ELECTROCARDIOGRAM SIGNAL BASED SUDDEN CARDIAC ARREST PREDICTION USING MACHINE LEARNING APPROACHES

L MURUKESAN LOGANATHAN

UNIVERSITY MALAYSIA PERLIS

2014

ELECTROCARDIOGRAM SIGNAL BASED SUDDEN CARDIAC ARREST PREDICTION USING MACHINE LEARNING APPROACHES

by

LMurukesan Loganathan
(1330610899)

A thesis submitted in fulfillment of the requirements for the degree of Master in Science (Mechatronic Engineering)

School of Mechatronic Engineering
UNIVERSITI MALAYSIA PERLIS

2014

ACKNOWLEDGMENT

Foremost, I would like to thank God for being with me, guiding me and giving wisdom, understanding and courage in every phase of this research work. Without His grace, completing M.Sc in Mechatronic would have remain only a dream for me.

I am grateful to the support and supervision of Dr. M. Murugappan, who brought in this opportunity, for his suggestion, guidance, encouragement and challenges throughout this work from its beginning. Indeed, his inputs and ideas were of immense help in the making of this thesis.

I extend thanks to my co-supervisor, Prof.Madya.Dr. Mohd Iqbal bin Omar, for the valuable comments and encouragement that helped me through this work. I also thanks Dr. Saravanan Krishnan from Hospital Pulau Pinang for his advice and support during initial stages of this thesis. In addition, I am grateful to Mr. Mohd Razif bin Omar, for his assistance in finishing the ECG DAQ circuit.

It is an honor to thank my fellow colleagues and members of intelligent signal processing (ISP) cluster who helped me by giving moral supports and ideas. Specifically, I would like to thank Mr R.Yuvaraj and Dr V.Vickneswaran for being there for me whenever I had doubts and problems in my research work.

At this junction, I would like to appreciate my family and friends whose prayers and wishes made a difference. Specifically, I would like to thank my parents (Mr.H.Loganathan and Mrs. Muniammal) for having faith in me to complete my research work during this two-year time. I also grateful to them for providing me with financial support whenever I requested. I also take this time to thank my future wife M.Prashanthini for providing me with moral support whenever I am stressed because of the research.

TABLE OF CONTENTS

LIST OF TABLES

algorithm with previous research work

LIST OF FIGURES

LIST OF ABBREVIATIONS

AED	-	Angle of Elevation or Depression
AF	-	Atrial Fibrillation
AMI	-	Acute Myocardial Infarction
ANOVA	-	Analysis of Variance
ANS	-	Autonomous Nervous System
ApEN	-	Approximate Entropy
AV	-	Atrioventricular
BMI	-	Body Mass Index
BP	-	Blood Pressure
BPM	-	Beats Per Minute
CD	-	Correlation Dimension
Ctree	-	Classification Tree
CFS	-	Correlation-based Feature Selection
CPR	-	Cardiopulmonary Resuscitation
CPVT	-	Catecholaminergic Polymorphic Ventricular Tachycardia
CVD	-	Cardiovascular Disease
DAD	-	Delayed After-Depolarization
DFA	-	Detrended Fluctuation Analysis
DWT	-	Discreet Wavelet Transform
EAD	-	Early After-Depolarization
ECG	-	Electrocardiogram
EDA	-	Exploratory Data Analysis
EPS	-	Electrophysiological Study
FFT	-	Fast Fourier Transform
GUI	-	Graphical User Interface
HCM	-	Hypertrophy Cardiomyopathy
HE	-	Hurst exponent
HF	-	High Frequency
HOS	-	Higher Order Statistics
HRV	-	Heart Rate Variability
HRVTi	-	Heart Rate Variability Index

