修 士 論 文 の 和 文 要 旨

研究科・専攻	大学院 情報理工学研究科 情報・ネットワーク工学専攻		
氏　　名	FRIAS PIMENTEL GILSON ALEXANDER フリーアス　ピメンテル　ヒルソン　アレキサンダー	学籍番号	1731173
論 文 題 目	反辞書確率モデルを用いた多クラス不整脈の分類 (Multi-type arrhythmia classification with Antidictionary probability models)		

要　　旨

現代社会では全世界的なレベルで高齢化が急速に進んでおり，それに伴い，死因が悪性新生物に次いで２番目に高い心疾患についても，心臓病理学上のさまざまな不規則性や異変の早期発見の必要性やその手段の開発がより一層求められてきている．本研究では，近年データ圧縮分野で提唱された反辞書符号化法を利用した新しい不整脈の検知ならびにその分類手法を提案し，その手法を組み込んだ無線ウェアラブル端末による常時心電図監視システムについて考察する．

代表的な不整脈パターンである心室性期外収縮（PVC）と心房性期外収縮（PAC）において，出現しない特徴的なパターンの統計的な性質を調べることによって，それらの分類を行なった．ここで，出現しない特徴的なパターンとは，それ自身を少しでも短くすると，もとのデータには出現してしまうという性質をもっており，極小禁止語とも呼ばれる．極小禁止語の抽出には，MIT-BIH より公開提供されている心電図データベースを利用している．このデータベースは専門医による診断結果が付されており，どの心電波形が不整脈であるかを正確に知ることができる．ところで，正常な心電図データから抽出された極小禁止語には，PVC や PAC を含む不整脈パターンが含まれていると考えられる．そこで，本研究では，抽出した極小禁止語から正常な心電図データの変動を表す有限状態遷移確率モデルを作成した．この確率モデルは，観測データに追従しながら内部状態が推移して行くだけでなく，状態が遷移する確率も合わせて計算するので，不整脈パターンの出現を，非常に小さな確率をとる状態遷移の生起として捉えることができる．

状態遷移確率モデルを作成するにあたっては，まず１サンプル 11 ビットで表現される心電図データを量子化するために，データの差分化による頻度分布のシェイピングや量子化レベル数について実験検討を行い，モデルの簡略化，実装にあたってのメモリ量の削減を図った．次に，提案法の性能評価のために，MIT-BIH で提供されている不整脈を含む複数の被験者の心電図データについて，心室期外収縮に対する検知率の実験を行なったところ，平均して，既存の不整脈検出法とほぼ同程度の，感度（sensitivity）97.53%，特異度（specificity）93.89% の結果が得られた．さらに，IPhone 端末のエミュレータに提案方式プログラムを組み込んだシミュレーション実験を行ない，必要とされる計算リソースについて調査した．その結果は，CPU 占有率が 65%，メモリ使用量 30.5M バイトであり，提案方式はウェアラブル端末へ十分実装可能であることが示された．加えて，先に示した PVC の二項分類問題に加えて，PVC に PAC を加えた多値分類問題を考え，それぞれのパターンを特徴付ける極小禁止語について考察を行なった．

令和元年　修士論文

Multi-type arrhythmia classification with Antidictionary probability models

電気通信大学大学院情報理工学研究科
情報・ネットワーク工学専攻
電子情報学プログラム

学籍番号: 1731173
Gilson Frías Pimentel

指導教員
森田啓義教授
笠井裕之教授

令和元年 7 月 29 日

Contents

Chapter 1

Introduction

1.1 Research Background

Wearable devices are increasingly been adopted for the ambulatory monitoring of human vital signs [6]. Currently and widely available wearable technologies platforms provide the capabilities of collecting and storing health and fitness related data, empowering patients with the capability of doing self assessments about their health condition [7].

With the proliferation of cardiovascular diseases within the global aging population, more preventive measures must be enforced in order to reduce the number of fatalities due to heart disease. One key application within this domain is the monitoring of **Electrocardiogram** (**ECG**) signals for the early detection and treatment of a broad range of cardiac diseases [6].

Wearables devices destined for the task of ECG monitoring offer an unique opportunity by incorporating anomaly detection capabilities on the monitored signals. This, however, can be a challenge due to the fact that they are usually resources limited devices thus difficulting the implementation of very complex detection techniques.

The algorithm for **D**ata **C**ompression with **A**ntidictionaries (**DCA**) was first introduced by *Crochemore et al* [5]. It makes use of the set of patterns that never appear on a source data set to effectively predict redundant symbols.

The feasibility of the DCA for the compression of ECG signals has been previously studied, and it was also shown that the DCA method can be used for the detection of irregular heart beat patterns [16]. It has been also demonstrated that the algorithm can be implemented effectively in relative small memory footprints, making it an attractive alternative for wearable and mobile ECG monitoring applications.

The algorithm constructs *finite-state* probabilistic models with the forbidden patterns obtained from the antidictionaries. The presence of expected patterns within the signals causes the algorithm to output a low and steady **C**ompression **R**atio (**CR**), while the appearance of forbidden patterns such as those that occur on arrhythmias causes an abrupt increase on the CR, thus enabling the DCA algorithm to be suitable for detection tasks.

It has been also shown that by translating the domain of the ECG distribution into a finite set of small integers it is possible to implement the detection algorithm with the use of less memory resources while maintaining an acceptable performance. This was done by the implementation of differentiation and a quantization stages on the signal processing chain that in effect redefines the signal in a restricted alphabet set [8].

1.2 Research Purpose

In this study, further exploration on the potential of the use of the antidictionary coding algorithm for irregular heartbeats detection on quantized

ECG signals is undertaken.

The main objectives of the research project are listed next:

- The consolidation of a process for the redefinition of the ECG signal distribution and the evaluation of its effects on the associated probabilistic models used by the algorithm.
- Analysis of the impact in computational resources that Finite State Machine probabilistic models constructed with the antidictionary coding scheme can have in resources constrained devices such as mobile devices.
- Examination of the possibility of extending the capabilities of the detection algorithm from a binary classification problem into a multi-class classification problem.

1.3 Document organization

Contents are presented in the following order:

- Chapter 2 presents a brief description about the the anatomical structure of the the heart, the characteristics of the ECG signal and arrhythmia.
- Chapter 3 introduces the core concepts related with the antidictionary coding algorithm and the constuction of Finite State Probability Models.
- Chapter 4 describes the processes of *differentiation* and *quantization* for redefining the ECG distribution.

- Chapter 5 describes the set of experimental procedures carried for offline processing of the ECG signals, detailing the process involved in the construction of the probabilistic models and the corresponding results for the detection of Premature Ventricular Contractions.

- Chapter 6 describes the set of experimental procedures and corresponding results obtained after porting pre-trained Finite State Chapter probability models into a mobile platform.

- Chapter 7 explores the statistics on Minimal Forbidden Word patters distribution for quantized ECG signals, characterizing them according to their correlation with Normal heartbeats, Premature Ventricular Contractions and Premature Atrial Contractions in order to establish the basis for a multi-class arrhythmia classification approach.

- Chapter 8 contains the main conclusions derived from the research project alongside some considerations for further research endeavours.

Chapter 2

The heart, ECG signals and Premature Contractions

The heart is the organ responsible of pumping the blood throughout the body. It is composed of two superior chambers, the atria, and two lower chambers, the ventricles. Synchronous functioning of the chambers is vital for proper functioning of the cardiovascular system. During normal operation, the atria are activated first to fill the ventricles with blood. Later, the ventricles are activated to pump the blood out of the heart.

The Electrocardiogram (ECG) is a signal that represents the electrical activity of the heart. Usually measured directly on the body's surface, the ECG waveform is mainly composed of five characteristic components denoted by the letters P, Q, R, S and T. Each component corresponds to a change in electrical activity in the heart and the corresponding movement of the cardiac muscle. Sometimes an U component is also present following the T peak. [17]

Figure 2.1 shows the anatomical structure of the heart and a normal sinus ECG waveform. The electrical conduction system of the heart is also

highlighted.

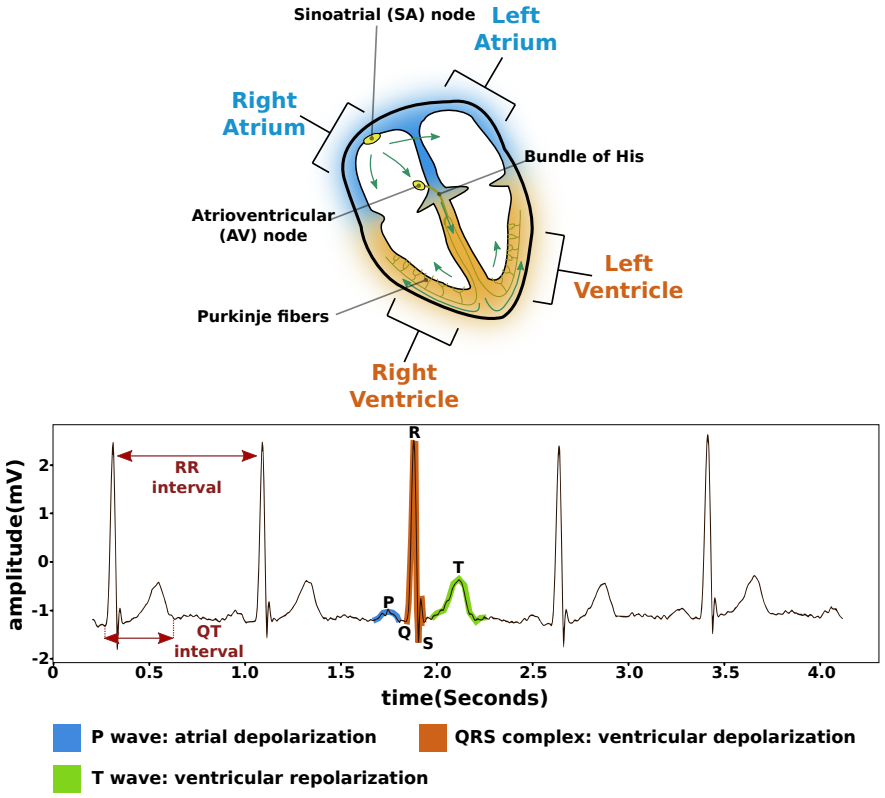


Figure 2.1: The anatomical and electrical structure of the heart (top) and a normal ECG signal (bottom) with its main peaks and valleys displayed. Green arrows on the heart's figure represent the normal flow of electrical impulses, from the Sinoatrial node down to the Purkinje fibers. The contoured peaks and valleys in the ECG waveform describe the main structure of a heartbeat.

To achieve the proper sequential functioning of the cardiac muscle, the cardiac cycle start from the Sinoatrial (SA) node where an electrical impulse is generated and then conducted through the Atrioventricular (AV) node and then to the ventricles trough the Bundle of His and the Purkinje fibers.

An arrhythmia is a disturbance of the rate, rhythm or pattern of the cardiac cycle [22]. Premature Beats are heartbeats that occur earlier in the

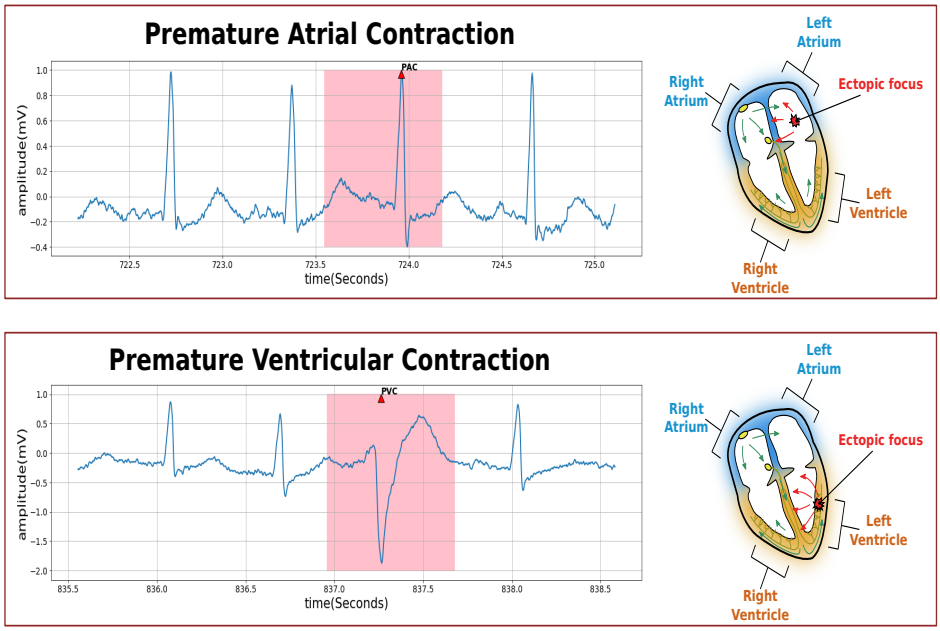


Figure 2.2: Two ECG waveforms showing 4 heartbeats each. In the top one, a Premature Atrial Contraction occurs in the third heartbeat. On the waveform on the bottom, a Premature Ventricular Contraction also occurs in the third heartbeat. The heart regions (ectopic focus) were the irregular pulses are triggered are also shown.

cardiac cycle due to the presence of ectopic focus, that is, segments of tissue that spontaneously fire impulses due to some underlying pathology [3].

