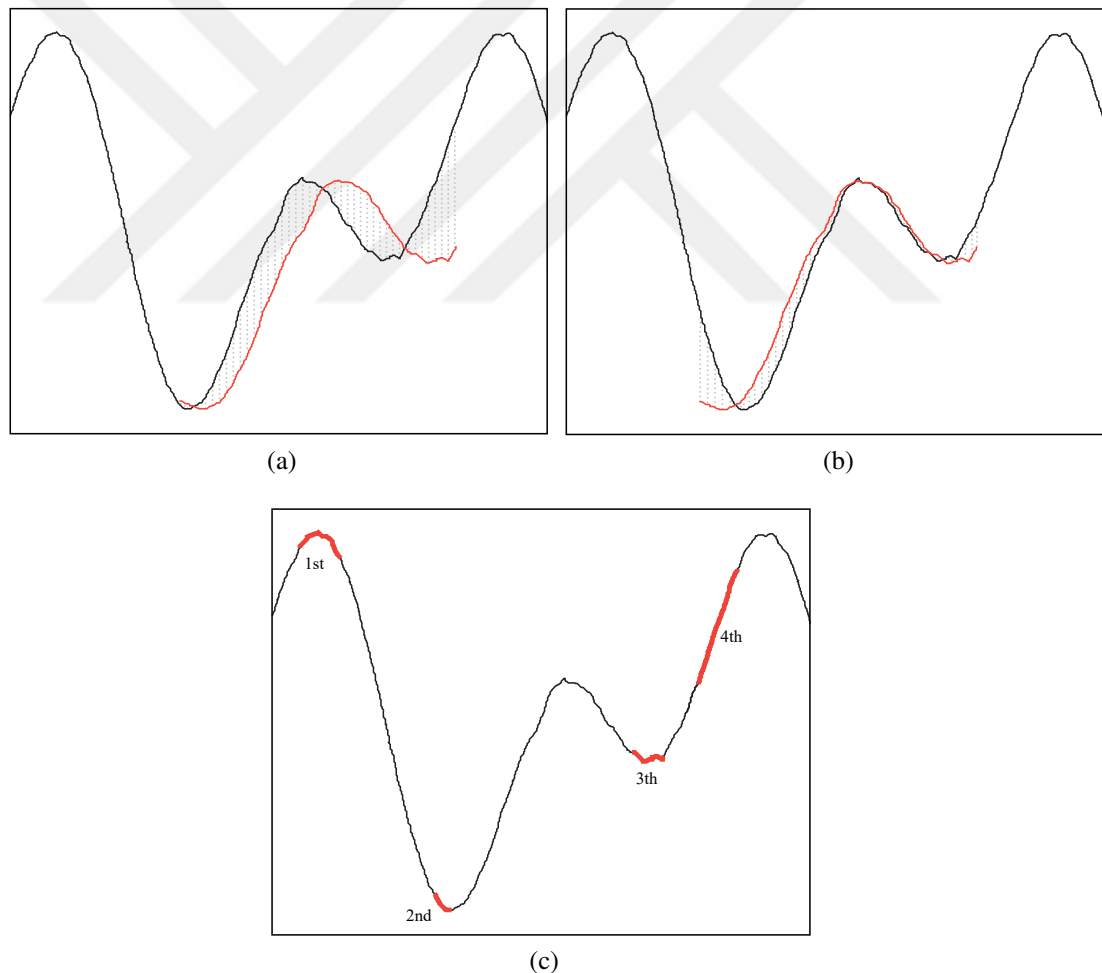
Figure 2.9 : Illustration of the (a) search for the best matching location with (b) the smallest Euclidean distance of candidate shapelet S in time series instance T. An example dictionary of the selected shapelets (c) ordered according to their information gain (adapted from [28]).

A great deal of research has been devoted to accelerate the aforementioned shapelet transform which is referred to as "brute-force shapelet transform" hereinafter. Among them are the logical shapelets [82], fast shapelets [83], ultra-fast shapelets [84] and random-shapelet discovery [85]. In this thesis, we concentrate on the so-called "fast shapelet transform" [83], the fastest variant to the best of our knowledge, along with the brute-force shapelet transform. For the relatively high computational complexity of the brute-force shapelet transform, the most prominent weakness of the algorithm, the proposed remedies are reducing either the number of instances in the training set or the sampling rate of these instances.

As part of the latest efforts put into further speeding up the training phase of the brute-force shapelet transform, the fast shapelet transform is introduced in [83]. In this approach, whole dataset is projected onto a discrete and low-dimensional representation referred to as Symbolic Aggregate approXimation (SAX) [86]. The distance metrics defined for this projected version of the time series instances are claimed to lower bound the same metrics used for the original time series instances. Before projecting the instances into discrete symbols with SAX, a preliminary step called Piecewise Aggregate Approximation (PAA), [87], is taken. PAA divides the entire duration of an instance into a predefined number of window frames of equal support. Each of these window frames is assigned the average value of the sample points falling inside. Fig. 2.10 shows this approximation on an example time series instance.
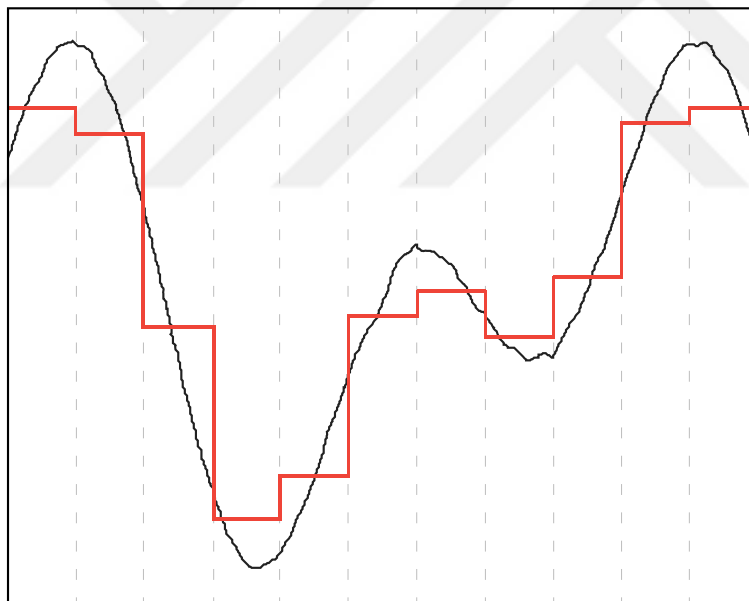
Figure 2.10 : PAA of an example time series with a linear combination of box basis functions (adapted from [87]).

The values of PAA are assumed to follow Gaussian distribution. The so-called "alphabet size" parameter denoting the number of distinct symbols representing the time series instances is then used to fix thresholds along the distribution such that the resulting area under the corresponding segments of the Gaussian curve are equal to each other. Fig. 2.11 illustrates how the PAA-transformed instances are further mapped into discrete symbols called SAX.

16
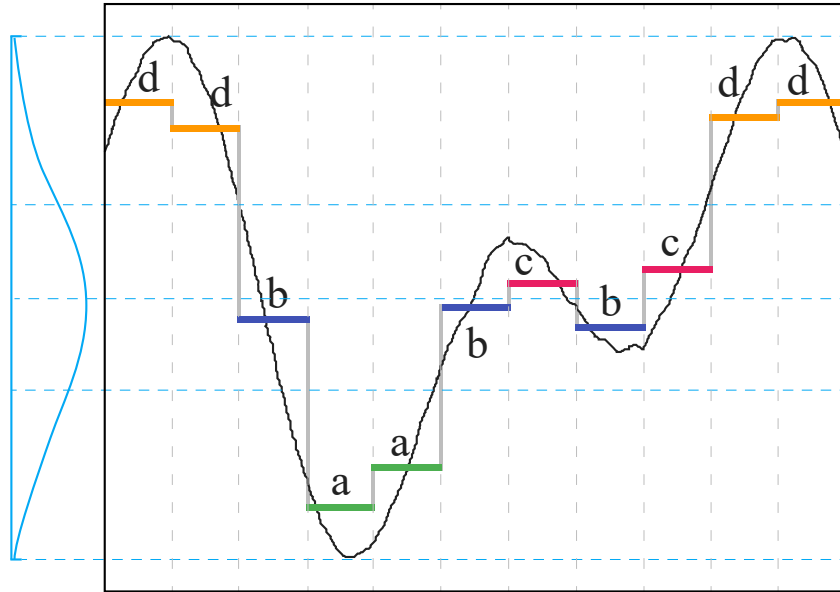
Figure 2.11 : PAA approximation of an example time series with the corresponding SAX symbols for an exemplary alphabet size of 4 (adapted from [87]).

As the original time series instances are projected onto the symbol space of SAX, the distance between each projected instance pair is stored in and used from a look-up table that lower-bounds the Euclidean distance [86]. Using several dataset applications, SAX representation is shown to successfully represent the general morphologies of the original datasets. Nevertheless, there exists no guarantee that the best performing shapelets of its brute-force counterpart can always be chosen. This can simply be attributed to the representational heuristic incorporated by the fast shapelet transform.

Shapelets, in a general context, are proposed to be a time series data mining "primitive" that can be used to assess similarity based on small common shapes occurring at arbitrary locations of a time series. It is often emphasized that shapelet transform produces interpretable outcomes enabling a better understanding of the data in hand [28]. In this thesis, we find shapelets worthy of inspection, since its operating domain is temporal and the primitives are extracted from the dataset itself. However, the temporal location of a shapelet is disregarded by the very nature of the shapelet extraction process, which may hinder the interpretation of the results particularly in the medical domain.

CHAPTER 3

METHOD

In this chapter, the proposed Window-based Time Series Characterization (WTC) method is explained in a stepwise manner. Initially, Section 3.1 describes the necessity of the pre-processing performed on the time series instances and how important this preliminary step is to the proposed method. Following that, Section 3.2 entails the proposed method step by step by elucidating its components that constitute a basis for it. The mathematical expressions for the major components are also predicated with pseudocodes for a further grasp of the computational complexity calculated at the end of this chapter. **Table 3.1** summarizes the notation presumed throughout this chapter. The detailed definitions are given in the following subsections.

3.1 Dataset Pre-processing

The proposed characterization method requires processing of the local features in the time series of interest. For that reason, time series instances are registered both in time and signal amplitude. Registration in time refers to aligning time series to a fixed and known position such as the time instant at which a certain event is expected to occur for a single-cycle phenomenon (e.g. ex-vivo action potential recordings) or the epoch of a cycle for a cyclic phenomenon (e.g. ECG and phonocardiogram recordings). Registration in signal amplitude, on the other hand, facilitates time series analysis to be made irrespective of the offset and scale of the recorded signal which may well depend on the experimental setup. We denote each instance in a time series dataset S by S_k where $k \in [1, K]$ and K denotes the total number of instances available. For the rest of this section, we presume the whole dataset S is registered to obtain the dataset T whose instances are denoted by T_k. Since registration is a data-dependent step, we describe how to obtain T from S in Chapter 4, where AP and ECG time series datasets used for the validation of the proposed method are introduced.

3.2 Proposed Analysis

For notational simplicity, we explain the proposed WTC method for a single class label throughout this chapter. We note that WTC can be independently run for each class for a multiple class scenario as in Chapter 4 where datasets with two class labels are examined.

Table 3.1 : Symbols in alphabetical order.

Symbol	Explanation
A	The maximum amplitude of the time series instances in T
α_j	The trajectory-based weight of SCB_j based on $I_{j,k}$
β_j	The distance-based weight of SCB_j
c	The PMF resolution
$\mathbb{D}$	PMF of the deterministic distribution with atom at zero
$\mathcal{D}(.,.)$	Distance between two PMFs
$d_{j,k}$	The Euclidean distance of time series instance T_k to T^A within SCB_j
γ_j	The weight of the distance-based component of the similarity score for SCB_j
$I_{j,k}$	The indicator of inclusion for time series instance T_k within SCB_j, where $I_{j,k} \in \{0, 1\}$
J	Number of shape confidence bands
$l^\star$	Cut-off index in DCT domain
K	Number of time series instances
M	Number of perceptually important time points (PIPs) in the class representative time series
N	Number of sample points in a time series instance
p	Confidence level for the confidence interval
$\mathbb{P}_j$	Distance PMF for SCB_j
r	PIP oversampling ratio
S	Time series training set
SCB_j	j^{th} shape confidence band
Z_j	The ensemble average similarity score for SCB_j
$Z_{j,k}$	The individual similarity score of time series instance T_k for SCB_j
$Z_{j,k}^{(d)}$	Distance-based component of the similarity score of time series instance T_k for SCB_j
$Z_{j,k}^{(t)}$	Trajectory-based component of the similarity score of time series instance T_k for SCB_j
T	Registered time series instances in the training set
T^A	Class representative average time series
T_k	k^{th} time series instance in the registered training set T
T^L	Lower bound of the confidence interval with the level p around the class representative time series T^A
T^U	Upper bound of the confidence interval with the level p around the class representative time series T^A
$\mathbb{U}$	PMF of the uniform distribution with support $[0, \sqrt{w}A]$
w	Number of sample points in an W_j except for the last one
W_j	j^{th} time window

3.2.1 Determining the Local Time Window

Even after registration, time series instances exhibit some degree of inter-subject variation. Averaging registered time series instances makes the proposed method robust against noise and distortions stemming from trajectory variability and possible artifacts [88]. In this respect, we average out the registered time series instances, T_k, in order to obtain a class representative average time series denoted by T^A as follows:

$$T^A(n) = \frac{1}{K} \sum_{k=1}^{K} T_k(n) \qquad (3.1)$$

where n denotes the sample point index of the time series. Restricting the averaging step to time series instances bearing the same class label emphasizes local dynamics and their significance in defining the class behavior.

The resulting representative trajectory constitutes the base time series from which the local features of the time series are extracted. Obtaining local features requires determination of suitable time windows, preferably to be extracted from the time series itself. We define the domain of locality as non-overlapping and contiguous time windows denoted by W_j of fixed length w, where $j \in \{1, 2, \dots, \lceil N/w \rceil\}$ and N being the total number of sample points in T^A. We note that these time windows span the entire time series and the last time window, $W_{\lceil N/w \rceil}$, may have a length smaller than w. In order to choose w, we resort to the so-called "Perceptually Important Points" (PIPs) [89], [90] and well-known Discrete Cosine Transform (DCT) as heuristic tools. Next, we briefly review the concept of PIP and then describe how it is complemented by DCT to choose w.
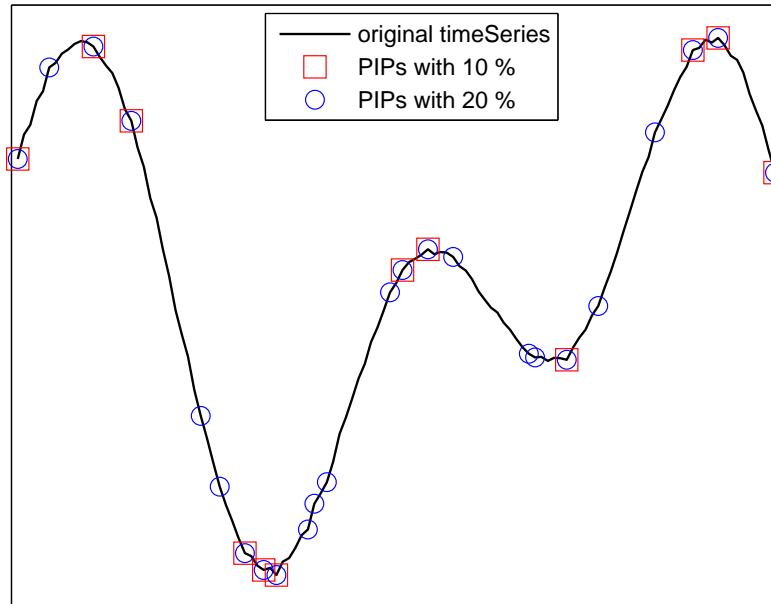
Figure 3.1 : Sample points marked on a generic time series determined as the first 10% and 20% of the PIPs.

21

PIPs are defined to be the observational points in a time series that have relatively more important influences on human vision [91]. PIP approach is useful in dimensionality reduction while indicating a hierarchy among the data points based on their visual importance. The identification of PIPs is as follows [90]: Given that the first and the last data points in the time series are identified as the first two important points in the so-called importance hierarchy, the next PIP is found to be the data point having the largest distance to them. Recursively, each candidate PIP at a particular iteration is the data point with the largest distance to its two adjacent PIPs determined so far. PIP selection process continues until all data points in the time series are exhausted. In the original work, the distance between PIPs is proposed to be measured by any of the Euclidean, vertical and perpendicular distances. In this study, we employ the Euclidean distance. Fig. 3.1 shows the first 10% and 20% of the PIPs found on a generic time series.

In our context, determining PIPs in a time series fits closely to the aim of extracting class-representative patterns from time series dataset. Essentially, the PIP algorithm sorts the data points of a given time series of total length N based on their perceptual importance. The procedure is summarized in the steps of Algorithm 1.

Algorithm 1 PIPIndices (T^A)

1: $PIPIndices \leftarrow \{1, N\}$
2: $TimeSeriesIndices \leftarrow \{2, 3, \ldots, N-1\}$
3: **for all** $n \in TimeSeriesIndices$ **do**
4:      find n such that
5:      $\sqrt{(n - n_1)^2 + (T(n) - T(n_1))^2} + \sqrt{(n - n_2)^2 + (T(n) - T(n_2))^2}$
6:      is maximum where $n_1 < n < n_2$
7:      for maximum $n_1 \in PIPIndices$ AND minimum $n_2 \in PIPIndices$
8: **end for**
9: append n to $PIPIndices$
10: remove n from $TimeSeriesIndices$
11: $PIPIndices \leftarrow$ SORTPIPINDICES($PIPIndices$)
12: **return** $PIPIndices$

Next, we choose the "most important" M data points out of those N points. For this purpose, it is assumed that $M = \lfloor N/r \rfloor$, where r denotes the oversampling ratio of the signal that is defined as follows:

$$r = f_s / f_N \tag{3.2}$$

The terms f_s and f_N in Equation 3.2 are the sampling and the Nyquist rates of the signal of interest, respectively. We employ the well-known one-dimensional DCT to calculate $r = N/l^\star$ where $r \in [1, N]$ and $l^\star \in \{1, 2, \ldots, N\}$. The term $l^\star$ is the lowest (cut-off) frequency index in the DCT domain representation of the signal below which δ percentage of its total energy is contained. Let DT^A denote the one-dimensional DCT of the time series T^A, then $l^\star$ can be found with the following expression:

$$\underset{l^\star}{\arg\min} \frac{\sum_{l=1}^{l^\star} |DT^A{}_l|^2}{\sum_{l=1}^{N} |DT^A{}_l|^2} \geq \delta \tag{3.3}$$

The best choice for the threshold parameter δ is explored in Chapter 4 together with the remaining parameters of our proposed method.

Extracting M number of PIPs from the time series, we finally set w to the maximum horizontal length between the consecutive PIPs so that at least one PIP is guaranteed to appear in W_j for all j. The corresponding steps are detailed in Algorithm 2.

Algorithm 2 LocalTimeWindows (P, r)

1: $maxDist \leftarrow 0$
2: **for all** $n \in \{1, 2, \ldots, N - 1\}$ **do**
3:     **if** $(P(n + 1) - P(n)) > maxDist$ **then**
4:         $maxDist \leftarrow P(n + 1) - P(n)$
5:     **end if**
6: **end for**
7: $w \leftarrow maxDist$
8: $J \leftarrow \lceil N/w \rceil$
9: **for all** $j \in \{1, 2, \ldots, J\}$ **do**
10:     $n_{j1} \leftarrow (j - 1) \times w + 1$
11:     **if** $j < J$ **then**
12:         $n_{j2} \leftarrow j \times w$
13:     **else**
14:         $n_{j2} \leftarrow N$
15:     **end if**
16:     $W_j \leftarrow \{n_{j1}, n_{j1+1}, \ldots, n_{j2}\}$
17: **end for**
18: **return** $\overline{W}$

3.2.2 Distance-based Similarity

Determining local time windows W_j, we calculate a distance-based similarity score for each. The Euclidean distance of a particular time series instance k to T^A within the support of W_j is denoted by $d_{j,k} = \|T^A(n) - T_k(n)\|$, where $n \in W_j$ and $d_{j,k} \in [0, \sqrt{w}A]$. Distance-based weight β_j belonging to W_j is expressed as follows:

$$\beta_j = \frac{\sum_{k=1}^{K} \mu_{j,k}}{\sum_{j=1}^{J} \sum_{k=1}^{K} \mu_{j,k}} \tag{3.4}$$

where,

$$\mu_{j,k} = \sqrt{w}A - d_{j,k} \tag{3.5}$$

The term $\mu_{j,k}$ in Equation 3.5 falling in the interval $[0, \sqrt{w}A]$ corresponds to the reward associated with the proximity of the time series instance T_k to T^A for the implied time window W_j. The term $\mu_{j,k}$ provides a standardized scale for all local time windows spanning the time series. Thus, they can be weighted against each other as in Equation 3.4. We note that, the closer the time series instances to T^A for a particular W_j, the higher the associated weight becomes. The distance-based similarity score denoted by $Z_{j,k}^{(d)}$ for time series instance T_k and time window W_j is calculated based on $\mu_{j,k}$ as follows:

$$Z_{j,k}^{(d)} = \beta_j \frac{\mu_{j,k}}{\sqrt{w}A} \tag{3.6}$$

23

3.2.3 Trajectory-based Similarity

We complement and reinforce the distance-based similarity with the so-called trajectory-based similarity. The average class representative time series is used to obtain a similarity band for characterizing the class representative trajectory. Assuming each sample point of the registered time series to follow Gaussian distribution as in [86], we form an upper (lower) time series denoted by T^U (T^L) whose sample points are higher (lower) than those the average time series T^A dictated by a confidence level, p. Consequently, we define the aforementioned similarity band as the one bounded by T^U and T^L which is then partitioned into time localized bandlets called Shape Confidence Bands (SCBs). Each SCB denoted by SCB_j is related to the corresponding W_j as follows:

$$SCB_j := \{(n, y) | n \in W_j \wedge y \in \mathbb{R} \wedge T^L(n) < y < T^U(n).\} \tag{3.7}$$

The procedure for extracting SCBs are summarized in the pseudocode Algorithm 3.

Algorithm 3 extractSCBs $(T^U, T^L, \overline{W})$

1: $J \leftarrow |W|$
2: **for all** $j \in \{1, 2, \ldots, J\}$ **do**
3:      $SCB_j \leftarrow (T^U(n), T^L(n))$ such that $n \in \{W_j\}$
4: **end for**
5: **return** $\overline{SCB}$

SCB_j is essential in deciding whether the temporally coinciding part of a time series of interest aligns with the class representative average time series within W_j. The rationale of SCBs is to quantify the similarity of trajectory by restricting the orientation of an instance to be aligned with that of the average time series T^A allowing some margin for error. For that purpose, each subsequence of a time series is tested for being completely residing within the boundaries of the corresponding SCB_j. Let $I_{j,k}$ be the indicator of inclusion for time series instance T_k within SCB_j, which is defined as follows:

$$I_{j,k} = \begin{cases} 1, & T_k(n) \in SCB_j \ \forall \ n \in W_j, \\ 0, & \text{otherwise.} \end{cases} \tag{3.8}$$

Finally, we express the trajectory-based similarity score denoted by $Z_{j,k}^{(t)}$ for each time series instance T_k and SCB_j as follows:

$$Z_{j,k}^{(t)} = I_{j,k}\alpha_j \tag{3.9}$$

where,

$$\alpha_j = \frac{\sum_{k=1}^{K} I_{j,k}}{\sum_{j=1}^{J} \sum_{k=1}^{K} I_{j,k}} \tag{3.10}$$

The term α_j in Equation 3.10 is defined as the trajectory-based weight of SCB_j calculated over all time series instances which reside within SCB_j. We note that the higher
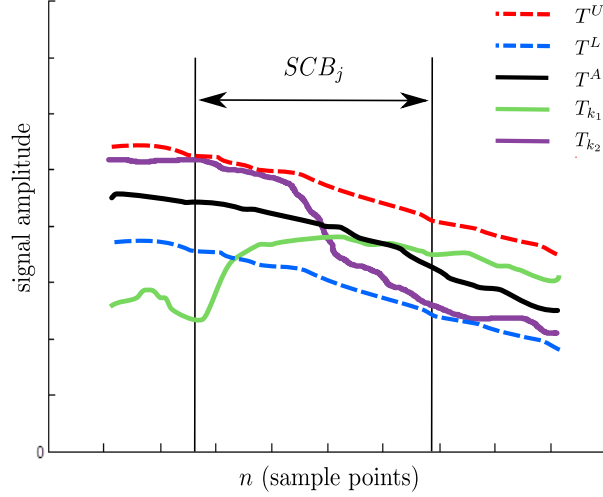
Figure 3.2 : An example case with two time series instances T_{k_1} and T_{k_2} with similar Euclidean distances from the representative average time series T^A yet following different trajectories.

the number of time series instances follow the trajectory of SCB_j, the higher the associated weight becomes. In Fig. 3.2, we give an example scenario emphasizing the difference between distance and trajectory metrics discussed so far. Although time series instances T_{k_1} and T_{k_2} are of similar distances to T^A, unlike T_{k_2} the trajectory of T_{k_1} falls out of the band SCB_j.

3.2.4 Overall Similarity

Linearly combining the trajectory-based and the distance-based similarity scores $Z_{j,k}^{(t)}$ and $Z_{j,k}^{(d)}$, we devise a single similarity score of time series T_k for each SCB_j denoted by $Z_{j,k}$ as follows:

$$Z_{j,k} = (1 - \gamma_j)Z_{j,k}^{(t)} + \gamma_j Z_{j,k}^{(d)}, \tag{3.11}$$

where $(1 - \gamma_j)$ and γ_j are the associated weights of the trajectory-based and distance-based components, respectively. Thus, the average similarity score for each SCB_j is defined as follows:

$$Z_j = \sum_{k=1}^{K} Z_{j,k}. \tag{3.12}$$

The steps for calculating the overall similarity scores, Z, are denoted in Algorithm 4.

We propose to calculate γ_j based on the probability mass function (PMF) of time series distances within SCB_j denoted by $\mathbb{P}_j$, as in the following equation:

$$\gamma_j = \frac{\mathcal{D}(\mathbb{D}, \mathbb{P}_j)}{\mathcal{D}(\mathbb{D}, \mathbb{P}_j) + \mathcal{D}(\mathbb{U}, \mathbb{P}_j)}, \tag{3.13}$$

25

Algorithm 4 SimilarityScores ($\mathcal{T}, T^U, T^L, \bar{\alpha}, \bar{\beta}, \bar{\gamma}$)

1: $\overline{I} \leftarrow$ TrajectoryInclusionIndicators(T, T^U, T^L) ▷ Equation 3.8
2: $\overline{Z^{(t)}} \leftarrow$ TrajectoryBasedSS($\bar{\alpha}, \overline{I^{(t)}}$) ▷ Equation 3.9
3: $\overline{Z^{(d)}} \leftarrow$ DistanceBasedSS($\bar{\beta}, \overline{I^{(d)}}$) ▷ Equation 3.6
4: $\overline{Z} \leftarrow$ calculateSS($\bar{\gamma}, \overline{Z^{(t)}}, \overline{Z^{(d)}}$) ▷ Equation 3.12
5: **return** $\overline{Z}$

The terms $\mathbb{U}$ and $\mathbb{D}$ in Equation 3.13 are the PMFs corresponding to the uniform and deterministic distance distributions, respectively. $\mathcal{D}$ is the operator computing the distance between two PMFs based on the Algorithm 1 presented in [92]. The aforementioned PMFs are constructed with a signal level resolution denoted by c. Throughout this study, we set c to 0.001. The PMF $\mathbb{U}$ is defined over the support $[0, \sqrt{w}A]$ and $\mathbb{D}$ is an impulse at zero. Random variable associated with the PMF $\mathbb{U}$ ($\mathbb{D}$) has the maximum (minimum) entropy for its support. The rationale of Equation 3.13 is to assign a weight to the distance-based (trajectory-based) similarity score of SCB_j inversely proportional to the distance of $\mathbb{P}_j$ to $\mathbb{U}$ ($\mathbb{D}$). Trajectory-based similarity scores are calculated over the time series instances residing within each SCB_j. On the other hand, distance-based similarity scores are calculated over the entire set of time series instances which are more dispersed out of the confidence band of interest. For the extreme cases where $\mathbb{P}_j$ tends to $\mathbb{U}$ ($\mathbb{D}$), the distance-based (trajectory-based) similarity scores dominates the overall similarity score.

The overall algorithm is outlined in Algorithm 5 and the related source codes are accessible online via [93].

Algorithm 5 WTC ($\mathcal{T}$, p)

1: $[T^A, T^U, T^L] \leftarrow$ RepresentativeTimeSeries($\mathcal{T}, p$) ▷ Equation 3.1
2: $r \leftarrow$ PIPCompressionRatio(T^A) ▷ Equation 3.2
3: $P \leftarrow$ PIPIndices(T^A) ▷ [90]
4: $\overline{W} \leftarrow$ LocalTimeWindows(P, r) ▷ Section 3.2.1
5: $\overline{SCB} \leftarrow$ extractSCBs(T^U, T^L, W_j) ▷ Equation 3.7
6: $\bar{\alpha} \leftarrow$ SCBWeightsAlpha($\mathcal{T}, \overline{SCB}, \overline{W}$) ▷ Equation 3.10
7: $\bar{\beta} \leftarrow$ SCBWeightsBeta($\mathcal{T}, T^A, \overline{W}$) ▷ Equation 3.4
8: $\bar{\gamma} \leftarrow$ SCBWeightsGamma($\mathcal{T}, \overline{SCB}, \overline{W}$) ▷ Equation 3.13
9: $\overline{Z} \leftarrow$ SimilarityScores($\mathcal{T}, \bar{\alpha}, \bar{\beta}, \bar{\gamma}, \overline{W}$) ▷ Equation 3.12
10: **return** $[\overline{SCB}, \overline{Z}]$

A block diagram is depicted in Fig. 3.3 showing input parameters of WTC and its major steps with their respective interim and main outputs. One prominent advantage of Algorithm 5 is its relatively lower complexity, which we analyze next. Step 1 of Algorithm 5 involves point-wise mean and confidence level computations which have $O(KN)$ complexity. Steps 2 and 3 require the computation of DCT and the execution of the PIP algorithm over the single times series T^A, which are $O(N \log N)$ when N is a power of 2 ($O(N^2)$ otherwise) [94] and $O(N^2)$ [90], respectively. Steps 4 and 5 have obviously $O(N)$ complexity. Suggested by the corresponding equations, Steps 6, 7, 8 and 9 are all of $O(KN)$ complexity. We conclude Algorithm 5 to have an $O(KN + N^2)$ quadratic complexity summing up the individual complexities of its constituent steps. We note that, it is possible to improve the efficiency of Step 3, the most computationally intensive part of the overall algorithm, by abandoning the related iterations as soon as finding the $100/r$ percentage of PIPs.