HRT	-	Heart Rate Turbulence
ICD	-	Implantable Cardioverter Defibrillator
ICM	-	Ischemic Cardioyopathy
IDWT	-	Inverse Discreet Wavelet Transform
IIR	-	Infinite Impulse Response
KNN	-	K-Nearest Neighbor
LBBB	-	Left Bundle Branch Block
LDL	-	Low Density Lipoprotein
LF	-	Low Frequency
LLE	-	Largest Lyapunov Exponent
LVEF	-	Left Ventricular Ejection Fraction
LVH	-	Left Ventricular Hypertropy
MI	-	Myocardial Infarction
MSNA	-	Muscle Sympathetic Nerve Activity
MTWA	-	Millivolt T Wave Alternans
NFC	-	Neuro-Fuzzy Classifier
NPV	-	Negative Predictive Value
NSR	-	Normal Sinus Rhythm
NYHA	-	New York Heart Association
PCA	-	Principal Component Analysis
PF	-	Purkinje Fibers
PNN	-	Probabilistic Neural Network
PPV	-	Positive Predictive Value
PVB	-	Premature Ventricular Beat
RBBB	-	Right Bundle Branch Block
$R\text{-}T_{end}$	-	R wave until end of T wave
SampEN	-	Sample Entropy
SA	-	Sino-Atrial
SCA	-	Sudden Cardiac Arrest
SCD	-	Sudden Cardiac Death
SDNN	-	Standard Deviation of all NN interval
SDANN5	-	Standard Deviation of the Averages of NN intervals in all 5 min segments of the entire recording

SDANN	-	Square root of the mean of the sum of the squares of Differences between Adjacent NN intervals
SFC	-	Subtractive Fuzzy Clustering
SFS	-	Sequential Forward Selection
SVM	-	Support Vector Machine
VEB	-	Ventricular Ectopic Beats
VF	-	Ventricular Fibrillation
VT	-	Ventricular Tachyarrhythmia

LIST OF SYMBOLS

θ^o	-	Angle of Elevation/Depression
L	-	Cost function
f_c	-	Cut-off frequency
α_{ft}	-	Fourier Transform pair of t
f	-	Frequency
τ_d	-	Frequency dependent time delay
μ	-	Group's mean
Hz	-	Hertz
C	-	Hyper-parameter that trades-off the effects of minimizing the empirical risk against maximizing the margin
n	-	Index
N	-	Length of data
c_{m}	-	Mean or median interval
ms	-	Millisecond
mV	-	Millivolt
$\overline{X}$	-	Mean
N_b	-	Number of beats
n_O	-	Order of polynomial
τ_p	-	Parameter controlling position of Gaussian window along the time axis
Δt	-	Sampling period of time series
α_s	-	Scaling component
l	-	Scale size
α_i, α_N	-	Set of Langrange multipliers
ζ_i	-	Slack variable
σ_s	-	Smoothing parameter
τ	-	Threshold

t	-	Time
σ^2	-	Variance
σ_w	-	Width of RBF kernel
σ	-	Window width

Ramalan Serangan Jantung Secara Mendadak Berdasarkan Isyarat Elektrokardiogram Menggunakan Kaedah Pembelajaran Mesin