Two common types of premature beats are Premature Atrial Contractions (PAC) and Premature Ventricular Contractions (PVC). Figure 2.2 shows the ECG waveform for those type of arrhythmia and the the hypothetical location of the ectopic focus in the heart.

Premature Atrial Contractions produce a change in the heart rate and present an altered P-wave that usually overlaps with the T-wave of the precedent heartbeat. Premature Ventricular Contractions also produce a change in the heart rate and can be identified by larger QRS complexes [3].

The occurrence of premature contractions can be an important indicator of the presence of an underlying hearth disease. Continuous ECG monitoring constitute a vital tool for early identifying such premature beats.

Chapter 3

Finite State Machines (FSM) probability models with antidictionaries

3.1 Overview of Antidictionary Coding Theory

For a string $\boldsymbol{x}^n = x_1 x_2 \dots x_n$ of length n defined over a finite alphabet of integers $\Sigma_m = \{0, 1, \dots, m-1\}$, the set $\mathcal{D}(\boldsymbol{x}^n)$ called the dictionary of $\boldsymbol{x}^n$ is considered as the set that contains all substrings of $\boldsymbol{x}^n$ including the null string $\boldsymbol{\lambda}$ of length zero. In contrast, the antidictionary $\mathcal{A}(\boldsymbol{x}^n)$ of $\boldsymbol{x}^n$ is defined as the set of *minimal* strings that never appear in $\boldsymbol{x}^n$. An element $\boldsymbol{v} = v_1 v_2 \dots v_k$ in $A(\boldsymbol{x}^n)$ is called Minimal Forbidden Word(MFW) which must satisfy the following three conditions:

1. $\boldsymbol{v} \notin \mathcal{D}(\boldsymbol{x}^n)$

2. A one-symbol shorter prefix of $\boldsymbol{v}$, defined as $p(\boldsymbol{v}) = v_1 v_2 ... v_{k-1}$, must

be contained in $\mathcal{D}(\boldsymbol{x}^n)$.

3. A one-symbol shorter suffix of $\boldsymbol{v}$, defined as $s(\boldsymbol{v}) = v_2 v_3 ... v_k$, must be also contained in $\mathcal{D}(\boldsymbol{x}^n)$.

Consider a short string $\boldsymbol{w} = 2210010$ over the alphabet $\Sigma_3 = \{0, 1, 2\}$. The antidictionary $\mathcal{A}(\boldsymbol{w})$ of $\boldsymbol{w}$ is given by the set of all MFWs of $\boldsymbol{w}$, thus $\mathcal{A}(\boldsymbol{w}) = \{02, 000, 11, 12, 101, 20, 222, 0100\}$.

A suffix trie structure $T(\boldsymbol{w})$ can be used to obtain $\mathcal{A}(\boldsymbol{w})$ [14]. The string path defined by the path covered from the root r of the trie to a given node s, here denoted by $\mathcal{V}(s)$, corresponds to a suffix of $\boldsymbol{w}$. Here, it is important to differentiate two types of nodes in the trie structure: *implicit* nodes have one child node while *explicit* nodes have at least two child nodes.

Given $T(\boldsymbol{w})$, an extended trie $T_{ex}(\boldsymbol{w})$ is formed by adding new child nodes to all *implicit* nodes on $T(\boldsymbol{w})$. To find all elements in $\mathcal{A}(\boldsymbol{w})$ it is sufficient to consider the strings defined by the newly added nodes on the leaf with the shortest distance to the root and all internal nodes [14].

Figure 3.1 shows the extended suffix trie $T_{ex}(\boldsymbol{w})$ for string $\boldsymbol{w}$. On the figure, nodes are identified by an integer number j, with the root r being the node with $j = 0$.

A string $\mathcal{V}(s)$ defined by an extended node s in $T_{ex}(\boldsymbol{w})$ is an MFW if it satisfies the three conditions mentioned before. In the case of Figure 3.1, strings $\mathcal{V}(24) = 0100$ and $\mathcal{V}(34) = 222$ are examples of valid MFWs.

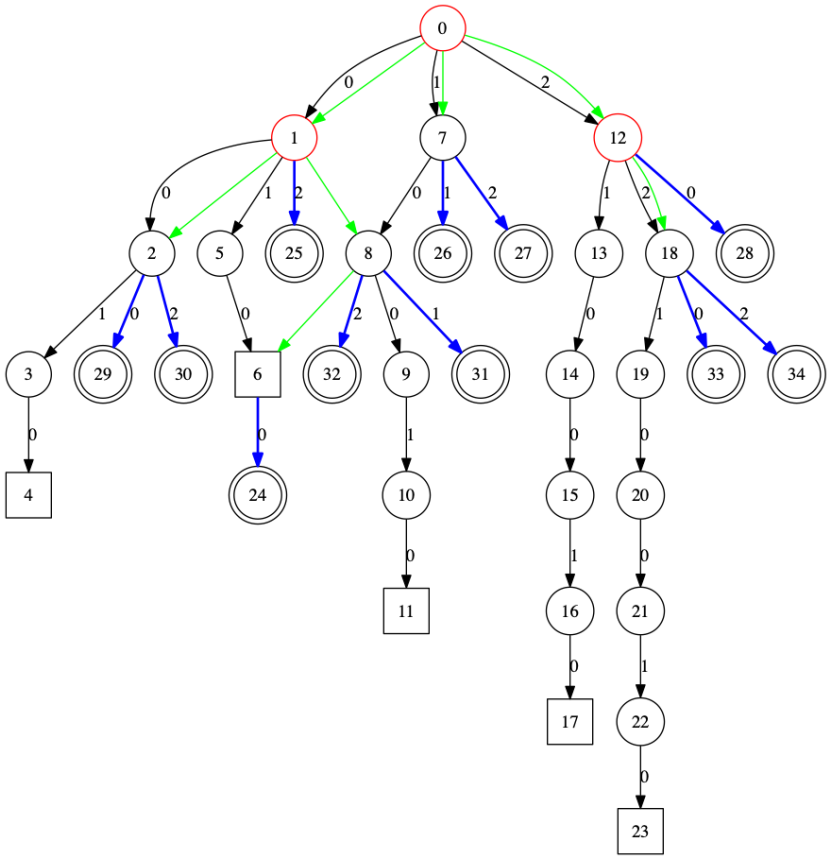


Figure 3.1: Extended suffix trie for string $\boldsymbol{w} = 2210010$. Black edges denote the path of the original suffix trie while blue edges signify the added edges pointing to the extended nodes. Green edges correspond to *MF-link* pointers [12].

For the efficient computation of $\mathcal{A}(\boldsymbol{w})$, a special pointer structure named *MF-link* is employed. An *MF-link* points an internal node u in $T(\boldsymbol{w})$ to a node v such as the string path of v equals the string path of u plus a single prefix symbol a, that is, $\mathcal{V}(v) = a\mathcal{V}(u)$ for $a \in \Sigma$ [13] [14].

3.2 Finite State Machine (FSM) models

A probabilistic model can be built for the assessment of irregular pattern occurrence following the Data Compression using Antidictionary (DCA) algorithm [15]. A proper set of MFW is selected from the antidictionary set $\mathcal{A}(\boldsymbol{x}^n)$ of a string $\boldsymbol{x}^n$ for the construction of a *Finite State Machine (FSM)*. A probabilistic model is built on top of the FSM, such as that normally it would accept substrings of $\boldsymbol{x}^n$, but in the presence of an MFW it will transition into a terminal state.

Let's reconsider the string $\boldsymbol{w} = 2210010$ defined over $\Sigma_3 = \{0, 1, 2\}$, introduced on section 3.1. A subset of $\mathcal{A}(\boldsymbol{w})$, say $\mathcal{A}_s(\boldsymbol{w}) = \{02, 11, 20\}$, can then be used for the construction of a FSM probabilistic model. The FSM is displayed in Figure 3.2.

The FSM is composed by a total of seven states or nodes, with four states S_1, S_2, S_3 and S_4 being the internal states and the remaining three states R_1, R_2 and R_3 being the external states. Each internal state points to another state (or to itself) through an outgoing edge, in Figure 3.2 an edge is defined by an arrow and the accompanying symbol $c \in \Sigma_3$ that causes the transition to a new state.

3.3 Probabilistic Model Construction

Assume that the *ith* symbol of $\boldsymbol{x}^n (1 \leq i \leq n)$ is being processed by the FSM, and let's define the next state reached sequentially by x_i as s_i, where $1 \leq i < n$ and s_0 denotes the initial state of the FSM. Moreover, assume that the state sequence $s^m = s_0 s_1 \dots s_m$ is uniquely determined by the input string $\boldsymbol{x}^m = x_1 x_2 \dots x_m (1 \leq m \leq n)$. The probability of transitioning to a state specified by the next symbol on the sequence, $P(x_{i+1}|s_i)$, is given by

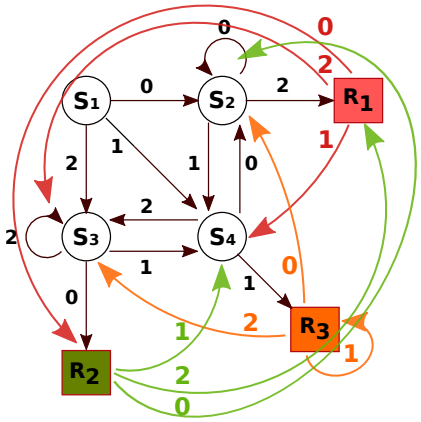


Figure 3.2: An FSM model for MFWs 02, 11 and 20. Coloured edges denote the corresponding transitions after a forbidden state is reached.

$$P(x_{i+1}|s_i) = \frac{N(x_{i+1}|s_i)}{\sum_{c\in\Sigma} N(c|s_i)}, \quad 0 \leq i < n \tag{3.1}$$

where $N(c|s_i)$ holds the number of times that a transition has occurred from state s_i with symbol c.

The FSM is devised to normally loop through internal nodes. In the case of a transition to a (forbidden) external state s_i through symbol x_i, the algorithm would point next to the node that covers the sequence $v = x_{i-s} \dots x_{i+1}$, where v coincides with the *Longest Common Prefix* with one MFW in $\mathcal{A}(\boldsymbol{w})$ and $(0 \leq s \leq i-1)$.

Chapter 4

Redefinition of the ECG distribution

In this section a dual process aimed to translate the binary ECG distribution into a distribution defined over an alphabet of small integers is presented.

4.1 Signal differentiation

The ECG signal usually contains non stationary features, such as individual RR and QT intervals between heartbeats [4]. Furthermore, noise and artifacts can significantly affect the shape of the probability distribution constructed from the ECG series [23]. Irregular arrhythmia heartbeats can also cause asymmetry and flattening on the distribution [21].

The irregular shape of the ECG amplitudes distribution can be appreciated on Figure 4.1. The left part shows a histogram computed from the amplitudes of an ECG record consisting of 650,000 samples (with 11 bit resolution).