26
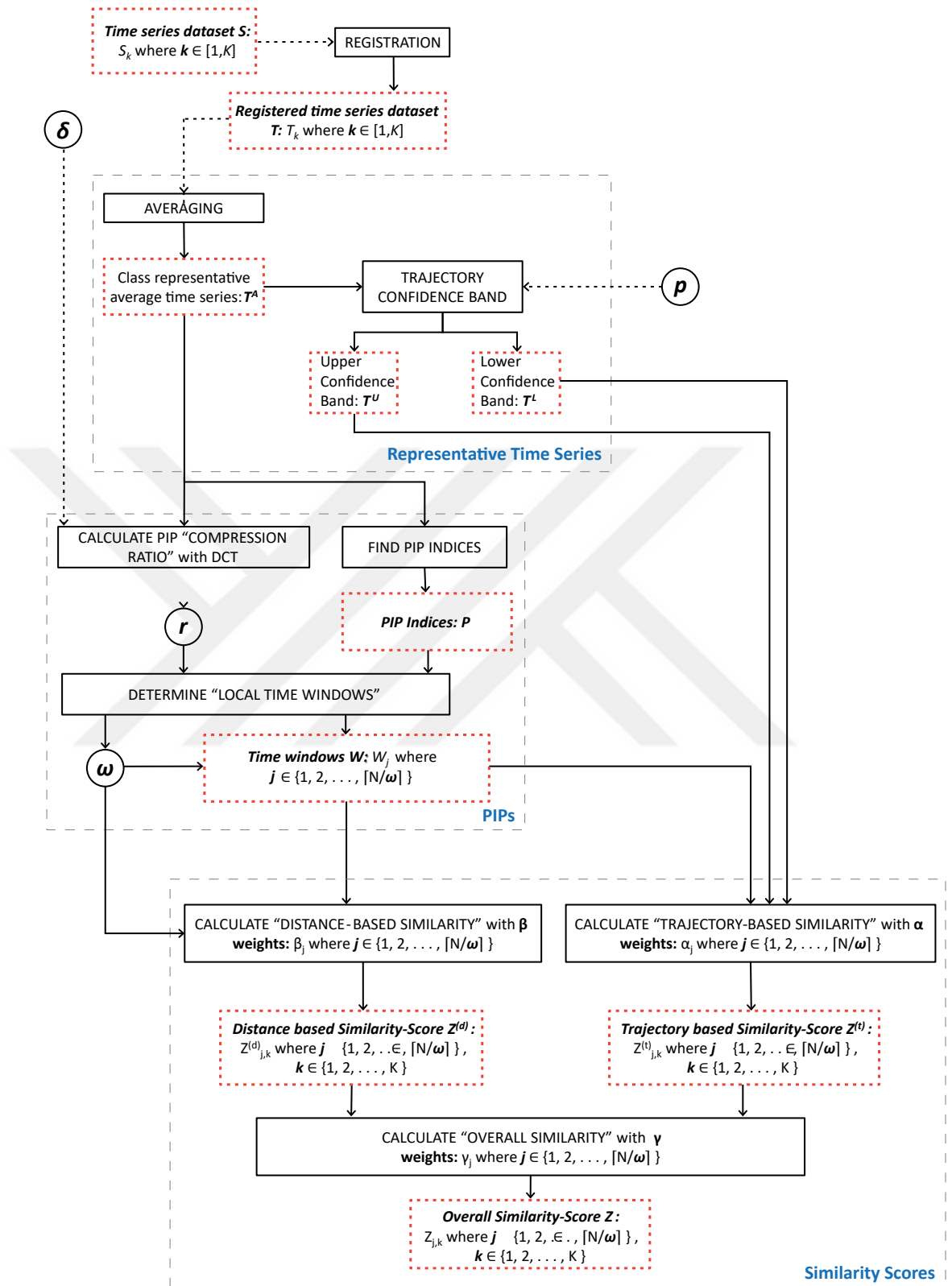
Figure 3.3 : A block diagram showing the major algorithmic steps of WTC

CHAPTER 4

EXPERIMENTS AND RESULTS

We evaluate the proposed WTC method in Sections 4.1 and 4.2 for AP and ECG time series datasets, respectively. The corresponding results are compared with those of the brute-force shapelet transform [28] and its accelerated variant, the fast shapelet transform [83]. The details on the operating principles of these shapelet-based algorithms are given in Chapter 2. The overall computational complexity of the brute-force shapelet transform is stated to be $O(K^2N^4)$ [82], [27]. To speed up the training time of the brute-force shapelet transform, the so-called fast shapelet transform is introduced [83]. In this approach, a complete dataset is projected to a lower-dimensional representation that is referred to as SAX [86] such that the entire shapelet discovery is completed in $O(KN^2)$ time at the expense of a degradation in performance in terms of accuracy. By avoiding exhaustive distance computations, the fast shapelet transform does not guarantee to reach the same set of shapelets as its brute-force counterpart. We refer the reader to Section 2.3 for further details. The proposed WTC method with $O(KN+N^2)$ complexity clearly outperforms (see Fig. 4.1) the family of shapelet methods in that regard. In this study, we employ an open-source Java solution [27] to obtain results for the brute-force shapelet transform and a proprietary C++ based implementation [83] for the fast shapelet transform.

A series of classification experiments are performed to facilitate a comparison of the proposed WTC and the benchmark methods (brute-force and fast shapelet transforms). The following methods within WEKA framework [95] are selected to maintain a broad coverage among the available classifiers: the well-known Naive Bayes classifier, J48 pruned tree (an implementation of the well-known C4.5 algorithm), random decision forest [96], adaptive boosting (AdaBoost.M1) [97], classification via regression (employing a type of decision tree with linear regression functions at the leaves [98]), bagging [99], multi-boosting (MultiBoostAB) [100], locally weighted learning (LWL) [101], partial decision tree classifier (PART) [102], ensemble of nested dichotomies (END) [103], decision stump [104], simple classification and regression tree (CART) [105], a proprietary algorithm termed as "ranking instances by maximizing the area under the ROC curve (RIMARC)" [106], Bayesian network learning, dagging [107], random subspace method [108], decision table majority classifier, ripple-down rule learner (RIDOR) [109], alternating decision tree (ADTree) [110], and random tree construction and multi-class alternating decision tree with a logit-boost strategy (LADTree). The similarity scores $Z_{j,k}$ and the shapelet distances to time series instances constitute transformed feature vectors for WTC and the benchmark methods, respectively. The classification accuracy results are obtained
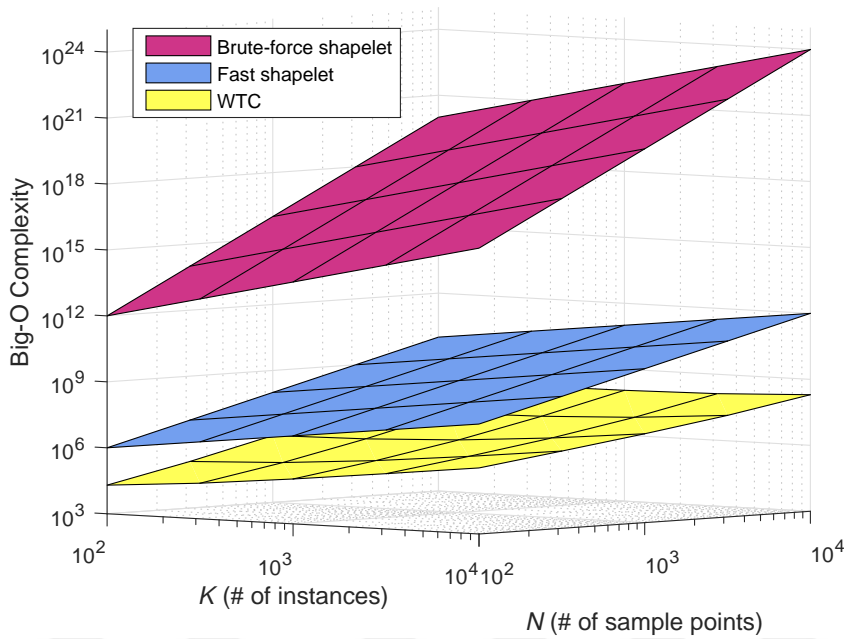
Figure 4.1 : Computational complexities of WTC (green), the fast shapelet (blue) and brute-force shapelet discovery methods.

with a 10-fold cross validation. In this Chapter, we describe the datasets and then explain the details of registration, evaluate the performance of the WTC method and compare with those of the benchmark methods for each studied dataset.

4.1 Cardiac Action Potential (AP) Dataset

As an application of the proposed method to a real-life dataset, we use AP recordings obtained from human right atrial biopsies [66].

4.1.1 AP Dataset Description

Within the scope of a project entitled "The European Network for Translational Research in Atrial Fibrillation (EUTRAF)" [111] that is funded by the European Community's Seventh Framework Programme, the AP dataset was collected in the period from January 2006 to February 2014. Each patient's written informed consent was obtained, and thus the study conforms to the Declaration of Helsinki and was approved by the ethics committee of Dresden University of Technology (No.EK790799). The time series instances of the dataset are clinically attributed to the class label AF if it is considered that a patient is in chronic atrial fibrillation (ICD-10 code I48.2) and to the class label SR with respect to control patients with a normal sinus rhythm. Patients with paroxysmal or intermittent AF are excluded. The presence of AF is confirmed with pre-operative ECGs throughout the data collection. Eligible AP recordings are obtained from a total of 341 unique patients that comprise 142 AF (aged from 47

years to 85 years with a mean of 72.29 years and 37.68% female, 62.32% male) and 199 SR (aged 22 years to 86 years with a mean of 67.12 years and 23.98% female, 76.02% male) patients. A common temporal sampling rate of 10 kHz is used for all recordings $K = 202$ ($K = 219$) that belong to the AF (SR) class.

4.1.2 AP Dataset Registration

AP time series instances are single event-triggered and acyclic, therefore require no segmentation for registration. We register all AF and SR instances in time according to the instant at which membrane potential of each time series peaks, i.e., end of phase 0. The time window length of a time series from its peak membrane potential to the resting membrane potential actually constitutes a differentiating factor between AF and SR classes due to ion channel activities under investigation. For that reason, AP registration is done with respect to the time course of the acquired signal by avoiding techniques such as re-sampling and dynamic time warping. Time domain alignment provides human-readability and brings localized time windows into prominence, thereby easing decision making in real-life practice. Time domain alignment is followed by truncating each time series instance based on members with the shortest support at each side of the peak membrane potential sample point. Hence, the common AP time series length N corresponds to 4303 sample points for both classes.

The next step comprises of registering signal amplitudes. With respect to the AP dataset, a single cycle begins with the voltage level that indicates a resting membrane potential and ends with almost the same ground value after an AP is triggered. Given this fact, the peak membrane potential occurring at the end of the rapid depolarization phase, i.e., phase 0, and the resting membrane potential phase, i.e., phase 4, of the experiment are considered as reference points for signal amplitude registration. Specifically, the registered time series T_k is obtained as follows:

$$T_k(n) = \frac{S_k(n) - S_k(N)}{\max\{S_k(n)\} - S_k(N)} \tag{4.1}$$

In Fig. 4.2, all registered instances of AF (blue) and SR (red) classes are shown.

Time series instances in Fig. 4.2 manifest substantial intra-class variability for both classes that corresponds to a coherent observation with the study in [112].

4.1.3 WTC Results for AP Dataset

In order to assess the performance of the proposed WTC method, the AP dataset is subjected to a training phase that is subsequently followed by a testing phase. The training set consists of 50 AF and 50 SR time series instances that are arbitrarily chosen. The same set is also used to train the brute-force shapelet transform and fast shapelet transform methods for benchmarking. A complementary set with 152 AF and 169 SR instances is spared to test all three methods.

31
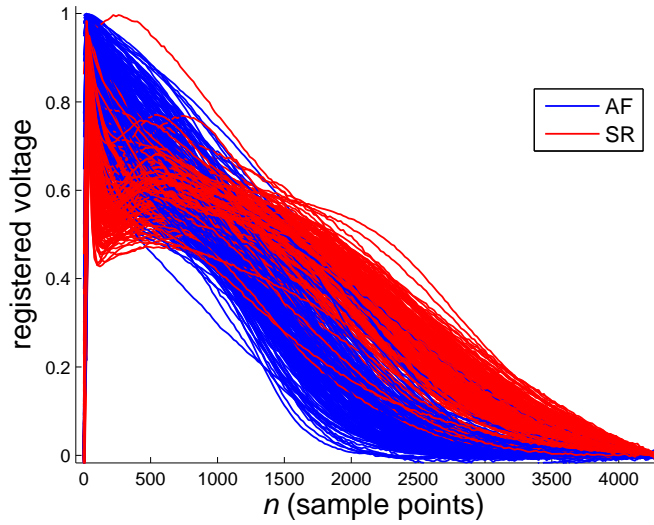
Figure 4.2 : Registered AP time series instances of AF (202 instances) and SR (219 instances) having 4303 sample points each.

In this section, we fix the energy threshold parameter δ and confidence level parameter p of the proposed WTC method as a preliminary step for benchmarking. For this purpose, we train the proposed WTC method for the Cartesian product of the sets $\delta \in \{0.9900, 0.9990, 0.9999\}$ and $p \in \{0.95, 0.99\}$.
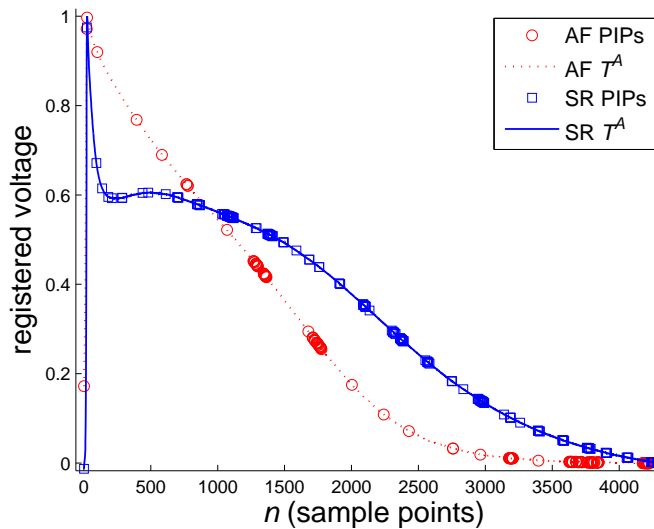
Figure 4.3 : Identified PIPs, with δ set to 0.9990, from the class representative average time series of AF and SR classes.

The resulting classification accuracies are tabulated in Table 4.1 for the aforementioned test set and various classification methods offered by WEKA. Except for a few entries in Table 4.1, WTC yields accuracies falling in 5% vicinity of each other which underlines relative insensitivity of WTC to its parameters. For the majority of the cases, the success rate of WTC exceeds 94%.

Table 4.1 : Individual and ensemble average classification accuracies of the proposed WTC method applied to AP dataset for different values of the energy threshold δ, the confidence level p and for the selected classifiers.

| | $\delta = 0.9900$ | | $\delta = 0.9990$ | | $\delta = 0.9999$ | |
	$p = 0.95$	$p = 0.99$	$p = 0.95$	$p = 0.99$	$p = 0.95$	$p = 0.99$
Naive Bayes	74.766%	81.620%	92.212%	92.212%	89.720%	90.031%
J48	91.589%	92.212%	93.770%	93.458%	93.458%	94.081%
Random forest	93.770%	93.770%	94.704%	94.704%	93.146%	94.081%
AdaBoost.M1	93.146%	93.146%	95.639%	95.639%	95.639%	95.950%
Classif. via regr.	95.327%	95.327%	95.327%	95.016%	93.458%	93.146%
Bagging	94.393%	94.393%	94.704%	94.704%	94.704%	94.704%
MultiBoostAB	93.458%	93.458%	94.704%	94.704%	94.704%	94.704%
LWL	90.654%	90.654%	94.081%	94.704%	94.393%	94.704%
PART	94.081%	94.081%	95.639%	95.639%	94.081%	93.770%
END	91.900%	92.212%	93.458%	93.146%	92.835%	93.458%
Decision stump	90.654%	90.654%	95.016%	95.016%	95.016%	95.016%
Simple CART	92.835%	92.835%	94.704%	94.704%	94.393%	94.704%
RIMARC	94.081%	94.081%	95.016%	95.016%	94.704%	94.704%
Bayes NET	92.835%	92.835%	93.458%	93.458%	94.081%	94.081%
Dagging	94.393%	93.770%	94.704%	94.081%	92.835%	94.704%
Random SubSpace	93.770%	93.458%	93.146%	94.081%	94.081%	93.458%
Decision Table	92.523%	93.770%	92.523%	92.212%	93.146%	93.458%
Ridor	92.523%	92.523%	94.704%	94.393%	93.770%	94.081%
ADTree	94.393%	94.081%	94.393%	94.393%	95.016%	95.016%
LAD Tree	93.146%	93.458%	94.393%	94.393%	95.016%	95.327%
Random Tree	90.343%	91.277%	94.704%	92.523%	93.146%	90.966%
AVERAGE	92.123%	92.553%	**94.333%**	94.200%	93.873%	94.007%

Based on the individual accuracies and the highest ensemble average accuracy (94.333%) presented in Table 4.1, we fix δ to 0.9990 and p to 0.95. This particular choice for δ results in oversampling ratio r equals to 27.0629 and 21.4078 for AF and SR, respectively.
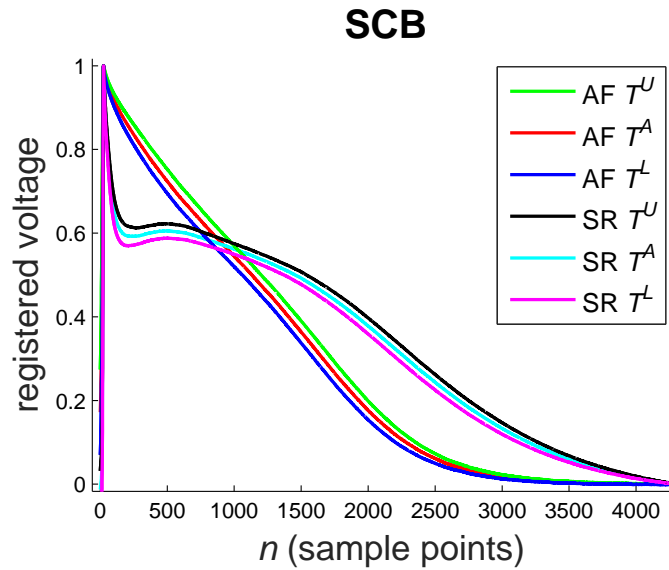
Figure 4.4 : Confidence bands with $p = 0.95$ overlaid for AF and SR classes.

33

In Fig. 4.3, selected PIPs from the class-representative average time series based on these values of r are shown for both AF and SR classes. For the other choices of δ, such as 0.9900 and 0.9999, we refer to the Appendix A.1 for the Figs. A.2 and A.3 depicting the corresponding PIPs.

The values of the localized time window length, w, which corresponds to the maximum horizontal separation between the adjacent PIPs, involve 329 and 169 sample points corresponding to 32.9 ms and 16.9 ms for AF and SR, respectively. The window length determination step is immediately followed by constructing confidence bands (see Fig. 4.4) with a chosen confidence level corresponding to $p = 0.95$. For an alternative construction of confidence bands with $p = 0.99$, we refer to Appendix A.1 for Fig. A.1.

Next, we segment these confidence bands as suggested by the local time windows W_j into SCBs for each class. The calculated weights, namely α_j, β_j and γ_j, for the corresponding W_j are depicted in Figs. 4.5a and 4.5b, for AF and SR, respectively.
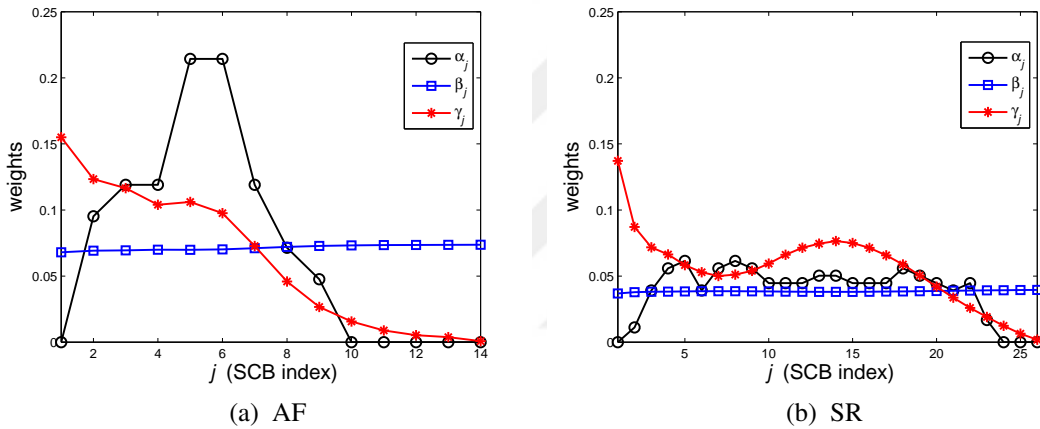
Figure 4.5 : The weights α_j, β_j and γ_j for SCBs in (a) AF and (b) SR.

With respect to both classes, the results indicate that the distance-based weights β_j are close to each other, and this implies that the rewards given based on the proximity to the class average for the corresponding SCB_j exhibit a similar trend. Trajectory-based weights α_j, on the other hand, vary more drastically for changing SCB_j and even decreasing to zero for higher indexed SCBs for both classes. This behavior indicates the failure of all time series instances to be completely encapsulated by the corresponding SCBs. The calculated weights γ_j are much closer to 0 than to 1, thereby indicating that they favour the trajectory-based component of the similarity score for both classes. In Appendix A.1, the weights of α_j, β_j and γ_j for SCBs constructed with the remaining combinations of $p \in \{0.95, 0.99\}$ and $\delta \in \{0.9900, 0.9990, 0.9999\}$ are also depicted in Figs. A.4a, A.4b to A.8a, A.8b.

In order to visualize the predictive power of WTC in terms of Z_j, color map plots are presented in Figs. 4.6a and 4.6b for AF and SR, respectively. These figures constitute the most interpretable output of WTC and show the similarity scores Z_j calculated by averaging over all AF and SR time series instances in the test set. The more reddish

$$(a) \ \text{AF} \qquad\qquad (b) \ \text{SR}$$
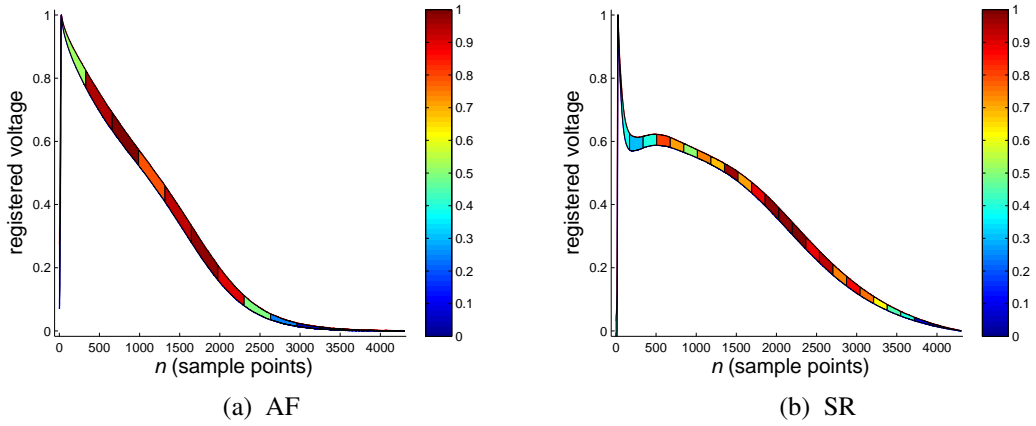
Figure 4.6 : Color map plots for the average values of Z_j for the test set.

tones imply regions with higher similarity, and thereby SCBs with higher predictive power. These figures also depict the double-sided confidence bands of the class representative average time series and the extracted SCB_j. For the purpose of completeness, the color maps constructed for AF and SR with the remaining combinations of $p \in \{0.95, 0.99\}$ and $\delta \in \{0.9900, 0.9990, 0.9999\}$ are depicted in Figs. A.9a, A.9b to A.13a, A.13b.

4.1.4   Comparative Results for AP Dataset

The following experiments are performed using the aforementioned training and test sets in Section 4.1.3. So as to achieve a meaningful level of parameter exploration for the brute-force shapelet transform in a reasonable time, a range for the shapelet length L (in units of sample points) is quantized as shown in Table 4.2. We note that the chosen values for L span a wider range when compared to the values reached by the proposed WTC method for its time localized window length, w, with the aid of DCT domain filtering. We resort to the default value of the Java implementation [27], which is 100, for the number of shapelets to be extracted.

As seen in Table 4.2, the brute-force shapelet transform yields varying test set classification accuracies for variations in L irrespective of the deployed classification method. The first 100 shapelets with $L = 400$ yielding the most favorable ensemble average accuracy level (93.057%) are shown in Fig.4.7 along with the class average time series, T^A, for both AF and SR. The figures reveal that the selected shapelets from different instances substantially overlap with each other for both classes. This observation is also interpreted as a redundancy in the computations since majority of subsequences corresponds to a narrow time interval.

Next, we evaluate the performance of the fast shapelet transform. The fast shapelet procedure acts as an extension to the brute-force shapelet transform method and builds an internal decision tree termed as a "fast shapelet tree" with nodes that are ordered based on the information gain that they individually offer. The resulting classification accuracies of the fast shapelet transformed test dataset are presented in Table 4.3. The

35

Table 4.2 : Individual and ensemble average classification accuracies of the brute-force shapelet transform applied to AP dataset for variations in shapelet length, *L*, and the selected classifiers.

	$L = 50$	$L = 100$	$L = 150$	$L = 200$	$L = 250$	$L = 300$	$L = 350$	$L = 400$
Naive Bayes	74.766%	86.293%	89.408%	90.654%	91.900%	92.835%	93.146%	93.458%
J48	85.358%	87.851%	90.966%	89.408%	94.393%	93.770%	92.835%	92.523%
Random forest	86.293%	90.654%	93.458%	92.835%	94.393%	95.639%	94.393%	94.704%
AdaBoost.M1	81.308%	88.785%	90.966%	91.277%	94.081%	94.081%	94.704%	95.016%
Classif. via regr.	83.489%	90.654%	90.966%	92.835%	95.327%	94.704%	94.081%	95.327%
Bagging	87.851%	90.654%	92.212%	92.212%	94.393%	92.835%	93.146%	93.770%
MultiBoostAB	84.424%	88.474%	91.589%	92.212%	94.393%	94.081%	95.016%	94.704%
LWL	81.620%	81.932%	88.474%	87.851%	91.277%	87.227%	87.227%	90.031%
PART	81.308%	90.031%	90.654%	89.408%	94.081%	92.212%	93.458%	93.770%
END	85.358%	87.851%	90.966%	89.408%	94.393%	93.770%	92.835%	92.523%
Decision stump	81.620%	81.620%	88.785%	88.162%	88.785%	89.097%	90.654%	89.097%
Simple CART	84.112%	86.916%	91.277%	89.097%	94.081%	91.900%	92.835%	92.835%
RIMARC	83.801%	90.654%	92.523%	91.900%	91.900%	93.458%	94.393%	94.393%
Bayes NET	82.866%	89.097%	90.966%	90.654%	90.031%	91.900%	91.589%	93.146%
Dagging	83.489%	92.212%	93.146%	92.523%	95.016%	95.950%	94.393%	95.950%
Random SubSpace	80.685%	88.785%	91.589%	90.966%	90.343%	91.589%	91.589%	93.458%
Decision Table	67.601%	72.586%	79.128%	78.505%	83.801%	77.882%	80.685%	83.178%
Ridor	84.424%	92.523%	90.966%	89.097%	94.393%	92.212%	94.704%	92.835%
ADTree	83.489%	89.720%	91.589%	91.277%	94.081%	95.016%	95.327%	96.573%
LAD Tree	87.851%	87.851%	90.966%	91.277%	94.393%	93.770%	94.704%	93.146%
Random Tree	80.685%	88.474%	88.785%	89.408%	90.031%	90.654%	92.835%	93.770%
AVERAGE	82.495%	87.791%	90.447%	90.046%	92.642%	92.123%	92.598%	**93.057%**
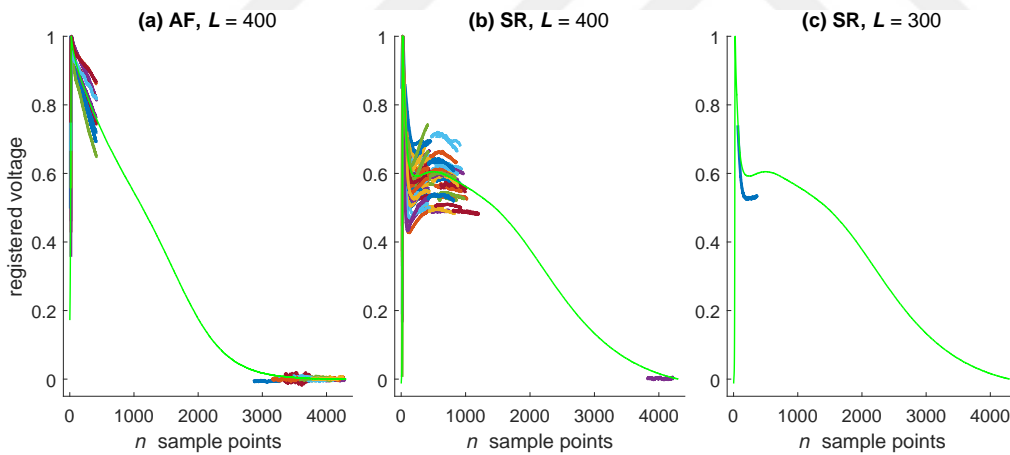
Figure 4.7 : Top 100 brute-force shapelets extracted with $L = 400$ from AF (a) and SR (b) instances, and the single fast shapelet (c) with $L = 300$ selected only among SR instances, overlaid with the associated class representative average time series T^A.

distance values obtained from the "fast shapelet tree" are considered as inputs for the listed classifiers of WEKA. It is assumed that *L* possesses the same range of values as those for the brute-force shapelet transform for the purpose of fairness.