ABSTRAK

Tesis ini memberi tumpuan kepada peramalan serangan jantung secara mendadak (SCA) dengan menggunakan kadar jantung kebolehubahan (HRV) dan isyarat elektrokardiogram (ECG). Kematian jantung secara mendadak (SCD) adalah penyakit jantung kritikal yang menyumbang kepada jutaan kematian setiap tahun. SCD berlaku apabila SCA tidak dirawat lebih daripada 10 minit. Oleh itu, dengan meramalkan kejadian SCA sebelum ia bermula atau mengenalpasti pesakit yang berisiko tinggi kepada SCA boleh menyelamatkan jutaan nyawa. Dua pangkalan data antarabangsa, iaitu MIT/BIH Sudden Cardiac Death (20 subjek) dan MIT/BIH Normal Sinus Rhythm (18 subjek) telah digunakan dalam penyelidikan ini. Kedua-dua pangkalan data ini mempunyai dua penunjuk ECG untuk merakam pesakit dalam keadaan terlentang. Di samping itu, isyarat HRV juga turut disediakan dalam pangkalan data ini. Dua segmen dari isyarat HRV telah digunakan dalam kajian ini. Segmen pertama adalah lima minit panjang dan ia telah disegmen dua minit sebelum permulaan 'ventricular fibrillation' (VF). Manakala, segmen kedua pula satu minit panjang dan ia disegmen lima minit sebelum permulaan VF. Untuk subjek normal, segmen ini telah di ambil secara rawak. Selain itu, segmen ini dilakukan masing-masing untuk mencapai dua dan lima minit ramalan kejadian SCA Kedua-dua segmen isyarat HRV ini telah dipra-proses untuk menyingkir and menginterpolasi degupan ektopik. Seterusnya, ciri-ciri masa dan bukan linear telah diekstrak. Selepas itu, isyarat HRV telah disingkirkan alirannya dan ciri-ciri frekuensi telah diekstrak. Kaedah pemilihan ciri-ciri adalah berbeza untuk setiap masa segmen. Untuk isyarat HRV lima minit, pemilihan ciri kaedah pemilihan kehadapan berterusan (SFS) telah digunakan manakala dalam analisis HRV satu minit, pemilihan ciri-ciri berdasarkan analisis komponen utama (PCA) dan pemilihan ciri-ciri berdasarkan korelasi (CFS) telah digunakan di samping SFS. Ciri-ciri optimum yang dipilih menggunakan kaedah-kaedah tadi telah dianalisa untuk kepentingan statistiknya menggunakan penganalisa varians (ANOVA). Seterusnya, empat pengelas pembelajaran mesin (mesin dorongan vektor (SVM), rangkain neural pembarangkalian (PNN), jiran K-terdekat (KNN) dan pokok klasifikasi) telah digunakan untuk peramalan dalam kedua-dua analisis. Manakala, satu minit EKG, lima minit sebelum permulaan VF telah disegmen dari pangkalan data. Kemudian, ia telah dipra-proses untuk menyingkirkan bunyi yang berlaku akibat ganguan talian kuasa dan frekuensi tinggi. Kaedah penyingkiran bunyi novel berdasarkan penjelmaan Stockwell (ST) telah digunakan untuk menyingkirkan bunyi bertenaga sifar. Seterusnya, segmen dari gelombang R sampai penghujung gelombang T ($R\text{-}T_{end}$) telah diekstrak dari setiap rakaman ECG. Dua kumpulan ciri-ciri (G1 dan G2) telah diekstrak dari segmen ECG yang novel ini. G1 terdiri daripada empat ciri-ciri bukan linear (eksponen Hurst, eksponen Lyapunov terbesar, penghampiran entropi dan entropi sampel) manakala G2 terdiri daripada empat ciri-ciri statistic perintah tinggi (purata, varians, 'skewness' dan kurtosis) dan satu ciri yang dicadang dalam tesis ini, sudut ketinggian/kemurungan (AED). Ciri AED yang dicadangkan ini didapati penting secara statistik (ANOVA) dengan $p < 0.05$. Dalam analisa ini, tiga pengelas (SVM, kluster penolakan kabur (SFC) dan pengelas neuro-kabur (NFC)) telah digunakan untuk meramal SCA. Melalui analisa-analisa ini, peramalan SCA dua minit dan lima minit dengan ketepatan maksimum 97.37% telah tercapai mengunakan isyarat HRV. Tambahan pula, ketepatan 100% tercapai dalam penganalisaan satu minit ECG. Ciri AED yang dicadang mencapai ketepatan 86.84% dalam peramalan.

Electrocardiogram Signal Based Sudden Cardiac Arrest Prediction Using Machine Learning Approaches