The irregularities in the shape of the ECG distribution can difficult the

computation of accurate probability models from the time series. A solution to overcome the lack of stationarity of the distribution consists on the use of a differentiation operation over successive samples in the time series. For a sampled ECG signal sequence denoted by $\boldsymbol{z}^n = z_1 z_2 \ldots z_n$ where $z_i \in \Sigma_{2048}, 1 \leq i \leq n$, the differentiation process that yields the output sequence $\boldsymbol{y}^n = y_1 y_2 \ldots y_n$ is stated as follows:

$$y_i = \begin{cases} z_i & i = 1 \\ z_i - z_{i-1} & 1 < i \leq n \end{cases} \tag{4.1}$$

where $|y_i| \leq 2047$ $(1 \leq i \leq n)$.

The amplitudes histogram built from $\boldsymbol{y}$ shows a shape that resembles the Laplace distribution although the former is a discrete distribution while the latter is a continuous one, with a considerable less amount of dispersion than the distribution of $\boldsymbol{z}$. On the right part of Figure 4.1, the histogram of the distribution obtained after the application of the differentiation process on the ECG signal is presented.

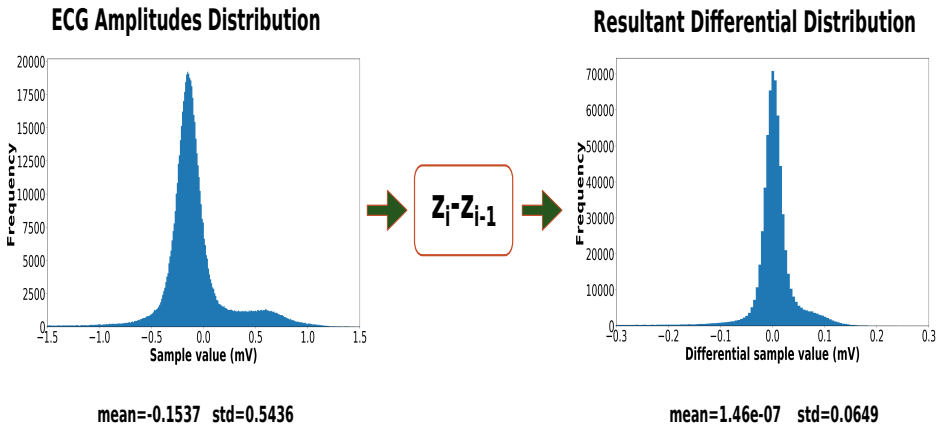


Figure 4.1: Histograms for raw ECG amplitudes (in millivolts) distribution (left) and the resultant differential amplitudes distribution after the differentiation process (right).

4.2 Signal quantization

In order to translate the information contained in the differential signals into a domain defined over a smaller alphabet set, a quantization stage is introduced. Let Q be the odd number of quantization levels under which the quantization will take place. Quantized symbols can be defined only on the set of integers $\{0, 1, \ldots, Q-1\}$.

The quantization process is carried out with a simple ranking system that assigns to each differential sample its corresponding quantized symbol depending on its differential amplitude value.

A sequence of quantized symbols of length n is denoted as a string $\boldsymbol{x}^n$ over Σ_Q where $m = Q$. The ith quantized symbol x_i on the sequence $\boldsymbol{x}^n = x_1 x_2 \ldots x_n \in \Sigma_Q^n$ can be obtained from the differential sequence $\boldsymbol{y}^n$ and the set of quantization parameters $\{q_0, q_1, \ldots, q_{Q-2}\}$ where q_l's $(0 \leq l \leq Q-2)$ are reals and $p_i < p_j \;\; (0 \leq i < j \leq Q-2)$.

The quantization rule is then given as follows:

$$x_i = \begin{cases} 0 & y_i \leq q_0, \\ l & q_{l-1} < y_i \leq q_l \\ Q-2 & y_i > q_{Q-2}. \end{cases} \tag{4.2}$$

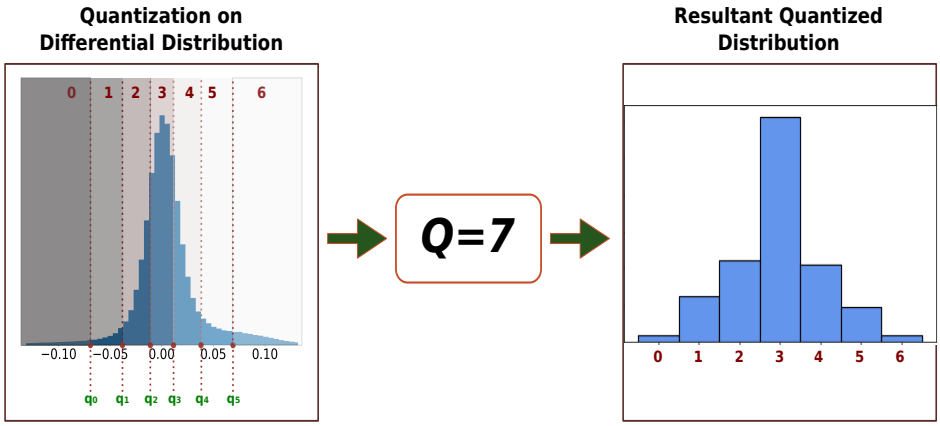


Figure 4.2: Quantization operation on a differential ECG distribution (left) and the resultant quantized distribution (right). The location of the quantization parameters and the corresponding interval for each symbol are highlighted.

The quantization procedure is illustrated on Figure 4.2 for a value of $Q = 7$.

Chapter 5

FSM model construction and accuracy assessment for PVC detection

This section presents the description of the experimental procedure undertaken with a set of ECG records for the detection of Premature Ventricular Contractions in an offline fashion.

The test signals used for the experiments were taken from the MIT-BIH Arrhythmia Database, a library of annotated ECG records commonly used for the evaluation of ECG arrhythmia detection algorithms. It consists of 48 records, each one comprising 30 minutes of ECG recording on two channels digitized at 360 samples per second per channel at an 11-bit resolution spanning a 10 mV dynamic range. Each ECG record contains annotations given by two or more cardiologists, thus providing a medical grade benchmark for the assessment of the quality of arrhythmia detectors [11, 9].

Figure 5.1 shows a schematic diagram of the proposed detection system. Multiple FSM probabilistic models were constructed for the accurate detec-

tion of PVC on the quantized ECG signals.

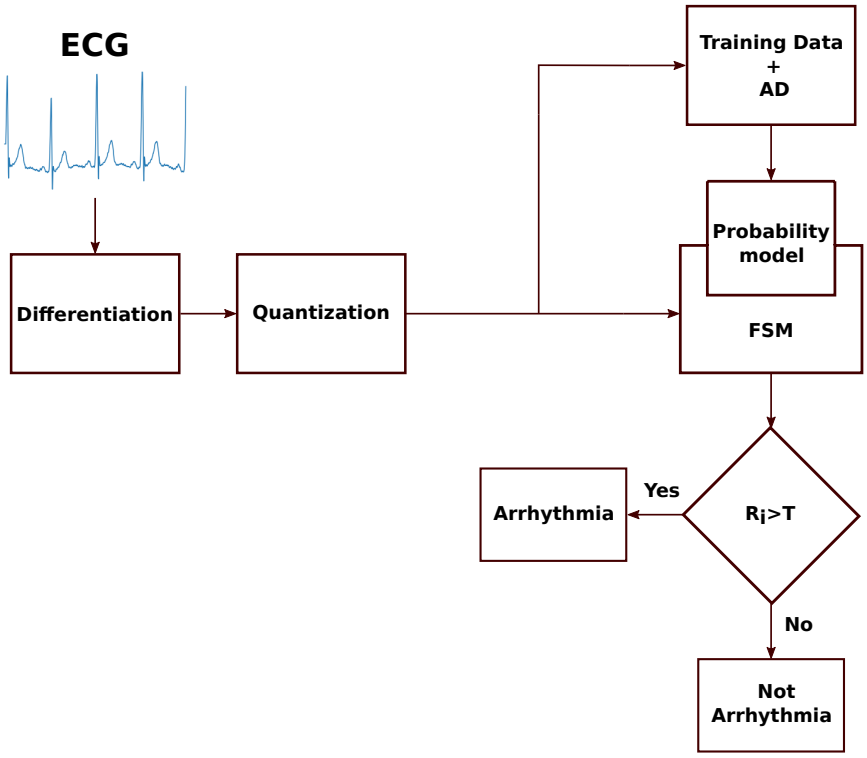


Figure 5.1: Schematic diagram of the detection system.

5.1 Training data and Antidictinary AD generation

Let $\mathcal{A}$ be the antidictionary set to be used in the encoding process. $\mathcal{A}$ is generated from a short segment of training ECG data in a preprocessing stage as follows:

Step 1. Let k be a certain positive integer denoting the total number of training files from which the antidictionary $\mathcal{A}$ will be constructed. Each training file $\boldsymbol{u}_i\,(1 \leq i \leq k)$ consists of 5 ECG waveforms (roughly between

3 and 5.5 seconds of ECG recording). Here, a waveform is defined as the portion of the signal covered by one R-R interval. For each training file $\boldsymbol{u}_i$ an antidictionary set $\mathcal{A}(\boldsymbol{u}_i)$ is constructed and the process results in the family of antidictionaries

$$\mathcal{A}_K = \{\mathcal{A}(\boldsymbol{u}_1), \mathcal{A}(\boldsymbol{u}_2), \ldots, \mathcal{A}(\boldsymbol{u}_k)\}\,.$$

The process is illustrated on Figure 5.2.

Step 2. A common antidictionary set $\mathcal{A}$ is conformed primarily by the set of MFWs that show a higher frequency of occurrence among all the generated antidictionaries $\mathcal{A}(\boldsymbol{u}_1), \mathcal{A}(\boldsymbol{u}_2), \ldots, \mathcal{A}(\boldsymbol{u}_k)$ in $\mathcal{A}_k$. Some MFWs are expected to appear constantly among the majority of the generated antidictionaries. However, the dynamic variations in the amplitude and periods of the training waveforms induce some variability on the frequency of occurrence of some MFWs. Given an MFW $\boldsymbol{w}$ in $\mathcal{A}_K$, the frequency of occurrence $f(\boldsymbol{w})$ on $\boldsymbol{w}$ is given by

$$f(\boldsymbol{w}) = |\{i \,|\boldsymbol{w} \in \mathcal{A}(\boldsymbol{u}_i), 1 \leq i \leq k\}|\,. \tag{5.1}$$

Step 3. The MFWs are next sorted based on those f values and the antidictionary set $\mathcal{A}$ can then be built with the MFWs that exhibit a relatively high frequency of occurrence. Experimental trials show that the MFWs with the higher frequency of occurrence are in general short strings of length one or two. Those short strings usually perform poorly when implemented in the FSM model in the detection scheme. In that sense, for the construction of $\mathcal{A}$, the constraint of choosing MFWs of length greater than or equal to 3 is imposed in order to achieve better performance in the detection algorithm.

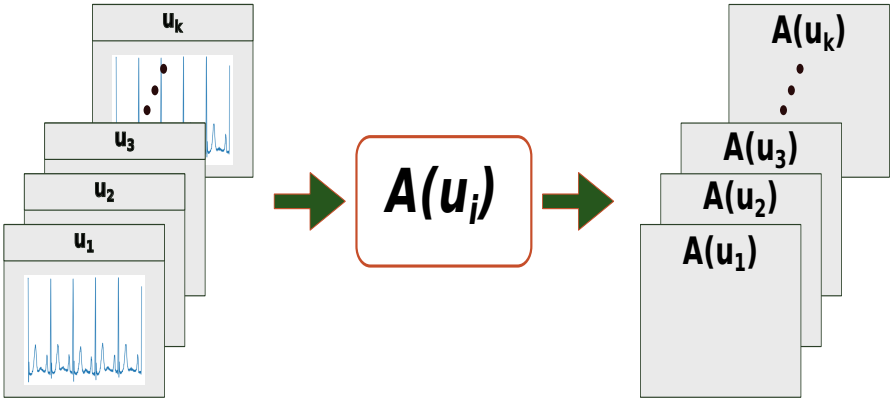


Figure 5.2: Antidictionary set construction from the ECG training files set.

5.2 FSM construction and detection criteria

With the appropriate set of MFWs picked out from the antidictionary set $\mathcal{A}$, the FSM can be build alongside the accompanying probability model.