According to Table 4.3, ensemble average classification accuracy (90.269%) peaks at $L = 300$ for which the single discovered shapelet among the instances of SR is shown in Fig. 4.7.

Table 4.3 : Individual and ensemble average classification accuracies of the fast shapelet transform applied to AP dataset for variations in shapelet length, L, and the selected classifiers.

	$L = 50$	$L = 100$	$L = 150$	$L = 200$	$L = 250$	$L = 300$	$L = 350$	$L = 400$
Naive Bayes	80.062%	81.308%	85.358%	86.604%	88.474%	90.654%	92.523%	88.474%
J48	78.193%	82.555%	86.293%	87.851%	88.474%	90.031%	90.966%	87.227%
Random forest	72.274%	82.866%	79.751%	81.620%	82.866%	87.227%	90.031%	78.816%
AdaBoost.M1	77.882%	80.997%	86.916%	87.851%	88.785%	90.654%	90.031%	86.604%
Classif. via regr.	79.439%	79.751%	85.358%	85.981%	89.720%	90.031%	90.966%	88.474%
Bagging	78.193%	84.112%	86.293%	87.227%	89.097%	91.277%	89.720%	87.851%
MultiBoostAB	77.570%	82.555%	87.851%	88.474%	88.785%	90.654%	89.720%	86.916%
LWL	78.816%	80.997%	85.047%	87.851%	88.785%	90.031%	89.408%	87.227%
PART	77.259%	82.555%	85.981%	87.851%	88.474%	90.343%	91.589%	87.227%
END	78.193%	82.555%	86.293%	87.851%	88.474%	90.654%	90.966%	87.227%
Decision Stump	78.816%	78.505%	87.227%	87.851%	88.785%	90.654%	90.343%	86.293%
Simple CART	78.505%	81.620%	86.604%	84.424%	89.097%	90.031%	90.343%	87.227%
RIMARC	80.062%	85.981%	87.227%	88.162%	90.654%	91.589%	93.458%	89.720%
Bayes NET	76.324%	82.866%	86.916%	87.851%	88.785%	90.654%	89.408%	86.604%
Dagging	78.505%	78.816%	78.193%	86.604%	89.408%	90.031%	90.654%	86.293%
Random SubSpace	77.570%	71.651%	86.916%	87.851%	88.785%	90.654%	84.735%	86.604%
Decision Table	78.816%	81.932%	86.916%	87.851%	88.785%	90.654%	88.474%	86.604%
Ridor	77.570%	82.243%	83.489%	85.358%	87.227%	90.031%	88.162%	85.670%
ADTree	77.259%	82.866%	84.735%	87.227%	88.474%	91.277%	90.654%	87.539%
LAD Tree	75.389%	82.243%	85.047%	85.981%	87.227%	91.277%	90.654%	87.851%
Random Tree	71.963%	80.685%	78.505%	81.620%	82.555%	87.227%	89.097%	79.128%
AVERAGE	77.555%	81.412%	85.091%	86.664%	88.177%	**90.269%**	90.090%	86.456%

The proposed WTC method is self-sufficient in predicting a time localized window length, w. However, results are also obtained by artificially enforcing a window length to reach complete coverage when compared to the two shapelet approaches. For this particular case, SCB extraction is performed by "by-passing" the computation part for w. In Table 4.4, the accuracies for variations in the enforced SCB lengths and for the selected classifiers are presented.

A comparison of Tables 4.1 and 4.4 for the columns corresponding to ($\delta = 0.9990$, $p = 0.95$) and $w = 200$ (also for $w = 150$), respectively, reveals the success of the PIP and DCT-based heuristic pursued by the proposed WTC method to determine the value of the local time window length w. Based on the results in Tables 4.1, 4.3 and 4.4, the classification accuracies calculated for WTC are around 4% higher than those of the fast shapelet transform. On the other hand, despite its exhaustive search with an impractical computational complexity, the peak ensemble average classification accuracy of the brute-force shapelet transform (93.057%) for $L = 400$ turns out to be slightly lower than that of WTC both for the enforced case with $w = 200$ (also for $w = 150$) (94.125%) and for the unenforced, i.e., the original case (94.333%).

4.2 ECG Dataset

In this section, an ECG dataset is examined to demonstrate the potential of the proposed WTC method from a clinical perspective. ECG is a highly clinically relevant

37

Table 4.4 : Individual and ensemble average classification accuracies of the proposed WTC method applied to AP dataset for varying enforced SCB length, *w*, and for the selected classifiers.

	w = 50	*w* = 100	*w* = 150	*w* = 200	*w* = 250	*w* = 300	*w* = 350	*w* = 400
Naive Bayes	91.277%	90.966%	92.212%	91.589%	91.277%	90.654%	90.654%	90.031%
J48	93.770%	90.654%	93.458%	93.770%	94.081%	94.393%	93.458%	93.146%
Random forest	94.081%	94.704%	94.393%	94.393%	94.081%	93.770%	94.081%	93.458%
AdaBoost.M1	94.704%	95.016%	95.950%	94.393%	95.639%	94.393%	96.262%	95.327%
Classif. via regr.	94.393%	95.016%	95.950%	96.885%	95.327%	95.327%	95.950%	95.639%
Bagging	94.704%	95.016%	95.016%	94.704%	94.393%	94.704%	94.393%	94.393%
MultiBoostAB	95.016%	95.016%	94.704%	95.016%	94.704%	94.704%	93.770%	93.458%
LWL	93.770%	94.704%	93.770%	94.393%	92.835%	94.081%	91.589%	91.900%
PART	93.146%	92.212%	91.589%	91.900%	94.081%	92.835%	94.704%	94.704%
END	93.458%	90.031%	93.458%	93.458%	94.081%	94.704%	93.146%	93.146%
Decision stump	94.393%	95.016%	93.770%	95.016%	94.081%	95.016%	91.900%	91.900%
Simple CART	95.327%	95.016%	94.081%	93.770%	93.458%	96.262%	93.146%	94.704%
RIMARC	95.327%	95.327%	95.327%	95.327%	95.327%	95.327%	95.327%	95.327%
Bayes NET	94.081%	93.770%	94.393%	94.081%	94.081%	94.081%	94.393%	94.081%
Dagging	93.458%	92.212%	94.393%	94.081%	91.900%	93.146%	92.835%	92.835%
Random SubSpace	93.770%	93.770%	94.393%	94.081%	93.770%	93.458%	94.393%	94.081%
Decision Table	94.081%	91.277%	90.343%	92.212%	93.146%	93.770%	92.523%	94.393%
Ridor	94.081%	91.900%	95.016%	94.704%	94.393%	94.081%	94.081%	94.393%
ADTree	93.146%	95.016%	96.573%	95.327%	94.704%	94.393%	94.393%	94.393%
LAD Tree	94.704%	94.704%	95.639%	94.704%	93.770%	94.704%	95.016%	94.393%
Random Tree	92.835%	90.654%	92.212%	92.835%	94.704%	92.212%	92.835%	92.835%
AVERAGE	93.977%	93.428%	**94.125%**	**94.125%**	93.992%	94.096%	93.755%	93.740%

and therefore one of the most studied biophysical signal types.

4.2.1   ECG Dataset Description

The ECG dataset used in this thesis is compiled from Physikalisch-Technische Bundesanstalt DataBase (PTBDB), a publicly available repository of physiological signals in PhysioNet [42]. PTBDB contains 549 12-lead ECG time series recordings from 290 unique patients. Each time series is digitized with 1 kHz sampling rate and 16-bit resolution over a signal amplitude range of ±16.384 mV. PTBDB contains instances with a wide variety of labels including myocardial infarction, cardiomyopathy/heart failure, dysrhythmia, myocardial hypertrophy and myocarditis. The largest subset of PTBDB is comprised of time series instances belonging to patients diagnosed with acute myocardial infarction (MI). There exist identified differences in the expression of P-QRS-T sequences of the ECG signal recorded from the precordial leads during MI [113]. Furthermore, ECG is known to contain noise originating from different sources [114],[115] and [116]. For this reason, we find this dataset suitable for testing robustness of the proposed WTC method against noise. Hereafter, the time series instances of this dataset will be referred to as MI and NR corresponding to patients diagnosed with acute MI and control patients with normal heart rhythms, respectively. In this context, out of 148 instances of MI patients, a sub-group of 40 unique patients (aged 37 to 85 with mean 61.03 and 22.50% female, 77.50% male) are randomly selected to obtain a balanced set against the 40 available NR patients (aged 17 to 81 with mean 45.59 and 25.00% female, 75.00% male). The record identifications of the
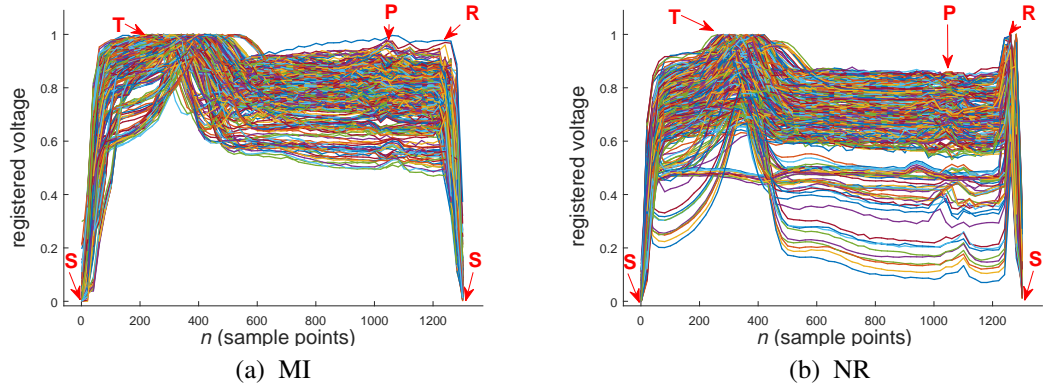
38

Figure 4.8 : Registered and annotated ECG lead v2 time series instances of (a) MI (40 instances) and (b) NR (40 instances) having 1308 sample points each.

selected instances are listed in Table B.1. For readers who are interested in further details on these patient instances as well as the localizations of their myocardial infarct, the medical records are accessible via PhysioBank ATM under the PTBDB "Record Description" section [117].

4.2.2 ECG Dataset Registration

All MI and NR time series instances are registered to prepare the same for the proposed method. Each cyclic ECG time series instance is first segmented into multiple single-cycle P-QRS-T sequences beginning from an S point to a subsequent point. In contrast to the acyclic recordings in the AP dataset, ECG recordings are self-repeating and their periodicities are determined by the heart rate of the patients. However, it is the signal variation among the P-QRS-T sequences as opposed to the heart rate variability that differentiates MI patients from NR patients. Hence, heart rate variability (i.e., length of the time series from an S point to a subsequent point) is compensated by incorporating a signal processing technique termed as re-sampling. Re-sampling essentially warps the specified time series, which constitutes a single-cycle ECG signal in this specific case. Each single-cycle P-QRS-T time series is re-sampled in order to extend its length to that of the longest one. It is assumed that $S_k(n)$ denotes the re-sampled single-cycle P-QRS-T time series, and the registered time series $T_k(n)$ is obtained by normalizing their amplitude as follows:

$$T_k(n) = \frac{S_k(n) - \min\{S_k(n)\}}{\max\{S_k(n)\} - \min\{S_k(n)\}}. \tag{4.2}$$

We study the three precordial leads of v2, v3 and v4 separately. In Figs. 4.8a and 4.8b, $T_k(n)$ of MI and NR classes are shown, respectively, for lead v2.

39

4.2.3 WTC Results for ECG Dataset

Out of 40 MI(NR) patients, 25 are reserved for the training set, and this leaves 15 patients for the test set. With respect to the training set, a single P-QRS-T sequence is allowed from each patient. However, multiple (either 3 or 4) sequences are collected from each patient to reach a total of 50 sequences for the test set.

As discussed in Section 4.1.3, the energy threshold δ and the confidence level p parameters are explored within a range of values to select the combination that yields the highest average classification with respect to the aforementioned 21 classification algorithms within WEKA. Classification accuracies for the lead v2 are presented in Table 4.5 for varying δ and p. Table 4.5 reveals that the average accuracy for lead v2 of ECG dataset peaks (98.095%) at the same parameter values ($\delta = 0.9990$ and $p = 0.95$) with the AP dataset in Section 4.1.3. Tables 4.6 and 4.7 show the individual accuracy values for varying parameters of δ and p, for leads v3 and v4 respectively. As seen form these tables, the aforementioned parameters $\delta = 0.9990$ and $p = 0.95$ are exceeded by at most 0.285% and 0.524% for leads v3 and v4 respectively. Therefore, we fix $\delta = 0.9990$ and $p = 0.95$ for the leads v3 and v4 as well proceeding with a single set of parameters for the rest of this thesis. With the chosen confidence level $p = 0.95$, the confidence bands constructed for the leads v2, v3, and v4 are shown in Figs. 4.9a, 4.9b and 4.9c, respectively.

Table 4.5 : Individual and ensemble average classification accuracies of the proposed WTC method applied to lead v2 of ECG dataset, for different values of the energy threshold δ, the confidence level p and for the selected classifiers.

	$\delta = 0.9900$		$\delta = 0.9990$		$\delta = 0.9999$	
	$p = 0.95$	$p = 0.99$	$p = 0.95$	$p = 0.99$	$p = 0.95$	$p = 0.99$
Naive Bayes	95.000%	99.000%	96.000%	95.000%	96.000%	94.000%
J48	97.000%	97.000%	98.000%	98.000%	98.000%	98.000%
Random forest	100.000%	100.000%	100.000%	99.000%	100.000%	99.000%
AdaBoost.M1	98.000%	98.000%	98.000%	98.000%	100.000%	100.000%
Classif. via regr.	99.000%	99.000%	100.000%	100.000%	94.000%	93.000%
Bagging	100.000%	99.000%	97.000%	97.000%	97.000%	97.000%
MultiBoostAB	100.000%	99.000%	98.000%	98.000%	98.000%	98.000%
LWL	93.000%	95.000%	97.000%	96.000%	99.000%	99.000%
PART	97.000%	97.000%	98.000%	98.000%	98.000%	98.000%
END	97.000%	97.000%	98.000%	98.000%	98.000%	98.000%
Decision stump	93.000%	93.000%	93.000%	93.000%	93.000%	93.000%
Simple CART	99.000%	99.000%	100.000%	100.000%	100.000%	100.000%
RIMARC	100.000%	100.000%	100.000%	100.000%	100.000%	100.000%
Bayes NET	100.000%	100.000%	98.000%	98.000%	98.000%	98.000%
Dagging	91.000%	94.000%	94.000%	94.000%	96.000%	90.000%
Random SubSpace	96.000%	96.000%	99.000%	98.000%	98.000%	98.000%
Decision Table	98.000%	99.000%	98.000%	97.000%	99.000%	99.000%
Ridor	99.000%	99.000%	100.000%	100.000%	100.000%	100.000%
ADTree	99.000%	99.000%	100.000%	100.000%	100.000%	100.000%
LAD Tree	99.000%	99.000%	100.000%	100.000%	100.000%	100.000%
Random Tree	98.000%	96.000%	98.000%	86.000%	92.000%	95.000%
AVERAGE	97.524%	97.810%	**98.095%**	97.286%	97.810%	97.476%

Table 4.6 : Individual and ensemble average classification accuracies of the proposed WTC method applied to lead v3 of ECG dataset, for different values of the energy threshold δ, the confidence level p and for the selected classifiers.

	$\delta = 0.9900$		$\delta = 0.9990$		$\delta = 0.9999$	
	$p = 0.95$	$p = 0.99$	$p = 0.95$	$p = 0.99$	$p = 0.95$	$p = 0.99$
Naive Bayes	95.000%	97.000%	92.000%	83.000%	88.000%	84.000%
J48	97.000%	97.000%	98.000%	98.000%	99.000%	99.000%
Random forest	99.000%	100.000%	99.000%	99.000%	100.000%	100.000%
AdaBoost.M1	98.000%	98.000%	99.000%	99.000%	100.000%	100.000%
Classif. via regr.	94.000%	95.000%	99.000%	99.000%	100.000%	100.000%
Bagging	96.000%	97.000%	99.000%	99.000%	99.000%	99.000%
MultiBoostAB	96.000%	94.000%	99.000%	99.000%	100.000%	100.000%
LWL	89.000%	92.000%	99.000%	99.000%	100.000%	100.000%
PART	97.000%	97.000%	98.000%	98.000%	99.000%	99.000%
END	97.000%	97.000%	98.000%	98.000%	99.000%	99.000%
Decision stump	86.000%	86.000%	99.000%	99.000%	100.000%	100.000%
Simple CART	94.000%	94.000%	99.000%	99.000%	100.000%	100.000%
RIMARC	99.000%	99.000%	99.000%	99.000%	99.000%	99.000%
Bayes NET	96.000%	96.000%	99.000%	99.000%	100.000%	100.000%
Dagging	94.000%	93.000%	93.000%	95.000%	96.000%	94.000%
Random SubSpace	95.000%	96.000%	99.000%	99.000%	99.000%	99.000%
Decision Table	97.000%	99.000%	99.000%	99.000%	99.000%	99.000%
Ridor	93.000%	93.000%	100.000%	99.000%	100.000%	100.000%
ADTree	98.000%	98.000%	100.000%	99.000%	100.000%	100.000%
LAD Tree	96.000%	96.000%	100.000%	97.000%	100.000%	100.000%
Random Tree	91.000%	90.000%	100.000%	90.000%	96.000%	96.000%
AVERAGE	95.095%	95.429%	98.429%	97.381%	**98.714%**	98.429%

For the sake of completeness, the confidence bands constructed with $p = 0.99$ are depicted in Appendix B.2 in Figs. B.1a, B.1b and B.1c for each ECG lead. In Figs. 4.10a, 4.10b and 4.10c, selected PIPs along with the class-representative average time series are shown for both MI (red) and NR (blue) classes for leads v2, v3 and v4, respectively.

For the remaining values of δ equals 0.9900 and 0.9999, the PIPs for each ECG lead are depicted in Appendix B.2 in Figs. B.2a, B.2b, B.2c and Figs. B.3a, B.3b, B.3c, respectively. Next, the confidence bands of each class are divided into SCBs as suggested by the local time windows W_j found after the PIP extraction step. Similar to the calculations made in Section 4.1.3, the weights α_j, β_j and γ_j are found for each W_j and depicted in Figs. 4.11a, 4.11c and 4.11e for MI and Figs. 4.11b, 4.11d and 4.11f for NR, respectively for leads v2, v3 and v4. In Appendix B.2, the weights α_j, β_j and γ_j of the corresponding SCBs constructed with the remaining combinations of $p \in \{0.95, 0.99\}$ and $\delta \in \{0.9900, 0.9990, 0.9999\}$ are depicted for MI and NR and for all leads in Figs. B.4a to B.4f, B.5a to B.5f, B.6a to B.6f, B.7a to B.7f and B.8a to B.8f.

The resulting similarity scores, Z_j, found with Equation 3.12 are shown in the form of color map plots for both MI and NR and for the three studied leads in Figs. 4.12a through 4.12f.

Table 4.7 : Individual and ensemble average classification accuracies of the proposed WTC method applied to lead v4 of ECG dataset, for different values of the energy threshold δ, the confidence level p and for the selected classifiers.

	$\delta = 0.9900$		$\delta = 0.9990$		$\delta = 0.9999$	
	$p = 0.95$	$p = 0.99$	$p = 0.95$	$p = 0.99$	$p = 0.95$	$p = 0.99$
Naive Bayes	97.000%	96.000%	96.000%	95.000%	91.000%	93.000%
J48	96.000%	94.000%	97.000%	97.000%	100.000%	100.000%
Random forest	98.000%	98.000%	99.000%	97.000%	100.00%	97.000%
AdaBoost.M1	98.000%	98.000%	97.000%	97.000%	99.000%	99.000%
Classif. via regr.	96.000%	96.000%	94.000%	94.000%	99.000%	97.000%
Bagging	95.000%	95.000%	95.000%	95.000%	96.000%	96.000%
MultiBoostAB	98.000%	98.000%	94.000%	94.000%	94.000%	96.000%
LWL	92.000%	93.000%	97.000%	97.000%	99.000%	99.000%
PART	96.000%	94.000%	97.000%	97.000%	100.000%	100.000%
END	96.000%	94.000%	97.000%	97.000%	100.000%	100.000%
Decision stump	92.000%	92.000%	94.000%	94.000%	97.000%	97.000%
Simple CART	96.000%	97.000%	95.000%	95.000%	96.000%	96.000%
RIMARC	99.000%	99.000%	99.000%	99.000%	99.000%	99.000%
Bayes NET	98.000%	99.000%	94.000%	94.000%	94.000%	96.000%
Dagging	88.000%	93.000%	94.000%	94.000%	90.000%	94.000%
Random SubSpace	98.000%	99.000%	93.000%	94.000%	94.000%	95.000%
Decision Table	96.000%	96.000%	97.000%	98.000%	93.000%	95.000%
Ridor	93.000%	93.000%	98.000%	98.000%	95.000%	95.000%
ADTree	98.000%	98.000%	99.000%	99.000%	98.000%	98.000%
LAD Tree	96.000%	96.000%	99.000%	99.000%	99.000%	99.000%
Random Tree	89.000%	87.000%	93.000%	92.000%	86.000%	88.000%
AVERAGE	95.476%	95.476%	96.095%	96.000%	96.143%	**96.619%**

These figures depict the degree of descriptiveness of the WTC method within each time window as mentioned in Section 4.1.3. The obvious difference in the thickness of the confidence bands when compared to those of the AP dataset indicates a higher variability of amplitude among the ECG instances.

As before, the color maps constructed for all leads of MI and NR with the remaining combinations of $p \in \{0.95, 0.99\}$ and $\delta \in \{0.9900, 0.9990, 0.9999\}$ are depicted in Figs. B.9a to B.9f, B.10a to B.10f, B.11a to B.11f, B.12a to B.12f and B.13a to B.13f, in Appendix B.2. The color maps relate to the weights in power of defining the intra-class similarity.

Characterizing intra-class similarity along the temporal dimension for each SCB, WTC also points out the morphological differences between the time series classes of interest. It is known in the prevalent cardiological literature that the earliest signs of acute MI include increased T-wave amplitude (defined as "hyper-acute") over the affected area. These so-called hyper-acute T-waves are most evident in the anterior precordial chest leads [68] (see Fig. 4.13a for the mentioned MI behavior in leads v2, v3 and v4). MI SCBs around the T-wave (Fig. 4.12a, 4.12c and 4.12e) are in relatively deeper red color. This observation aligns with the characteristic hyper-acute T-wave pattern expected to appear in the early phases of MI accompanied by the loss of R-wave amplitude in the anterior chest leads v2, v3 and v4 [40], [118].
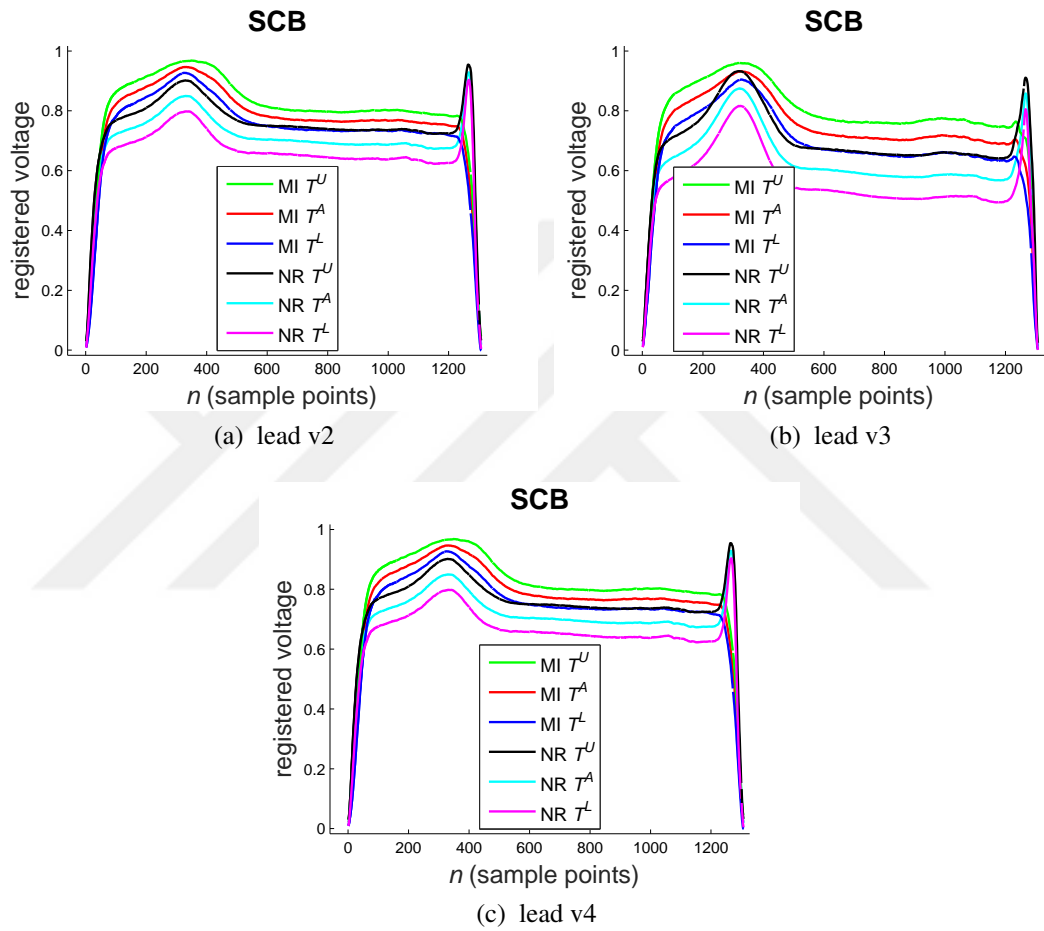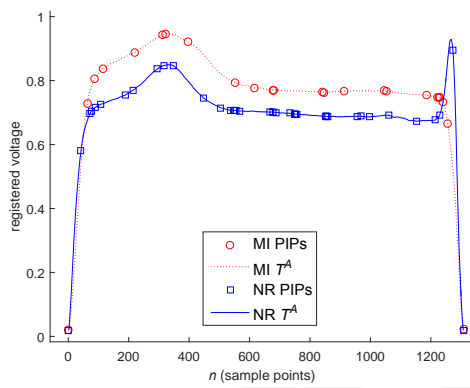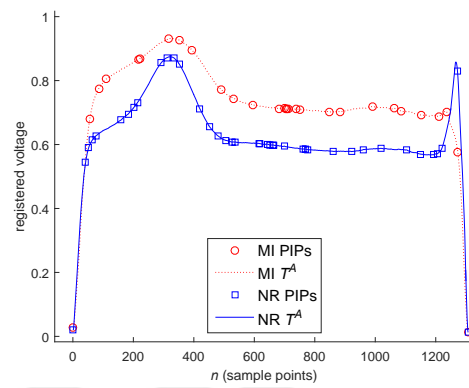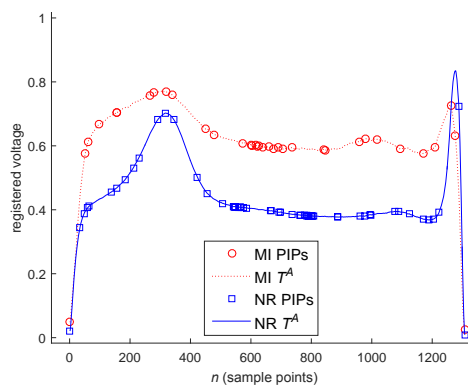
(a) lead v2

(b) lead v3

(c) lead v4

Figure 4.9 : Confidence bands with $p = 0.95$ overlaid for the MI and NR classes in (a)v2 (b)v3 and (c)v4 leads of ECG dataset.
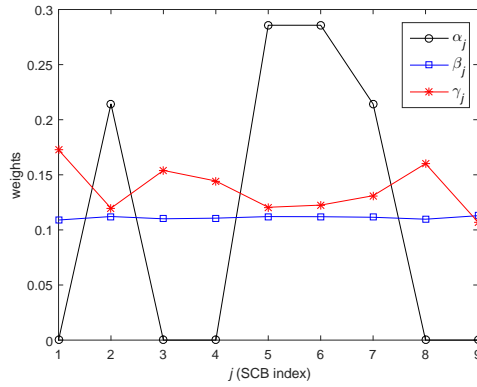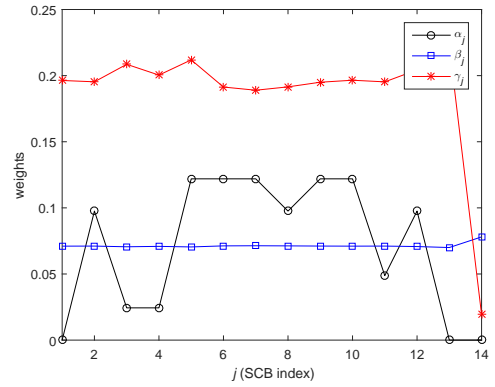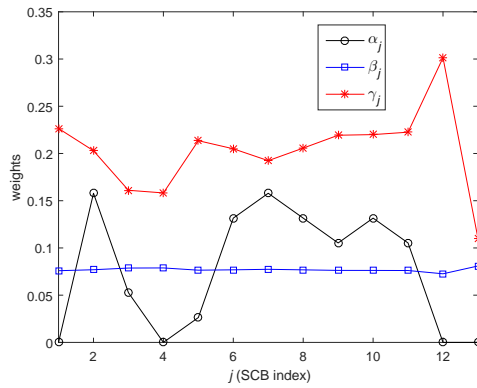
(a) lead v2