ABSTRACT

This thesis focuses on predicting occurrence of imminent sudden cardiac arrest (SCA) using heart rate variability (HRV) and electrocardiogram (ECG) signals. Sudden cardiac death (SCD) is a devastating cardiovascular disease that responsible for millions of deaths per year. SCD occurs when SCA went untreated for more than 10 minutes. Hence, predicting imminent SCA before its occurrence or identification of high-risk patients for SCD can save millions of lives. Two international databases, namely MIT/BIH Sudden Cardiac Death database (20 subjects) and MIT/BIH Normal Sinus Rhythm database (18 subjects) were used in this work. Both databases have two leads ECG recording of patients in supine condition. In addition, HRV signals are provided in these databases. Two segments of HRV signals were used in this work. First segment is five minutes long and it was segmented two minutes before the onset of ventricular fibrillation (VF). Consequently, second segment is one minute long and it was segmented five minutes before the onset of VF. As for normal subjects, these segmentations were done at random intervals. Besides, these segmentations were done to achieve two and five minute prediction of imminent SCA, respectively. Both HRV signal segments were pre-processed to remove and interpolate ectopic beats. Then, time and non-linear domain features were extracted. Next, HRV signals were detrended and frequency domain features were extracted. Feature selection method is different for each time segment. For features of five minutes HRV signal, sequential forward selection (SFS) was used to select optimal features while in one minute HRV analysis, feature selection using principal component analysis (PCA) and correlation based feature selection (CFS) were experimented in addition to SFS. Optimal features selected using each methods were analyzed for its statistical significance using analysis of variance (ANOVA) test. Based on literature, four machine learning classifiers (support vector machine (SVM), probabilistic neural network (PNN), K-nearest neighbour (KNN) and classification tree (CTree)) were used for prediction in both analyses. In contrast, one minute ECG, which is five minutes before the onset of VF, was extracted from the database. Then, it was pre-processed to eliminate power line interference and high frequency noises. S-Transform (ST) based novel noise removal method was used for removing zero energy noises. Then, segment from R wave until the end of T wave ($R\text{-}T_{end}$) was extracted from each ECG trace. Two groups of features (G1 and G2) were extracted from this novel ECG segment. G1 consists of four non-linear features (Hurst exponent, largest Lyapunov exponent, approximate entropy and sample entropy) while G2 consists of four higher order statistic features (mean, variance, skewness and kurtosis) and proposed angle of elevation/depression (AED) feature. The proposed AED feature is statistically significant (ANOVA) with $p < 0.05$. In this analysis, three classifiers (SVM, subtractive fuzzy clustering (SFC) and neuro-fuzzy classifier (NFC)) were used for SCA prediction. Through these analyses, maximum prediction accuracy of 97.37% was achieved in both two and five minutes SCA prediction using HRV signals. In addition, 100% prediction accuracy was produced in one-minute ECG analysis. The proposed AED feature produced 86.84% prediction accuracy.

CHAPTER 1

INTRODUCTION

1.1 Research Background

Early prognosis of fatal heart diseases (such as myocardial infarction, coronary artery disease and hypertrophy cardiomyopathy) is one of the crucial problems faced by modern society. These fatal cardiac diseases contribute millions of deaths worldwide in both developing and developed countries. Among these diseases, sudden cardiac death (SCD) is the most severe cardiovascular disease (CVD) since it alone contributes to almost 50% of cardiovascular mortalities (Heikki, Castellanos, & Myerburg, 2001); it is a major cause for 20% of total mortality in industrialized world (Wellens et al., 2014). Besides, patients who had survived SCD with the help of resuscitation where re-hospitalized within one year and 40% of them died within two years of period (Unit, 2013). Cardiac arrhythmias were found to contribute 90% of SCD (Martínez-Rubio, Bayés-Genís, Guindo, & Bayés, 1999). There are various definitions used by researchers to classify a cardiovascular death as SCD. However, European Society of Cardiology Task Force on Sudden Cardiac Death has suggested the definition of SCD as: “natural death due to cardiac causes, heralded by abrupt loss of consciousness within one hour of the onset of acute symptoms; preexisting heart disease may have been known to be present, but the time and the mode of death are unexpected” (Priori et al., 2001).

Researchers have proposed several SCD risk factors over the past several decades. These risk factors often categorized into three groups for sake of simplicity namely, substrate, modulator and trigger (Straus, 2005). Factors that damage the normal structure of the myocardium are known as substrate. Substrates are the more stable risk factors among these three groups. The effects left by substrate factors (heart failure, myocardial infarction) are permanent thus; it creates a surrounding in which ventricular fibrillation (VF) can readily occur (Myerburg, 1997). Besides, risk factor that temporarily increases the risk of SCD such as drugs, plaque rupture and electrolyte disturbances are known as modulators. Finally, triggers are the critically timed premature stimulus such as ventricular extra systole that triggers the VF.

In general, physiological signals based SCD risk markers are divided into two groups namely, first group is based on electrocardiogram (ECG) signals and second group contains rest of related risk markers. High resting heart rate (> 75 bpm) (bpm: beats per minute), limited heart rate increase during stress test (< 89 bpm) and sluggish heart rate recovery after stress test (< 25 bpm) are reported as SCD risk markers in current clinical practice (Jouven, Empana, Schwartz, & Desnos, 2005). Besides, corrected QT interval (QTc)>450 ms in men and >470 ms in women is associated with three-fold risk of SCD (Straus et al., 2006). In addition, prolonged T peak-T end duration in ECG lead V5 is considered as one of the SCD risk marker (Panikkath et al., 2011). Furthermore, T wave inversion, prolonged QRS duration and wide QRST angle were used to predict SCD (Tikkanen, Anttonen, & Juhani, 2009). Besides, various features from RR interval of normal and heart disease patients proved to be a significant SCD predictor (Ebrahimzadeh & Pooyan, 2011).