Given a set of MFWs defined over an alphabet Σ_Q, in the implementation of the FSM each state is modeled with two memory registers for each outgoing edge associated with symbol $c \in \Sigma_Q$. The first register acts as a pointer to the next state reached through symbol c and the second register is implemented as a counter that holds the number of transitions to the next state though c.

For the particular case of the FSM with $Q = 3$ on Figure 3.2, there are seven states in total with three outgoing edges per state. Thus, the number of memory registers necessary for the implementation of the FSM would be equal to 42.

Once the FSM model is constructed, the transition probabilities are calculated by performing a second pass in the training data, that is, by sequentially feeding the quantized sequence contained in each training file into the FSM model. After every node transition, the corresponding transition counter

would be updated.

Figure 5.3 exemplifies the memory map for node S_3 on the FSM introduced on Figure 3.2.

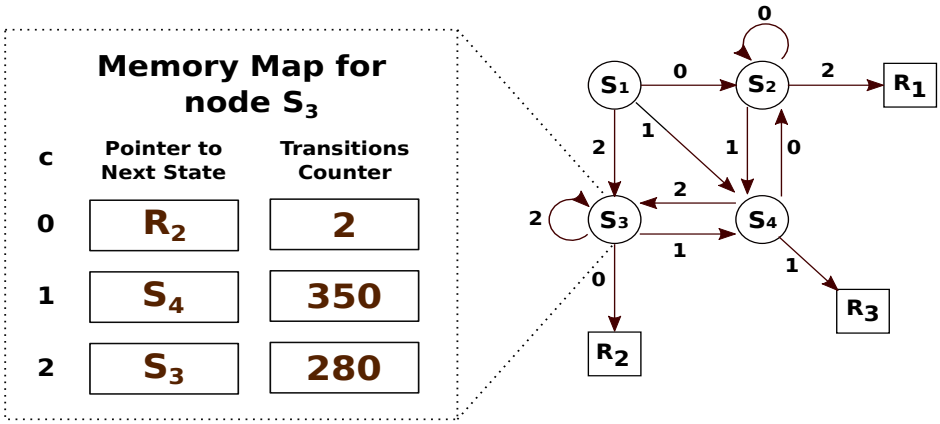


Figure 5.3: A table of memory registers corresponding to node S_3 on the FSM from Figure 3.2 .

Assume a string $\boldsymbol{x}^n = x_1 x_2 \dots x_n$ to be processed with the detection algorithm by means of a FSM constructed with the appropriate probabilistic model. For $1 \leq i \leq n$ and a given number $d > 0$, the *instantaneous* Compression Ratio R_i defined on a sliding window w_i of size d is given by

$$R_i = \begin{cases} \dfrac{1}{d} \displaystyle\sum_{k=i-d+2}^{i+1} \ln \frac{1}{P(x_k|s_{k-1})} & i - d > 0, \\ \\ \dfrac{1}{i} \displaystyle\sum_{k=2}^{i+1} \ln \frac{1}{P(x_k|s_{k-1})} & i - d \leq 0. \end{cases} \tag{5.2}$$

where $P(x_k|s_{k-1})$ is the transition probability defined in (3.1).

A proper threshold value T must be chosen in such a way that an instantaneous Compression Ratio R_i greater than T will signal the presence of

an arrhythmia pattern in the input string. Most experiments suggest that a threshold value between the range $[2.0, 2.8]$ can lead to the best results.

Figure 5.4 illustrates the case of a positive detection when the PCV located around 156 sec. time stamp causes an increase on the compression ratio, and thus surpassing the set threshold (2.5).

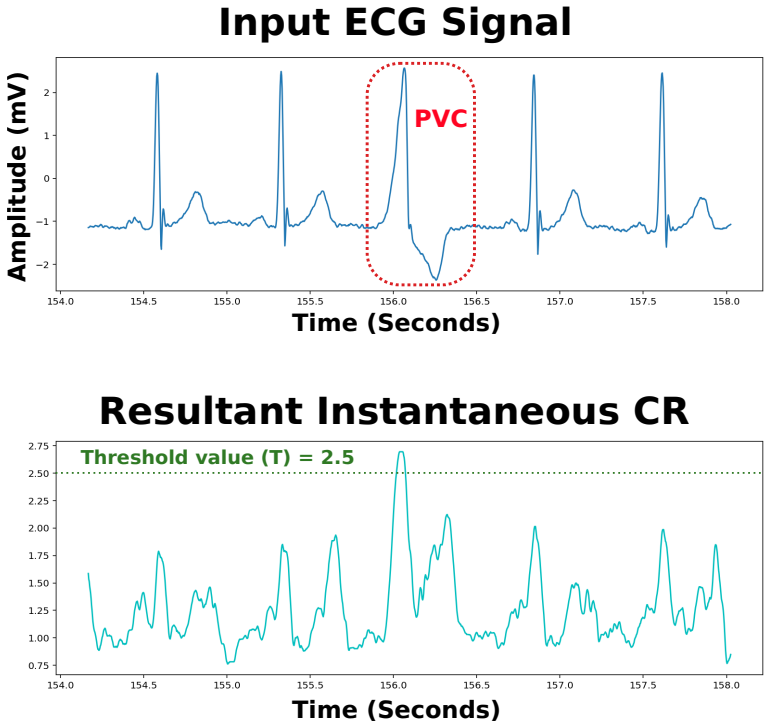


Figure 5.4: Positive detection of a PVC heartbeat. The top figure displays an ECG sequence containing one PVC heartbeat while the bottom figure shows how the Instantaneous Compression Ratio goes above the set threshold value (T=2.5), likely due to a forbidden pattern occurring within the PVC.

The accompanying annotations files from the MIT-BIH Arrhythmia Database where used for the posterior calculation of the detection evaluation metrics sensitivity and specificity. The sensitivity Se, or true positive rate, measures the ratio of true arrhythmia heartbeats detected while the speci-

ficity Sp measures the ratio of normal heartbeats identified as such by the algorithm. Both metrics are computed as

$$Se = \frac{TP}{TP + FN}, \; Sp = \frac{TN}{TN + FP} \tag{5.3}$$

were the True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) conditions are defined in the confusion matrix on Table 5.1.

Table 5.1: Confusion matrix

	Detected Abnormal heartbeat	**Detected Normal heartbeat**
Actual Abnormal heartbeat	TP	FN
Actual Normal heartbeat	FP	TN

5.3 Experimental procedure and results for offline FSM construction and PVC detection

In this section, the details of two different set of experiments for the detection of PVC are given alongside the obtained results in each case. A family of FSM probabilistic models was constructed in each experiment set.

The first group of experiments was performed with quantized ECG signlas under a range of different quantization levels Q, and the results give an insight into the most suitable quantization level and parameters.

The second group of experiments was carried on quantized ECG signals that were obtained with the quantization criteria learned from the results of the first set of experiments.

5.3.1 Preliminary experimentation for different quantization levels: assessment on optimal quantization parameters

Experimental procedure with multiple quantization leves

A set of experiments was carried out with 15 different records from the MIT-BIH database targeting the detection of PVC. Each record consists of a total of 650,000 samples.

In order to explore the performance of the detection algorithm under different quantization levels Q, the preliminary stage of experimentation consisted in the construction of FSM probabilistic models for the set of 5 different quanization levels $Q_s = \{3, 5, 7, 9, 11\}$.

For a given an ECG record, and a quantization level $Q \in Q_s$, the experimental procedures were performed as described next:

Step 1. The *differentiation* of the binary ECG signal $\boldsymbol{z}^n$ was performed to yield the differential time series $\boldsymbol{y}^n$, with $n = 650,000$.

Step 2. The *quantization* of the differential sequence $\boldsymbol{y}^n$ was then performed according to the picked value of Q, to yield the quantized sequence $\boldsymbol{x}^n$.

Step 3. A training set $\{\boldsymbol{u}_1, \boldsymbol{u}_2 \dots \boldsymbol{u}_{50}\}$ was form from 50 different segments of data from the quantized ECG sequence $\boldsymbol{x}^n$. Each segment contains the data of 5 contiguous RR intervals of the sequence. The corresponding information on the R peaks location within the ECG time series is easily

obtained from the MIT-BIH database annotation files.

Step 4. The set of antidictionaries $\mathcal{A}_{50} = \{\mathcal{A}(\boldsymbol{u}_1), \mathcal{A}(\boldsymbol{u}_2), \ldots, \mathcal{A}(\boldsymbol{u}_{50})\}$ was computed from the set of 50 training files.

Step 5. The frequency of occurrence $f(\boldsymbol{w})$ of every MFW $\boldsymbol{w}$ in $\mathcal{A}_{50}$ was then evaluated in addition to its string length $|\boldsymbol{w}|$ and a common antidictionary set $\mathcal{A}_c$ was formed with the 10 MFWs with the highest frequency $f(\boldsymbol{w})$ and a string length $|\boldsymbol{w}| \geq 3$.

Step 6. From the common antidictionary set $\mathcal{A}_c$, A family of 45 different FSM models $\{\boldsymbol{FSM}_1, \boldsymbol{FSM}_2, \ldots, \boldsymbol{FSM}_{45}\}$ was constructed taking into consideration all unique combinations of 2 MFWs in $\mathcal{A}_c$. The transition probabilities for each FSM model in the set were learnt by performing a pass through the data on the training set $\{\boldsymbol{u}_1, \boldsymbol{u}_2 \ldots \boldsymbol{u}_{50}\}$.

Step 7. For each model in $\{\boldsymbol{FSM}_1, \boldsymbol{FSM}_2, \ldots, \boldsymbol{FSM}_{45}\}$, the resultant instantaneous compression ratio R_i was calculated from the whole quantized ECG sequence $\boldsymbol{x}^n$, where $1 \leq i \leq n$ and $n = 650000$, as stated on equation 5.2. The used value for the size of the sliding window was $d = 25$ for all FSM models.

Step 8. The accuracy of each FSM model for PVC detection was calculated multiple times on a range of threshold values T in the closed interval $[1.8, 3.2]$ with increments of 0.01 units, for a total of 141 different evaluations. The performance metrics sensitivity Se and specificity Sp were calculated as described in section 5.2 of Chapter 5.

Step 9. The FSM model with the highest combination of Se and Sp was then selected as the best performing model, and the process was then repeated for the next quantization level in Q_s.

Results for FSM with multiple quantization leves

The obtained results for the FSM probability models constructed with multiple quantization levels are displayed in Table 5.2. The average performance values for each quantization level Q over all records, points out that the algorithm may perform at its best when $Q = 7$.

From observations made on the quantization parameters $\{q_0, q_1, \ldots, q_5\}$ used in the experiments for $Q = 7$, a rough approximation to the best quantization parameters is given as follows: consider the *percentile* P_r as the value on the ECG distribution below which a percentage $r(\%)$ of the samples is allocated. Then, the quantization parameters for quantization rule 4.2 are given as:

$$q_0 = P_{1.5}, \quad q_1 = P_{10}, \quad q_2 = P_{25}, \quad q_3 = P_{75}, \quad q_4 = P_{90}, \quad q_5 = P_{98.5} \quad (5.4)$$

5.3.2 Experimentation for FSM models with percentile quantization and $Q = 7$

Experimental Procedure with percentile quantization and $Q = 7$

The second stage of experimentation was undertaken with the quantization parameters given in 5.4. Similar to the previous round of experiments, the MIT-BIH database was employed as the source of the annotated ECG data. In particular, records 105, 201, 205, 215, 221 and 228 were used.

The experimental process is described in the following 9 steps:

Step 1. The *differentiation* of the binary ECG signal $\boldsymbol{z}^n$ was performed to yield the differential time series $\boldsymbol{y}^n$, with $n = 650000$.

Step 2. The *quantization* of the differential sequence $\boldsymbol{y}^n$ was made with

Table 5.2: Table of results (%) for the detection of PVC on 15 ECG records under 5 different quantization levels (Q). Sensitivity and Specificity values for PVC detection are given for each Q, with total averages in the last row [8].