(b) lead v3

(c) lead v4

Figure 4.10 : Identified PIPs from the MI and NR class representative average time series from (a)v2 (b)v3 and (c)v4 leads of ECG dataset.
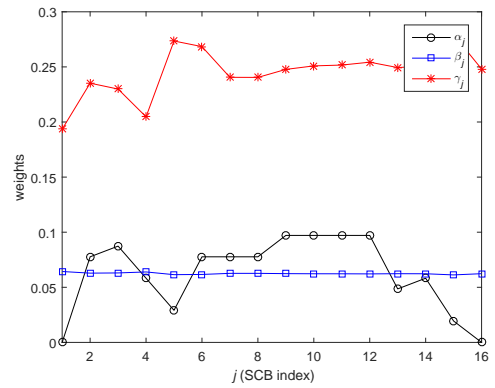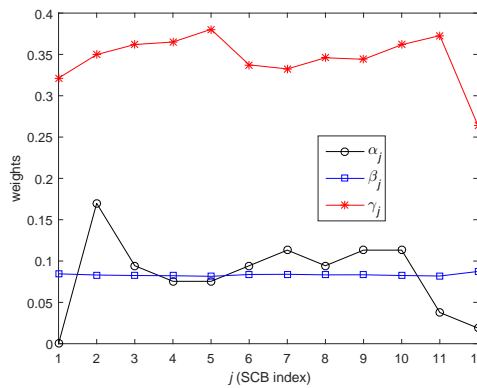
44

Figure 4.11 : The weights α_j, β_j and γ_j for SCBs in leads v2, v3 and v4 for (a,c,e) MI and (b,d,f) NR.
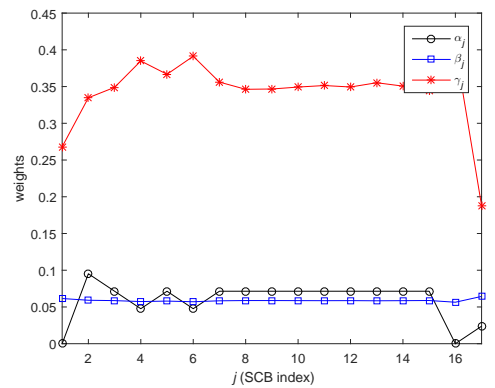
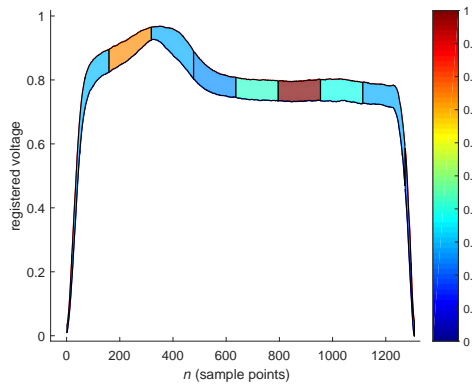(a) MI(lead v2)

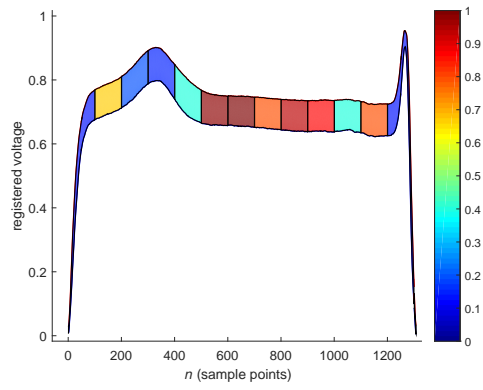(b) NR(lead v2)

(c) MI(lead v3)

(d) NR(lead v3)

(e) MI(lead v4)

(f) NR(lead v4)

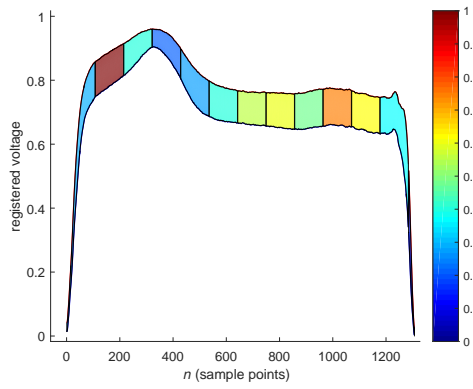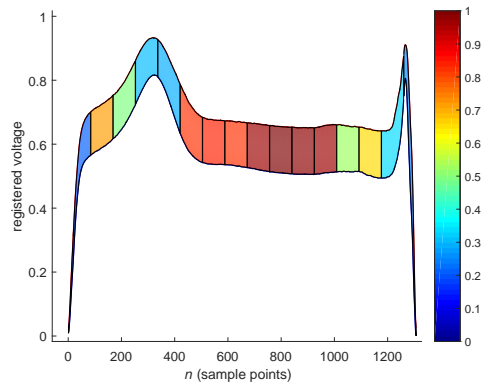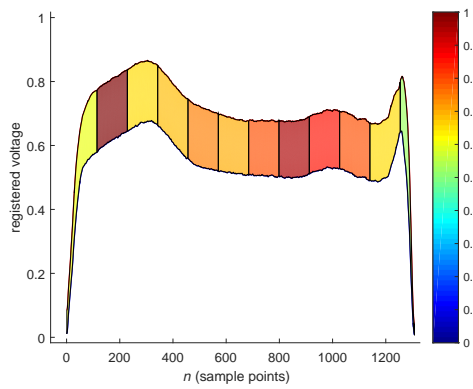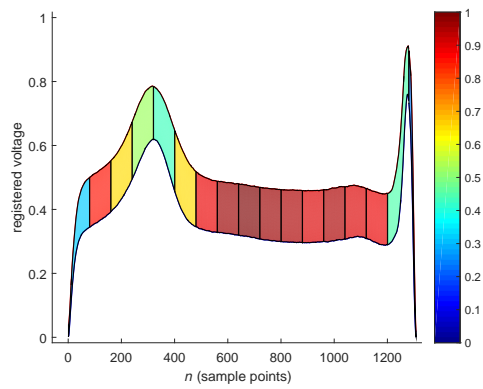Figure 4.12 : Color map plots for the average values of Z_j of leads v2, v3 and v4 for (a,c,e) MI and (b,d,f) NR.

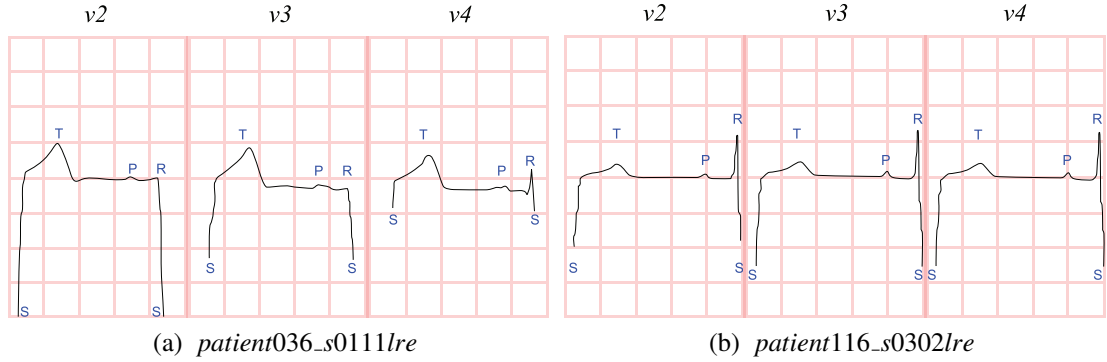(a) *patient*036_*s*0111*lre*   (b) *patient*116_*s*0302*lre*

Figure 4.13 : (a) Typical "hyper acute" T-waves in ECG recordings collected from precordial leads v2, v3 and v4 indicating the condition of MI during early stages (record id: *patient*036_*s*0111*lre*) and (b) waveforms from corresponding leads of a control patient (record id: *patient*116_*s*0302*lre*). Each unit cell represents 0.2 sec in time and 0.5 mV in amplitude. (Source: PhysioBank ATM)

4.2.4   Comparative Results for ECG Dataset

The following evaluations are performed separately on the leads v2, v3 and v4 of aforementioned training and test sets, comprising 25 MI and 25 NR and 15 MI and 15 NR, time series instances, respectively. As in Section 4.1.4, we sample a range for the desired shapelet length while obtaining results for the brute-force and fast shapelet transforms and limit the number of shapelets to be extracted to 100. Tables 4.8 and 4.9 tabulate the individual accuracies achieved by the chosen set of classification algorithms for the test set of lead v2 using features extracted by the brute-force and fast shapelet transform methods, respectively.

As shown in Table 4.8 and 4.9, the success of both the brute-force and fast shapelet transforms are highly sensitive to L. The first 100 shapelets with $L = 300$ that yield the highest ensemble average accuracy (94.381%) for the brute-force shapelet transform and lead v2 are depicted in Fig. 4.14 overlaid with the class average time series, T^A, for MI and NR. In Fig. 4.14, the fast shapelets of $L = 150$ with the highest ensemble average accuracy level (95.143%) are also shown. For the purpose of completeness, we present the resulting classification accuracies obtained for lead v3 and v4 in Tables 4.10 through 4.13 for the benchmark methods by noting that similar observations with lead v2 are performed.

Considering the accuracy results of the proposed WTC method presented in Section 4.2.3, it is apparent that the brute-force and fast shapelet transforms yield relatively lower accuracies for all three leads of the ECG dataset. The brute-force and fast shapelet transform methods attain accuracies of (94.381%, 95.190%, 95.190%) and (95.143%, 94.667%, 88.286%) for leads (v2, v3, v4), respectively, exceeded by those of the proposed WTC method which are (98.095%, 98.429%, 96.095%).

Although the proposed WTC method is self-sufficient in determining a proper local time window length, w, we also obtain results by artificially enforcing a window length in order to reach a complete coverage in comparisons with the two shapelet approaches. For this particular case, SCB extraction is performed by "by-passing" the computation part for w. In Tables 4.14, 4.15 and 4.16 the accuracies for varying

Table 4.8 : Individual and ensemble average classification accuracies of the brute-force shapelet transform applied to lead v2 of ECG dataset for variations in shapelet length, *L*, and the selected classifiers.

	$L = 50$	$L = 100$	$L = 150$	$L = 200$	$L = 250$	$L = 300$	$L = 350$	$L = 400$
Naive Bayes	92.000%	95.000%	94.000%	90.000%	92.000%	90.000%	90.000%	90.000%
J48	91.000%	93.000%	94.000%	93.000%	93.000%	96.000%	93.000%	94.000%
Random forest	92.000%	96.000%	97.000%	97.000%	98.000%	99.000%	98.000%	97.000%
AdaBoost.M1	91.000%	96.000%	97.000%	96.000%	98.000%	97.000%	95.000%	97.000%
Classif. via regr.	92.000%	95.000%	95.000%	91.000%	94.000%	95.000%	93.000%	94.000%
Bagging	92.000%	93.000%	91.000%	93.000%	94.000%	93.000%	93.000%	92.000%
MultiBoostAB	92.000%	93.000%	91.000%	93.000%	93.000%	93.000%	93.000%	93.000%
LWL	87.000%	94.000%	98.000%	95.000%	96.000%	90.000%	95.000%	94.000%
PART	92.000%	93.000%	94.000%	93.000%	94.000%	96.000%	93.000%	94.000%
END	91.000%	93.000%	94.000%	93.000%	93.000%	96.000%	93.000%	94.000%
Decision stump	85.000%	93.000%	95.000%	90.000%	94.000%	93.000%	92.000%	93.000%
Simple CART	90.000%	93.000%	95.000%	91.000%	96.000%	96.000%	92.000%	95.000%
RIMARC	95.000%	94.000%	94.000%	93.000%	94.000%	93.000%	93.000%	93.000%
Bayes NET	92.000%	93.000%	92.000%	93.000%	93.000%	93.000%	93.000%	93.000%
Dagging	91.000%	93.000%	93.000%	93.000%	93.000%	92.000%	89.000%	92.000%
Random SubSpace	92.000%	93.000%	93.000%	94.000%	93.000%	93.000%	92.000%	92.000%
Decision Table	83.000%	88.000%	84.000%	86.000%	86.000%	87.000%	87.000%	85.000%
Ridor	86.000%	93.000%	95.000%	92.000%	91.000%	96.000%	96.000%	95.000%
ADTree	91.000%	96.000%	97.000%	98.000%	98.000%	97.000%	98.000%	97.000%
LAD Tree	90.000%	97.000%	97.000%	96.000%	97.000%	98.000%	97.000%	98.000%
Random Tree	87.000%	90.000%	91.000%	94.000%	93.000%	99.000%	98.000%	96.000%
AVERAGE	90.190%	93.524%	93.857%	93.048%	93.952%	**94.381%**	93.476%	93.714%

Table 4.9 : Individual and ensemble average classification accuracies of the fast shapelet transform applied to lead v2 in ECG dataset for variations in shapelet length, *L*, and the selected classifiers.

	$L = 50$	$L = 100$	$L = 150$	$L = 200$	$L = 250$	$L = 300$	$L = 350$	$L = 400$
Naive Bayes	81.000%	90.000%	96.000%	91.000%	90.000%	86.000%	86.000%	93.000%
J48	82.000%	88.000%	95.000%	92.000%	92.000%	92.000%	92.000%	93.000%
Random forest	80.000%	85.000%	98.000%	92.000%	92.000%	93.000%	88.000%	87.000%
AdaBoost.M1	85.000%	89.000%	98.000%	91.000%	93.000%	92.000%	92.000%	93.000%
Classif. via regr.	81.000%	90.000%	92.000%	93.000%	93.000%	93.000%	93.000%	93.000%
Bagging	85.000%	90.000%	90.000%	93.000%	93.000%	93.000%	93.000%	94.000%
MultiBoostAB	85.000%	91.000%	93.000%	93.000%	93.000%	93.000%	93.000%	94.000%
LWL	80.000%	90.000%	98.000%	93.000%	93.000%	93.000%	92.000%	94.000%
PART	80.000%	88.000%	95.000%	92.000%	92.000%	92.000%	92.000%	93.000%
END	82.000%	88.000%	95.000%	92.000%	92.000%	92.000%	92.000%	93.000%
Decision stump	85.000%	90.000%	98.000%	93.000%	93.000%	93.000%	92.000%	94.000%
Simple CART	82.000%	90.000%	98.000%	91.000%	93.000%	94.000%	88.000%	94.000%
RIMARC	85.000%	92.000%	99.000%	93.000%	93.000%	93.000%	93.000%	94.000%
Bayes NET	85.000%	91.000%	91.000%	93.000%	93.000%	93.000%	93.000%	94.000%
Dagging	80.000%	88.000%	93.000%	93.000%	86.000%	83.000%	89.000%	91.000%
Random SubSpace	76.000%	87.000%	84.000%	84.000%	81.000%	93.000%	75.000%	84.000%
Decision Table	85.000%	90.000%	98.000%	93.000%	93.000%	93.000%	93.000%	94.000%
Ridor	83.000%	90.000%	98.000%	91.000%	90.000%	92.000%	91.000%	89.000%
ADTree	84.000%	85.000%	96.000%	92.000%	93.000%	95.000%	89.000%	88.000%
LAD Tree	81.000%	85.000%	97.000%	92.000%	93.000%	94.000%	90.000%	89.000%
Random Tree	83.000%	83.000%	96.000%	92.000%	92.000%	94.000%	88.000%	87.000%
AVERAGE	82.381%	88.571%	**95.143%**	91.857%	91.571%	92.190%	90.190%	91.667%

Table 4.10 : Individual and ensemble average classification accuracies of the brute-force shapelet transform applied to lead v3 in ECG dataset for variations in shapelet length, *L*, and the selected classifiers.

	$L = 50$	$L = 100$	$L = 150$	$L = 200$	$L = 250$	$L = 300$	$L = 350$	$L = 400$
Naive Bayes	95.000%	95.000%	96.000%	95.000%	96.000%	100.000%	100.000%	100.000%
J48	95.000%	90.000%	94.000%	99.000%	98.000%	97.000%	99.000%	99.000%
Random forest	96.000%	94.000%	97.000%	99.000%	100.000%	100.000%	100.000%	100.000%
AdaBoost.M1	97.000%	95.000%	96.000%	100.000%	98.000%	98.000%	98.000%	98.000%
Classif. via regr.	97.000%	93.000%	93.000%	100.000%	98.000%	98.000%	98.000%	98.000%
Bagging	95.000%	95.000%	94.000%	96.000%	96.000%	100.000%	100.000%	100.000%
MultiBoostAB	95.000%	95.000%	94.000%	96.000%	97.000%	100.000%	100.000%	100.000%
LWL	93.000%	90.000%	93.000%	100.000%	97.000%	100.000%	99.000%	99.000%
PART	95.000%	90.000%	94.000%	99.000%	98.000%	97.000%	99.000%	99.000%
END	95.000%	90.000%	94.000%	99.000%	98.000%	97.000%	99.000%	99.000%
Decision stump	96.000%	92.000%	93.000%	100.000%	98.000%	98.000%	98.000%	98.000%
Simple CART	96.000%	92.000%	93.000%	100.000%	98.000%	98.000%	98.000%	98.000%
RIMARC	98.000%	96.000%	96.000%	96.000%	97.000%	100.000%	100.000%	100.000%
Bayes NET	95.000%	95.000%	94.000%	96.000%	97.000%	100.000%	100.000%	100.000%
Dagging	90.000%	91.000%	84.000%	90.000%	90.000%	92.000%	89.000%	90.000%
Random SubSpace	95.000%	95.000%	96.000%	96.000%	97.000%	100.000%	100.000%	100.000%
Decision Table	85.000%	77.000%	85.000%	79.000%	71.000%	79.000%	83.000%	85.000%
Ridor	97.000%	91.000%	94.000%	99.000%	97.000%	98.000%	98.000%	98.000%
ADTree	97.000%	95.000%	94.000%	100.000%	98.000%	98.000%	98.000%	98.000%
LAD Tree	96.000%	92.000%	94.000%	100.000%	98.000%	98.000%	98.000%	98.000%
Random Tree	91.000%	91.000%	90.000%	94.000%	95.000%	99.000%	99.000%	98.000%
AVERAGE	94.714%	92.095%	93.238%	96.809%	95.809%	97.476%	97.762%	**97.857%**

Table 4.11 : Individual and ensemble average classification accuracies of the fast shapelet transform applied to lead v3 in ECG dataset for variations in shapelet length, *L*, and the selected classifiers.

	$L = 50$	$L = 100$	$L = 150$	$L = 200$	$L = 250$	$L = 300$	$L = 350$	$L = 400$
Naive Bayes	95.000%	92.000%	84.000%	81.000%	69.000%	81.000%	74.000%	76.000%
J48	94.000%	89.000%	82.000%	87.000%	77.000%	74.000%	74.000%	77.000%
Random forest	94.000%	87.000%	82.000%	91.000%	73.000%	68.000%	67.000%	72.000%
AdaBoost.M1	96.000%	91.000%	86.000%	87.000%	75.000%	80.000%	71.000%	78.000%
Classif. via regr.	95.000%	88.000%	81.000%	88.000%	76.000%	65.000%	73.000%	76.000%
Bagging	96.000%	88.000%	83.000%	87.000%	69.000%	78.000%	69.000%	71.000%
MultiBoostAB	97.000%	89.000%	82.000%	88.000%	73.000%	71.000%	70.000%	74.000%
LWL	97.000%	89.000%	82.000%	88.000%	72.000%	70.000%	74.000%	73.000%
PART	94.000%	89.000%	83.000%	86.000%	72.000%	74.000%	73.000%	77.000%
END	94.000%	89.000%	82.000%	87.000%	77.000%	74.000%	74.000%	77.000%
Decision stump	97.000%	89.000%	84.000%	88.000%	66.000%	69.000%	73.000%	76.000%
Simple CART	96.000%	91.000%	83.000%	88.000%	69.000%	76.000%	75.000%	76.000%
RIMARC	98.000%	94.000%	85.000%	93.000%	87.000%	86.000%	81.000%	81.000%
Bayes NET	97.000%	89.000%	83.000%	88.000%	71.000%	71.000%	73.000%	77.000%
Dagging	74.000%	73.000%	67.000%	73.000%	62.000%	60.000%	59.000%	67.000%
Random SubSpace	97.000%	71.000%	73.000%	86.000%	65.000%	71.000%	72.000%	73.000%
Decision Table	97.000%	89.000%	83.000%	88.000%	75.000%	80.000%	73.000%	78.000%
Ridor	94.000%	88.000%	83.000%	83.000%	71.000%	76.000%	72.000%	72.000%
ADTree	96.000%	89.000%	84.000%	88.000%	80.000%	75.000%	74.000%	71.000%
LAD Tree	96.000%	90.000%	83.000%	88.000%	71.000%	79.000%	76.000%	72.000%
Random Tree	94.000%	89.000%	85.000%	87.000%	76.000%	75.000%	70.000%	70.000%
AVERAGE	**94.667%**	87.762%	81.905%	86.667%	72.667%	73.952%	72.238%	74.476%

Table 4.12 : Individual and ensemble average classification accuracies of the brute-force shapelet transform applied to lead v4 in ECG dataset for variations in shapelet length, *L*, and the selected classifiers.

	L = 50	*L* = 100	*L* = 150	*L* = 200	*L* = 250	*L* = 300	*L* = 350	*L* = 400
Naive Bayes	84.000%	74.000%	70.000%	80.000%	77.000%	89.000%	95.000%	95.000%
J48	97.000%	92.000%	88.000%	87.000%	94.000%	96.000%	99.000%	96.000%
Random forest	89.000%	92.000%	88.000%	90.000%	96.000%	99.000%	99.000%	99.000%
AdaBoost.M1	96.000%	94.000%	82.000%	84.000%	87.000%	99.000%	99.000%	97.000%
Classif. via regr.	94.000%	93.000%	83.000%	85.000%	88.000%	93.000%	99.000%	96.000%
Bagging	73.000%	72.000%	81.000%	79.000%	82.000%	84.000%	90.000%	90.000%
MultiBoostAB	89.000%	74.000%	78.000%	78.000%	80.000%	86.000%	91.000%	91.000%
LWL	98.000%	91.000%	85.000%	84.000%	90.000%	94.000%	100.000%	95.000%
PART	97.000%	92.000%	87.000%	89.000%	92.000%	96.000%	99.000%	96.000%
END	97.000%	92.000%	88.000%	87.000%	94.000%	96.000%	99.000%	96.000%
Decision stump	98.000%	91.000%	83.000%	84.000%	89.000%	93.000%	99.000%	96.000%
Simple CART	98.000%	94.000%	83.000%	84.000%	89.000%	97.000%	99.000%	96.000%
RIMARC	91.000%	94.000%	86.000%	84.000%	85.000%	87.000%	95.000%	95.000%
Bayes NET	78.000%	74.000%	80.000%	78.000%	81.000%	86.000%	92.000%	91.000%
Dagging	72.000%	74.000%	76.000%	81.000%	80.000%	83.000%	85.000%	86.000%
Random SubSpace	77.000%	74.000%	79.000%	77.000%	81.000%	85.000%	89.000%	91.000%
Decision Table	91.000%	67.000%	72.000%	72.000%	74.000%	76.000%	83.000%	81.000%
Ridor	96.000%	93.000%	84.000%	88.000%	89.000%	92.000%	99.000%	94.000%
ADTree	96.000%	92.000%	88.000%	90.000%	93.000%	97.000%	99.000%	97.000%
LAD Tree	95.000%	91.000%	86.000%	87.000%	95.000%	97.000%	99.000%	98.000%
Random Tree	79.000%	71.000%	73.000%	77.000%	70.000%	91.000%	90.000%	95.000%
AVERAGE	89.762%	84.810%	81.905%	83.095%	86.000%	91.238%	**95.190%**	93.857%

Table 4.13 : Individual and ensemble average classification accuracies of the fast shapelet transform applied to lead v4 in ECG dataset for variations in shapelet length, *L*, and the selected classifiers.

|	*L* = 50	*L* = 100	*L* = 150	*L* = 200	*L* = 250	*L* = 300	*L* = 350	*L* = 400
Naive Bayes	58.000%	71.000%	88.000%	81.000%	85.000%	77.000%	77.000%	77.000%
J48	58.000%	75.000%	90.000%	79.000%	84.000%	74.000%	81.000%	82.000%
Random forest	59.000%	70.000%	91.000%	71.000%	81.000%	65.000%	64.000%	70.000%
AdaBoost.M1	64.000%	70.000%	87.000%	81.000%	86.000%	74.000%	77.000%	83.000%
Classif. via regr.	57.000%	72.000%	90.000%	81.000%	81.000%	75.000%	80.000%	83.000%
Bagging	53.000%	70.000%	88.000%	76.000%	83.000%	70.000%	79.000%	83.000%
MultiBoostAB	49.000%	70.000%	88.000%	77.000%	82.000%	74.000%	80.000%	83.000%
LWL	60.000%	74.000%	90.000%	76.000%	87.000%	73.000%	80.000%	83.000%
PART	58.000%	75.000%	90.000%	78.000%	83.000%	74.000%	81.000%	82.000%
END	58.000%	75.000%	90.000%	79.000%	84.000%	74.000%	81.000%	82.000%
Decision stump	55.000%	74.000%	90.000%	78.000%	87.000%	75.000%	80.000%	83.000%
Simple CART	61.000%	74.000%	88.000%	79.000%	82.000%	72.000%	80.000%	83.000%
RIMARC	77.000%	78.000%	97.000%	86.000%	89.000%	80.000%	83.000%	85.000%
Bayes NET	49.000%	68.000%	87.000%	78.000%	84.000%	75.000%	80.000%	83.000%
Dagging	57.000%	68.000%	82.000%	80.000%	80.000%	74.000%	77.000%	75.000%
Random SubSpace	49.000%	67.000%	78.000%	74.000%	76.000%	75.000%	80.000%	83.000%
Decision Table	58.000%	74.000%	87.000%	79.000%	87.000%	75.000%	80.000%	83.000%
Ridor	58.000%	71.000%	81.000%	77.000%	87.000%	72.000%	71.000%	81.000%
ADTree	68.000%	73.000%	93.000%	85.000%	77.000%	71.000%	77.000%	82.000%
LAD Tree	64.000%	73.000%	93.000%	82.000%	82.000%	72.000%	75.000%	81.000%
Random Tree	50.000%	70.000%	86.000%	70.000%	69.000%	72.000%	70.000%	67.000%
AVERAGE	58.095%	72.000%	**88.286%**	78.429%	82.667%	73.476%	77.762%	80.667%

50
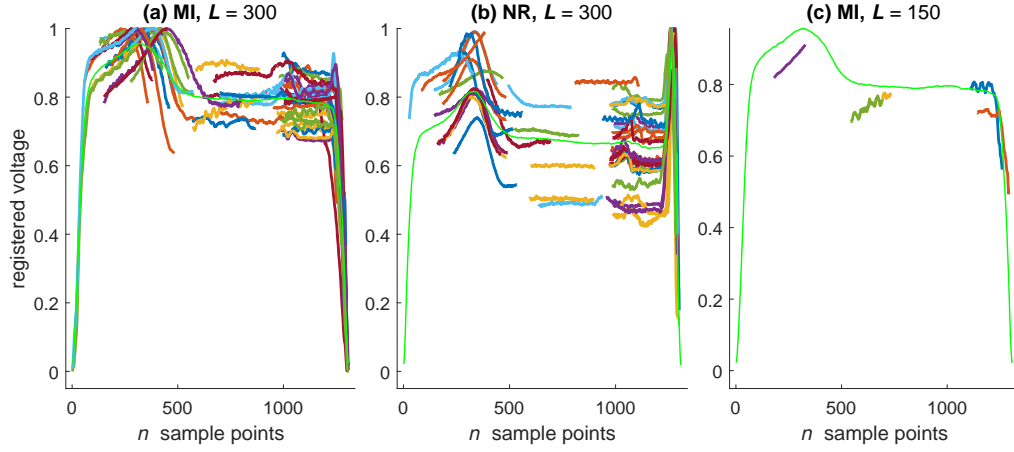
Figure 4.14 : Top 100 brute-force shapelets extracted with $L = 300$ from (a) MI and (b) NR instances, and (c) the fast shapelets with $L = 150$ selected only among MI instances, overlaid with the associated class representative average time series T^A of lead v2.

enforced SCB lengths and for the selected classifiers are presented.

Similar to the results for AP in Section 4.1, the enforced window length of WTC performs better than the shapelet-based methods.

4.3   Statistical Evaluation

We conclude the performance evaluation of the proposed WTC method by presenting a critical difference (CD) diagram [119]. The CD diagram is a representation that enables statistical evaluation of multiple methods over multiple datasets. Briefly, methods of interest are ranked for each dataset to obtain corresponding rank vectors that are subsequently averaged to determine average ranks. Methods whose rank differences exceed the value of CD are termed as "critically different". The CD is calculated as follows:

$$CD = q_\alpha \sqrt{\frac{F(F + 1)}{6R}}, \tag{4.3}$$

where F, R and q_α denote the number of methods, number of datasets and "critical value", respectively. In our present study, we use $F = 3$ methods, $R = 4$ datasets and $q_{0.1}$ which is equal to 2.052 for a two-tailed Nemenyi test. We refer the reader to [119] for further details about the CD diagram. In this analysis, we rank methods based on their average classification accuracies taken over 21 classification algorithms. We favor the benchmark methods by choosing their "best" L yielding the highest average classification accuracy for each dataset as tabulated in Table 4.17.

Table 4.14 : Individual and ensemble average classification accuracies of the proposed WTC method applied to lead v2 of the ECG dataset for varying enforced SCB length, w, and for the selected classifiers.