The impact of SCD in various populations is shown in Table 1.1. Current medical practice is to prioritize group 3 patients since they are the most obvious candidates or victims for SCD. However, percentage of fatality due to SCD in group 3 is merely 13% of total SCD mortalities compared to 45% and 40% from group 1 and 2.

This led the researchers to investigate variety of methods to identify high-risk individual from all groups.

Table 1.1: Four groups of people that contribute to SCD and their current predictability status (Wellens et al., 2014)

Groups	Description	Contribution to all SCD (%)	SCD Predictability
1	Not diagnosed with heart disease	45	Poor
2	With history of heart disease: LVEF > 40%	40	Limited
3	With history of heart disease: LVEF < 40%	13	Possible
4	Arrhythmic disease by genetic substrate	2	Limited

LVEF=Left Ventricular Ejection Fraction

There are some specific risk markers being correlated to group two, three and four patients in Table 1.1. Heart rate turbulence (HRT), deceleration capacity and microvolt T wave alternans (MTWA) produced high specificity in group 2 patients (Bauer et al., 2009). MTWA also found to have high negative predictive value (NPV) in group 3 patients (Merchant et al., 2012). Finally, in group 4 patients, specifically for long QT (delayed Q wave and T wave interval) syndrome and Brugada syndrome, RR interval and QT interval (interval from Q wave to T wave) were established as risk factors (Moss et al., 1991).

There are numerous non-ECG risk markers proposed by researcher for SCD stratification. Left ventricular ejection fraction (LVEF) which is less than 40% is considered as a threshold to separate low and high-risk SCD patients (Rouleau, Talajic, Sussex, Potvin, & Warnica, 1996). Hence, LVEF became the most important clinical parameter in identifying high-risk patients. Besides, cardiovascular disease risk scores such as Framingham and QRISK (since it is based on QResearch database) were used to calculate the probability of generating cardiac events using age, gender, low-density lipoprotein (LDL) cholesterol, body mass index (BMI) and blood pressure (BP) (Laslett et al., 2012). Other than that, classification of patients using New York heart association (NYHA) functional class (I, II, III and IV) is used to prioritize high-risk groups. Class IV patients are diagnosed with severe heart dysfunctions while Class I is patients with very mild heart dysfunctions. Furthermore, electrophysiological study (EPS), an invasive method is used to assess the electrical conduction system of the heart. Patients with positive EPS that implanted with implantable cardioverter defibrillator (ICD) are found to be less susceptible to SCD (Buxton et al., 1999). Besides, there are nuclear studies and cardiac angiography that are used for SCD and other cardiovascular disease assessments.

Both these groups (ECG and non-ECG) of risk markers have several limitations as follow. Among the ECG based risk markers, HRT and MTWA were found to be more promising in identifying high-risk SCD patients. Yet, these two parameters cannot be assessed on patients with atrial fibrillation (AF) (which by itself act as a risk predictor for SCD) (Wellens et al., 2014). Besides, there are lack of dynamic risk assessment studies over time for these ECG risk factors. In addition, specific risk marker for SCD based on ECG is yet to be discovered. Whereas, LVEF and NYHA functional classes are the widely used non-ECG based risk marker for SCD (Wellens et al., 2014). However, the sensitivity and specificity of these risk factors are considered marginal in medical point of view.

Drugs and ICDs are used in current medical practice as preventative measure of SCD. Drugs such as Beta-blockers are used to reduce the occurrence of ventricular ectopic beats (VEB), thus reduce the probability of SCD (Ellison et al., 2002). Besides, the safety and efficacy of the drug makes it first choice in protecting patients against SCD. Furthermore, amiodarone blocks the potassium repolarization and increases the re-entry wavelength, which found to oppose the occurrence of ventricular arrhythmias.