	Q=3		Q=5		Q=7		Q=9		Q=11	
Record	Sen.	Spec.	Sen.	Spec.	Sen.	Spec.	Sen.	Spec.	Sen.	Spec.
105	4.87	97.26	95.12	91.01	100	76.56	100	87.45	100	48.85
106	1.73	99.40	94.42	99.20	92.50	93.09	98.07	93.43	94.80	94.09
114	4.65	100	97.67	86.75	95.34	83.18	83.72	83.13	97.67	60.82
116	3.66	99.13	98.16	99.21	98.16	94.43	95.41	93.22	97.24	97.30
119	8.78	99.61	100	100	99.54	99.54	100	99.44	99.77	99.48
200	11.98	99.65	72.76	84.27	80.38	79.23	87.65	82.32	87.16	81.46
201	5.05	99.87	94.54	95.94	96.46	97.53	72.22	96.18	86.36	86.70
203	3.82	98.97	69.59	63.02	72.07	78.84	70.04	67.69	55.18	82.99
205	33.80	98.56	76.05	89.80	94.36	97.51	98.59	96.69	98.59	93.27
210	23.19	99.71	77.31	79.15	94.32	66.85	72.68	76.22	60.30	91.63
215	3.04	99.78	96.34	93.77	90.85	96.24	84.14	93.52	95.73	86.38
219	9.37	99.85	89.06	93.99	70.31	95.82	84.37	97.83	93.75	98.27
221	74.24	40.22	9.59	97.83	96.46	98.03	98.48	97.63	99.24	90.94
228	0.82	99.70	96.68	94.84	97.79	96.44	97.23	95.26	96.40	96.20
233	1.44	99.50	75.21	93.67	85.92	97.21	85.07	86.72	89.89	90.53
Averages	**12.69**	**95.41**	**82.83**	**90.83**	**90.96**	**90.03**	**82.90**	**89.78**	**90.13**	**86.59**

the set of quantization parameters $\{q_0, q_1, \ldots, q_5\}$ obtained as stated on expression 5.4 and $Q = 7$.

For the determination of the set of quantization parameters, the percentiles values where calculated from the distribution of differential ECG samples obtained within the first minute of recording (corresponding to 21,600 samples for records from the MIT-BIH Arrhythmia Database).

Step 3. A training set $\{\boldsymbol{u}_1, \boldsymbol{u}_2 \ldots \boldsymbol{u}_{50}\}$ was formed from 50 different segments of data from the quantized ECG sequence $\boldsymbol{x}^n$. Each segment contains the data of 5 contiguous RR intervals of the sequence.

Step 4. The set of antidictionaries $\mathcal{A}_{50} = \{\mathcal{A}(\boldsymbol{u}_1), \mathcal{A}(\boldsymbol{u}_2), \ldots, \mathcal{A}(\boldsymbol{u}_{50})\}$ was computed from the set of 50 training files.

Step 5. The frequency of occurrence $f(\boldsymbol{w})$ of every MFW $\boldsymbol{w}$ in $\mathcal{A}_{50}$ was then evaluated in addition to its string length $|\boldsymbol{w}|$ and a common anti-

dictionary set $\mathcal{A}_c$ was formed with the 20 MFW with the highest frequency $f(\boldsymbol{w})$ and a string length $|\boldsymbol{w}|= 3$.

Step 6. From the common antidictionary set $\mathcal{A}_c$, A family of 190 different FSM models $\{\boldsymbol{FSM}_1, \boldsymbol{FSM}_2, \ldots, \boldsymbol{FSM}_{190}\}$ was constructed taking into consideration all unique combinations of 2 MFW in $\mathcal{A}_c$. The transition probabilities for each FSM model in the set were learnt by performing a pass through the data on the training set $\{\boldsymbol{u}_1, \boldsymbol{u}_2 \ldots \boldsymbol{u}_{50}\}$.

Step 7. For each model in $\{\boldsymbol{FSM}_1, \boldsymbol{FSM}_2, \ldots, \boldsymbol{FSM}_{190}\}$, the resultant instantaneous compression ratio R_i was calculated from the whole quantized ECG sequence $\boldsymbol{x}^n$, where $1 \leq i \leq n$ and $n = 650,000$, as stated on equation 5.2. The used value for the size of the sliding window was $d = 25$ for all FSM models.

Step 8. The accuracy of each FSM model for PVC detection was calculated multiple times on a range of threshold values T in the closed interval $[1.8, 3.2]$ with increments of 0.01 units, for a total of 141 different evaluations. The performance metrics sensitivity Se and specificity Sp were calculated as described in section 5.2 of Chapter 5.

Step 9. The FSM model with the highest combination of Se and Sp was then selected as the best performing model.

In some cases the FSM architecture models constructed with 2 MFW did not performed well enough (Se and Sp values below 90%), so bigger models were then implemented. In those cases, the FSM architecture was increased from 2 MFW to 4 MFW, taking into consideration the MFW pair $\mathcal{A}_t = \{\boldsymbol{w}_1, \boldsymbol{w}_2\}$ from the top performing FSM model.

To build the expanded FSM models, the conditions depicted on steps **5** and **6** for the construction of the common antidictionary $\mathcal{A}_c$ and family of

FSM models were modified as is described next:

Updated Step 5. With the MFW pair $\mathcal{A}_t = \{\boldsymbol{w}_1, \boldsymbol{w}_2\}$, look for a set of 6 new MFW $\mathcal{A}_n = \{\boldsymbol{w}_3, \boldsymbol{w}_4, \ldots, \boldsymbol{w}_8\}$ such as any MFW $\boldsymbol{w}$ in $\mathcal{A}_n$ satisfy the three conditions

1. has a string length $|\boldsymbol{w}| > 3$.
2. displays highest appearance frequency $f(\boldsymbol{w})$.
3. shares a two symbol common prefix with $\boldsymbol{w}_1$ or $\boldsymbol{w}_2$.

The new common antidictionary set $\mathcal{A}_c$ is then arranged as

$$\mathcal{A}_c = \mathcal{A}_t \cup \mathcal{A}_n = \{\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_8\}$$

.

Updated Step 6. From the new common antidictionary set $\mathcal{A}_c$, construct a family of 70 different FSM models $\{\boldsymbol{FSM}_1, \boldsymbol{FSM}_2, \ldots, \boldsymbol{FSM}_{70}\}$ from all possible different combinations of 4 MFWs in $\mathcal{A}_c$. The transition probabilities for each FSM model are acquired by performing a pass through the data on the training set $\{\boldsymbol{u}_1, \boldsymbol{u}_2 \ldots \boldsymbol{u}_{50}\}$.

After building the new set of FSM probability models, their performance is evaluated following steps **7** through **9**.

Results for FSM probability models with percentile quantization and $Q = 7$

Table 5.3 shows the performance results for the set of records 105, 201, 205, 215, 221, 228. The results for records 105, 205 and 228 correspond to cases with FSM models of 2 MFWs, while in the cases of records 201, 215 and 221 bigger models of 4 MFWs were obtained to improve detection accuracy.

Table 5.4 shows a comparative outline of the average metrics (sensitivity and specificity) obtained with the proposed detection method and the metrics achieved by other methods available on literature.

Table 5.3: Table of results (%) for the detection of PVC on 6 ECG records from the MIT-BIH database.

Record	**Sen.**	**Spec.**
105	100	94.77
201	98.98	98.40
205	97.18	99.57
215	93.29	79.74
221	97.97	94.43
228	97.79	96.44
Average	**97.53**	**93.89**

Table 5.4: Comparison of the proposed method with other arrhythmia detectors. Results are given on percentages (%).

Algorithm	**Sensitivity**	**Specificity**
Proposed method	**97.53**	**93.89**
(Ota et al, 2013)[16]	97.9	98.6
(Ittatinut et al, 2013)[10]	91.05	99.55
(Adnane et al, 2013)[1]	97.21	98.67
(Alajlan et al, 2014)[2]	100	93.71

The proposed detection algorithm achieves comparable values of Sensitivity with the other approaches highlighted on Table 5.4. The average Specificity, however, is quite low in comparison with the other methods, this most likely due to the fact that no previous treatment for noise removal on the test signals was performed in the experimental process. It is expected

that improvements on Specificity would follow with the use of appropriate methods for noise removal.

Table 5.5 shows a comparative description of the antidictionaries and FSM implementation characteristics for the proposed detection system and the detection system for binary ECG signals proposed in [16]. For the calculation of the antidictionary size values a byte has been assigned to describe every quantized symbol that conform an MFW.

Table 5.5: Antidictionaries and FSM characteristics for quantized signals from six different ECG records. Results obtained previously in [16] are given for comparison purposes.

Results obtained on (Ota et al, 2013) [16]						
ECG record	**105**	**201**	**205**	**215**	**221**	**228**
number of MFWs	281	90	56	178	85	189
AD size (bits)	3,586	996	436	1,792	988	2,316
FSM size (kB)	24.2	6.5	2.6	11.5	6.5	15.4

Results obtained with the proposed method						
ECG record	**105**	**201**	**205**	**215**	**221**	**228**
number of MFWs	2	4	2	4	4	2
AD size (bits)	48	120	48	144	112	48
FSM size (kB)	**1.568**	**3.136**	**1.568**	**3.584**	**2.464**	**1.568**

Figure 5.5 displays the Receiver Operating Characteristics (ROC) curves for three different FSM built for record 228, evaluated on a range of threshold values T raging from 1.8 to 3.2 on increments of 0.01 units. The set of MFWs for FSM Model-1 is $AD_1 = \{656, 513\}$ while the antidictionary sets for the remaining two models (FSM Model-2 and FSM Model-3) are $AD_2 = \{656, 5351\}$ and $AD_2 = \{013, 514\}$, respectively.

A common MFW, 656, can be found on the antidictionary set of the two best performing FSM models (FSM Model-1 and FSM Model-2), thus suggesting that MFW very likely corresponds to a forbidden pattern within the PVCs morphology.

Figure 5.6 shows the corresponding architecture for each FSM model.