	$w = 50$	$w = 100$	$w = 150$	$w = 200$	$w = 250$	$w = 300$	$w = 350$	$w = 400$
Naive Bayes	95.000%	98.000%	97.000%	96.000%	97.000%	95.000%	90.000%	95.000%
J48	99.000%	98.000%	96.000%	97.000%	97.000%	98.000%	95.000%	98.000%
Random forest	99.000%	99.000%	99.000%	100.000%	99.000%	99.000%	98.000%	99.000%
AdaBoost.M1	97.000%	99.000%	98.000%	98.000%	98.000%	97.000%	96.000%	98.000%
Classif. via regr.	95.000%	94.000%	97.000%	98.000%	100.000%	97.000%	94.000%	98.000%
Bagging	97.000%	97.000%	100.000%	100.000%	99.000%	100.000%	99.000%	99.000%
MultiBoostAB	97.000%	98.000%	100.000%	100.000%	100.000%	100.000%	100.000%	99.000%
LWL	96.000%	100.000%	99.000%	97.000%	99.000%	94.000%	89.000%	95.000%
PART	99.000%	98.000%	96.000%	97.000%	97.000%	98.000%	95.000%	98.000%
END	99.000%	98.000%	96.000%	97.000%	97.000%	98.000%	95.000%	98.000%
Decision stump	93.000%	93.000%	92.000%	93.000%	93.000%	90.000%	88.000%	92.000%
Simple CART	100.000%	100.000%	99.000%	98.000%	100.000%	97.000%	95.000%	97.000%
RIMARC	100.000%	100.000%	100.000%	100.000%	100.000%	100.000%	100.000%	100.000%
Bayes NET	97.000%	98.000%	100.000%	100.000%	100.000%	100.000%	100.000%	99.000%
Dagging	92.000%	92.000%	96.000%	98.000%	94.000%	94.000%	91.000%	90.000%
Random SubSpace	96.000%	96.000%	99.000%	100.000%	100.000%	100.000%	98.000%	98.000%
Decision Table	98.000%	96.000%	100.000%	98.000%	99.000%	98.000%	99.000%	99.000%
Ridor	100.000%	100.000%	98.000%	96.000%	100.000%	97.000%	95.000%	97.000%
ADTree	100.000%	100.000%	99.000%	98.000%	100.000%	97.000%	95.000%	97.000%
LAD Tree	100.000%	100.000%	99.000%	99.000%	100.000%	97.000%	96.000%	97.000%
Random Tree	94.000%	98.000%	99.000%	97.000%	97.000%	98.000%	98.000%	99.000%
AVERAGE	97.286%	97.714%	98.048%	97.952%	**98.381%**	97.333%	95.524%	97.238%

Table 4.15 : Individual and ensemble average classification accuracies of the proposed WTC method applied to lead v3 of the ECG dataset for varying enforced SCB length, w, and for the selected classifiers.

	$w = 50$	$w = 100$	$w = 150$	$w = 200$	$w = 250$	$w = 300$	$w = 350$	$w = 400$
Naive Bayes	86.000%	92.000%	96.000%	97.000%	95.000%	97.000%	97.000%	97.000%
J48	97.000%	100.000%	99.000%	96.000%	97.000%	98.000%	99.000%	97.000%
Random forest	99.000%	100.000%	99.000%	99.000%	99.000%	100.000%	99.000%	99.000%
AdaBoost.M1	98.000%	99.000%	99.000%	97.000%	100.000%	99.000%	98.000%	98.000%
Classif. via regr.	98.000%	95.000%	99.000%	94.000%	97.000%	96.000%	97.000%	95.000%
Bagging	98.000%	98.000%	97.000%	99.000%	98.000%	97.000%	99.000%	98.000%
MultiBoostAB	99.000%	100.000%	98.000%	98.000%	99.000%	97.000%	97.000%	97.000%
LWL	98.000%	95.000%	94.000%	95.000%	96.000%	91.000%	94.000%	94.000%
PART	97.000%	100.000%	99.000%	96.000%	97.000%	98.000%	99.000%	97.000%
END	97.000%	100.000%	99.000%	96.000%	97.000%	98.000%	99.000%	97.000%
Decision stump	98.000%	95.000%	93.000%	95.000%	96.000%	88.000%	94.000%	88.000%
Simple CART	98.000%	96.000%	97.000%	94.000%	96.000%	97.000%	98.000%	97.000%
RIMARC	99.000%	97.000%	97.000%	99.000%	100.000%	97.000%	97.000%	97.000%
Bayes NET	99.000%	100.000%	98.000%	98.000%	99.000%	97.000%	96.000%	97.000%
Dagging	95.000%	93.000%	93.000%	94.000%	94.000%	94.000%	95.000%	97.000%
Random SubSpace	99.000%	98.000%	97.000%	100.000%	98.000%	96.000%	96.000%	97.000%
Decision Table	97.000%	98.000%	97.000%	94.000%	97.000%	97.000%	96.000%	98.000%
Ridor	98.000%	97.000%	100.000%	94.000%	99.000%	95.000%	98.000%	93.000%
ADTree	98.000%	98.000%	99.000%	99.000%	99.000%	98.000%	97.000%	98.000%
LAD Tree	98.000%	97.000%	98.000%	97.000%	98.000%	99.000%	98.000%	98.000%
Random Tree	95.000%	93.000%	91.000%	94.000%	97.000%	90.000%	92.000%	97.000%
AVERAGE	97.190%	97.190%	97.095%	96.429%	**97.524%**	96.143%	96.905%	96.476%

Table 4.16 : Individual and ensemble average classification accuracies of the proposed WTC method applied to lead v4 of the ECG dataset for varying enforced SCB length, w, and for the selected classifiers.

	$w = 50$	$w = 100$	$w = 150$	$w = 200$	$w = 250$	$w = 300$	$w = 350$	$w = 400$
Naive Bayes	93.000%	92.000%	93.000%	97.000%	97.000%	97.000%	97.000%	97.000%
J48	99.000%	98.000%	98.000%	98.000%	100.000%	96.000%	97.000%	98.000%
Random forest	100.00%	98.000%	97.000%	99.000%	98.000%	97.000%	97.000%	99.000%
AdaBoost.M1	99.000%	100.000%	99.000%	98.000%	98.000%	94.000%	98.000%	98.000%
Classif. via regr.	97.000%	99.000%	95.000%	97.000%	97.000%	93.000%	92.000%	95.000%
Bagging	95.000%	96.000%	97.000%	97.000%	98.000%	95.000%	98.000%	98.000%
MultiBoostAB	96.000%	97.000%	97.000%	99.000%	98.000%	98.000%	98.000%	98.000%
LWL	98.000%	99.000%	95.000%	99.000%	100.000%	96.000%	96.000%	94.000%
PART	99.000%	98.000%	98.000%	98.000%	100.000%	96.000%	96.000%	98.000%
END	99.000%	98.000%	98.000%	98.000%	100.000%	96.000%	97.000%	98.000%
Decision stump	97.000%	97.000%	95.000%	94.000%	97.000%	88.000%	91.000%	91.000%
Simple CART	96.000%	96.000%	95.000%	95.000%	96.000%	94.000%	89.000%	94.000%
RIMARC	100.000%	100.000%	100.000%	100.000%	100.000%	100.000%	100.000%	97.000%
Bayes NET	96.000%	99.000%	97.000%	99.000%	98.000%	98.000%	98.000%	98.000%
Dagging	87.000%	90.000%	94.000%	90.000%	96.000%	95.000%	94.000%	97.000%
Random SubSpace	96.000%	98.000%	96.000%	98.000%	98.000%	94.000%	98.000%	95.000%
Decision Table	97.000%	97.000%	93.000%	98.000%	98.000%	94.000%	97.000%	93.000%
Ridor	93.000%	94.000%	97.000%	96.000%	98.000%	91.000%	93.000%	94.000%
ADTree	99.000%	100.000%	99.000%	96.000%	97.000%	95.000%	98.000%	98.000%
LAD Tree	98.000%	100.000%	99.000%	99.000%	98.000%	96.000%	95.000%	98.000%
Random Tree	91.000%	92.000%	94.000%	93.000%	94.000%	81.000%	90.000%	94.000%
AVERAGE	96.429%	97.048%	96.476%	97.048%	**97.905%**	94.476%	95.667%	96.286%

Table 4.17 : Values of the shapelet length, L, used for critical difference diagram for the brute-force and fast shapelet transforms.

	Brute-force shapelet	Fast shapelet
AP	400	300
ECG lead v2	300	150
ECG lead v3	400	50
ECG lead v4	350	150

The resulting CD diagram is shown in Fig 4.15. The horizontal axis in the diagram represents the average ranks of each feature extraction method. The average rank based on classification accuracy improves from left to right.
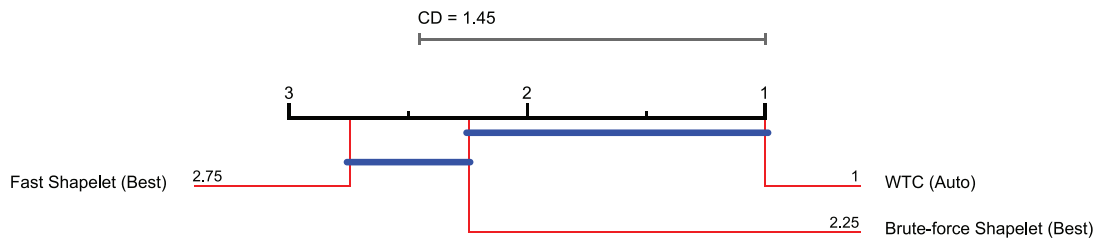
Figure 4.15 : Critical difference diagram based on the average classification accuracies achieved by WTC, the brute-force and fast shapelet transforms.

The CD is marked above the axis and corresponds to an indicator in rank magnitude that is required for the compared methods to differ such that they are termed as "critically different". The connecting blue lines in the diagram represent the groups of methods that are not critically different. As far as the datasets of this study are concerned, the proposed WTC method performs "critically better" than the fast shapelet transform. Although WTC has a higher average rank than the brute-force shapelet transform, the difference is not critically different.

CHAPTER 5

DISCUSSION AND CONCLUSION

In this study, we propose a feature extraction termed as WTC to characterize the signals in the form of time series. WTC is applied to a cardiac AP dataset with labeled SR and AF patients and three precordial leads of an ECG dataset that consist of control subjects and patients diagnosed with acute MI. Extracted feature vectors from these datasets are then examined for their classification accuracies that yield favorable results. Subsequently, brute-force and fast shapelet transforms are used to compare the performance of WTC in terms of predictive accuracy and computational complexity. Critical Difference (CD) analysis is performed on the datasets to reveal that WTC is "critically better" than fast shapelet transform. Although WTC and brute-force shapelet transforms are not "critically different", the performance of the former is slightly better.

Analog signal acquisition and digitization involve thermal and quantization noise and biophysical signals that are not exceptions in this aspect. Noise caused by motion artifacts resulting from electrode, probe, sensor, equipment, and even from patient movements constitutes disturbances in the desired signal in addition to thermal noise. Furthermore, almost all biophysical signals exhibit individualized polymorphism among instances to a certain extent as mentioned by [120] for ECGs, and this may hamper interpretation. WTC constructs the class representative time series as an ensemble average of all instances sharing a common class label and devises a mean trajectory that the population statistically follows within a confidence bound to mitigate the noise originated effects and variability among individual instances. Hence, as opposed to directly extracting features from the instances themselves, WTC uses a class representative trajectory for this purpose. The study findings indicate that this approach yields favorable results for AP and ECG datasets, and the latter is recognized as a signal that is usually contaminated with high levels of noise [114], [115], [116].

The DCT representation of the time series in conjunction with the extraction of perceptually important points (PIPs) constitutes a central part of the proposed WTC method since it allows the determination of a proper window length for local features. Briefly, DCT is utilized as a robust heuristic to feed a cut-off percentage to select the PIPs from the class representative average time series. As a widely used digital processing technique [121], DCT serves for the purpose of eliminating redundancy in the processed data. We note that, WTC determines the time windows and also assigns scores to indicating an order based on their descriptive power.

The presented results from completely different datasets reveal that the values of the

energy threshold (δ) and the confidence level (p) parameters that yield favorable results for the WTC method are almost the same, thereby demonstrating its consistency. In contrast to the length parameter (L) of the shapelet-based methods, the parameters of the WTC method are unitless and relatively less sensitive to the dataset of interest. Moreover, the accuracy cost of deviation from the values is relatively less when compared to those involved in the benchmark methods. The results of the study indicate that it is possible for prospective users of the WTC method to use the recommended values of the parameters ($\delta = 0.9990$ and $p = 0.95$) for different datasets without engaging in the parameter exploration step.

Execution time is an important criterion for online applications and especially for those involving big datasets. The $O(KN + N^2)$ complexity of the proposed WTC method outperforms the benchmark methods of brute-force, [28] and fast shapelet, [83] transforms in terms of computational complexity. In addition to the time-consuming training phase of shapelet-based transforms, [28], the discovered shapelets may not belong to the time windows associated with the expected morphologies of the class label of interest. Unconstrained selection of subsequences based solely on their minimum distance to other subsequences from arbitrary time intervals may hinder accurate assessment of the underlying dynamics. Moreover, the temporal location of a subsequence is typically associated with a specific importance for clinical interpretation that may be disregarded by the very nature of time series shapelets. For instance, if a shapelet is extracted from a temporal interval between sample points $n1$ and $n2$ on a time series instance and matched with another location between $n3$ and $n4$ as depicted in Fig. 5.1, then the ordinality of the sequence is lost. Additionally, the necessity of providing the shapelet length and total number of shapelets requires preliminary knowledge as well as forming assumptions about the dataset of interest.
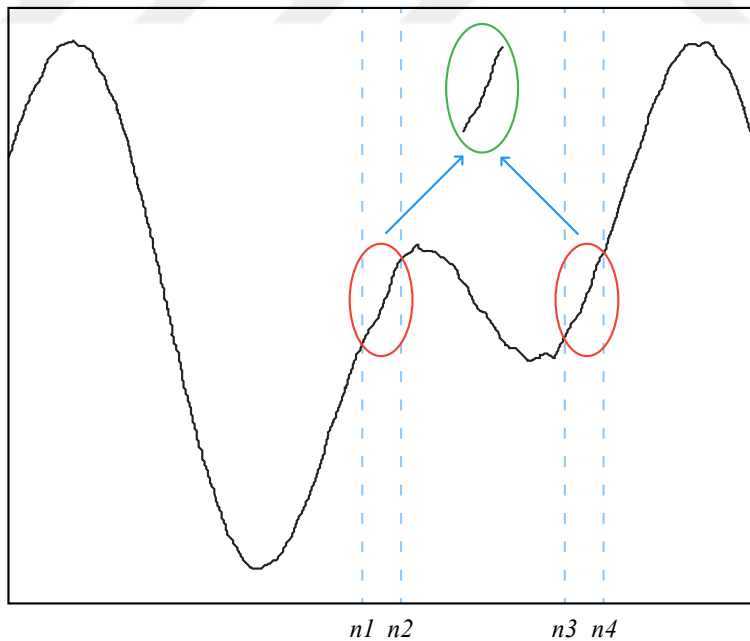
Figure 5.1 : An extracted shapelet from an example time series matching to multiple time windows ([$n1$, $n2$] and [$n3$, $n4$]) along the entire time series instance.

Our study indicates that time windows of WTC that coincide with phases 2 and 3

of the AP signal (Fig. 2.1) are most predictive for both classes. The plateau of AP (phase 2) is governed by the balance between calcium influx through L-type Ca^{2+} channels and potassium efflux via a plethora of K^+ channels, whereas late repolarization (phase 3) is dominated by K^+ current (I_{K1}) through inward rectifier channels (compare Fig. 2.2). The more positive potentials during early repolarization and the rapid final repolarization in AF correspond to reduced transient outward current and enhanced I_{K1}, respectively as induced by remodeling in chronic AF [36]. The high predictive power of phases 2 and 3 as shown in the present study (with reference to color maps in Figs. 4.6a and 4.6b) is in good agreement with a previous study on human atrial APs that introduced the analysis parameter "plateau potential" (PLT_{20}) that is defined as the average membrane potential of a time window between 20% and 30% of the AP duration at 90% of repolarization (APD_{90}) [122]. In fact, PLT_{20} (mV) and APD_{90} (ms) were ranked 5th and 1st, respectively, out of 62 features in classifying an instance as positive for AF [66]. The ability to classify an AP as AF or SR allows the formation of conclusions with respect to the underlying changes in ion channel properties and is of significance to predict antiarrhythmic drug action [123]. The findings on AP dataset do not point to a clinical outcome, but they hold an experimental relevance. Thus, it provides good real-life biophysical data to test the discriminative power of the WTC method aided by the preserved locality of the extracted features.

In addition to the AP dataset, analysis of the three chest leads of an ECG dataset is suitable for demonstrating the usability of the proposed method for clinical applications. With respect to the MI that corresponds to the condition of interest, certain leads in ECG recordings are expected to undergo predictable morphological and temporal changes including an elevation of the ST-segment (during STEMI) [68] and is usually preceded by subtle changes in the T-wave shape and size [124] and succeeded by a loss of the R-wave amplitude [40]. The distribution of the similarity scores across the time windows in Figs. 4.12a, 4.12c and 4.12e is coherent with the aforementioned morphologies expected to be observed in anterior chest leads of v2, v3 and v4, respectively. On the other hand, the resulting shapelets of the brute-force shapelet transform for lead v2 of MI time series instances (Fig. 4.14) do not particularly highlight any clinically relevant time windows. In a manner similar to the AP dataset, the resulting classification accuracies of brute-force shapelet transform for the ECG dataset (Tables 4.8, 4.10 and 4.12) show considerable variation with respect to the input parameter of shapelet length.

We believe WTC has potential clinical benefits including its integration related to emergency, ambulatory, inpatient, and home care online expert systems that commonly suffer from the misinterpretation of different symptoms [125]. Cases demanding urgent decision making, such as MI, require a digital assessment of the available measurements and decision support for triage. Early diagnosis and determining the severity of MI is vital for the patients, [126]. Prior to in-hospital interventions, it is advantageous to perform a preliminary analysis of the prehospital ECGs through mobile expert systems to alert patients, caregivers, or primary healthcare professionals [118]. Recently, ECG devices are equipped with various diagnostic detection mechanisms such as ST-segment monitoring to detect silent ischemia. Specifically for the case of ischemia, it is possible for the course of the disease to change rapidly and therefore a single randomly performed ECG may be insufficient to describe its extent. In order to devise a reasonable plan for treatment, extant studies indicate that

a continuum of ECGs should be assessed [127]. In addition to pointing out the incidence of a disorder, WTC is also suitable for distinguishing the various types of morphologies. For example, it is necessary to differentiate tall and symmetrical T waves that are observed during hyperkalemia from broad and rather skewed hyperacute T waves that typically occur in the early stages (within the first 30 minutes) of STEMI [128]. Furthermore, WTC can also contribute to the education of medical professionals in the course of their training. The "human readability" aspect of the algorithm allows practitioners to recognize patterns more easily and study underlying dynamics and possible "cause-and-effect" relations that require emphasis. Moreover, its modest computational complexity is such that it can be integrated with any kind of mobile or centralized framework as well as stand-alone offline units that require the rapid examination of densely sampled time series data.

The findings of this study indicate that WTC emerges a robust, fast, interpretable, and accurate time series feature extraction method based on the results presented for the atrial AP and ECG time series datasets. In addition to achieving relatively higher classification accuracies and lower execution times, the proposed method highlights the underlying dynamics that are typical of the inspected class label. WTC yields results in accordance with extant studies concerning atrial AP and ECG.

In conclusion, WTC emerges as a promising analysis tool with an unbiased parameter determination for biomedical researchers to conduct objective data mining analysis. Future studies will focus on the application of WTC to other biophysical time series such as heart auscultation, spirogram (for assessing pulmonary disabilities), electromyogram and electroneurogram recordings, event-related brain potentials, and regular measurements of blood pressure and blood sugar (for detecting hypo/hyperglycemic periods).

REFERENCES

[1] Ben Y Reis and Kenneth D Mandl. "Time series modeling for syndromic surveillance". In: *BMC Medical Informatics and Decision Making* 3.1 (2003), p. 2.

[2] Curtis E Kennedy and James P Turley. "Time series analysis as input for clinical predictive modeling: Modeling cardiac arrest in a pediatric ICU". In: *Theoretical Biology and Medical Modelling* 8.1 (2011), p. 40.

[3] Nabil Channouf, Pierre L'Ecuyer, Armann Ingolfsson, and Athanassios N Avramidis. "The application of forecasting techniques to modeling emergency medical system calls in Calgary, Alberta". In: *Health Care Management Science* 10.1 (2007), pp. 25–45.

[4] Sujit Kumar Sahoo, Wenmiao Lu, Sintiani Dewi Teddy, Desok Kim, and Mengling Feng. "Detection of atrial fibrillation from non-episodic ECG data: A review of methods". In: *33rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE. 2011, pp. 4992–4995.

[5] Ireneous N Soyiri and Daniel D Reidpath. "An overview of health forecasting". In: *Environmental Health and Preventive Medicine* 18.1 (2013), pp. 1–9.

[6] Tak-chung Fu. "A review on time series data mining". In: *Engineering Applications of Artificial Intelligence* 24.1 (2011), pp. 164–181.

[7] Femand S Cohen, Zhaohui Huang, and Zhengwei Yang. "Curve recognition using B-spline representation". In: *Applications of Computer Vision, Proceedings, 1992., IEEE Workshop on*. IEEE. 1992, pp. 213–220.

[8] Yue Wang and Earn Khwang Teoh. "A novel 2D shape matching algorithm based on B-spline modeling". In: *Image Processing, 2004. ICIP'04. 2004 International Conference on*. Vol. 1. IEEE. 2004, pp. 409–412.

[9] Jack M Rogers. "Modeling the cardiac action potential using B-spline surfaces". In: *Biomedical Engineering, IEEE Transactions on* 47.6 (2000), pp. 784–791.

[10] Tet Toe and Tang Van To. "Curve Matching by using B-spline Curves". In: *AU Journal of Technology* 5.3 (2002).

[11] Fabio Badilini, Arthur J Moss, and Edward L Titlebaum. "Cubic spline baseline estimation in ambulatory ECG recordings for the measurement of ST segment displacements". In: *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. Vol. 13. IEEE. 1991, pp. 584–585.

[12] Dusan Cakmakov and Emilija Celakoska. "Estimation of curve similarity using turning functions". In: *International Journal of Applied Mathematics* 15 (2004), pp. 403–416.

[13] Ferrari Marques and Guillermo Gutierrez. "Shape representation for object correspondence based on sub-graph matching and Fourier descriptors". In: *Signal Processing Conference (EUSIPCO 1998), 9th European*. IEEE. 1998, pp. 1–4.

[14] William J Rucklidge. "Locating objects using the Hausdorff distance". In: *Computer Vision, 1995. Proceedings., 5th International Conference on*. IEEE. 1995, pp. 457–464.

[15] Richard A Davis and Irene Hueter. "The convex hull of consecutive pairs of observations from some time series models". In: *Extremes* 16.4 (2013), pp. 487–505.

[16] Yong Se Kim. "Recognition of form features using convex decomposition". In: *Computer-Aided Design* 24.9 (1992), pp. 461–476.

[17] Vadim Shapiro. "A convex deficiency tree algorithm for curved polygons". In: *International Journal of Computational Geometry & Applications* 11.02 (2001), pp. 215–238.

[18] Markus Peura and Jukka Iivarinen. "Efficiency of simple shape descriptors". In: *Aspects of Visual Form* (1997), pp. 443–451.

[19] Nicholas I Fisher. *Statistical analysis of circular data*. Cambridge University Press, 1995.

[20] Miroslav Koprnicky, Maher Ahmed, and Mohamed Kamel. "Contour description through set operations on dynamic reference shapes". In: *International Conference Image Analysis and Recognition*. Springer. 2004, pp. 400–407.

[21] Farzin Mokhtarian, Sadegh Abbasi, and Josef Kittler. "Efficient and Robust Retrieval by Shape Content". In: *Image Databases and Multi-Media Search* 8 (1998), p. 51.

[22] Dengsheng Zhang and Guojun Lu. "Review of shape representation and description techniques". In: *Pattern Recognition* 37.1 (2004), pp. 1–19.

[23] Elias Masry. "Multivariate local polynomial regression for time series: Uniform strong consistency and rates". In: *Journal of Time Series Analysis* 17.6 (1996), pp. 571–599.

[24] Stefanos Manganaris. *Supervised classification with temporal data*. Vanderbilt University, 1997.

[25] Zitao Liu, Lei Wu, and Milos Hauskrecht. "Modeling clinical time series using Gaussian process sequences". In: *SIAM International Conference on Data Mining*. SIAM. 2013, pp. 623–631.

[26] S Roberts, M Osborne, M Ebden, S Reece, N Gibson, and S Aigrain. "Gaussian processes for time-series modelling". In: *Philosophical Transactions of*

the Royal Society of London A: Mathematical, Physical and Engineering Sciences 371.1984 (2013), p. 20110550.

[27] Jon Hills, Jason Lines, Edgaras Baranauskas, James Mapp, and Anthony Bagnall. "Classification of time series by shapelet transformation". In: *Data Mining and Knowledge Discovery* 28.4 (2014), pp. 851–881.

[28] Lexiang Ye and Eamonn Keogh. "Time series shapelets: a novel technique that allows accurate, interpretable and fast classification". In: *Data Mining and Knowledge Discovery* 22.1-2 (2011), pp. 149–182.

[29] Yonghong Peng, Peter A Flach, Pavel Brazdil, and Carlos Soares. "Decision tree-based data characterization for meta-learning". In: *IDDM-2002* (2002), p. 111.

[30] Ricardo BC Prudencio and Teresa B Ludermir. "Meta-learning approaches to selecting time series models". In: *Neurocomputing* 61 (2004), pp. 121–137.

[31] Maide Bucolo, Federica Di Grazia, Francesca Sapuppo, and Maria C Virzi. "A new approach for nonlinear time series characterization, DivA". In: *Control and Automation, 2008 16th Mediterranean Conference on*. IEEE. 2008, pp. 1284–1289.

[32] Selmo Eduardo Rodrigues and Ginalber Luiz de Oliveira Serra. "An evolving algorithm based on unobservable components neuro-fuzzy model for time series forecasting". In: *Evolving and Adaptive Intelligent Systems (EAIS), 2016 IEEE Conference on*. IEEE. 2016, pp. 122–129.

[33] Pierre-Francois Marteau and Gildas Menier. "Adaptive multiresolution and dedicated elastic matching in linear time complexity for time series data mining". In: *Intelligent Systems Design and Applications, 2006. ISDA'06. Sixth International Conference on*. Vol. 1. IEEE. 2006, pp. 700–706.

[34] Richard Povinelli. "Identifying temporal patterns for characterization and prediction of financial time series events". In: *Temporal, Spatial, and Spatio-Temporal Data Mining* (2001), pp. 46–61.

[35] David J Odde and Helen M Buettner. "Time series characterization of simulated microtubule dynamics in the nerve growth cone". In: *Annals of Biomedical Engineering* 23.3 (1995), pp. 268–286.

[36] Dobromir Dobrev and Ursula Ravens. "Remodeling of cardiomyocyte ion channels in human atrial fibrillation". In: *Basic Research in Cardiology* 98.3 (2003), pp. 137–148.

[37] Huhe Dai, Shouda Jiang, and Ye Li. "Atrial activity extraction from single lead ECG recordings: Evaluation of two novel methods". In: *Computers in Biology and Medicine* 43.3 (2013), pp. 176–183.

[38] PH Steg, Stefan K James, Dan Atar, Luigi P Badano, Carina Blomstrom Lundqvist, Michael A Borger, Carlo Di Mario, Kenneth Dickstein, Gregory Ducrocq, Francisco Fernandez-Aviles, Freek Verheugt, and Lars Wallentin. "ESC Guidelines for the management of acute myocardial infarction in pa-

tients presenting with ST-segment elevationThe Task Force on the management of ST-segment elevation acute myocardial infarction of the European Society of Cardiology (ESC)". In: *European Heart Journal* 33.20 (2012), pp. 2569–2619.

[39] Patrick T O'Gara, Frederick G Kushner, Deborah D Ascheim, Donald E Casey, Mina K Chung, James A De Lemos, Steven M Ettinger, James C Fang, Francis M Fesmire, Barry A Franklin, Christopher B Granger, Harlan M Krumholz, Jane A Linderbaum, David A Morrow, L Kristin Newby, Joseph P Ornato, Narith Ou, Martha J Radford, Jacqueline E Tamis-Holland, Carl L Tommaso, Cynthia M Tracy, Y Joseph Woo, and David X Zhao. "2013 ACCF/AHA Guideline for the management of ST-elevation myocardial infarction: A report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines". In: *Journal of the American College of Cardiology* 61.4 (2013), e78–e140.

[40] Jose Victor Nable and William Brady. "The evolution of electrocardiographic changes in ST-segment elevation myocardial infarction". In: *The American Journal of Emergency Medicine* 27.6 (2009), pp. 734–746.

[41] 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". In: *Circulation* 101.23 (2000), e215–e220. URL: https://physionet.org/.

[42] R Bousseljot, D Kreiseler, and A Schnabel. "Nutzung der EKG-Signaldatenbank CARDIODAT der PTB über das Internet". In: *Biomedizinische Technik/Biomedical Engineering* 40.s1 (1995), pp. 317–318. URL: https://physionet.org/physiobank/database/ptbdb/.

[43] Josef Melcr, Daniel Bonhenry, Stepan Timr, and Pavel Jungwirth. "Transmembrane potential modeling: comparison between methods of constant electric field and ion imbalance". In: *Journal of chemical theory and computation* 12.5 (2016), pp. 2418–2425.

[44] Ove Sten-Knudsen. *Biological membranes: theory of transport, potentials and electric impulses*. Cambridge University Press, 2002.

[45] Richard Klabunde. *Cardiovascular physiology concepts*. Lippincott Williams & Wilkins, 2011.

[46] Alan Lloyd Hodgkin. "The ionic basis of electrical activity in nerve and muscle". In: *Biological Reviews* 26.4 (1951), pp. 339–409.

[47] Yoram Rudy. "Molecular basis of cardiac action potential repolarization". In: *Annals of the New York Academy of Sciences* 1123.1 (2008), pp. 113–118.

[48] Arthur Keith and Martin Flack. "The form and nature of the muscular connections between the primary divisions of the vertebrate heart". In: *Journal of anatomy and physiology* 41.Pt 3 (1907), p. 172.

[49]   Stefan Kurtenbach, Sarah Kurtenbach, and Georg Zoidl. "Gap junction modulation and its implications for heart function". In: *Frontiers in physiology* 5 (2014), p. 82.

[50]   Donald M Bers. "Cardiac excitation–contraction coupling". In: *Nature* 415.6868 (2002), pp. 198–205.

[51]   Masayuki Yazawa, Brian Hsueh, Xiaolin Jia, Anca M Pasca, Jonathan A Bernstein, Joachim Hallmayer, and Ricardo E Dolmetsch. "Using induced pluripotent stem cells to investigate cardiac phenotypes in Timothy syndrome". In: *Nature* 471.7337 (2011), pp. 230–234.

[52]   *The action potential in typical cardiomyocytes by Grigoriy Ikonnikov and Eric Wong*. 2005. URL: http://www.pathophys.org/physiology-of-cardiac-conduction-and-contractility/#Electrophysiology.

[53]   Andre G Kleber and Yoram Rudy. "Basic mechanisms of cardiac impulse propagation and associated arrhythmias". In: *Physiological reviews* 84.2 (2004), pp. 431–488.

[54]   GE Hall and AC Guyton. *Textbook of Medical Physiology. Sounders Elsevier, Philadelphia (2011)*. Tech. rep. ISBN 978-1-4160-4574-8.