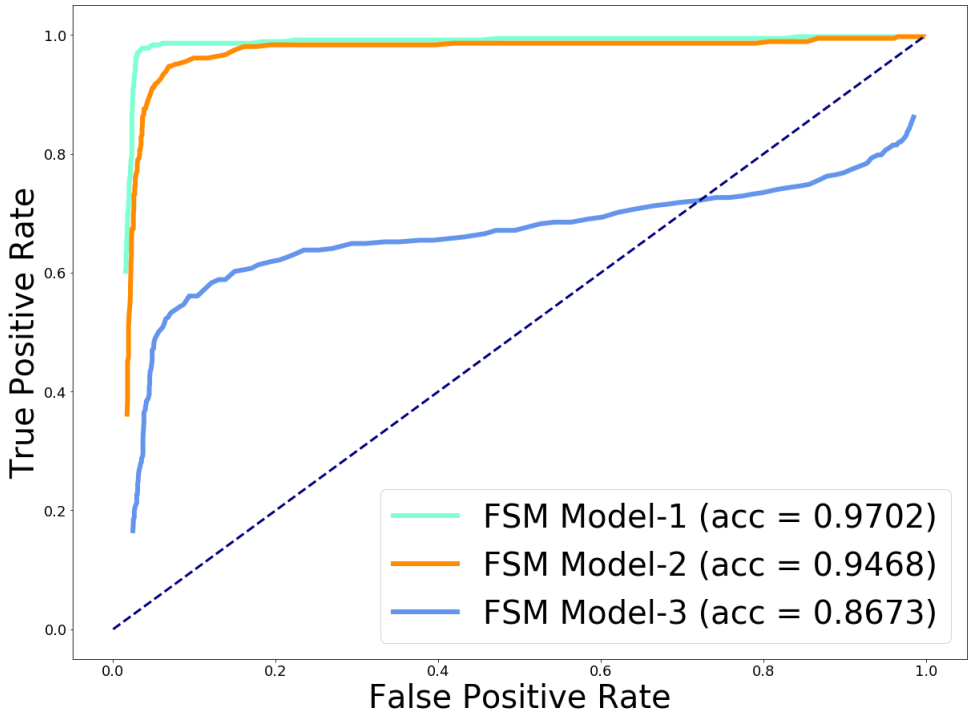
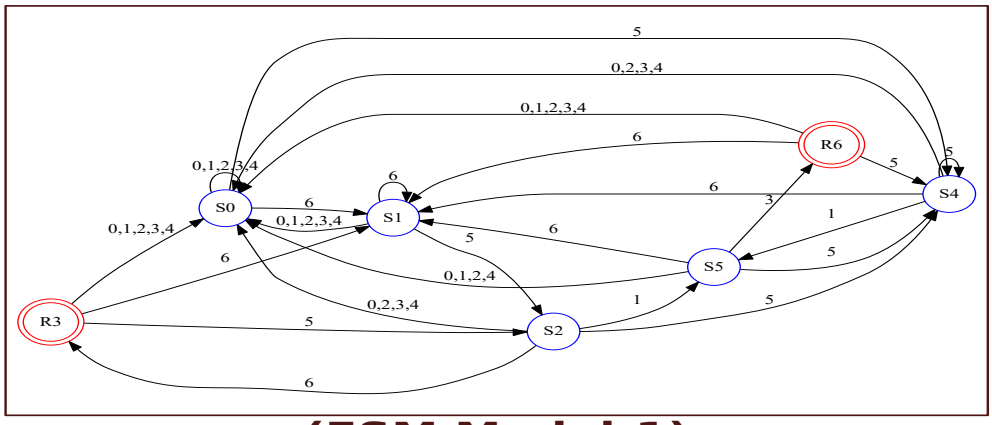
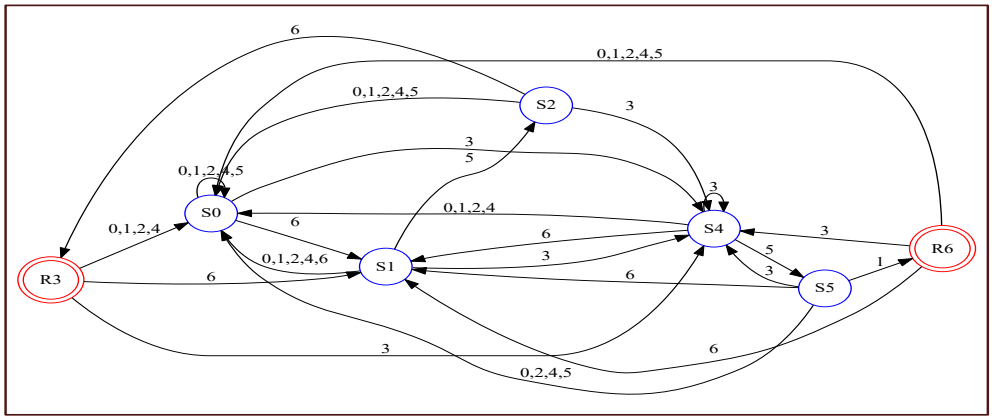


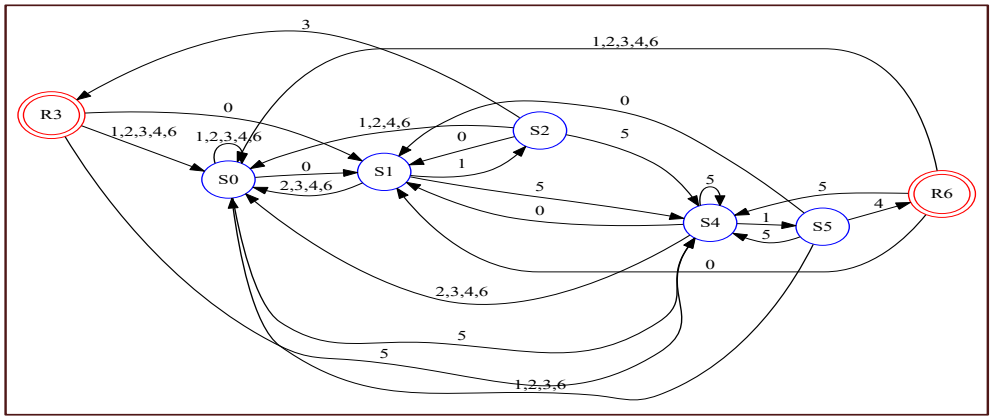
Figure 5.5: Receiver Operating Characteristics curves for three different FSM trained to process record 228 under a wide range of threshold values (T). Maximum accuracy values and the corresponding threshold values for each model are given as follows: FSM Model-1 achieving 97.02% accuracy at $T = 2.56$, FSM Model-2 reaching 94.68 % at $T = 2.67$ and FSM Model-3 with 86.73% accuracy at $T = 2.63$.



(FSM Model-1)



(FSM Model-2)



(FSM Model-3)

Figure 5.6: FSM models architectures. External states are represented in red, while internal states are blue.

Chapter 6

FSM Implementation and the evaluation of computational resources usage fo online processing on a mobile platform

A third stage of experimentation consisted on the port of the detection algorithm into a mobile environment for the evaluation of performance at online operation. The experimental setup is described on Figure 6.1. It basically consists of an emulated wearable ECG sensor for the transmission of ECG samples and a companion mobile application.

A pre-trained FSM model has been ported into the mobile application, in addition to a differentiation and quantization stages for the processing of a stream of ECG samples on real time. Continuing with the same methodology used on off-line experimentation, the records from the MIT-BIH Arrhythmia databased have been employed for testing the algorithm.

For the wireless transmission of the ECG samples from the emulated

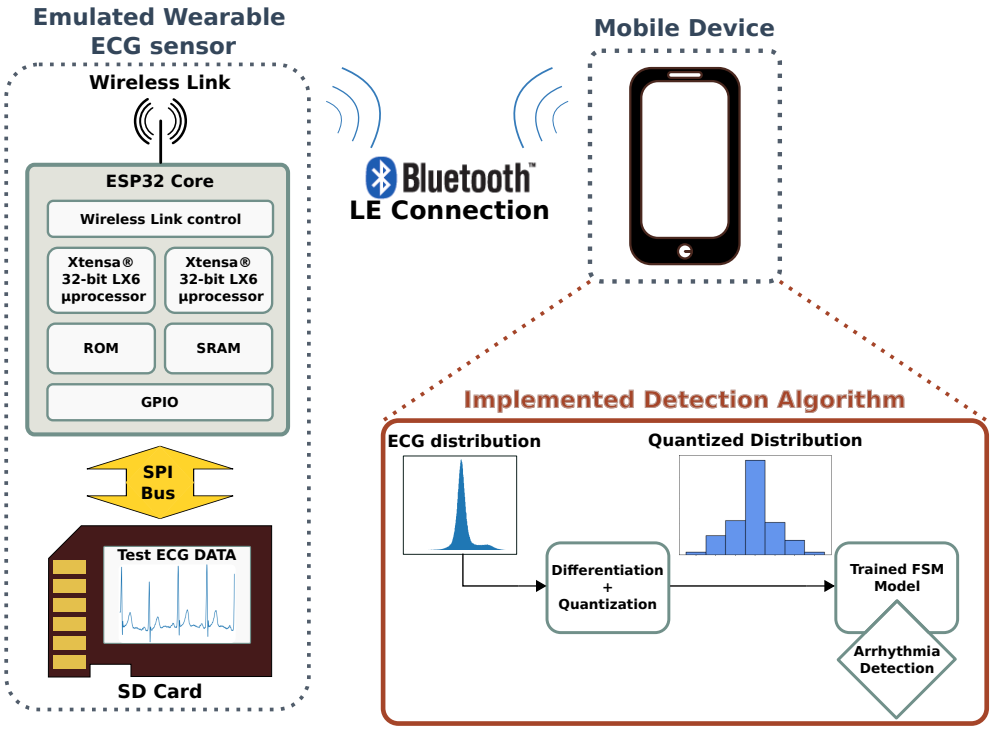


Figure 6.1: Schematic digram of the experimental setup used for the evaluation of the detection algorithm on a mobile platform.

sensor in an efficient and versatile fashion, the Bluetooth Low Energy (LE) wireless communication protocol has been used.

6.1 Bluetooth Low Energy stack implementation

Introduced on the 4.0 version of the Bluetooth specification, the Bluetooth LE standard enables low power communication for short distances, making it ideal for Internet of Things (IoT) applications. The Generic Attribute Profile (GATT) specification provides guidelines on standard data transfer

procedures and formats for Bluetooth LE devices[18].

A custom Bluetooth LE profile has been designed to transmit the ECG samples. Figure 6.2 displays a conceptual view of the devised profile. The profile specifies a *Service*, that is, a collection of related data structures called *characteristics*. The *ECG Measurement Characteristic* holds the actual sample values for two ECG samples at a given time. Samples are represented in a 16 bits (11 bits resolution) unsigned integer format each.

As shown on Figure 6.2, the characteristic is granted with a set of properties that specify the set of allowed operations that can be undertaken with the . Of particular importance is the *NOTIFY* property which informs the client side (mobile phone) of the availability of new samples to be transmitted. Each new notification is pushed every 6 ms, for a data rate of 333 samples/second.

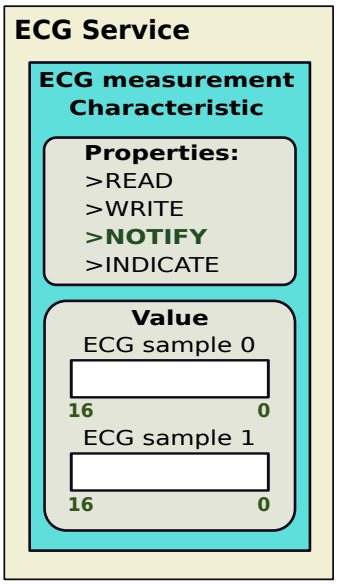


Figure 6.2: Conceptual view of the constructed Bluetooth LE profile. The profile implemented in both the emulated ECG sensor and the mobile application, holds a characteristic with the data pertaining two 16 bits ECG samples.

6.2 Wearable ECG hardware device characteristics

In order to simulate the characteristics of a wearable ECG sensor, a custom hardware arrangement has been configured. The device is capable of handling the wireless transmissions of the ECG samples obtained from the annotated files on the MIT-BIH Arrhythmia database. This allows the possibility of not only transmitting regular heartbeat patterns, but also arrhythmia heartbeats, simulating under different scenarios.

The virtual wearable monitor is based around the ESP32-WROOM-32 Microcontroller Unint from Espressif Systems[19]. Some of the main specifications of the device are:

- Dual Xtensa dual-core 32 bit processors.
- 80 to 240 MHz frequency of operation.
- 520 KiB SRAM memory.
- Bluetooth v4.2 and Wifi 802.11 b/g/n communication.

The ECG samples are read from binary files stored on a SD card with the use of one of the multiple on-board Serial Peripheral Interface (SPI) buses available on the ESP32-WROOM-32 device. The ESP32 (SoC) device supports a broad range of open software initiatives, such as the Arduino Open Software project. For this particular project, the whole configuration of the ESP32 core device has been carried out on top of the SPI and Bluetooth LE libraries freely provided by Espressif Systems and the open source community[20].

6.3 Mobile application deployment

6.3.1 Application architecture

The development of the mobile application has been carried on the IOS mobile operating system. The object oriented approach of the Swift programming language has facilitated the implementation of the quantization and FSM related data structures in conjunction with the Bluetooth LE related functionality in a relative short development time.

Key frameworks employed in the application architecture are listed next:

- *Core Bluetooth* (establish and manage Bluetooth LE connections).
- *Charts* (ECG and Compression Ratio signals plotting).
- *SigmaSwiftStatistics* (statistical features calculation).
- *CSVImporter* (management of text files in Comma Separated Values format).

An overview of a schematic flowchart that describes the application structure and functionality is given on Figure 6.3. The client side of the Bluetooth connection is implemented in the application.

Arriving samples are feed to a queue data structure that in turns updates a plot on the User Interface view. In the same fashion, another queue is responsible of holding the corresponding Compression Ratio samples of the already processed samples and updating a secondary graph with the Compression Ratio signal. Both graphs are refreshed for every new 10 samples feed to the queues structures.

The *Differentiator* class (abbreviated as *Diff.* on Figure 6.3) performs the subtraction of contiguous samples.