[55]   Peter G Danias, Todd A Caulfield, Marilyn J Weigner, David I Silverman, and Warren J Manning. "Likelihood of spontaneous conversion of atrial fibrillation to sinus rhythm". In: *Journal of the American College of Cardiology* 31.3 (1998), pp. 588–592.

[56]   Bünyamin Yavuz, Naim Ata, Emre Oto, Deniz Katircioglu-Öztürk, Kudret Aytemir, Banu Evranos, Rasim Koselerli, Emre Ertugay, Abdulkadir Burkan, Emrah Ertugay, Christ P Gale, A John Camm, and Ali Oto. "Demographics, treatment and outcomes of atrial fibrillation in a developing country: the population-based TuRkish Atrial Fibrillation (TRAF) cohort." In: *Europace: European pacing, arrhythmias, and cardiac electrophysiology: journal of the working groups on cardiac pacing, arrhythmias, and cardiac cellular electrophysiology of the European Society of Cardiology* (2017).

[57]   Paulus Kirchhof, Angelo Auricchio, Jeroen Bax, Harry Crijns, John Camm, Hans-Christoph Diener, Andreas Goette, Gerd Hindricks, Stefan Hohnloser, Lukas Kappenberger, Karl-Heinz Kuck, Gregory YH Lip, Bertil Olsson, Thomas Meinertz, Silvia Priori, Ursula Ravens, Gerhard Steinbeck, Elisabeth Svernhage, Jan Tijssen, Alphons Vincent, and Günter Breithardt. "Outcome parameters for trials in atrial fibrillation". In: *Europace* (2007).

[58]   Massimo Zoni-Berisso, Fabrizio Lercari, Tiziana Carazza, and Stefano Domenicucci. "Epidemiology of atrial fibrillation: European perspective". In: *Clin Epidemiol* 6.213 (2014), e220.

[59]   Centers for Disease Control and Prevention. "Atrial fibrillation fact sheet". In: *Fact Sheets and At-a-Glance Reports* (2010).

[60]   Valentin Fuster, Lars E Rydén, David S Cannom, Harry J Crijns, Anne B Curtis, Kenneth A Ellenbogen, Jonathan L Halperin, Jean-Yves Le Heuzey,

G Neal Kay, James E Lowe, S Bertil Olsson, Eric N Prystowsky, Juan Luis Tamargo, and Samuel Wann. In: *Journal of the American College of Cardiology* 48.4 (2006), pp. 854–906.

[61]   Caleb Ferguson, Sally C Inglis, Phillip J Newton, Sandy Middleton, Peter S Macdonald, and Patricia M Davidson. "Atrial fibrillation: Stroke prevention in focus". In: *Australian Critical Care* 27.2 (2014), pp. 92–98.

[62]   MANFRED Thurmann and James G Janney. "The diagnostic importance of fibrillatory wave size". In: *Circulation* 25.6 (1962), pp. 991–994.

[63]   Galen S Wagner, Peter Macfarlane, Hein Wellens, Mark Josephson, Anton Gorgels, David M Mirvis, Olle Pahlm, Borys Surawicz, Paul Kligfield, Rory Childers, and Leonard S Gettes. "AHA/ACCF/HRS recommendations for the standardization and interpretation of the electrocardiogram". In: *Circulation* 119.10 (2009), e262–e270.

[64]   *ECG material by James Heilman, MD - Own work, CC BY-SA 3.0*. 2011. URL: `https://commons.wikimedia.org/w/index.php?curid=19126125`.

[65]   Ulrich Schotten, Sander Verheule, Paulus Kirchhof, and Andreas Goette. "Pathophysiological mechanisms of atrial fibrillation: A translational appraisal". In: *Physiological Reviews* 91.1 (2011), pp. 265–325.

[66]   Ursula Ravens, Deniz Katircioglu-Öztürk, Erich Wettwer, Torsten Christ, Dobromir Dobrev, Niels Voigt, Claire Poulet, Simone Loose, Jana Simon, Agnes Stein, Klaus Matschke, Michael Knaut, Emre Oto, Ali Oto, and H. Altay Güvenir. "Application of the RIMARC algorithm to a large data set of action potentials and clinical parameters for risk prediction of atrial fibrillation". In: *Medical & Biological Engineering & Computing* 53.3 (2015), pp. 263–273.

[67]   Thomas N Martin, Bjoern A Groenning, Heather M Murray, Tracey Steedman, John E Foster, Alex T Elliot, Henry J Dargie, Ronald H Selvester, Olle Pahlm, and Galen S Wagner. "ST-Segment Deviation Analysis of the Admission 12-Lead Electrocardiogram as an Aid to Early Diagnosis of Acute Myocardial Infarction With a Cardiac Magnetic Resonance Imaging Gold Standard". In: *Journal of the American College of Cardiology* 50.11 (2007), pp. 1021–1028.

[68]   Kristian Thygesen, Joseph S Alpert, Allan S Jaffe, Maarten L Simoons, Bernard R Chaitman, and Harvey D White. "Third universal definition of myocardial infarction". In: *Circulation* 126.16 (2012), pp. 2020–2035.

[69]   Elizabeth G Nabel and Eugene Braunwald. "A tale of coronary artery disease and myocardial infarction". In: *New England Journal of Medicine* 366.1 (2012), pp. 54–63.

[70]   Antoni Bayés de Luna, Galen Wagner, Yochai Birnbaum, Kjell Nikus, Miguel Fiol, Anton Gorgels, Juan Cinca, Peter M Clemmensen, Olle Pahlm, Samuel Sclarovsky, Shlomo Stern, Hein Wellens, and Wojciech Zareba. "A New terminology for left ventricular walls and location of myocardial infarcts that

present Q wave based on the standard of cardiac magnetic resonance imaging". In: *Circulation* 114.16 (2006), pp. 1755–1760.

[71]   Wei Liang, Yinlong Zhang, Jindong Tan, and Yang Li. "A novel approach to ECG classification based upon two-layered HMMs in body sensor networks". In: *Sensors* 14.4 (2014), pp. 5994–6011.

[72]   *ECG leads by The University of Nottingham, Division of Nursing, Cardiology Teaching Package*. 2017. URL: http://www.nottingham.ac.uk/nursing/practice/resources/cardiology/function/index.php.

[73]   Anthony J Trevor, Bertram G Katzung, Susan B Masters, and Marieke Kruidering-Hall. *Pharmacology Examination & Board Review*. McGraw-Hill Medical New York, 2010.

[74]   Maria Sejersten Ripa. *The ECG as decision support in STEMI*. Maria Sejersten Ripa, 2011.

[75]   Marco Roffi, Carlo Patrono, Jean-Philippe Collet, Christian Mueller, Marco Valgimigli, Felicita Andreotti, Jeroen J Bax, Michael A Borger, Carlos Brotons, Derek P Chew, Baris Gencer, Gerd Hasenfuss, Keld Kjeldsen, Patrizio Lancellotti, Ulf Landmesser, Julinda Mehilli, Debabrata Mukherjee, Robert F. Storey, and Stephan Windecker. "2015 ESC Guidelines for the management of acute coronary syndromes in patients presenting without persistent ST-segment elevation". In: *European heart journal* (2015), ehv320.

[76]   Gan-Xin Yan, Ramarao S Lankipalli, James F Burke, Simone Musco, and Peter R Kowey. "Ventricular repolarization components on the electrocardiogram: cellular basis and clinical significance". In: *Journal of the American College of Cardiology* 42.3 (2003), pp. 401–409.

[77]   Chris De Zwaan, Frits WHM Bär, and Hein JJ Wellens. "Characteristic electrocardiographic pattern indicating a critical stenosis high in left anterior descending coronary artery in patients admitted because of impending myocardial infarction". In: *American heart journal* 103.4 (1982), pp. 730–736.

[78]   Ezra A Amsterdam, Nanette K Wenger, Ralph G Brindis, Donald E Casey, Theodore G Ganiats, David R Holmes, Allan S Jaffe, Hani Jneid, Rosemary F Kelly, Michael C Kontos, Glenn N Levine, Philip R Liebson, Debabrata Mukherjee, Eric D Peterson, Marc S Sabatine, Richard W Smalling, and Susan J Zieman. "2014 AHA/ACC guideline for the management of patients with non-ST-elevation acute coronary syndromes". In: *Circulation* (2014), CIR–0000000000000134.

[79]   Giora Landesberg, Morris Mosseri, Yehuda Wolf, Yellena Vesselov, and Charles Weissman. "Perioperative Myocardial Ischemia and InfarctionIdentification by Continuous $12 - lead$ Electrocardiogram with Online ST-segment Monitoring". In: *The Journal of the American Society of Anesthesiologists* 96.2 (2002), pp. 264–270.

[80]   Diego di Bernardo and Alan Murray. "Medical physics: Explaining the T-wave shape in the ECG". In: *Nature* 403.6765 (2000), pp. 40–40.

[81] Tetsuro Emori and Charles Antzelevitch. "Cellular basis for complex T waves and arrhythmic activity following combined IKr and IKs block". In: *Journal of cardiovascular electrophysiology* 12.12 (2001), pp. 1369–1378.

[82] 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*. ACM. 2011, pp. 1154–1162.

[83] Thanawin Rakthanmanon and Eamonn Keogh. "Fast shapelets: A scalable algorithm for discovering time series shapelets". In: *Proceedings of the 13th SIAM Conference on Data Mining (SDM)*. SIAM. 2013, pp. 668–676.

[84] Martin Wistuba, Josif Grabocka, and Lars Schmidt-Thieme. "Ultra-Fast Shapelets for Time Series Classification". In: *CoRR* abs/1503.05018 (2015).

[85] Xavier Renard, Maria Rifqi, Walid Erray, and Marcin Detyniecki. "Random-shapelet: an algorithm for fast shapelet discovery". In: *IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. 2015.

[86] Jessica Lin, Eamonn Keogh, Li Wei, and Stefano Lonardi. "Experiencing SAX: a novel symbolic representation of time series". In: *Data Mining and Knowledge Discovery* 15.2 (2007), pp. 107–144.

[87] Eamonn Keogh, Kaushik Chakrabarti, Michael Pazzani, and Sharad Mehrotra. "Dimensionality reduction for fast similarity search in large time series databases". In: *Knowledge and information Systems* 3.3 (2001), pp. 263–286.

[88] Saeid Soheily-Khah, Ahlame Douzal-Chouakria, and Eric Gaussier. "Progressive and Iterative Approaches for Time Series Averaging". In: *Proceedings of AALTD 2015* (2015), p. 123.

[89] Tak-chung Fu, Fu-lai Chung, Ka-yan Kwok, and Chak-man Ng. "Stock time series visualization based on data point importance". In: *Engineering Applications of Artificial Intelligence* 21.8 (2008), pp. 1217–1232.

[90] Fu-Lai Chung, Tak-Chung Fu, R Luk, and V Ng. "Flexible time series pattern matching based on Perceptually Important Points". In: *International Joint Conference on Artificial Intelligence IJCAI-01 Workshop on Learning from Temporal and Spatial Data*. 2001, pp. 1–7.

[91] Gao-Zhan Yu, Hong Peng, and Qi-Lun Zheng. "Pattern distance of time series based on segmentation by important points". In: *Machine Learning and Cybernetics, 2005. Proceedings of 2005 International Conference on*. Vol. 3. IEEE. 2005, pp. 1563–1567.

[92] Sung-Hyuk Cha and Sargur N Srihari. "On measuring the distance between histograms". In: *Pattern Recognition* 35.6 (2002), pp. 1355–1370.

[93] *WTC Source Codes, GitHub Repository*. 2017. URL: https://github.com/denizkatircioglu/WTC.

[94] M Narasimha and A Peterson. "On the computation of the discrete cosine transform". In: *IEEE Transactions on Communications* 26.6 (1978), pp. 934–936.

[95] *Weka 3: Data Mining Software in Java.* 2015. URL: `http://www.cs.waikato.ac.nz/ml/weka/index.html/`.

[96] Leo Breiman. "Random forests". In: *Machine Learning* 45.1 (2001), pp. 5–32.

[97] Yoav Freund and Robert E Schapire. "Experiments with a new boosting algorithm". In: *ICML*. Vol. 96. 1996, pp. 148–156.

[98] Eibe Frank, Yong Wang, Stuart Inglis, Geoffrey Holmes, and Ian H Witten. "Using model trees for classification". In: *Machine Learning* 32.1 (1998), pp. 63–76.

[99] Leo Breiman. "Bagging predictors". In: *Machine Learning* 24.2 (1996), pp. 123–140.

[100] Geoffrey I. Webb. "MultiBoosting: A Technique for Combining Boosting and Wagging". In: *Machine Learning* Vol.40.No.2 (2000).

[101] Christopher G Atkeson, Andrew W Moore, and Stefan Schaal. "Locally weighted learning for control". In: *Lazy Learning*. Springer, 1997, pp. 75–113.

[102] Eibe Frank and Ian H Witten. "Generating accurate rule sets without global optimization". In: (1998).

[103] Eibe Frank and Stefan Kramer. "Ensembles of nested dichotomies for multiclass problems". In: *Proceedings of the 21st International Conference on Machine Learning*. ACM. 2004, p. 39.

[104] Wayne Iba and Pat Langley. "Induction of one-level decision trees". In: *Proceedings of the 9th International Conference on Machine Learning*. 1992, pp. 233–240.

[105] Leo Breiman, Jerome Friedman, Charles J Stone, and Richard A Olshen. *Classification and regression trees*. CRC press, 1984.

[106] H Altay Güvenir and Murat Kurtcephe. "Ranking instances by maximizing the area under ROC curve". In: *IEEE Transactions on Knowledge and Data Engineering* 25.10 (2013), pp. 2356–2366.

[107] Kai Ming Ting and Ian H Witten. "Stacking Bagged and Dagged Models". In: *14th International Conference on Machine Learning*. 1997.

[108] Tin Kam Ho. "The random subspace method for constructing decision forests". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20.8 (1998), pp. 832–844.

[109] Paul Compton and R Jansen. "Knowledge in context: A strategy for expert system maintenance". In: *Australian Joint Conference on Artificial Intelligence*. Springer. 1988, pp. 292–306.

[110] Yoav Freund and Llew Mason. "The alternating decision tree learning algorithm". In: *16th International Conference on Machine Learning (ICML)*. Vol. 99. 1999, pp. 124–133.

[111] *"European Network for Translational Research in Atrial Fibrillation (EU-TRAF)*. 2015. URL: http://www.eutraf.eu/.

[112] C Sánchez, A Bueno-Orovio, E Wettwer, S Loose, J Simon, U Ravens, E Pueyo, and B Rodriguez. "Inter-subject variability in human atrial action potential in sinus rhythm versus chronic atrial fibrillation." In: *PLoS ONE* 9.8 (2013), e105897–e105897.

[113] Francis Morris and William J Brady. "ABC of clinical electrocardiography: Acute myocardial infarction, Part I". In: *BMJ* 324.7341 (2002), pp. 831–834.

[114] Jannatul Robaiat Mou, Sheikh Md Rabiul Islam, Xu Huang, and Keng Liang Ou. "Noise Removal and QRS Detection of ECG Signal". In: *Journal of Biomedical Engineering and Medical Imaging* 3.4 (2016), p. 01.

[115] Suranai Poungponsri and Xiao-Hua Yu. "An adaptive filtering approach for electrocardiogram (ECG) signal noise reduction using neural networks". In: *Neurocomputing* 117 (2013), pp. 206–213.

[116] Udit Satija, Barathram Ramkumar, and M Sabarimalai Manikandan. "Low-complexity detection and classification of ECG noises for automated ECG analysis system". In: *International Conference on Signal Processing and Communication*. IEEE. 2016, pp. 1–5.

[117] *PhysioBank ATM: PTB Diagnostic ECG Database*. 2000. URL: https://www.physionet.org/cgi-bin/atm/ATM?database=ptbdb\&tool=describe_record.

[118] Michael P Somers, William J Brady, Andrew D Perron, and Amal Mattu. "The prominant T wave: Electrocardiographic differential diagnosis". In: *The American Journal of Emergency Medicine* 20.3 (2002), pp. 243–251.

[119] Janez Demšar. "Statistical comparisons of classifiers over multiple data sets". In: *Journal of Machine Learning Research* 7.Jan (2006), pp. 1–30.

[120] Bin Liu, Jikui Liu, Guoqing Wang, Kun Huang, Fan Li, Yang Zheng, Youxi Luo, and Fengfeng Zhou. "A novel electrocardiogram parameterization algorithm and its application in myocardial infarction detection". In: *Computers in Biology and Medicine* 61 (2015), pp. 178–184.

[121] K Ramamohan Rao and Ping Yip. *Discrete cosine transform: algorithms, advantages, applications*. Academic Press, 2014.

[122] Erich Wettwer, Ottó Hála, Torsten Christ, Jürgen F Heubach, Dobromir Dobrev, Michael Knaut, András Varró, and Ursula Ravens. "Role of IKur in controlling action potential shape and contractility in the human atrium influence of chronic atrial fibrillation". In: *Circulation* 110.16 (2004), pp. 2299–2306.

[123] Ursula Ravens, Erich Wettwer, and Ottó Hála. "Pharmacological modulation of ion channels and transporters". In: *Cell Calcium* 35.6 (2004), pp. 575–582.

[124] William Dressler and Roesler Hugo. "High T waves in the earliest stage of myocardial infarction". In: *American Heart Journal* 34.5 (1947), pp. 627–645.

[125] Kurt Kroenke and A David Mangelsdorff. "Common symptoms in ambulatory care: incidence, evaluation, therapy, and outcome". In: *The American journal of medicine* 86.3 (1989), pp. 262–266.

[126] Samuel D Turnipseed, Aaron E Bair, J Douglas Kirk, Deborah B Diercks, Poroshat Tabar, and Ezra A Amsterdam. "Electrocardiogram differentiation of benign early repolarization versus acute myocardial infarction by emergency physicians and cardiologists". In: *Academic emergency medicine* 13.9 (2006), pp. 961–966.

[127] Christopher P Cannon. *Management of acute coronary syndromes*. Springer Science & Business Media, 1998.

[128] Carlos Van Mieghem, Marc Sabbe, and Daniel Knockaert. "The clinical value of the ECG in noncardiac conditions". In: *Chest Journal* 125.4 (2004), pp. 1561–1576.

APPENDIX A

Extended Results for AP Dataset

A.1   Figures for Cardiac Action Potential (AP) Dataset
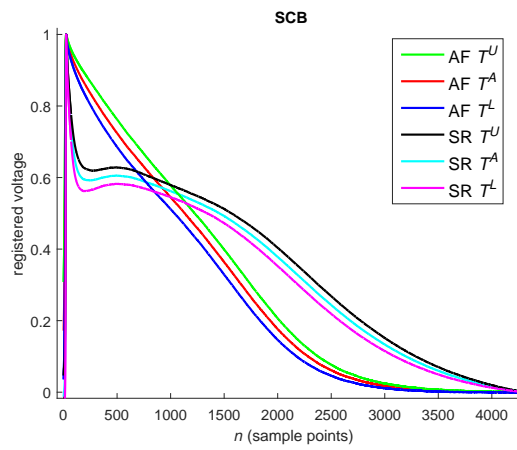
Figure A.1 : Confidence bands with $p = 0.99$ overlaid for AF and SR classes.
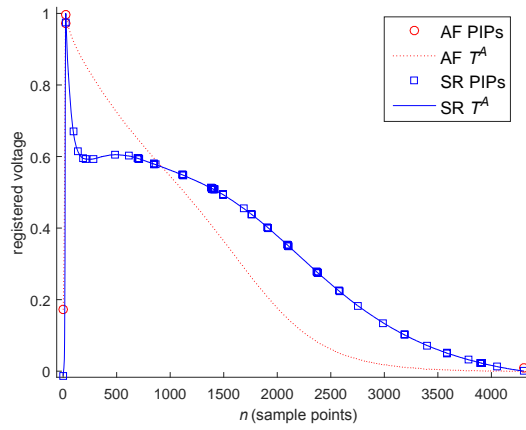
Figure A.2 : Identified PIPs with $\delta = 0.9900$ from the class representative average time series of AF and SR classes.
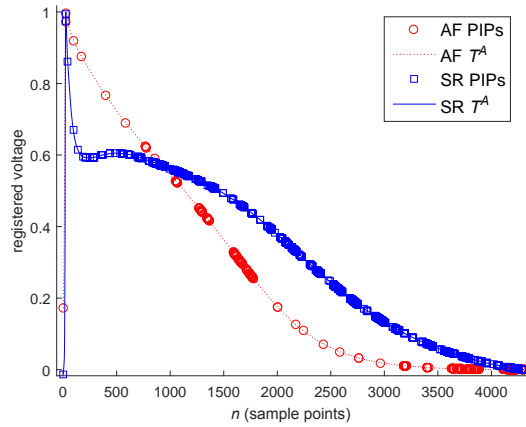
Figure A.3 : Identified PIPs with $\delta = 0.9999$ from the class representative average time series of AF and SR classes.
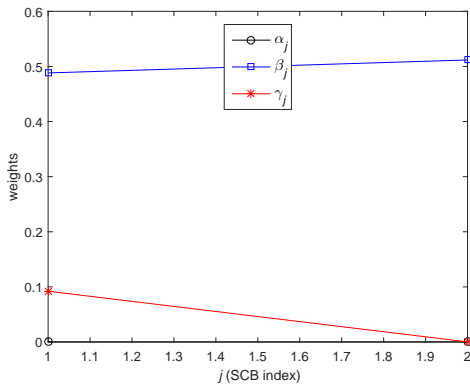
(a) AF
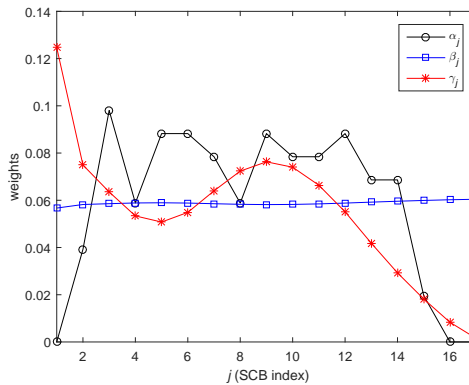
(b) SR

Figure A.4 : The weights α_j, β_j and γ_j calculated with $p = 0.95$ and $\delta = 0.9900$ for for SCBs in (a) AF and (b) SR.

72
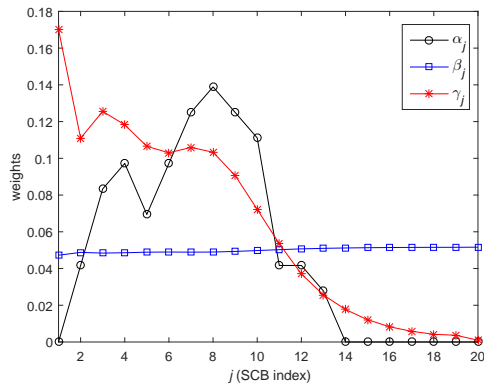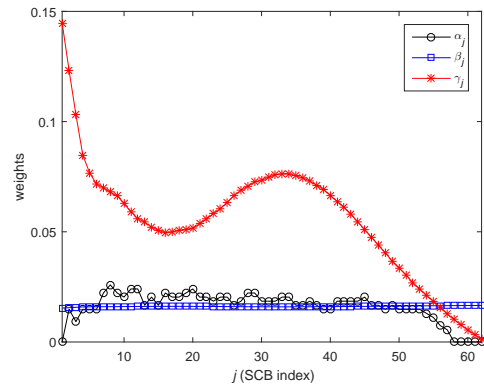
(a) AF   (b) SR

Figure A.5 : The weights α_j, β_j and γ_j calculated with $p = 0.95$ and $\delta = 0.9999$ for for SCBs in (a) AF and (b) SR.
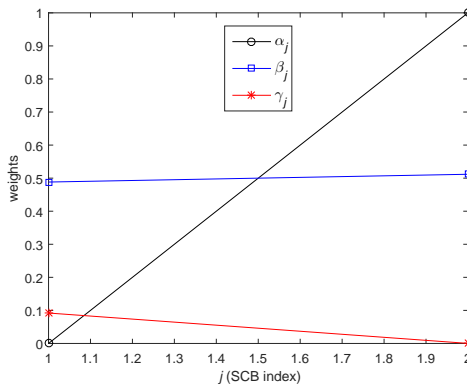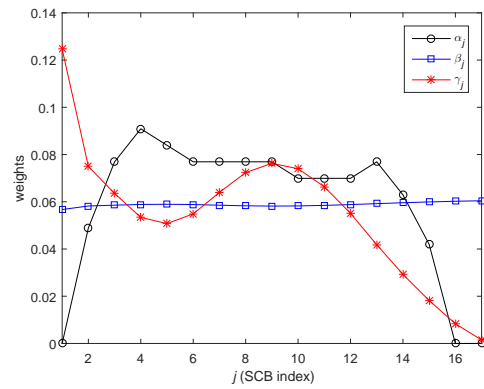
(a) AF   (b) SR

Figure A.6 : The weights α_j, β_j and γ_j calculated with $p = 0.99$ and $\delta = 0.9900$ for for SCBs in (a) AF and (b) SR.
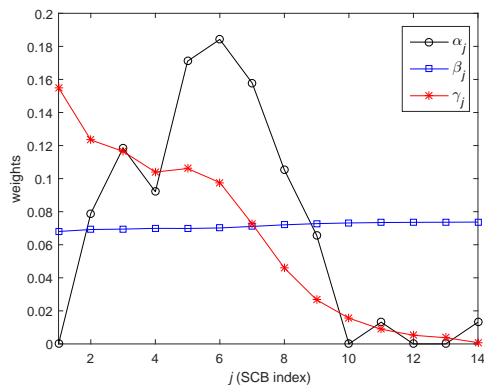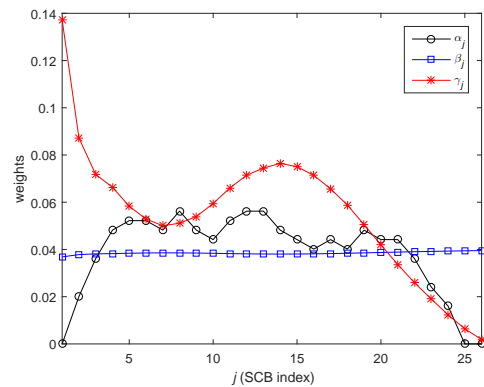
(a) AF   (b) SR

Figure A.7 : The weights α_j, β_j and γ_j calculated with $p = 0.99$ and $\delta = 0.9990$ for for SCBs in (a) AF and (b) SR.
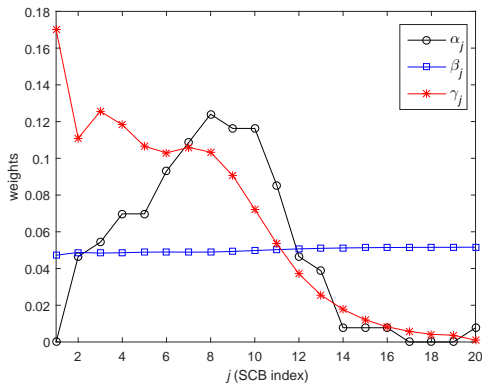
(a) AF
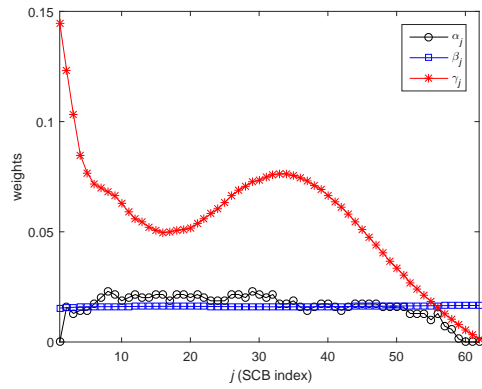
(b) SR

Figure A.8 : The weights α_j, β_j and γ_j calculated with $p = 0.99$ and $\delta = 0.9999$ for for SCBs in (a) AF and (b) SR.
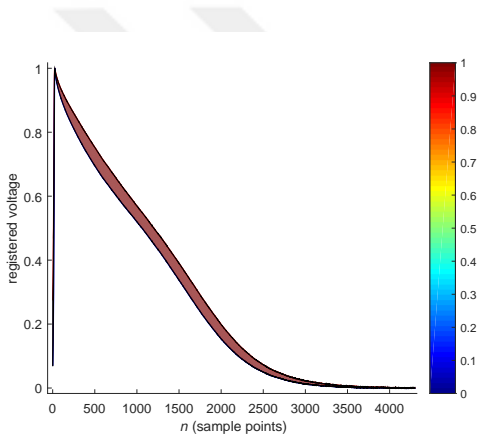
(a) AF
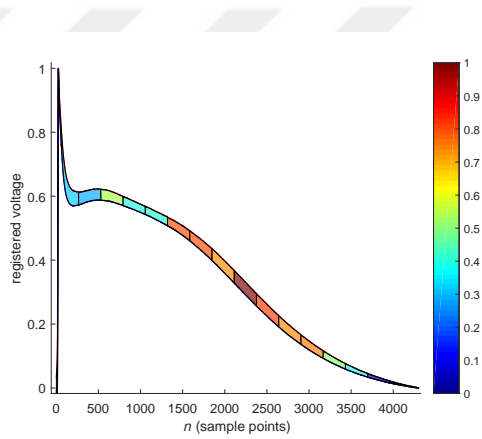
(b) SR

Figure A.9 : Color map plots for the average values of Z_j calculated with $p = 0.95$ and $\delta = 0.9900$ for the test set.
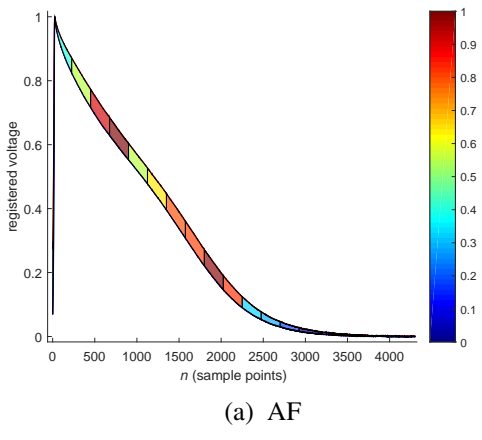
(a) AF
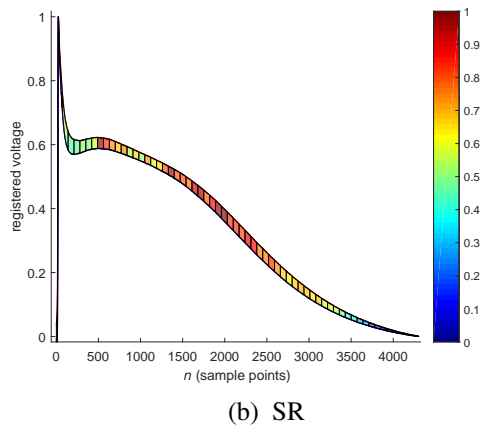
(b) SR