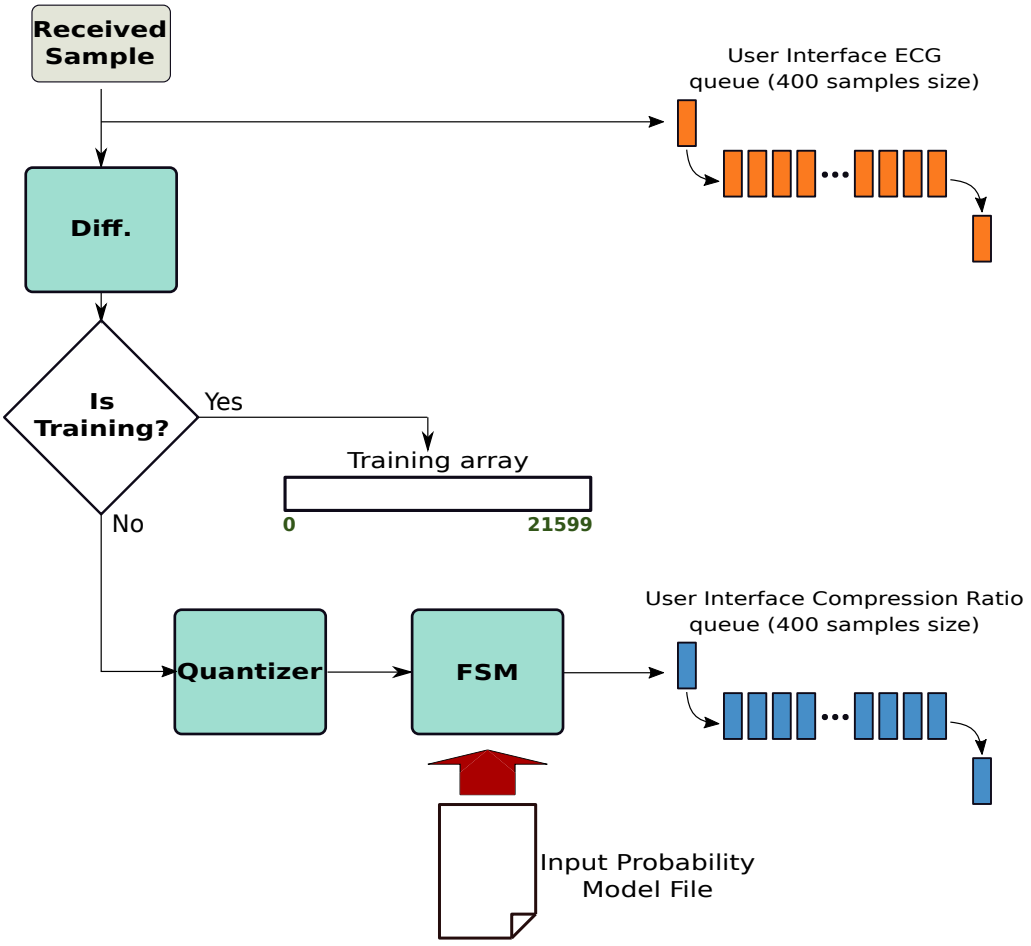


Figure 6.3: IOS application flowchart.

A training process is carried out with one minute of ECG recording, 21600 samples for records of the MIT-BIH database sampled at 360 Hz. An array structure store the incoming differential samples during the training process. After the array gets full, a calculation of the quantization parameters over the samples distribution on the array is performed as stated in expression 5.4.

The *Quantizer* class is updated with set of learned quantization parameters, and quantization is then performed following rule 4.2. The *FSM* class

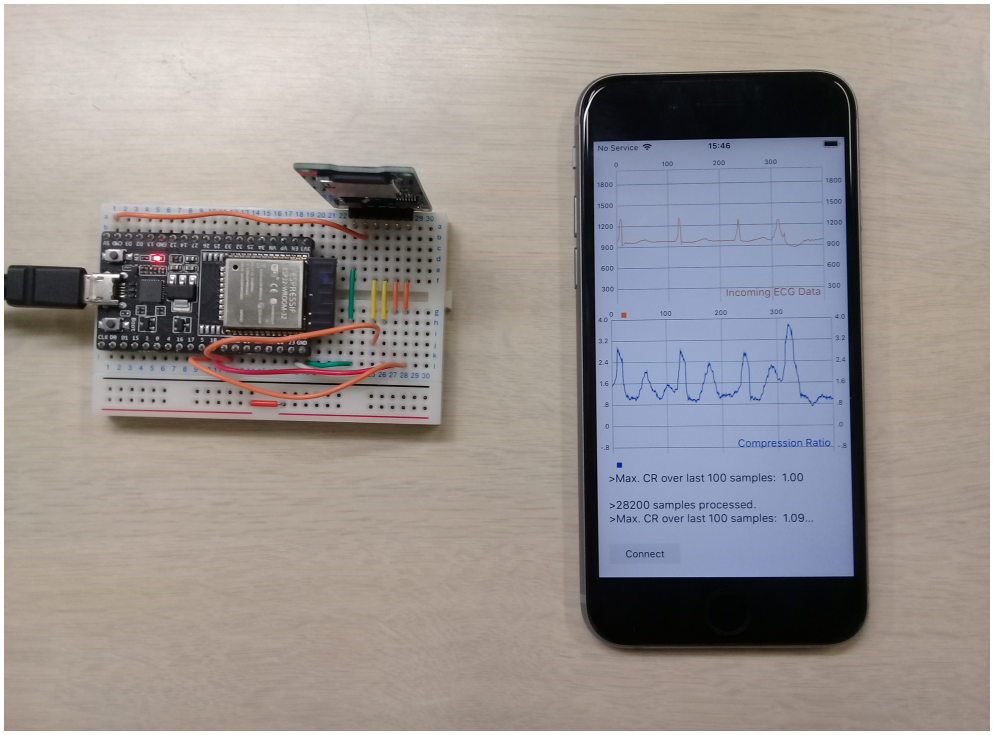


Figure 6.4: The experimental setup for the online evaluation of the trained FSM models. The emulated ECG sensor (on the left) transmits the ECG samples through Bluetooth LE while in the mobile application the quantization and posterior processing in the FSM is effectuated to produce the output CR waveform displayed on the phone scrreen.

holds the architecture and probabilistic features of a pre-trained Finite State Machine from the set of experiments described on section 5.3.3. The constituent nodes on the *FSM* class are implemented as described on section 5.2, using 16 bits signed integer registers for the memory units.

Figure 6.4 displays the working mobile application alongside the wireless ECG sensor. An USB connection was used to upload the firmware on the ESP32-WROOM-32 unit and to power it during operation.

6.3.2 Application benchmarks

The mobile application performance was evaluated with Apple's Xcode development environment and tested on an Iphone 6s device.

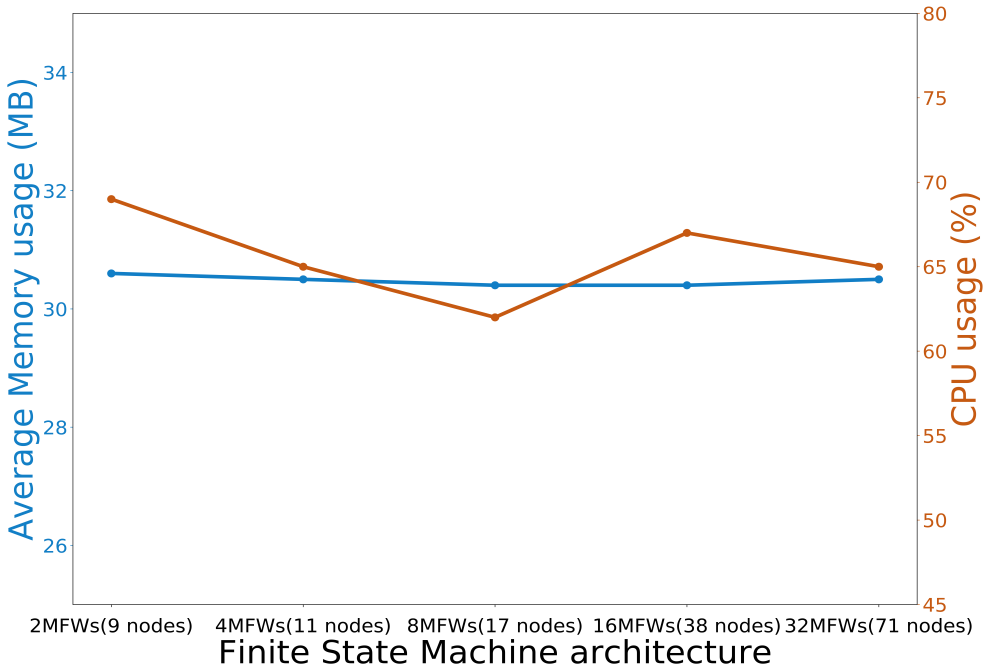


Figure 6.5: Mobile application benchmark statistics. CPU (brown line) and memory (blue line) usage are displayed for FSM models varying in increased size (from left to right).

The performance of the application was evaluated using Xcode's memory profiler and benchmark tools. Figure 6.5 shows the percentage of CPU and Memory resources used by the application for 5 different FSM models. The models are given in increasing order of size. The application ran in one thread demanding between 60% and 70% of one of the two cores resources in the Iphone device. Likewise, an average of 30.5 Mb of memory resources were spent for all models.

Chapter 7

Analysis on MFWs statistics for Normal, PVC and PAC heartbeats

7.0.1 Limitations on the thresholding method for a multi-class arrhythmia classification approach

As shown in Chapter 3, the proposed algorithm can perform reasonably well for the detection of Premature Ventricular Contractions as long as a proper threshold value (T) is chosen. The thresholding method, however, has some considerable limitations when implemented in a multi-class classification approach.

Figure 7.1 gives an illustrative example of the resultant instantaneous Compression Ratio R_i obtained after processing an ECG signal. The signal in question contains three normal heartbeats, one PAC and one PVC.

On Figure 7.1, the PVC heartbeat produces a pronounced peak in the CR signal, while the opposite happens for the PAC heartbeat. The reason for

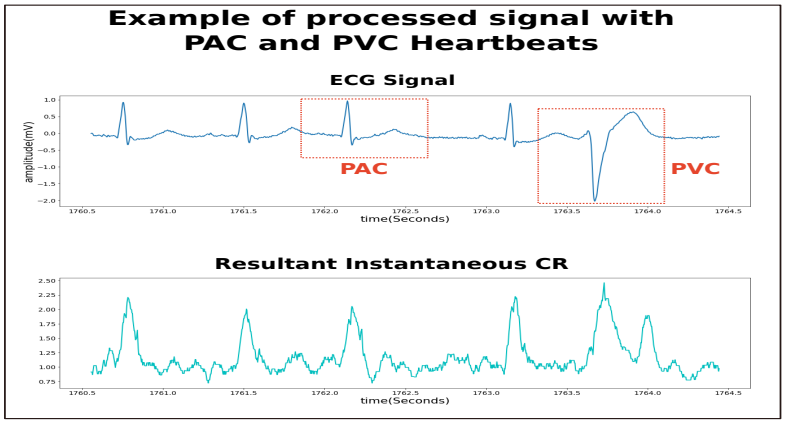


Figure 7.1: Failed detection of a PAC heartbeat with thresholding method.

this lays in the fact that the PVC morphology greatly differs from the Normal and PAC heartbeats, thus increasing the chances of transitions to forbidden states with low transition probabilities within the constructed FSM.

PAC heartbeats present themselves with an overlap in T and P waves on contiguous heartbeats. The relative small change in the resultant ECG morphology usually do not translate in comparable increments in the CR signal.

A solution to the problem relays on the use of multiple threshold values to discriminate between the three types of heartbeat. In that sense, a primary threshold value is set to binary discriminate from PVC heartbeats and others types of heartbeat. In those cases when the primary threshold is not reached, a second evaluation is made under a secondary threshold value to further classify the heartbeats into Normal or PAC heartbeats.

The following section tries to take a different approach for the possible classification of heartbeats in a multi-class scenario. By taking a closer look at the MFW statistics and their morphology, the basis are set for a multi-class heartbeat classification approach based on antidictionaries statistical

features.

7.0.2 Minimal Forbidden Word analysis on Normal, PVC and PAC heartbeats

In order to extend the capabilities of the antidictionary coding algorithm for a multi-class detection problem, a survey on the characteristic of the different MFWs that can be linked to each type of heartbeat has been performed.