Figure A.10 : Color map plots for the average values of Z_j calculated with $p = 0.95$ and $\delta = 0.9999$ for the test set.
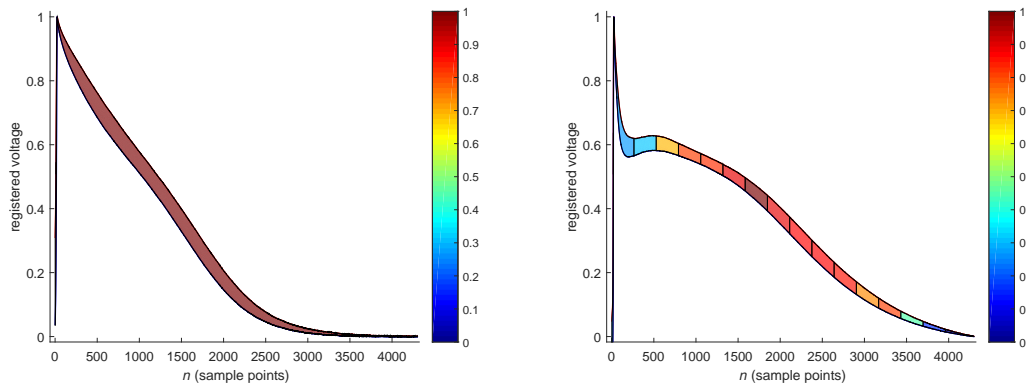
(a) AF

(b) SR

Figure A.11 : Color map plots for the average values of Z_j calculated with $p = 0.99$ and $\delta = 0.9900$ for the test set.
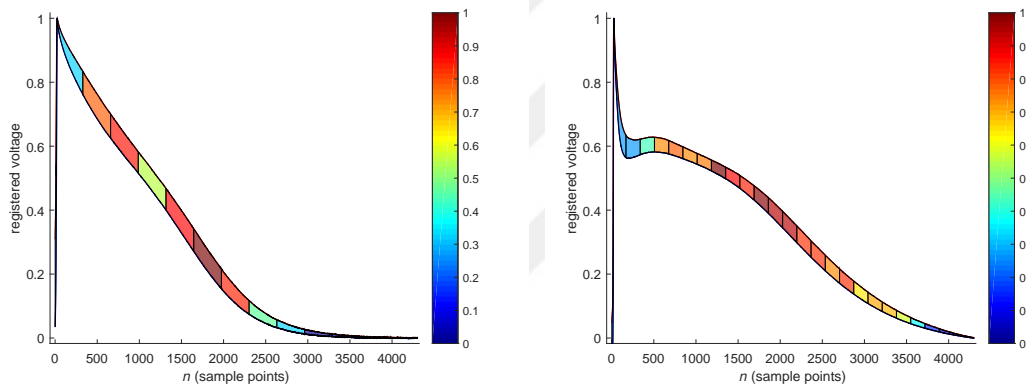
(a) AF

(b) SR

Figure A.12 : Color map plots for the average values of Z_j calculated with $p = 0.99$ and $\delta = 0.9990$ for the test set.
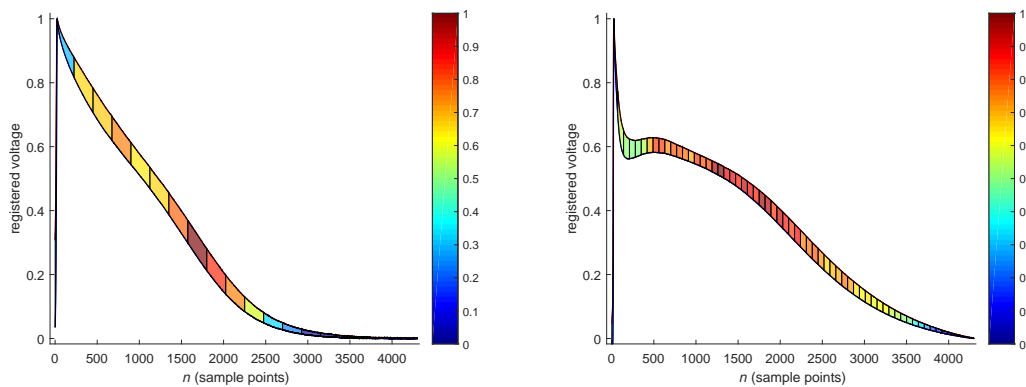
(a) AF

(b) SR

Figure A.13 : Color map plots for the average values of Z_j calculated with $p = 0.99$ and $\delta = 0.9999$ for the test set.

APPENDIX B

Extended Results for ECG Dataset

B.1   PTBDB ECG Dataset Instances

Table B.1 : Record identifications for the time series instances analyzed in Chapter 4 from PTB Diagnostic ECG Database (https://physionet.org/cgi-bin/atm/ATM)

| MI PATIENTS	CONTROL SUBJECTS
patient004-s0020are	patient104-s0306lre
patient005-s0021are	patient116-s0302lre
patient006-s0022lre	patient117-s0291lre
patient007-s0026lre	patient121-s0311lre
patient010-s0036lre	patient122-s0312lre
patient012-s0043lre	patient156-s0299lre
patient013-s0045lre	patient165-s0322lre
patient019-s0077lre	patient166-s0275lre
patient024-s0084lre	patient169-s0328lre
patient025-s0087lre	patient170-s0274lre
patient026-s0088lre	patient172-s0304lre
patient027-s0089lre	patient173-s0305lre
patient029-s0122lre	patient174-s0300lre
patient032-s0102lre	patient180-s0374lre
patient034-s0109lre	patient182-s0308lre
patient036-s0111lre	patient184-s0363lre
patient039-s0129lre	patient198-s0402lre
patient042-s0135lre	patient233-s0457-re
patient046-s0156lre	patient235-s0461-re
patient048-s0171lre	patient236-s0462-re
patient058-s0216lre	patient238-s0466-re
patient061-s0210lre	patient239-s0467-re
patient063-s0214lre	patient240-s0468-re
patient075-s0242lre	patient241-s0469-re
patient076-s0250lre	patient242-s0471-re
patient082-s0267lre	patient243-s0472-re
patient091-s0357lre	patient245-s0474-re
patient094-s0370lre	patient246-s0478-re
patient096-s0379lre	patient247-s0479-re
patient098-s0398lre	patient248-s0481-re
patient099-s0387lre	patient251-s0486-re
patient101-s0400lre	patient252-s0487-re
patient120-s0331lre	patient255-s0491-re
patient138-s0005-re	patient260-s0496-re
patient139-s0223-re	patient263-s0499-re
patient211-s0433-re	patient266-s0502-re
patient223-s0446-re	patient267-s0504-re
patient265-s0501-re	patient276-s0526-re
patient268-s0505-re	patient277-s0527-re
patient270-s0507-re	patient284-s0551-re

B.2 Figures for ECG Dataset

(a) lead v2
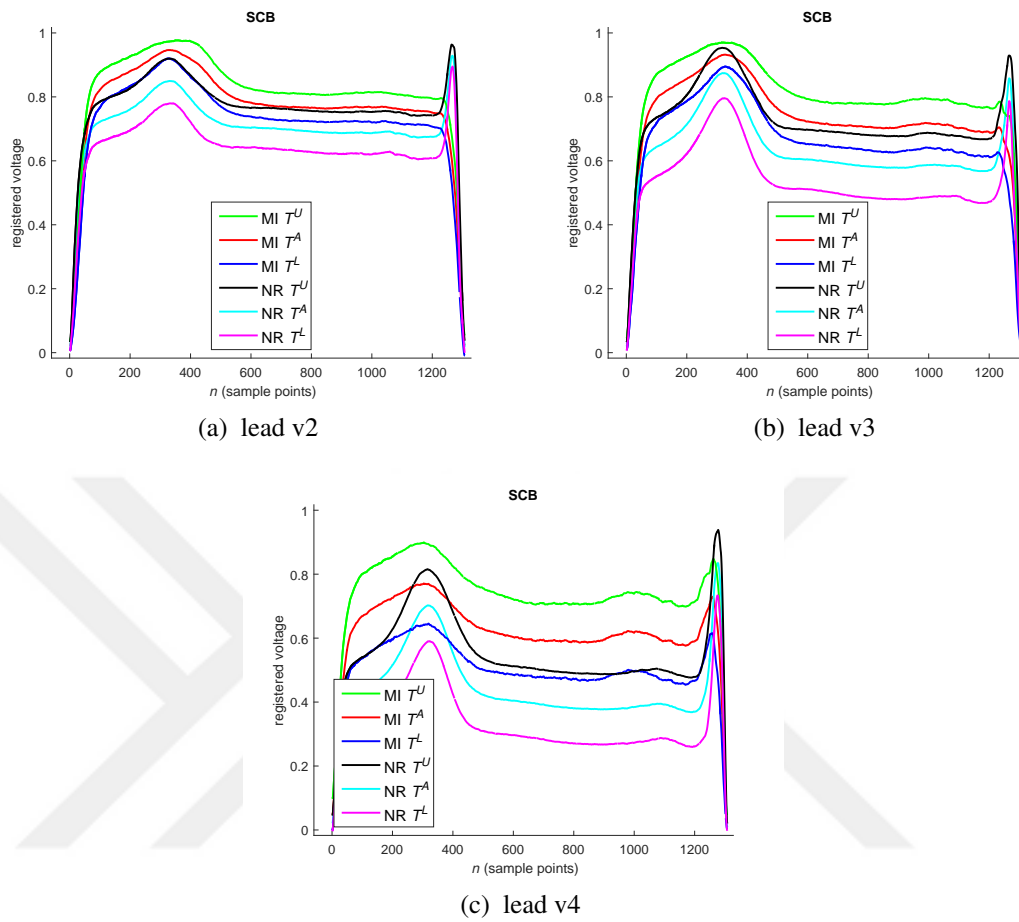
(b) lead v3

(c) lead v4

Figure B.1 : Confidence bands with $p = 0.99$ overlaid for the MI and NR classes in (a)v2 (b)v3 and (c)v4 leads of ECG dataset.
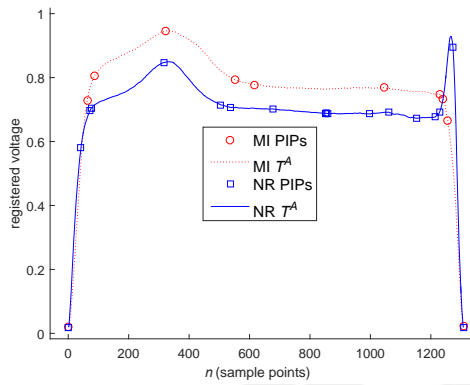
(a) lead v2
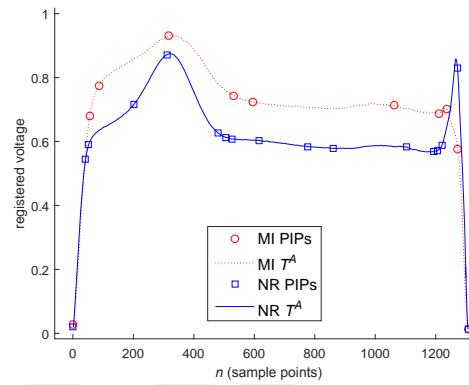
(b) lead v3
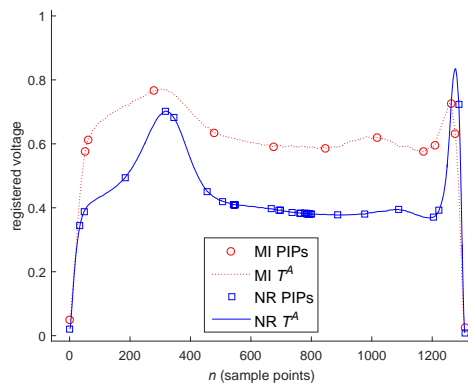
(c) lead v4

Figure B.2 : Identified PIPs from the MI and NR class representative average time series with δ = 0.9900 from (a)v2 (b)v3 and (c)v4 leads of ECG dataset.

(a) lead v2
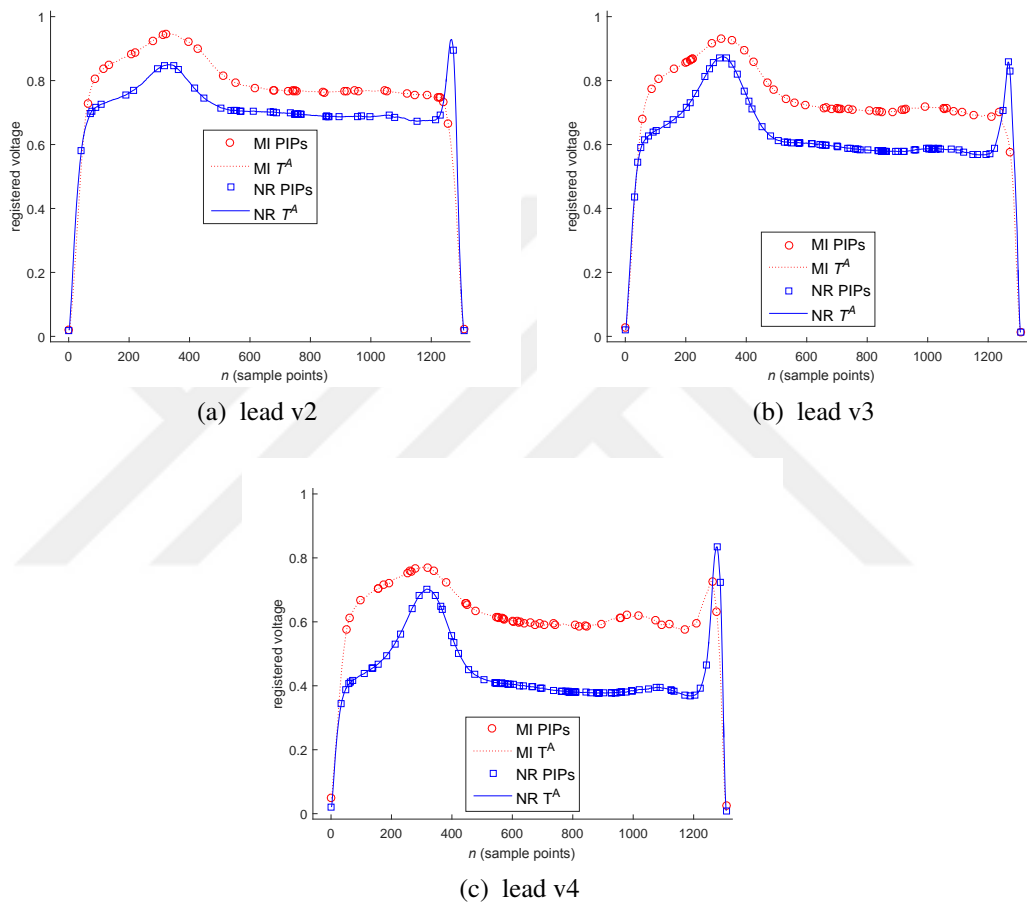
(b) lead v3

(c) lead v4

Figure B.3 : Identified PIPs from the MI and NR class representative average time series with δ = 0.9999 from (a)v2 (b)v3 and (c)v4 leads of ECG dataset.
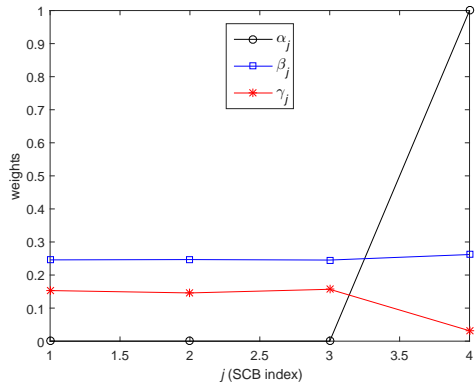
(a) MI(lead v2)
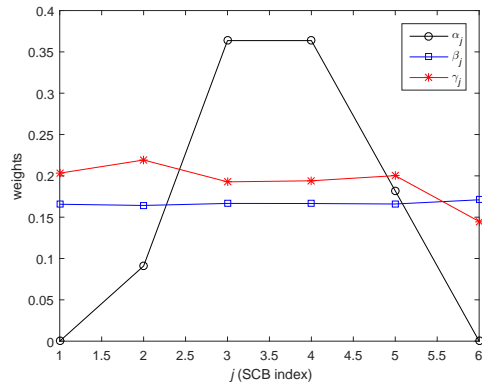
(b) NR(lead v2)
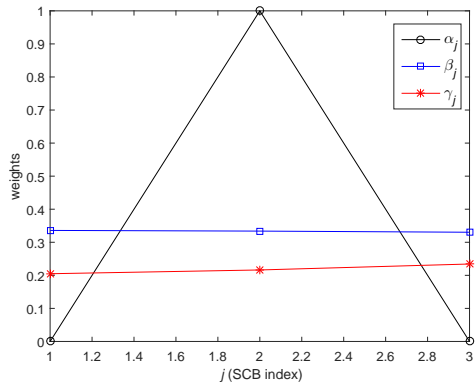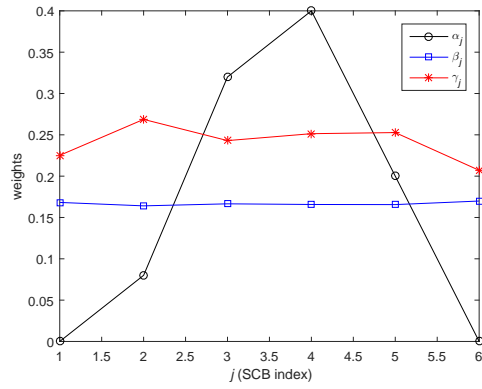
(c) MI(lead v3)
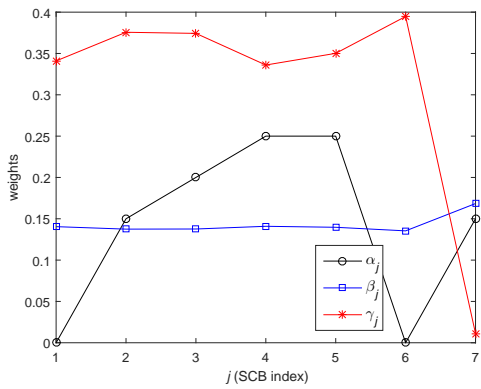
(d) NR(lead v3)
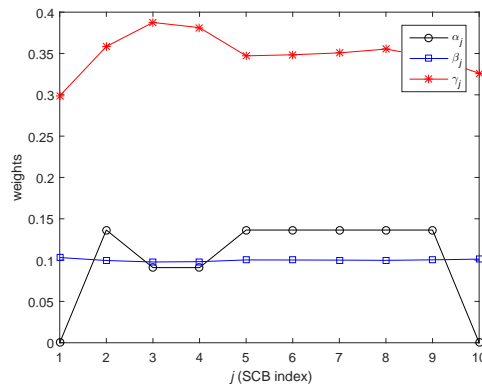
(e) MI(lead v4)

(f) NR(lead v4)

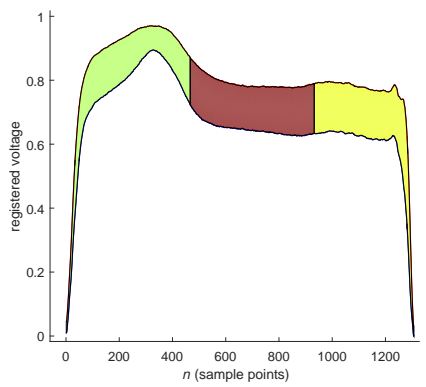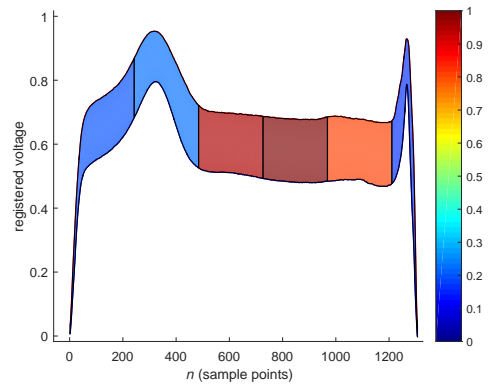Figure B.4 : The weights α_j, β_j and γ_j calculated with $p = 0.95$ and $\delta = 0.9900$ for SCBs in leads v2, v3 and v4 for (a,c,e) MI and (b,d,f) NR.
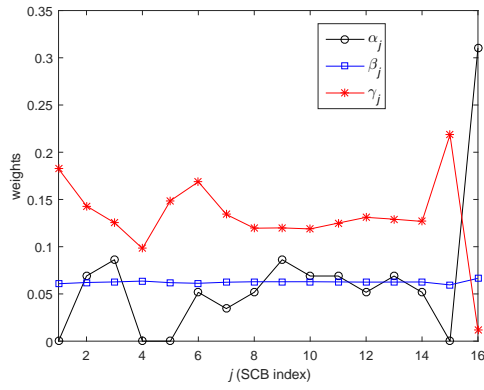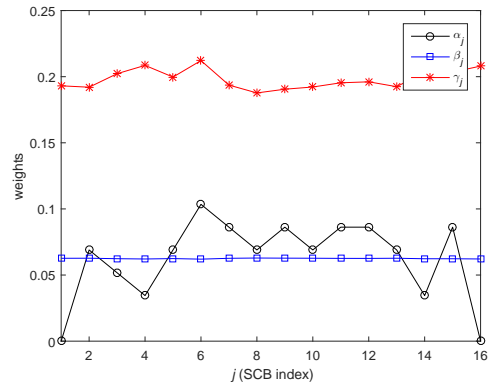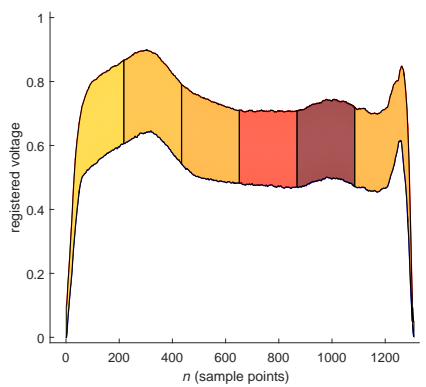
(a) MI(lead v2)
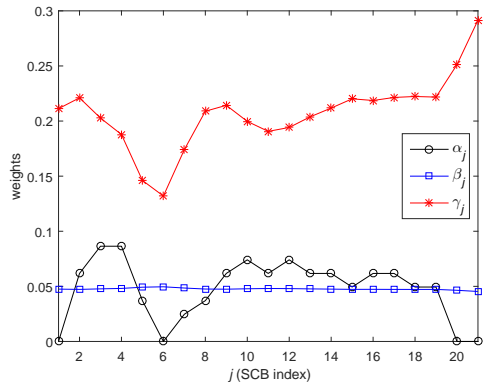
(b) NR(lead v2)
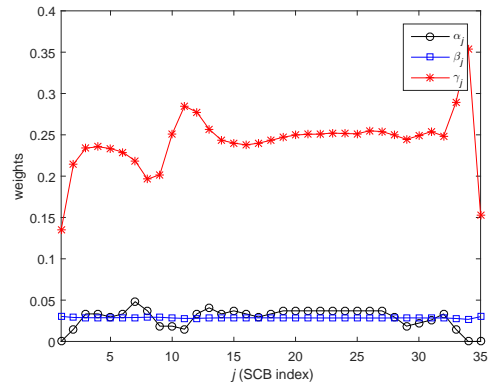
(c) MI(lead v3)
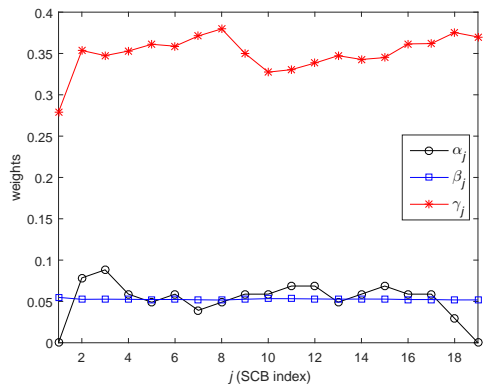
(d) NR(lead v3)
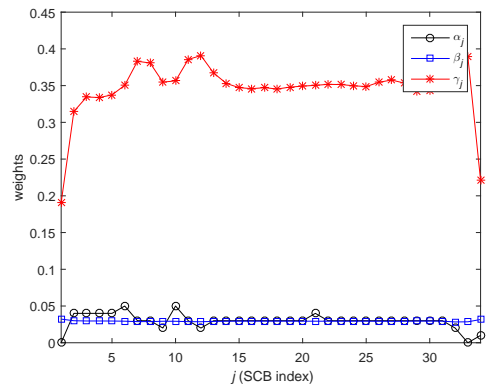
(e) MI(lead v4)

(f) NR(lead v4)

Figure B.5 : The weights α_j, β_j and γ_j calculated with $p = 0.95$ and $\delta = 0.9999$ for SCBs in leads v2, v3 and v4 for (a,c,e) MI and (b,d,f) NR.
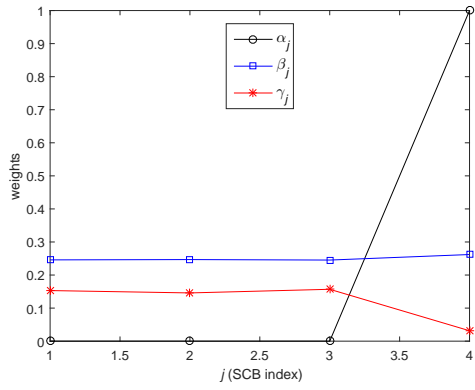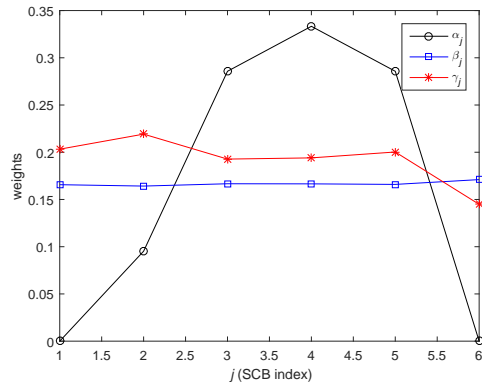
(a) MI(lead v2)
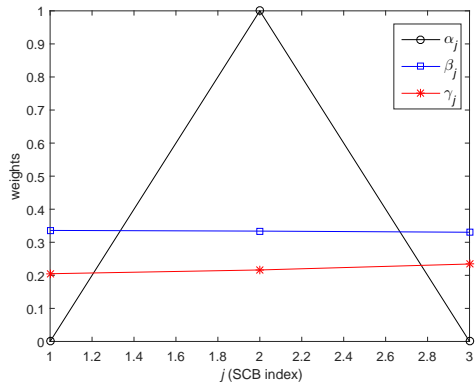
(b) NR(lead v2)
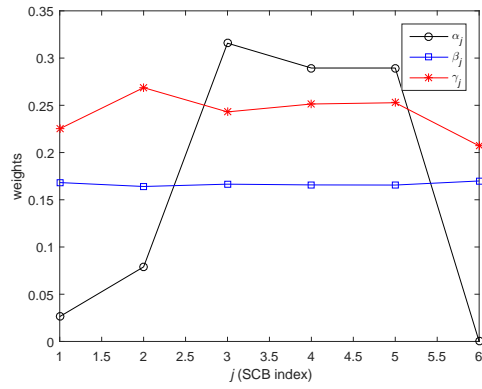
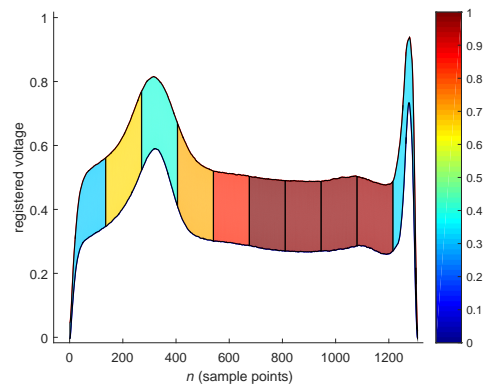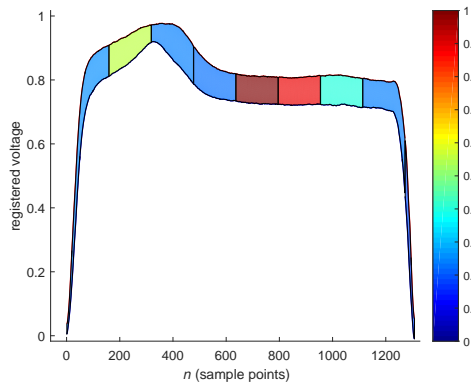(c) MI(lead v3)
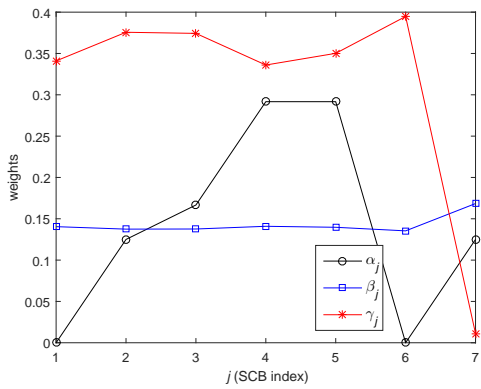
(d) NR(lead v3)
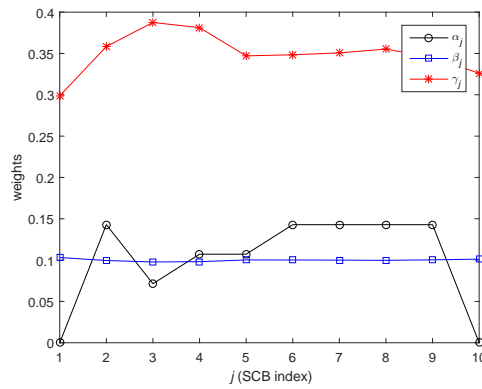
(e) MI(lead v4)

(f) NR(lead v4)

Figure B.6 : The weights α_j, β_j and γ_j calculated with $p = 0.99$ and $\delta = 0.9900$ for SCBs in leads v2, v3 and v4 for (a,c,e) MI and (b,d,f) NR.
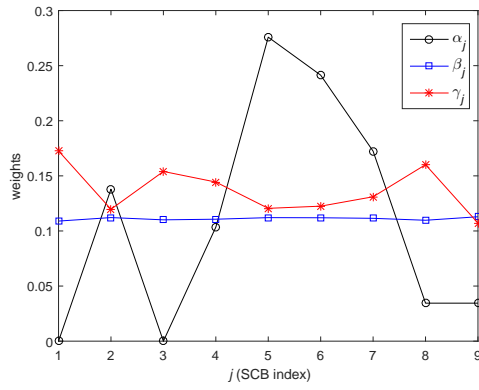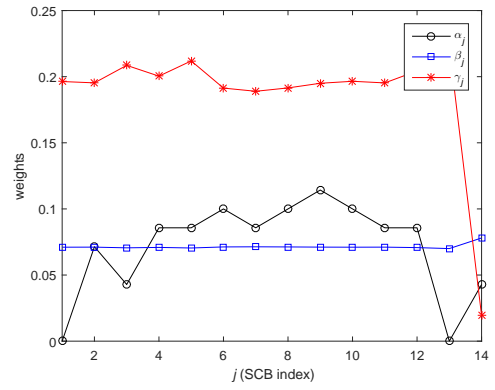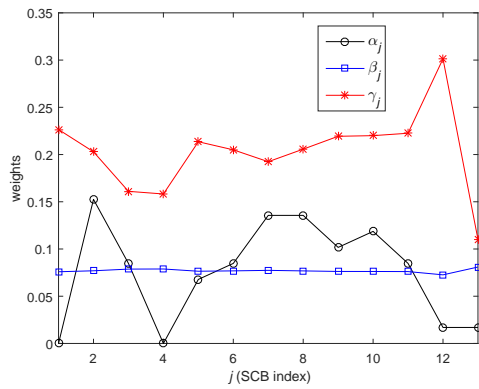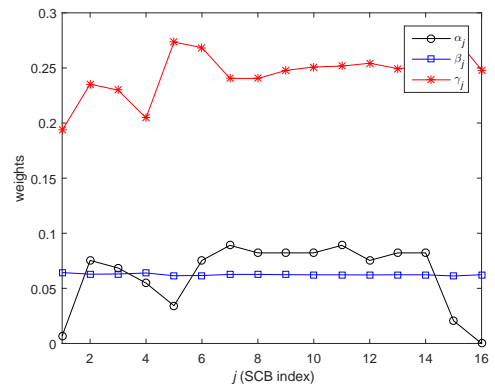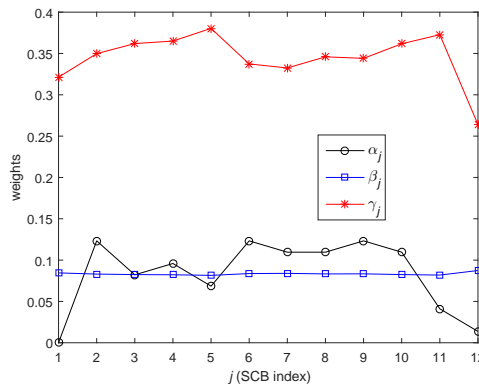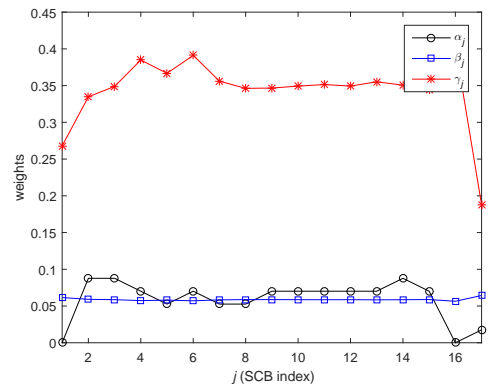
84

Figure B.7 : The weights α_j, β_j and γ_j calculated with $p = 0.99$ and $\delta = 0.9990$ for SCBs in leads v2, v3 and v4 for (a,c,e) MI and (b,d,f) NR.
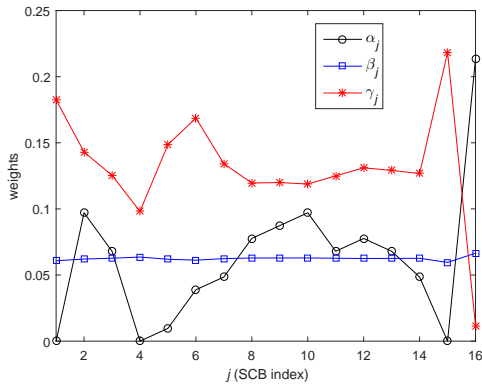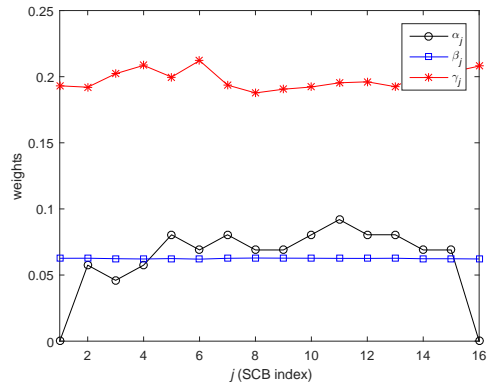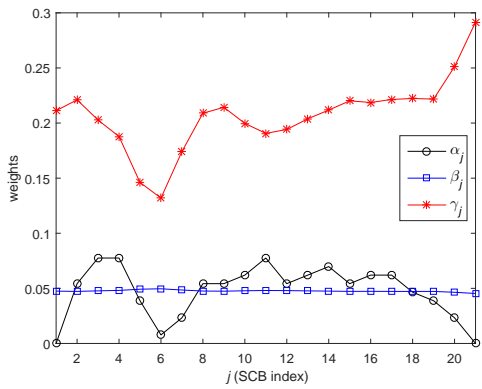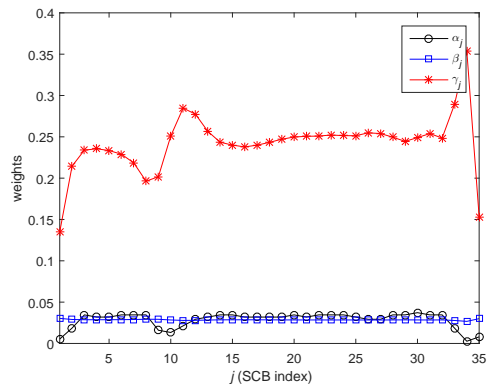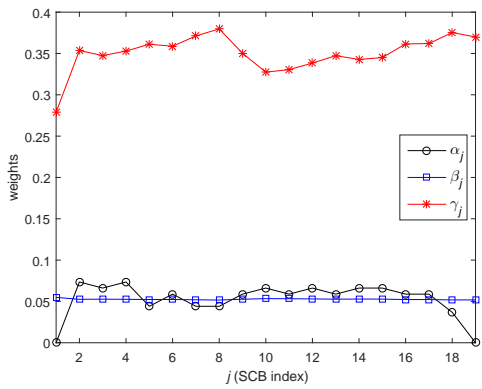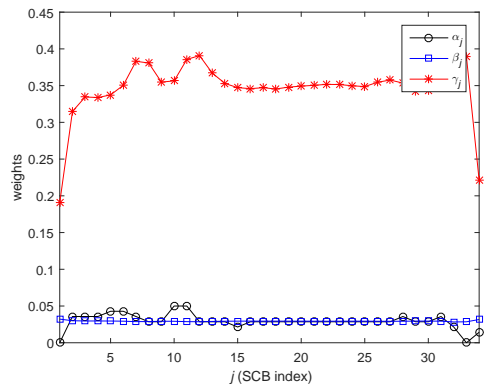
Figure B.8 : The weights α_j, β_j and γ_j calculated with $p = 0.99$ and $\delta = 0.9999$ for SCBs in leads v2, v3 and v4 for (a,c,e) MI and (b,d,f) NR.
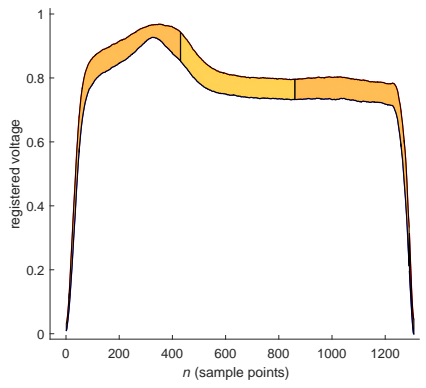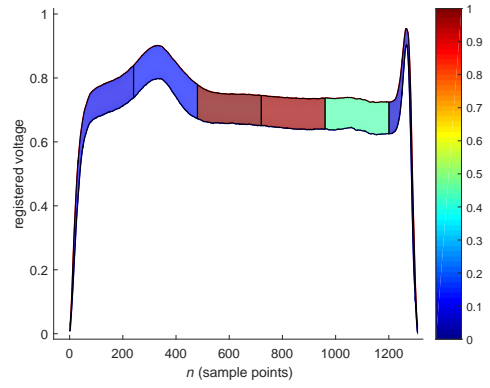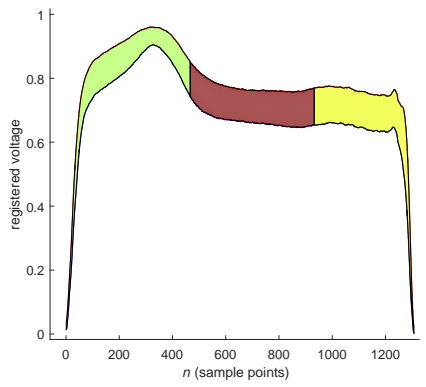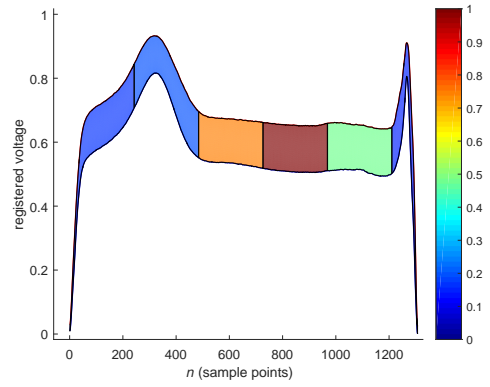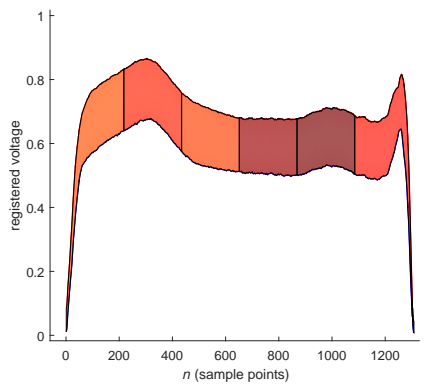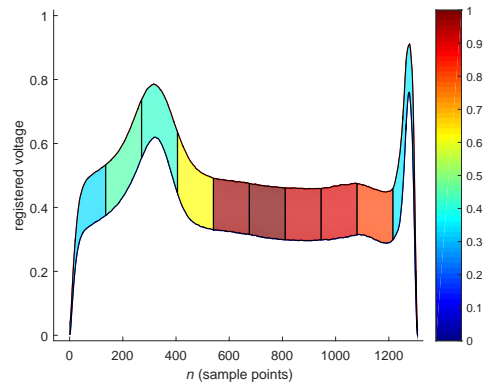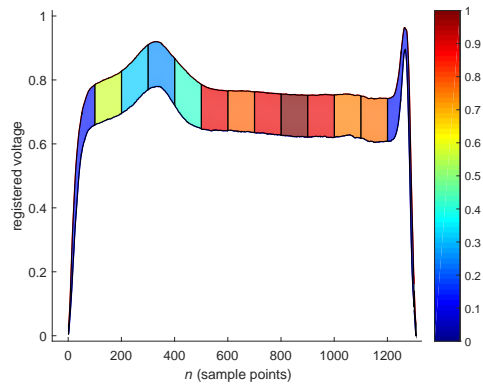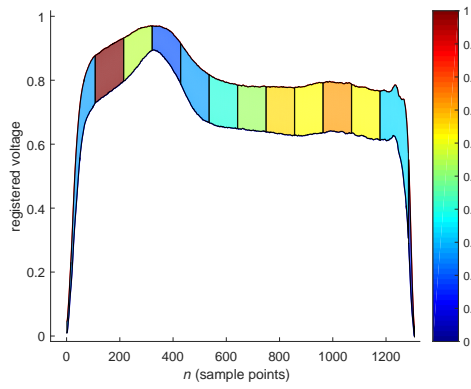
Figure B.9 : Color map plots for the average values of Z_j calculated with $p = 0.95$ and $\delta = 0.9900$ for the leads v2, v3 and v4 for (a,c,e) MI and (b,d,f) NR.
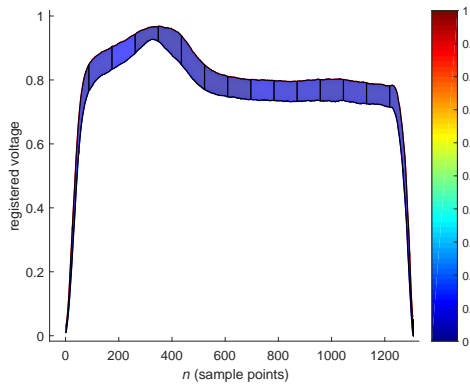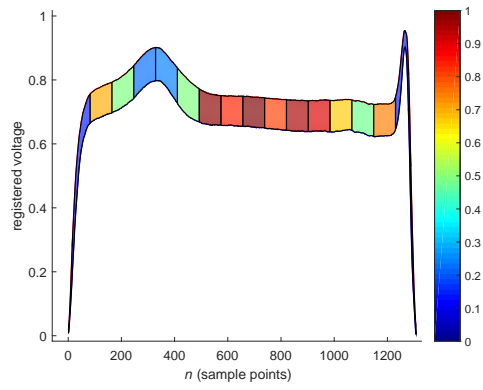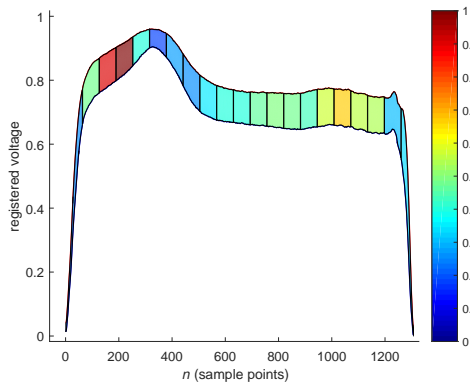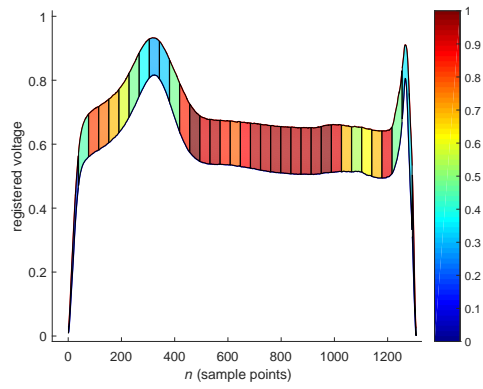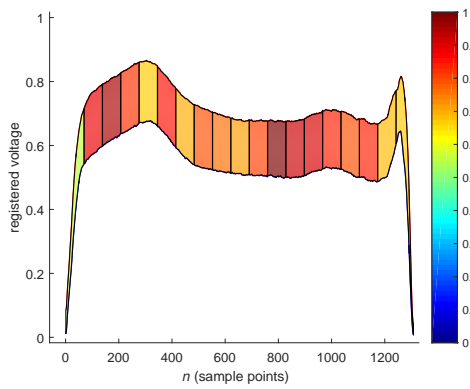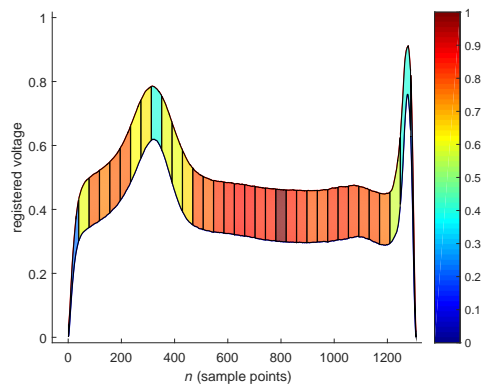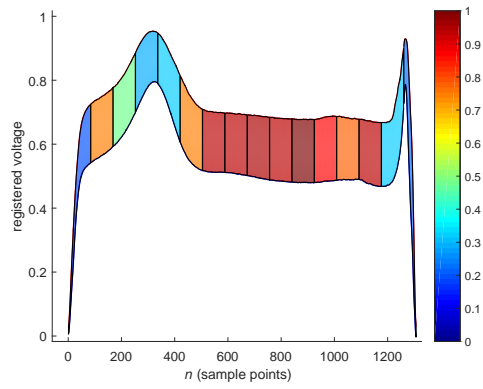
Figure B.10 : Color map plots for the average values of Z_j calculated with $p = 0.95$ and $\delta = 0.9999$ for the leads v2, v3 and v4 for (a,c,e) MI and (b,d,f) NR.
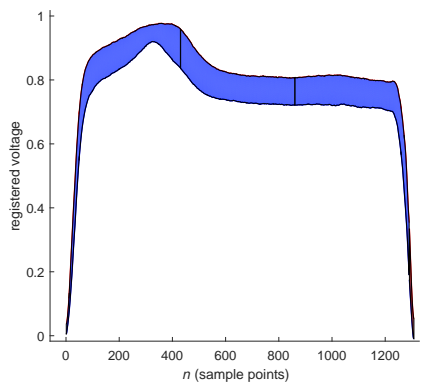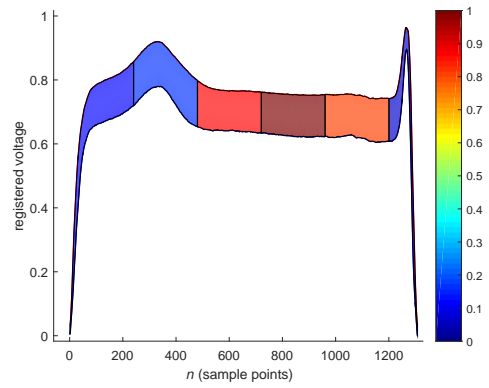
Figure B.11 : Color map plots for the average values of Z_j calculated with $p = 0.99$ and $\delta = 0.9900$ for the leads v2, v3 and v4 for (a,c,e) MI and (b,d,f) NR.
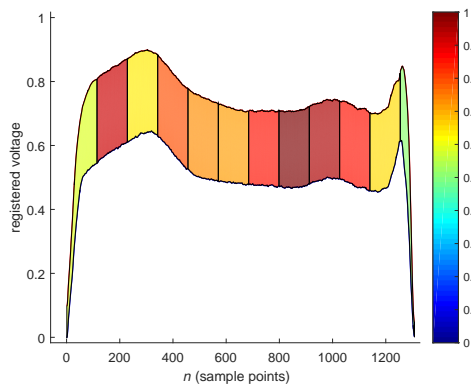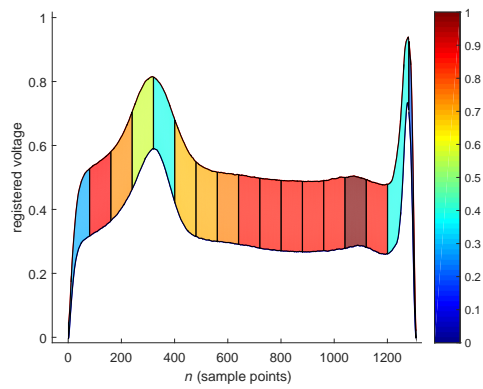
(a) MI(lead v2)

(b) NR(lead v2)
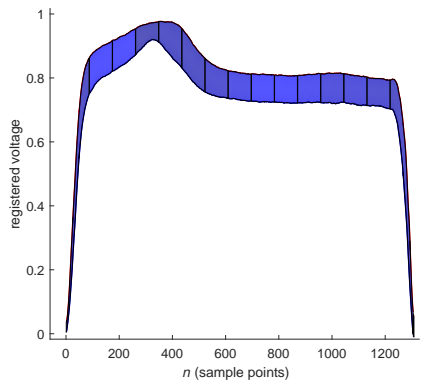
(c) MI(lead v3)
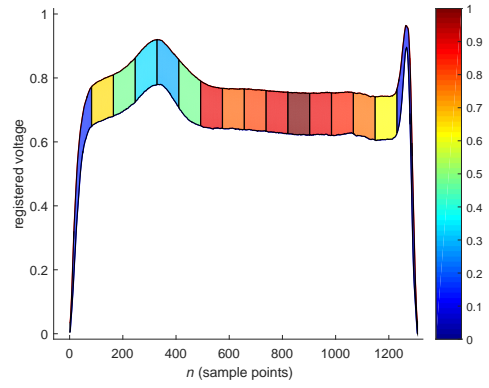
(d) NR(lead v3)

(e) MI(lead v4)

(f) NR(lead v4)

Figure B.12 : Color map plots for the average values of Z_j calculated with $p = 0.99$ and $\delta = 0.9990$ for the leads v2, v3 and v4 for (a,c,e) MI and (b,d,f) NR.
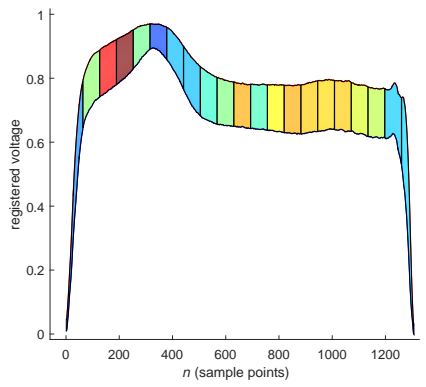
90

Figure B.13 : Color map plots for the average values of Z_j calculated with $p = 0.99$ and $\delta = 0.9999$ for the leads v2, v3 and v4 for (a,c,e) MI and (b,d,f) NR.
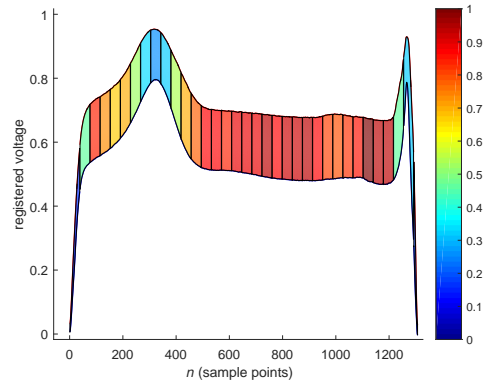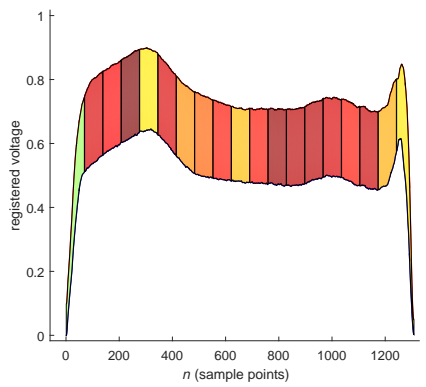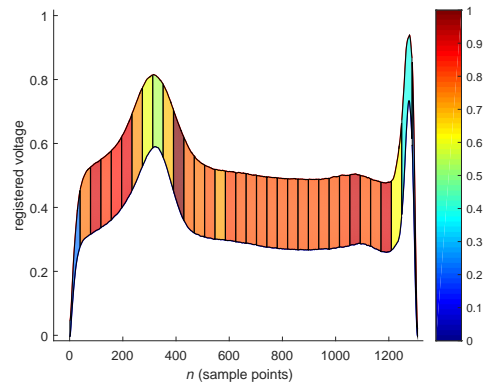
CURRICULUM VITAE

PERSONAL INFORMATION

Surname, Name: Katırcıoğlu Öztürk, Deniz

Nationality: Republic of Turkey (TC)

E-mail Address: denizkatircioglu@gmail.com

EDUCATION

Degree	Institution	Year of Graduation
M.Sc.	METU Graduate School of Informatics, Medical Informatics	2009
B.Sc.	METU Faculty of Engineering, Computer Engineering	2005
High School	METU Development Foundation Schools - Ankara	2001

PROFESSIONAL EXPERIENCE

Enrollment	Company	Year
Software R&D Director	Medical Information Technology Solutions (MITS)	2009 - Present
Sr Software Development Engineer	Entegre Enformasyon Sistemleri Ltd (EES)	2008 - 2009
Software Development Engineer	TCHEALTH Bilgi Teknolojileri AS	2005 - 2008

PUBLICATIONS

International Journal Publications

1. B. Yavuz, N. Ata, E. Oto, **D. Katircioglu-Öztürk,** K. Aytemir, B. Evranos, R. Koselerli, E. Ertugay, A. Burkan, E. Ertugay, C.P. Gale, A.J. Camm, A. Oto "**Demographics, treatment and outcomes of atrial fibrillation in a developing country: the population-based TuRkish Atrial Fibrillation (TRAF) cohort**, *EP-Europace*, Feb. 2017.

2. S. Okutucu, **D. Katircioglu-Öztürk**, E. Oto, H.A. Güvenir, E. Karaagaoglu, A. Oto, T. Meinertz, A. Goette "**Data mining experiments on the Angiotensin II-Antagonist in Paroxysmal Atrial Fibrillation (ANTIPAF-AFNET 2) trial: exposing the invisible**, *EP-Europace*, Oct. 2016.

3. E. Oto, S. Okutucu, **D. Katircioglu-Öztürk**, H.A. Güvenir, E. Karaagaoglu, M. Borggrefe, G. Breithardt, A. Goette, U. Ravens, G. Steinbeck, K. Wegscheider, A. Oto, P. Kirchhof "**Predictors of sinus rhythm after electrical cardioversion of atrial fibrillation: results from a data mining project on the Flec-SL trial data set**, *EP-Europace*, Jul. 2016.

4. U. Ravens*, **D. Katircioglu-Öztürk***, E. Wettwer, T. Christ, D. Dobrev, N. Voigt, C. Poulet, S. Loose, J. Simon, A. Stein, K. Matschke, M. Knaut, E. Oto, A. Oto*, H.A. Güvenir* "**Application of the RIMARC algorithm to a large data set of action potentials and clinical parameters for risk prediction of atrial fibrillation**, *Medical & Biological Engineering & Computing*, Mar. 2015 (*: equal contribution).

International Conference Publications

1. U. Canpolat, K. Aytemir, E. Oto, **D. Katircioglu**, B. Evranos, N. Ata, R. Koselerli, Y.G. Ilhan, A.J. Camm, A. Oto "**Mortality in valvular and non-valvular atrial fibrillation: insights from a nationwide cohort analysis-TRAF (Turkish AF database)**, *European Heart Journal*, Sep. 2014.

Abstracts

1. N. Ata, E. Oto, B. Evranos, S. Okutucu, B. Yavuz, **D.K. Öztürk**, K. Aytemir, A. Birkan, H. Kiliçarslan, Y.G. Ilhan, J. Camm, A. Oto "**OP-244 Warfarin Use in Patients with Nonvalvular Atrial Fibrillation: Analysis From AA Nationwide Cohort-TRAF (Turkish Atrial Fibrillation Data Base)**, *The American Journal of Cardiology*, Apr. 2014.

2. B. Evranos, E. Oto, **D.K. Öztürk**, B. Yavuz, S. Okutucu, K. Aytemir, E. Ertugay, M. Kuruca, A. Bürkan, Y.G. Ilhan, J. Camm, A. Oto "**OP-278 Time to Heart Failure in Patients with Atrial Fibrillation: Analysis from a A Nationwide Cohort-TRAF (TuRkish Atrial Fibrillation Data Base)**, *The American Journal of Cardiology*, Apr. 2014.

3. B. Yavuz, N. Ata, E. Oto, K. Aytemir, S. Okutucu, B. Evranos, **D.K. Ozturk**, R. Koselerli, E. Ertugay, Y.G. Ilhan, A.J. Camm, A. Oto "**OP-243 Results of Anticoagulation in Patients with CHADS2 Score of 0 and CHA2DS2VASC Score of 1; An Investigation Using the National Turkish Atrial Fibriallation (TRAF) Registry**, *The American Journal of Cardiology*, Apr. 2014.

4. E. Oto, B. Evranos, N. Ata, B. Yavuz, S. Okutucu, K. Aytemir, **D.K. Öztürk**, E. Ertugay, H. Kiliçarslan, Y.G. Ilhan, J. Camm, A. Oto "**OP-242 Time to Ischemic Stroke/Embolism in Patients with Atrial Fibrillation: Analysis from a A Nationwide Cohort-TRAF (TuRkish Atrial Fibrillation Data Base)**, *The American Journal of Cardiology*, Apr. 2014.

5. A. Oto, N. Ata, B. Yavuz, E. Oto, K. Aytemir, **D.K. Ozturk**, B. Evranos, A. Burkan, E. Ertugay, H. Kiliçarslan, M. Kuruca, Y.G. Ilhan "**OP-237 Real World patients with Atrial Fibrillation: A Nationwide Cohort Study-TRAF (Turkish Atrial Fibrillation Data Base)**, *The American Journal of Cardiology*, Apr. 2014.

6. N. Ata, E. Oto, B. Yavuz, B. Evranos, **D. Katircioglu Öztürk**, U. Canpolat, K. Aytemir, E. Ertugay, M. Kuruca, Y.G. Ilhan, J. Camm, A. Oto "**OP-238 TRAF-Characteristicsof Non-valvular Atrial Fibrillation in patients with Cronic Renal Failure**, *The American Journal of Cardiology*, Apr. 2014.

7. B. Evranos, K. Aytemir, B. Yavuz, E. Oto, N. Ata, **D. Katircioglu Öztürk**, E. Karaagaoglu, A. Güvenir, E. Ertugay, R. Koselerli, A. Bürkan, M. Kuruca, Y.G. Ilhan, J.A. Camm, A. Oto "**Hospitalization for Atrial Fibrillation Increases in the Elderly: Recent Analysis From TuRkish Atrial Fibrillation Data Base**, *Circulation*, Nov. 2013.

8. N. Ata, B. Yavuz, E. Oto, **D. Katircioglu Öztürk**, K. Aytemir, E. Karaismailoglu, A. GÜvenir, E. Ertugay, R. Koselerli, A. BÜrkan, M. Kuruca, Y.G. Ilhan, J. Camm, A. Oto "**Characteristics of Atrial Fibrillation in Hemodialysis Patients**, *Circulation*, Nov. 2013.