Using records 114, 200 and 201 from the MIT-BIH database, individual MFW statistics for each type of heartbeat were collected. For each record, a common antidictionary set $\mathcal{A}_c$ was constructed from the quantized ECG signal $\boldsymbol{x}^n$, following the procedure described on section 5.3.1 of Chapter 5, picking up the 300 most common MFW acording to their frequency $f(\boldsymbol{w})$ values.

From the ECG annotation files, a set that contains the sample location of the R peak of every heartbeat within the ECG time series can be expressed as $\boldsymbol{R_{SET}} = \{\boldsymbol{R_1}, \boldsymbol{R_2}, \ldots, \boldsymbol{R_s}\}$, were s is a positive integer and its value varies from record to record.

Given the location of the R peak of some heartbeat $\boldsymbol{R_v}$ in $\boldsymbol{R_{SET}}$, it is possible to delimit the heartbeat in terms of a fragment of quantized signal $\boldsymbol{x}_l^h = \{x_l, x_{l+1}, \ldots x_h\}$ with $\boldsymbol{x}_l^h \subset \boldsymbol{x}^n$.

The boundaries (l, h) of the segment of quantized signal $\boldsymbol{x}_l^h$ are given as

$$l = \boldsymbol{R_{v-1}} + \frac{1}{2}(\boldsymbol{R_v} - \boldsymbol{R_{v-1}}), \quad h = \boldsymbol{R_v} + \frac{1}{2}(\boldsymbol{R_{v+1}} - \boldsymbol{R_v}) \tag{7.1}$$

where $1 < l < h < n$.

With every heartbeat delimited in terms of the quantized signal $\boldsymbol{x}^n$ and with the aid of the corresponding annotation label for each heartbeat, it

is possible to categorize the set of MFWs found within the confines of the quantized sequences.

Figure 7.2 shows an ECG signal with a scatter plot of a set of MFW found through the signal quantized sequence. Three different kinds of heartbeats are considered: 'Normal', 'PAC' and 'PVC'.

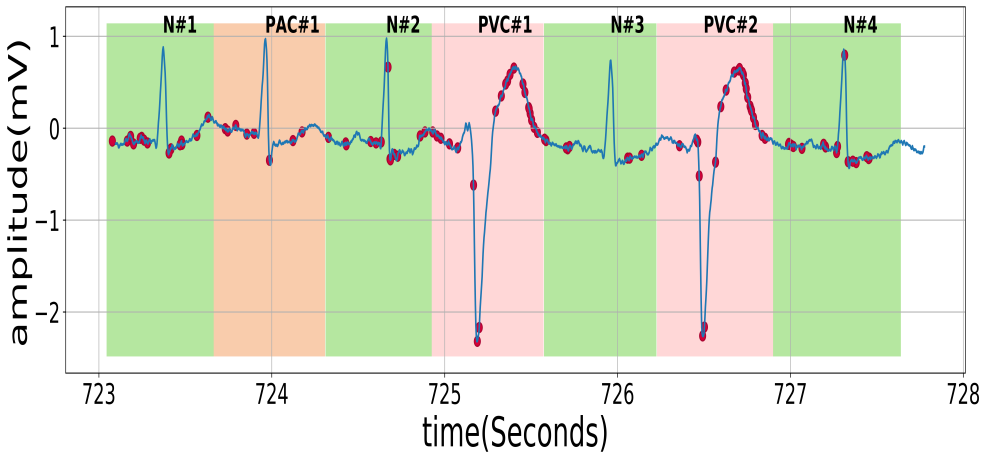


Figure 7.2: ECG signal displaying the localization (red markers) of MFW found in the quantized signal. Three different types of heartbeats are shown, with Normal heartbeats identified with the 'N' label (green background), PVC heartbeats with the 'PVC' label (pink background) and one PAC displaying the 'PAC' label (orange background).

For a given MFW $\boldsymbol{w}$, lets define the frequency of occurrence within the set of 'Normal' labelled quantized sequences as $f_N(\boldsymbol{w})$. Likewise, $f_{PAC}(\boldsymbol{w})$ and $f_{PVC}(\boldsymbol{w})$ correspond to the frequencies of occurrence of the MFW on quantized sequences labelled as 'PAC' and 'PVC', respectively.

For ECG records 114, 200 and 201, a survey on the MFW that display the maximum values of $f_N(\boldsymbol{w})$, $f_{PAC}(\boldsymbol{w})$ and $f_{PVC}(\boldsymbol{w})$ has been performed yielding the results shown on Table 7.1.

Table 7.1: MFW occurring most frequently within the quantized sequences of three different types of heartbeats.

	114			**200**			**201**		
	N	**PAC**	**PVC**	**N**	**PAC**	**PVC**	**N**	**PAC**	**PVC**
MFW	*1135*	*1355*	*565*	*0112*	*144*	*22222*	*2312, 4224*	*11111*	*11111, 2111, 55554*
f-value (%)	23.90	50.00	100	100	30	88.37	18.46	83.33	100

The obtained results give an insight into the differences of occurring MFWs withing the quantized ECG sequence.

Chapter 8

Conclusion and future works

The study on the antidictionary coding theory has shown that the implementation of the quantization and differentiation procedures on the ECG time series can lead to the construction of Finite State Machine probabilistic models that require less memory resources for their implementation in comparison to models built on signals defined over a binary alphabet.

The constructed models on the experimental process achieved relative high average levels of sensitivity (97.53%), while suffering on the specificity metric (93.89%) due to the lack of noise treatment of the test signals.

The low requirements of computational resources for the trained FSM probability models was confirmed with the port of a pre-trained FSM model into a mobile application. Average performance results of CPU (65%) and memory (30.5 MB) usage for online processing of ECG samples transmitted through BLuetooht LE confirm the antidictionary coding algorithm as a viable option for arrhythmia detection on mobile and wearable devices.

A look into the statistical features of MFWs on quantized segments of ECG signal show that it is possible to classify the MFWs according to the type of heartbeat in which they are most likely to appear. This sets the bases

for extending the binary classification scheme based on antidictionaries for a multi-class perspective.

8.1 Future works

From the obtained results and the difficulties encountered during the experimental process, some key points of improvement have been identified. From those points, it is possible to comment on future task to further improve the detection algorithm performance:

- The execution of further experiments including denoising techniques to further reduce the rate of false positive detection.
- The increment of the FSM models sizes on future experiments in order to increase the rate of true positive detection of abnormal heartbeats and to potentially include multiple types of MFWs corresponding to the diverse kinds of arrhythmias.
- To look for the optimum threshold value T from the data already collected from the multiple set of experiments.
- The implementation of the antidictionary construction algorithm in the IOS mobile application to effitiently construct the antidictionaries from the transmitted ECG signals on the fly.
- To dive in more specific details about the characteristic forbidden patterns that appear in each type of heartbeat. The use of string similarity metrics between the set of MFWs could be a valuable tool for their proper classification.

- The experimentation with more sophisticated methods for the evaluation of the changes in the Instantaneous Compresio Ration signal to better discriminate between the different types of heartbeats.

Chapter 9

Acknowledgements

I must start this section by expressing my most deep thanks to the people of Japan for having allowed me to come to this wonderful country to do science. In the same way, I would like to express my most sincere gratitude to Professor Hiroyoshi Morita for his continuous support and invaluable guidance and expertise. I feel extremely fortunate for the opportunity that was put in my hands to do research in his laboratory.

I also express my gratitude to Professors Hiroyuki Kasai and Akiko Manada for their useful and opportune feedback and support.

I extend my thanks to all the wonderful friends that I made during my stay in Japan. I will always cherish all those pleasant moments we shared.

Finally, I want to say thank you to all my family, to my mother Angela, my sister Claret, aunts Doris and Andrea and all my dear cousins. They have been always there for me and I will never forget them.

Bibliography

[1] Adnane, M., Belouchrani, A.: Premature ventricular contraction arrhythmia detection using wavelet coefficients. In: 2013 8th International Workshop on Systems, Signal Processing and their Applications (WoSSPA). pp. 170–173 (May 2013). https://doi.org/10.1109/WoSSPA.2013.6602356

[2] Alajlan, N., Bazi, Y., Melgani, F., Malek, S., Bencherif, M.A.: Detection of premature ventricular contraction arrhythmias in electrocardiogram signals with kernel methods. Signal, Image and Video Processing **8**(5), 931–942 (2014). https://doi.org/10.1007/s11760-012-0339-8

[3] Baura, G.D.: Medical Device Technologies Medical Device Technologies A Systems Based Overview Using Engineering Standards. Academic Press (2012)

[4] Clifford, G.D.: Ecg statistics , noise , artifacts , and missing data (2006)

[5] Crochemore, M., Mignosi, F., Restivo, A., Salemi, S.: Text compression using antidictionaries. In: Springer-Verlag (ed.) International Conference onAutomata, Languages an Programming. vol. 1644, pp. pp.261–270 (1999), `https://hal-upec-upem.archives-ouvertes.fr/hal-00619991/document`

[6] Diaz, D., Cunha, J.P.S.: Review Wearable Health Devices-Vital Sign Monitoring, Systems and Technologies. Sensors **18**(2414) (Aug 2018)

[7] Documentation, A.D.: Healthkit (2019), `https://developer.apple.com/documentation/healthkit`

[8] G. Frias, H. Morita, T.O.: Anomaly detection on quantized ECG signals by the use of antidictionary coding. Proceedings of the 41st Symposium on Information Theory and its Applications (Dec 2018)

[9] Goldberger, A.L., Amaral, L., et al.: Physiobank, PhysioToolkit, and Physionet: Components of a New Research Resource for Complex Physiologic Signals. AHA - Circulation **101 23**, E215–20 (June 2000)

[10] Ittatirut, S., Lek-uthai, A., Teeramongkonrasmee, A.: Detection of premature ventricular contraction for real-time applications. In: 2013 10th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology. pp. 1–5 (May 2013). https://doi.org/10.1109/ECTICon.2013.6559531

[11] Moody, G.B., Mark, R.G.: The impact of the MIT-BIH Arrhythmia Database. IEEE Engineering in Medicine and Biology Magazine **20**, 45–50 (June 2001)

[12] Morita, H.: Mfw-st repository, `https://github.com/hiroyoshi917/MFW-ST`

[13] Morita, H.: Overview: Antidictionary and its applications. In: Proc. 11th Asia-Europe Workshop on Concepts in Information theory. p. 10 (2019)

[14] Ota, T., Morita, H.: On the construction of an antidictionary of a binary string with linear complexity. IEEE ISIT pp. 2343–2347 (July 2006)

[15] Ota, T., Morita, H.: On-Line Electrocardiogram Lossless Compression Using Antidictionary Codes for a Finite Alphabet. IEICE Trans. on Information and Systems **E93-D**(12), 3384–3391 (Dec 2010)

[16] Ota, T., Morita, H., de Lind van Wijngaarden, A.J.: Real-Time and Memory-Efficient Arrhythmia Detection in ECG Monitors Using Antidictionary Coding. IEICE Fundamentals **E96-A**(12), 2343–2350 (Dec 2013)

[17] Rajni, Kaur, I.: Electrocardiogram Signal Analysis - An Overview. International Journal of Computer Applications **84**(7), 22–25 (2013)

[18] SIG), T.B.S.I.G.B.: Gatt services, `https://www.bluetooth.com/specifications/gatt/services`

[19] Systems, E.: Esp32-wroom-32 datasheet, `https://www.espressif.com/sites/default/files/documentation/esp32-wroom-32_datasheet_en.pdf`

[20] Systems, E.: Repository: Arduino core for esp32 wifi chip, `https://github.com/espressif/arduino-esp32`

[21] Tziakouri, M., et al.: Classification of AF and Other Arrhythmias from a Short Segment of ECG Using Dynamic Time Warping. Computer in Cardiology **44** (2017)

[22] University, T.: Cellular basis for cardiac arrhythmias, `http://tmedweb.tulane.edu/pharmwiki/doku.php/cellular_basis_for_arrhythmias`

[23] Zhao, Z., Zhang, Y.: SQI Quality Evaluation Mechanism of Single-Lead ECG Signal Based on Simple Heuristic Fusion and Fuzzy Comprehensive Evaluation. Frontiers in Physiology **9**(727) (June 2018)